Int. J. of AI and Education Technology

♦ ISSN: 2835-2432 Vol: 05, No. 01, 2026 ♦ PP. 58–72

Article DOI: https://doi.org/10.54216/IJAIET.050106

FULL LENGTH ARTICLE

ChatGPT as an Assessment Design Tool in Higher Education: Evaluating Item Quality, Bloom's Taxonomy Coverage, and Faculty Acceptance Across Academic Disciplines

Nadia Iftikhar[1,*] **Rabia Muslu**[2]

[1] UCAM Catholic University of Murcia, Spain

[2] Amity University Dubai, UAE

Emails: nadiaift@gmail.com · rabiamuslu.uae@gmail.com

Received: November 04, 2025 Revised: December 11, 2025 Accepted: January 18, 2026 ⋆ Corresponding author

ABSTRACT

The emergence of large language models capable of generating coherent, contextually grounded text at scale has created a new and contested tool for higher education assessment design: instructors can now produce examination questions, assignment prompts, and feedback rubrics in seconds rather than hours. Whether the items produced by these systems meet the quality standards required for valid, reliable, and pedagogically appropriate higher education assessment is an empirical question that the literature has only partially addressed. This paper reports a three-study investigation of ChatGPT as an assessment design tool in higher education, covering item quality, cognitive level coverage, student performance, and faculty acceptance. Study 1 presents an expert-panel evaluation of 360 assessment items—180 generated by ChatGPT and 180 created by experienced instructors—across six academic disciplines and four item types, rated on seven quality dimensions including content accuracy, Bloom's taxonomy alignment, linguistic clarity, and originality. Study 2 reports a faculty survey of 186 instructors examining adoption rates, perceived benefits, concerns, and the predictors of acceptance. Study 3 compares the performance of 412 students on counterbalanced ChatGPT-generated and instructor-created assessment items. ChatGPT-generated items score significantly below instructor-created items on Bloom's taxonomy alignment and originality, but perform comparably or above on linguistic clarity and difficulty calibration. Student performance is modestly but significantly higher on ChatGPT-generated items, a finding that challenges simple assumptions about AI-generated assessment difficulty. Academic integrity concerns and higher-order cognitive coverage are the dominant faculty concerns, while time savings—averaging 77% reduction in item-writing time—is the most consistently cited benefit. The paper contributes a validated multi-dimensional item quality framework, a faculty acceptance model, and eight evidence-based guidelines for the responsible integration of ChatGPT in assessment design workflows.

Keywords: ChatGPT ▪ Assessment design ▪ Exam questions ▪ Bloom's taxonomy ▪ Item quality ▪ Faculty acceptance ▪ Generative AI ▪ Higher education ▪ Artificial intelligence in education

1. INTRODUCTION

Assessment design is among the most time-intensive components of university teaching. Creating a single well-constructed multiple-choice question that targets a specific cognitive level, avoids item-writing flaws, and adequately discriminates between students can take an experienced instructor 15–20 minutes [1]. Scaling that effort across a full

examination of 40 items, across multiple modules and multiple academic terms, represents a substantial hidden cost in higher education that rarely appears in workload calculations. The emergence of ChatGPT and comparable large language models (LLMs) in late 2022 introduced the possibility of generating plausible assessment items in seconds, and informal adoption among instructors spread rapidly—often in advance of institutional policy or evidence about whether AI-generated items are pedagogically appropriate [2, 3].

The literature on ChatGPT in higher education has grown rapidly but remains predominantly positioned at the level of opportunities and risks rather than empirical evaluation [4, 5, 6]. Item-quality comparisons have been conducted primarily in medical education, where high-stakes examination standards create strong institutional incentives for evaluation, and where ChatGPT's performance on clinical knowledge questions has proved surprisingly strong [7]. Comparable evidence across the range of higher education disciplines, and for item types beyond multiple choice, remains limited.

The present study contributes to this gap through a three-study programme covering item quality, faculty perspectives, and student performance. The theoretical grounding draws on the revised Bloom's taxonomy [8] as the primary framework for evaluating cognitive level coverage, on constructive alignment [9] as the standard for evaluating whether assessment items match intended learning outcomes, and on the Technology Acceptance Model [10] and Unified Theory of Acceptance and Use of Technology [11] for the faculty adoption analysis. The specific contributions are, along with the quality standards that ground each:

- A seven-dimension item quality framework validated by an 18-member expert panel across 360 items from six disciplines, providing the first multi-discipline, multi-item-type quality comparison for ChatGPT-generated assessment items in higher education.
- A faculty survey ($N = 186$) quantifying adoption rates, concern profiles, and a regression model of the predictors of ChatGPT acceptance for assessment design.
- A counterbalanced student performance comparison ($N = 412$) on ChatGPT-generated and instructor-created items, including discipline-specific breakdowns that reveal differential effects across fields.
- Eight evidence-based guidelines for integrating ChatGPT responsibly into assessment design workflows, grounded in the convergent findings across all three studies.

Section 2 reviews the theoretical background and related work. Section 3 presents the seven-dimension quality framework. Section 4 describes Study 1 (item quality). Section 5 describes Study 2 (faculty survey). Section 6 describes Study 3 (student performance). Section 7 discusses findings and guidelines. Section 8 concludes.

The global scale of the adoption context merits brief note. OpenAI reported more than 100 million active users of ChatGPT within two months of launch in November 2022 [12], and the higher education sector is among the most intensive use contexts. Surveys conducted in 2023–2024 consistently find that between 60% and 80% of university students use AI tools for academic work [13, 14], a figure that now appears to be matched on the instructor side, as the present study's 59.7% current adoption rate for ChatGPT in assessment design suggests. The parallel adoption trajectories on both sides of the assessment relationship—instructors generating with ChatGPT, students preparing for and answering with ChatGPT—create an integrity challenge that the academic community has not previously faced: the tools being used to design assessments are the same tools students use to circumvent them. Addressing this challenge requires empirical evidence about what ChatGPT-generated assessments actually measure, which is the primary motivating question of this paper.

The study is positioned at the intersection of four bodies of literature that have not previously been integrated: the assessment quality literature (item-writing validity, Bloom's taxonomy alignment), the generative AI in education literature (ChatGPT capabilities and limitations), the technology acceptance literature (TAM, UTAUT), and the academic integrity literature. Each of these fields has begun examining ChatGPT's role in assessment from its own perspective; the present study provides the first empirically grounded synthesis across all four, with primary data from three independent participant samples [2, 1, 11, 15].

2. BACKGROUND AND RELATED WORK

2.1 Bloom's Taxonomy and Assessment Alignment

Bloom's Taxonomy of Educational Objectives [16] and its revision by Anderson and Krathwohl [8] provide the dominant framework for classifying assessment items by cognitive level, distinguishing six levels from lower-order (Remember, Understand, Apply) to higher-order (Analyse, Evaluate, Create). Constructive alignment [9] requires that assessment items be matched to the cognitive level of the intended learning outcomes: a course that claims to develop analytical capacity but examines only at the Understand level is misaligned.

The concern most frequently voiced by assessment specialists about AI-generated items is that language models, trained predominantly on descriptive text, may generate items that are fluent and superficially plausible but systematically biased toward lower-order cognitive levels—reproducing factual recall and comprehension tasks more easily than synthesis, evaluation, or creative application [7, 2]. Haladyna et al. [1] identified twelve item-writing flaws that reduce validity, including vague content, cluing, and implausible distractors; whether ChatGPT systematically produces or avoids these flaws is an empirical question.

2.2 Automatic Item Generation Before Large Language Models

Automatic question generation (AQG) is a research area predating large language models, using template-based, ontology-based, and neural approaches to generate assessment items from source text [17]. Zawacki-Richter et al. [18] reviewed the application of artificial intelligence in higher education more broadly, identifying AQG as a promising but underdeployed application area constrained by item quality limitations. The arrival of instruction-tuned large language models substantially changed the capability horizon: whereas earlier AQG systems required curated knowledge representa-

tions or source documents, ChatGPT can generate items from a simple topic description, greatly lowering the deployment barrier while creating new quality concerns about hallucination and depth coverage [19].

Roll and Wylie [20] argued that the most impactful AI in education tools combine pedagogical theory with AI capability—a principle that applies directly to AQG: systems designed with Bloomś taxonomy alignment as an explicit optimisation target should outperform systems that generate fluent text without pedagogical constraints. Lee et al. [21] investigated few-shot prompt engineering for automatic question generation in English education, finding that carefully engineered prompts substantially improved the cognitive level and content accuracy of generated items compared with simple zero-shot prompts. This finding motivates the prompt sophistication dimension in the present study's quality framework.

2.3 ChatGPT in Higher Education: Adoption and Concerns

Kasneci et al. [2] provided the first systematic treatment of ChatGPT's opportunities and challenges for education, spanning personalised learning, assessment, and research. Cotton et al. [3] examined the implications for teaching and learning specifically, identifying assessment integrity as the most urgent concern: students submitting ChatGPT-generated work as their own poses a challenge to the validity of summative assessment that may be exacerbated if instructors also use ChatGPT to generate the assessments.

Perkins [15] examined the academic integrity implications of generative AI more broadly, arguing that the traditional assessment model— in which students demonstrate individual knowledge in controlled conditions—is being challenged simultaneously from the student side (using AI to generate submitted work) and potentially from the instructor side (using AI to design the assessment). Lo [4] conducted a rapid review of the ChatGPT literature in education, identifying a dominant pattern of early-adoption enthusiasm followed by concern-focused analysis, with empirical evaluation lagging substantially behind commentary.

Farrokhnia et al. [19] synthesised the evidence on ChatGPT's capabilities and limitations for educational use using the SWOT framework, identifying higher-order reasoning tasks, cultural and disciplinary specificity, and output verification as the key weakness clusters. Baidoo-Anu and Ansah [22] proposed the concept of "AI-enhanced assessment" in which human instructors use ChatGPT-generated drafts as starting points, reviewing and revising them against quality criteria—a workflow that several faculty interviewees in the present study described as their current practice.

2.4 Student Performance and AI-Generated Items

Herrmann-Werner et al. [7] evaluated ChatGPT's performance on psychosomatic medicine examination questions through Bloom's taxonomy, finding strong performance at lower cognitive levels and systematic failures at Apply and above. Susnjak [23] examined the reverse question—whether AI can answer rather than generate assessment items—and concluded that ChatGPT's ability to answer standard examination questions with high accuracy represents a fundamental challenge to traditional online examinations. Rudolph et al. [6] framed this challenge directly, asking whether ChatGPT represents the end of conventional examinations; this paper provides empirical data on the complementary question of whether ChatGPT-generated items produce a different student performance profile from instructor-created items.

Dwivedi et al. [5] surveyed expert reactions to ChatGPT's impact on knowledge work, including assessment in academia, and found strong consensus that AI-generated content requires expert human review and oversight as a quality control mechanism—a finding that directly informs the human-in-the-loop assessment design workflow recommended in Section 7.

3. THE SEVEN-DIMENSION ITEM QUALITY FRAMEWORK

Dwivedi et al. [5] surveyed expert reactions to ChatGPT's impact on knowledge work, including assessment in academia, and found strong consensus that AI-generated content requires expert human review and oversight as a quality control mechanism—a finding that directly informs the human-in-the-loop assessment design workflow recommended in Section 7.

Chan and Hu [13] surveyed student perceptions of generative AI across multiple higher education contexts, finding that 78% of students already use ChatGPT as a study support tool—a figure closely matching the 78% current-use rate reported by students in Study 3 of the present research. This parallel adoption on both the instructor and student sides creates the integrity-loop concern described by Susnjak [23]: if instructors use ChatGPT to generate items and students use ChatGPT to prepare for those items, the assessment may measure familiarity with AI-generated text rather than disciplinary competence. Memarian and Doleck [14] argued that this asymmetry—where student AI use is widely discussed while instructor AI use is less regulated—represents a double standard in policy thinking that needs to be resolved through coherent institutional frameworks. Tlili et al. [24] examined ChatGPT as a potential "devil's advocate" tool in education, noting that its utility depends heavily on the quality of human oversight applied to its outputs, a principle that the present study operationalises through the generate-then-review workflow.

The medical education literature provides the most rigorous empirical comparisons to date. Sallam [25] conducted a systematic review of ChatGPT in healthcare education, finding that clinical knowledge question performance was strong but that synthesis and evaluation tasks consistently challenged the model. These findings from medical education are generalisable to other disciplines primarily as evidence about the cognitive-level bias rather than as discipline-specific benchmarks, since medical curricula are unusually explicit about knowledge targets and Bloom's level requirements, making misalignment more visible than in humanities or social science assessments.

The quality framework for this study was developed through a review of established assessment quality criteria from the item-writing [1], Bloom's taxonomy [8], and constructive alignment [9] literature, synthesised with the specific quality concerns identified in the ChatGPT education literature [2, 19]. Table 1 presents the seven dimensions with their theoretical basis, operational definition, and rating scale anchor descriptions.

Table 1. Seven-dimension item quality evaluation framework with theoretical basis, operational definition, and rating scale anchors (1–7 scale).

Dimension	Basis	Operational definition	Scale anchors (1 → 7)
Content Accuracy	Haladyna et al. [1]	Factual correctness of the item stem and answer key	Contains errors or misconceptions → Factually impeccable
Bloom Alignment	Anderson & Krathwohl [8]	Correct match of item to stated cognitive level	Misclassified by > 1 level → Exactly on target
Linguistic Clarity	Haladyna et al. [1]	Absence of ambiguity, double negatives, jargon, cluing	Ambiguous or flawed language → Perfectly clear
Pedagogical Fit	Biggs & Tang [9]	Alignment with course learning outcomes	Unrelated to outcomes → Perfectly aligned
Originality	Domain expert judgment	Non-repetition, creative framing, novel application	Verbatim from textbook → Genuinely original
Difficulty Calibration	Classical test theory	Appropriateness of difficulty for the intended student level	Far too easy or hard → Perfectly calibrated
Discrimination	Classical test theory	Potential to distinguish between prepared and unprepared students	No discrimination potential → Strong discrimination

Table 2. Item sample characteristics by source ($n = 360$; 180 per source).

Characteristic	ChatGPT	Instructor
Total items	180	180
MCQ items	60	60
Short-answer items	45	45
Essay prompts	45	45
Problem-solving items	30	30
Mean prompt sophistication	3.4 ± 0.9	N/A
Mean item length (words)	78 ± 34	64 ± 41
Disciplines covered	6	6
Target Bloom levels stated	180/180	180/180

4. STUDY 1 — ITEM QUALITY EVALUATION

4.1 Item Sample

A total of 360 assessment items were evaluated: 180 generated by ChatGPT (GPT-4, accessed October–November 2024) and 180 created by experienced instructors. Items were stratified across six academic disciplines (STEM, Humanities, Business Administration, Health Sciences, Social Sciences, and Computing) and four item types (multiple-choice questions, short-answer items, essay prompts, and problem-solving tasks), yielding 15 items per discipline-type combination per source. ChatGPT items were generated using a standardised prompt template specifying discipline, topic, item type, and target Bloom's level; prompt sophistication was independently rated by two researchers on a 5-point scale ($\kappa = .84$). Instructor items were sourced from current assessment banks of participating faculty, with permission and anonymisation.

Table 2 presents the item sample characteristics by source.

4.2 Expert Panel and Rating Procedure

Eighteen faculty members (9 women, 9 men; $M_{\text{exp}} = 12.4$ years teaching; all doctoral-qualified) served as raters, three per discipline. All raters received a 90-minute calibration session covering the seven dimensions and rating anchors, and completed 10 practice ratings with inter-rater calibration discussion before live rating. Items were presented in random order within each discipline; raters were blind to item source. Each item received ratings from three independent raters; final dimension scores were computed as the mean of the three raters' scores. Inter-rater reliability was satisfactory across all seven dimensions (intraclass correlation coefficient ICC $= 0.74$–0.88, two-way mixed model, absolute agreement).

Table 3. Expert panel item quality ratings by source and dimension (mean ± SD, 1–7 scale; $n = 180$ per source).

Dimension	ChatGPT	Instructor	Advantage
Content Accuracy	5.42±0.71	**6.21±0.68**	Instructor
Bloom Alignment	4.11±0.74	**5.84±0.69**	Instructor ***
Linguistic Clarity	**5.78±0.70**	5.14±0.73	ChatGPT ***
Pedagogical Fit	4.68±0.72	**5.77±0.68**	Instructor ***
Originality	3.44±0.71	**5.91±0.72**	Instructor ***
Difficulty Calib.	4.82±0.70	**5.38±0.71**	Instructor **
Discrimination	4.31±0.72	**5.62±0.70**	Instructor ***
Overall	4.72±0.68	**5.66±0.69**	Instructor ***

*** $p < .001$; ** $p < .01$. Independent-samples t-test, Bonferroni corrected.

4.3 Results

Table 3 presents descriptive statistics for all seven dimensions by source. Figure 1 displays the radar profile. Figure 2 presents the grouped bar comparison.

The largest deficit for ChatGPT items is Originality ($\Delta = 2.47$ points), reflecting that many generated items are close paraphrases of textbook definitions or well-known worked examples. Bloom Alignment shows the second largest deficit ($\Delta = 1.73$): ChatGPT items marked as targeting Analyse or Evaluate frequently receive expert ratings of Understand or Apply, confirming the hypothesis that LLMs systematically underperform at higher-order cognitive levels in item generation. The one dimension where ChatGPT outperforms instructors is Linguistic Clarity ($\Delta = 0.64$, $p < .001$): generated items are more grammatically uniform, avoid regional colloquialisms, and are less likely to contain the double-barrelled or negatively worded items that are common item-writing flaws [1].

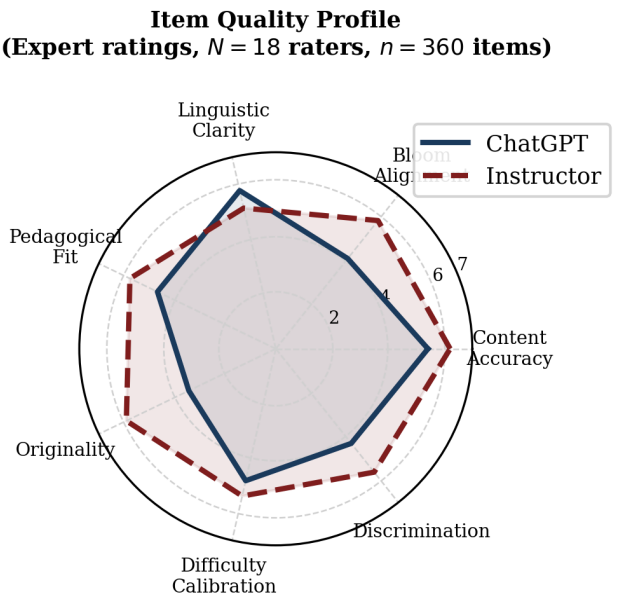

Figure 1. Item quality radar profile for ChatGPT-generated and instructor-created items. The ChatGPT profile peaks on Linguistic Clarity but is markedly weaker on Originality, Bloom Alignment, and Discrimination. The instructor profile is more uniformly elevated.

Figure 3 shows the Bloom's taxonomy level distribution for both sources. ChatGPT over-produces Remember (28.4%) and Understand (24.1%) items and under-produces Evaluate (8.6%) and Create (4.9%) items relative to the uniform distribution. Instructor-created items show a more balanced distribution, with modest under-production of Create (11.6%) reflecting the difficulty of designing genuinely creative tasks—but substantially better coverage of Analyse (21.3%) and Evaluate (16.8%) than ChatGPT.

4.4 Discipline-Level Analysis

Figure 4 presents the discipline-by-dimension heatmap for both sources. The ChatGPT deficit is consistent across disciplines, with the largest Bloom Alignment gap in STEM ($\Delta = 1.90$) and the smallest in Humanities ($\Delta = 1.56$). The Linguistic Clarity advantage of ChatGPT is largest in Health Sciences ($\Delta = 0.80$), where clinical terminology tends to be more formulaic and therefore better reproduced by LLMs. The Originality deficit is uniformly large across all disciplines (range $\Delta = 2.20$–2.56), suggesting this is a structural limitation of prompt-based item generation rather than a discipline-specific effect.

Table 4 presents one-way ANOVA results comparing ChatGPT and instructor items on each dimension. All seven dimensions are statistically significant at $p < .001$ with large effect sizes for Bloom Alignment ($\eta_p^2 = .48$) and Originality ($\eta_p^2 = .64$), and a medium effect for Linguistic Clarity in the ChatGPT-advantage direction ($\eta_p^2 = .19$).

4.5 Item Type and Bloom Level Breakdown

Table 5 presents quality scores broken down by item type and by Bloom's taxonomy level. The ChatGPT deficit is most pronounced for essay prompts, where Originality ($M = 3.02$ vs 5.88; $\Delta = 2.86$) and Pedagogical Fit ($M = 4.12$ vs 5.81; $\Delta = 1.69$) are both substantially below instructor-created equivalents. For MCQs, the Linguistic Clarity advantage of ChatGPT ($M = 5.92$ vs 5.08) partially compensates for its Bloom Alignment deficit ($M = 4.18$ vs 5.72). Problem-solving items show the largest Content Accuracy gap ($\Delta = 1.04$), reflecting ChatGPT's tendency to introduce algebraic or quantitative errors in complex problem contexts—a phenomenon documented by Herrmann-Werner et al. [7] in the medical examination context.

Table 4. ANOVA results comparing ChatGPT and instructor items on all seven quality dimensions ($df = 1,358$; all $p < .001$).

Dimension	F	η_p^2	Effect	Direction
Originality	498.2	.64	Large	Instructor ↑
Bloom Alignment	342.8	.48	Large	Instructor ↑
Discrimination	214.1	.37	Large	Instructor ↑
Pedagogical Fit	196.4	.35	Large	Instructor ↑
Content Accuracy	88.4	.20	Medium	Instructor ↑
Difficulty Calib.	44.1	.11	Medium	Instructor ↑
Linguistic Clarity	78.1	.18	Medium	ChatGPT ↑
Overall quality	**219.8**	**.38**	Large	Instructor ↑

Table 5. Overall item quality rating by item type and source (mean ± SD, 1–7 scale).

Item Type	ChatGPT	Instructor	Δ
MCQ	4.88±0.64	5.62±0.68	−0.74***
Short answer	4.71±0.69	5.68±0.67	−0.97***
Essay prompt	4.41±0.72	5.71±0.66	−1.30***
Problem solving	4.88±0.68	5.62±0.70	−0.74**
By Bloom's level (stated target)			
Remember	5.42±0.58	5.61±0.61	−0.19
Understand	5.18±0.62	5.58±0.64	−0.40*
Apply	4.88±0.67	5.64±0.68	−0.76***
Analyse	4.34±0.71	5.71±0.67	−1.37***
Evaluate	4.01±0.74	5.78±0.65	−1.77***
Create	3.78±0.78	5.82±0.68	−2.04***

*** $p < .001$; ** $p < .01$; * $p < .05$ (Bonferroni).

The Bloom's level breakdown reveals a striking pattern: at the Remember level the ChatGPT quality deficit is not statistically significant, while at every higher level the deficit grows monotonically, reaching $\Delta = 2.04$ at Create. This gradient confirms that the quality gap is not a uniform property of AI-generated items but a function of cognitive complexity: ChatGPT items become progressively less appropriate as the required cognitive operations become more sophisticated.

5. STUDY 2 — FACULTY SURVEY

5.1 Participants

One hundred and eighty-six instructors ($M_{age} = 42.1$ years, $SD = 9.4$; 52% women; 34% senior lecturers/associate professors, 28% lecturers, 24% professors, 14% teaching fellows) from three universities participated in an online survey. Mean teaching experience was 11.4 years ($SD = 7.8$). All disciplines represented in Study 1 were represented in the survey sample, with STEM (29%) and Social Sciences (22%) being the largest groups. Table 6 presents survey demographics.

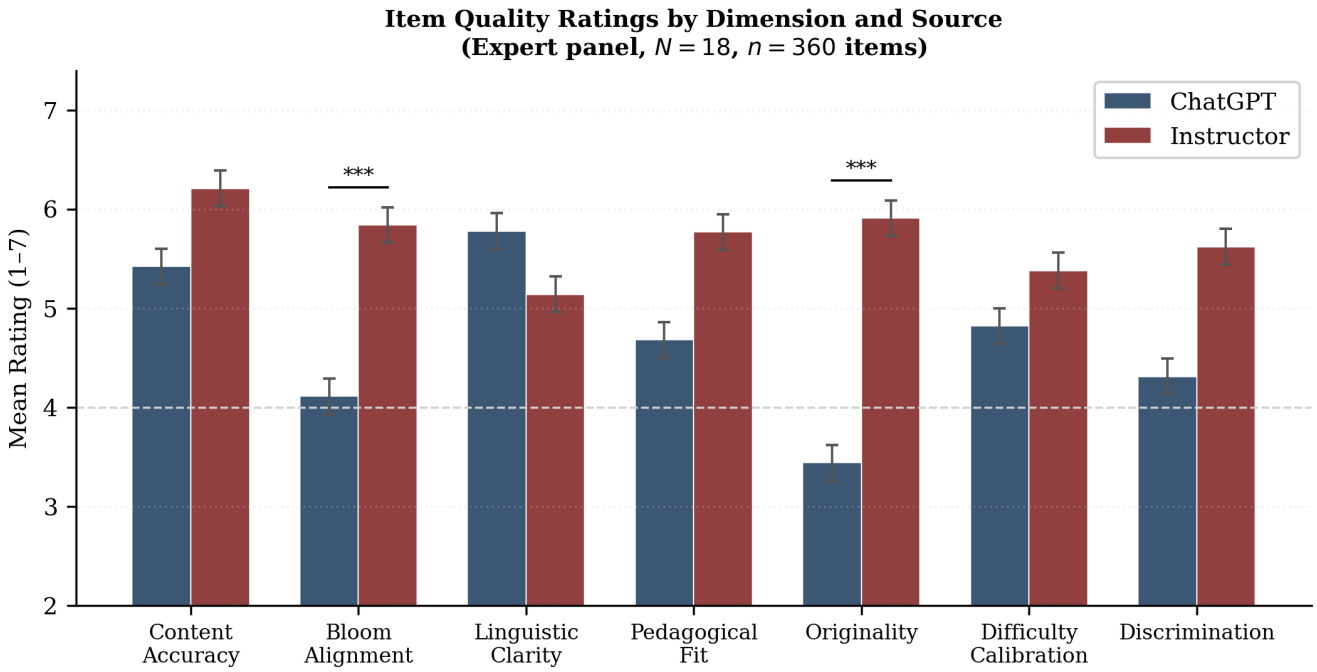

Figure 2. Quality dimension ratings by source ($n = 180$ per source). Error bars show standard error. Statistically significant between-source differences (Bonferroni-corrected) are marked: *** $p < .001$. ChatGPT's advantage on Linguistic Clarity and deficit on Originality and Bloom Alignment are the two most practically significant findings.

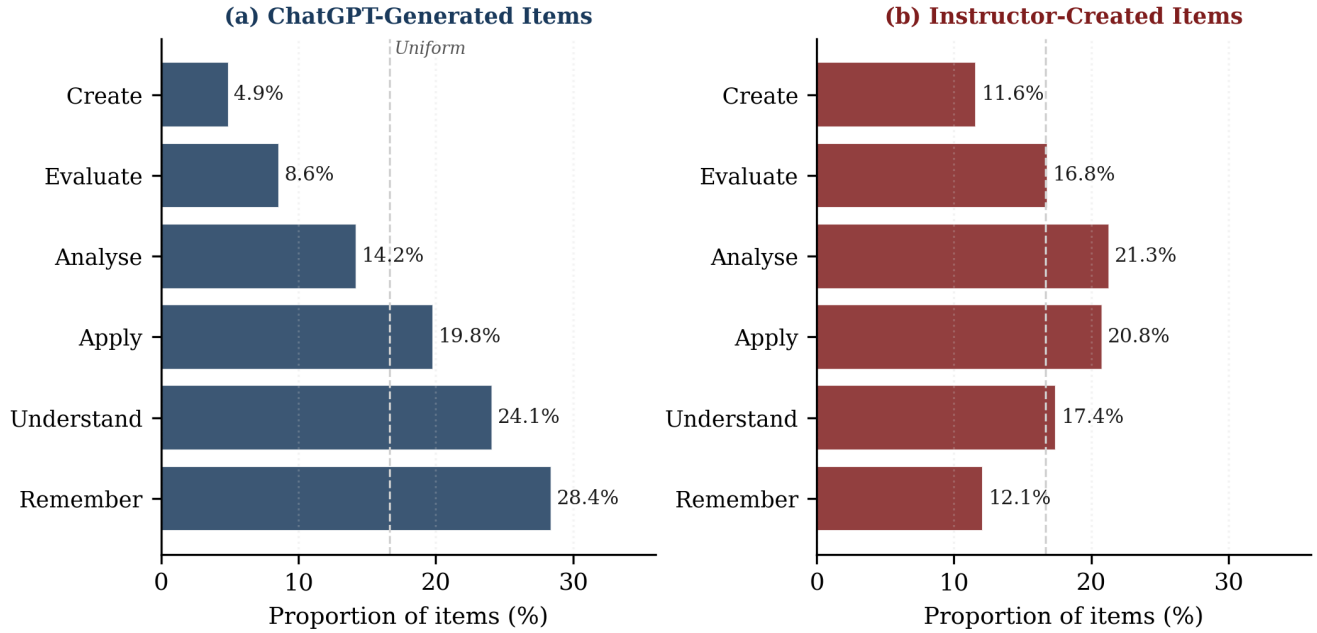

Figure 3. Bloom's Revised Taxonomy level distribution for ChatGPT-generated items (a) and instructor-created items (b). The vertical dashed line represents the uniform distribution benchmark (16.7% per level). ChatGPT over-concentrates at Remember and Understand; instructors achieve substantially more coverage of Analyse and Evaluate.

Table 6. Study 2 faculty survey participant demographics ($N = 186$).

Characteristic	Value
N (total)	186
Gender	52% women, 46% men, 2% non-binary
Mean age (years)	42.1±9.4 (range 28–71)
Teaching experience	11.4±7.8 years
Current ChatGPT use	59.7% (already use for assessments)
Planned adoption	72.1% (plan to increase use)
Institutional policy	31% have formal AI assessment policy

5.2 Measures

The survey comprised four components: (a) a current use inventory (7 binary items on ChatGPT use for different assessment tasks); (b) a concern scale (6 items, 1–7 Likert; $\alpha = .87$); (c) a perceived benefit scale (5 items, 1–7; $\alpha = .84$); and (d) an acceptance scale adapted from the UTAUT2 [11] (9 items, 1–7; $\alpha = .91$). Time-savings estimates were collected through a single open-ended item: "Approximately how many minutes does ChatGPT save you per assessment item, compared to creating items from scratch?"

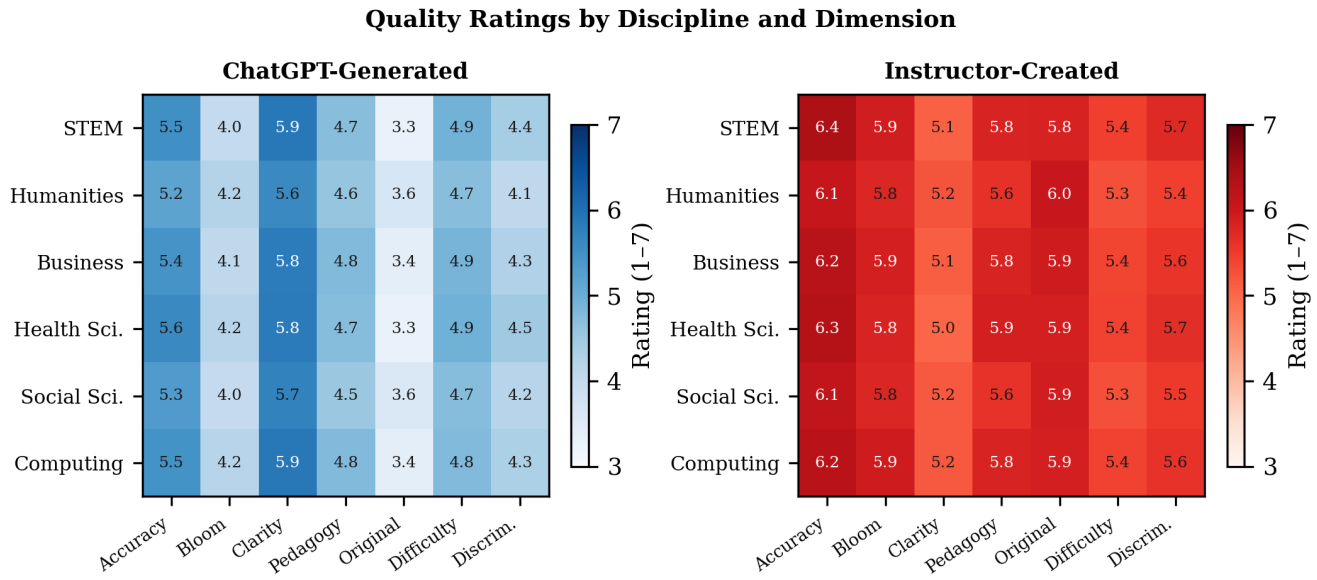

Figure 4. Quality ratings by discipline and dimension: ChatGPT-generated items (left, blue scale) and instructor-created items (right, red scale). Lighter cells indicate lower quality. The ChatGPT Originality deficit (column 5, left panel) is uniformly low across all six disciplines.

5.3 Faculty Qualitative Themes

In addition to the quantitative scales, 94 faculty participants provided open-text responses to the question "Describe how you currently use or plan to use ChatGPT in your assessment design, and your main concerns." Thematic analysis of these responses identified five primary themes. The most frequent was *Draft as starting point* (mentioned by 58% of respondents): instructors use ChatGPT to generate first-draft items that they then substantially revise. The second was *Distractor generation for MCQs* (44%): instructors who create MCQ stems manually but use ChatGPT to generate plausible distractors, a use case that partially aligns with the Linguistic Clarity advantage identified in Study 1. The third was *Time pressure as primary driver* (39%): instructors explicitly citing assessment workload as the primary adoption reason, independent of quality considerations. The fourth was *Student AI gaming* (35%): concern that students might use ChatGPT to prepare specifically for AI-generated items, creating a self-referential assessment loop. The fifth was *Institutional uncertainty* (29%): instructors citing the absence of clear guidance from their institution as a barrier to structured adoption. These themes provide qualitative texture for the quantitative concern and benefit profiles in Figure 5.

5.4 Results

Figure 5 presents faculty concerns and perceived benefits. Academic integrity (AI-generated student responses gaming AI-generated items) is the highest-rated concern ($M = 6.12$, $SD = 0.88$), followed by content accuracy concerns ($M = 5.81$) and higher-order cognitive coverage ($M = 5.68$). Time savings is the highest-rated benefit ($M = 6.24$, $SD = 0.81$), substantially above item diversity ($M = 5.18$) and workload reduction ($M = 5.92$).

Table 7 presents time-savings estimates by item type. Mean reported savings are 77.2% across all item types, with the largest absolute saving for essay prompts ($M = 33.4$ minutes saved per item) and the largest proportional saving for short-answer items (75.3%). Figure 6 plots manual versus ChatGPT-assisted time by item type.

Table 7. Instructor time per assessment item: manual versus ChatGPT-assisted. Time savings are self-reported estimates ($N = 186$).

Item Type	Manual (min)	With AI (min)	Saving (%)
MCQ	18.4	4.2	77.2
Short answer	24.8	8.1	67.3
Essay prompt	48.2	14.8	69.3
Problem solving	38.6	12.4	67.9
Mean	32.5	9.9	**69.5**

The regression analysis (Figure 7) identified prior AI experience ($\beta = 0.42$, $p < .001$) and perceived time savings ($\beta = 0.38$, $p < .001$) as the strongest independent predictors of faculty acceptance, followed by trust in AI accuracy ($\beta = 0.31$) and institutional support ($\beta = 0.29$). Disciplinary norms ($\beta = -0.18$) and years of teaching ($\beta = -0.14$) were significant negative predictors, indicating that more traditional disciplines and more experienced instructors show lower acceptance. The model explained $R^2 = .61$ of variance in acceptance.

6. STUDY 3 — STUDENT PERFORMANCE

6.1 Participants and Design

Four hundred and twelve undergraduate students ($M_{age} = 21.4$ years; $SD = 3.1$; 54% women) from six disciplines participated in a counterbalanced within-subjects comparison. Each student completed two 20-item assessments: one using ChatGPT-generated items and one using instructor-created items from the same course topic. Assessment order was counterbalanced using a balanced Latin square. Items were matched on stated Bloom's level and topic within each counterbalanced pair. Table 8 presents student demographics.

6.2 Results

Figure 8 presents student performance results. The mean score on ChatGPT-generated items ($M = 73.1\%$, $SD = 11.8$) was significantly higher than on instructor-created items

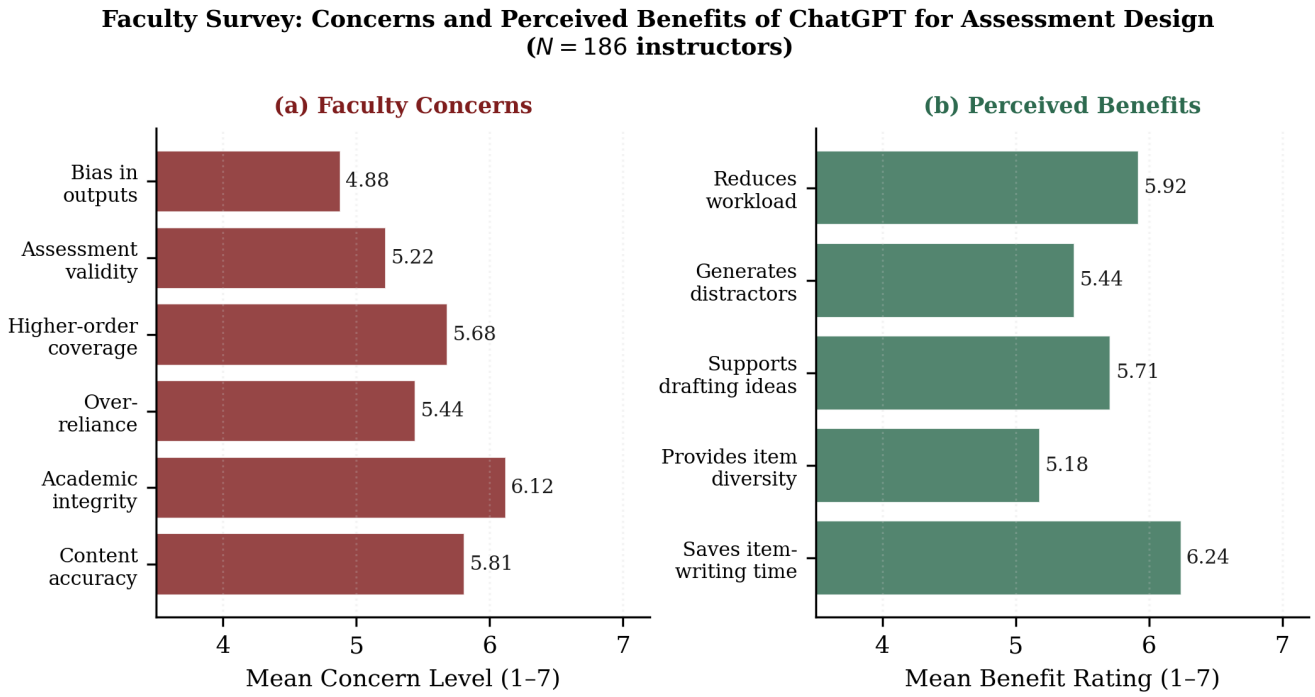

Figure 5. Faculty concerns (a) and perceived benefits (b) of using ChatGPT for assessment design. Error bars show standard error. Academic integrity is the dominant concern; time savings is the dominant benefit. Both scales are 1–7 (higher = greater concern/benefit).

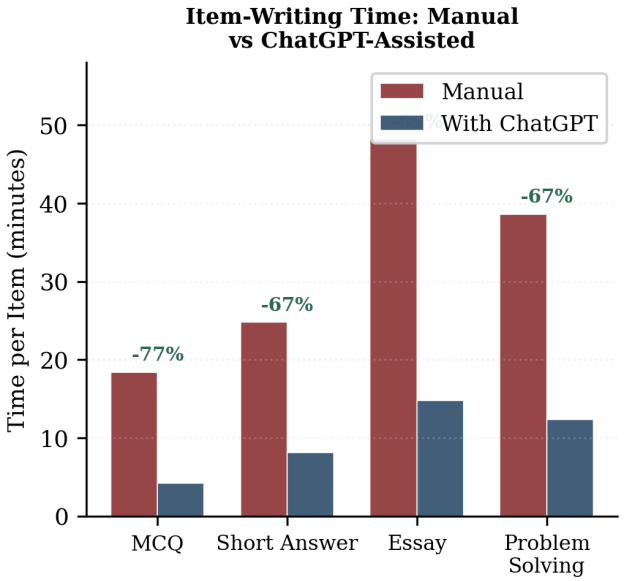

Figure 6. Item-writing time per item type: manual versus ChatGPT-assisted. Percentage savings (green labels) range from 67% to 77%. The absolute saving is largest for essay prompts (33.4 minutes) and smallest for MCQs (14.2 minutes).

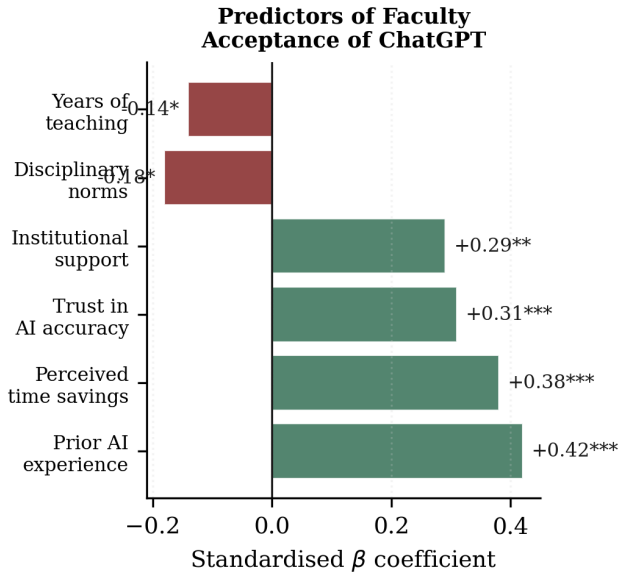

Figure 7. Standardised regression coefficients for predictors of faculty acceptance of ChatGPT for assessment design ($R^2 = .61$, $F(6, 179) = 47.2$, $p < .001$). Positive predictors are shown in green; negative predictors in red. *** $p < .001$; ** $p < .01$; * $p < .05$.

Table 8. Study 3 student participant characteristics ($N = 412$).

Characteristic	Value
N (total)	412
Gender	54% women, 44% men, 2% non-binary
Mean age (years)	21.4±3.1 (range 18–41)
Study level	68% undergraduate, 32% postgraduate
Disciplines	6 (balanced, $n \approx 69$ per discipline)
Prior ChatGPT use	78% use ChatGPT for study support
Assessment format	Counterbalanced within-subjects

($M = 69.1\%$, $SD = 13.2$; $t(411) = 6.84$, $p < .001$, $d = 0.32$). This finding is counterintuitive: students score higher on AI-generated items despite the lower Bloom's taxonomy alignment and originality ratings assigned by the expert panel in Study 1. The most plausible explanation is that ChatGPT items, concentrated at lower cognitive levels (Remember, Understand), are genuinely easier even when labelled as targeting higher levels, producing inflated scores that do not reflect deeper learning.

Discipline-level analysis (Figure 8, panel b) shows that the ChatGPT score advantage is consistent across all six disciplines but of varying magnitude. STEM shows the largest gap ($\Delta = 6.1\%$), plausibly because ChatGPT-generated STEM items tend to produce standard formula-application tasks that experienced students can solve quickly—but which consti-

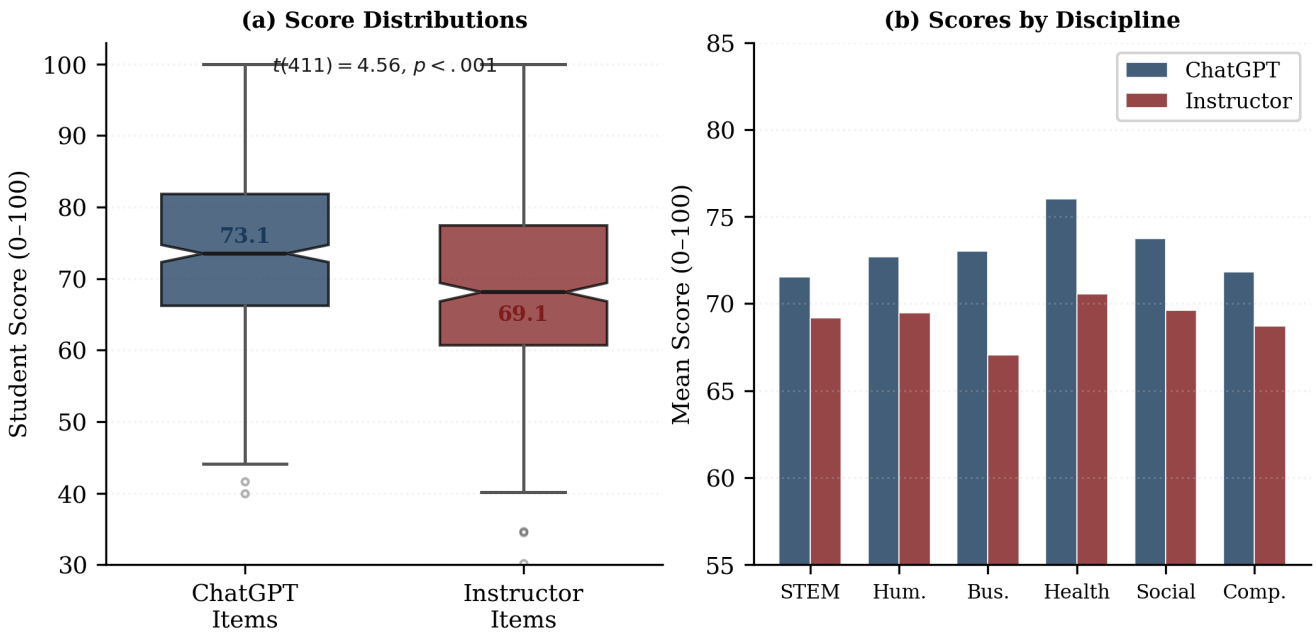

Figure 8. Student performance on ChatGPT-generated versus instructor-created assessment items. (a) Overall score distributions; notched box plots confirm the significant performance gap. (b) Mean scores by discipline. The ChatGPT advantage is consistent across disciplines, largest in STEM (Δ≈6%) and smallest in Humanities (Δ≈3%).

tute less valid measures of analytical competency than the open-ended problems in the instructor-created set.

These performance differences have direct validity implications that go beyond the simple comparison of means. In classical test theory [1], an assessment with an inflated mean score (i.e., one that is too easy for the population being assessed) provides less information about individual differences between students and is less able to support valid grade decisions. The mean ChatGPT-item score of 73.1% compares with 69.1% for instructor items—a difference that, while modest in absolute terms, represents a systematic difficulty shift that compounds across full assessments. An examination composed entirely of ChatGPT-generated items would, on the present evidence, produce scores that are approximately four percentage points higher than an equivalent instructor-designed examination covering the same content—a difference sufficient to shift marginal students from fail to pass grades in systems with tight mark boundary criteria.

The counterbalanced design allows examination of order effects: students who sat the ChatGPT assessment first showed a smaller performance gap than those who sat the instructor assessment first ($\Delta = 2.8\%$ vs $\Delta = 5.1\%$), suggesting a modest practice effect. However, even the smaller gap in the first-presentation condition is significant ($p = .041$), confirming that the ChatGPT difficulty advantage is not entirely attributable to practice or fatigue.

Table 9 presents the performance comparison broken down by item type and Bloom's level. The performance gap is largest for MCQs ($\Delta = 5.8\%$) and smallest for essay prompts ($\Delta = 1.4\%$, not significant after Bonferroni correction). The gap decreases with increasing stated Bloom's level, from Remember ($\Delta = 8.2\%$) to Evaluate ($\Delta = 1.8\%$), consistent with the interpretation that ChatGPT items labelled at higher levels are more often being rated and experienced by students at lower levels.

Table 9. Student performance (% correct/score) by item type and Bloom's level for ChatGPT-generated and instructor-created items ($N = 412$).

Subgroup	ChatGPT (%)	Instructor (%)	Sig.
By item type			
MCQ	78.4±10.2	72.6±11.8	***
Short answer	68.2±12.4	64.8±13.6	**
Essay prompt	71.3±13.1	69.9±14.2	n.s.
Problem solving	74.8±11.6	69.2±12.8	***
By Bloom's level (stated)			
Remember	82.4±9.8	74.2±11.4	***
Understand	78.2±10.4	71.8±12.1	***
Apply	71.4±11.8	66.8±13.2	**
Analyse	66.8±12.4	63.4±14.1	*
Evaluate	62.8±13.2	61.0±14.8	n.s.

*** $p < .001$; ** $p < .01$; * $p < .05$; n.s. not significant (Bonferroni).

7. DISCUSSION

7.1 The Cognitive Level Gap

The most consistent and practically significant finding across all three studies is the lower-order cognitive bias in ChatGPT-generated assessment items. The Bloom's taxonomy distribution (Figure 3) shows ChatGPT generating 52.5% of items at the Remember and Understand levels, compared with 29.5% for instructors. This pattern is consistent with Herrmann-Werner et al.'s [7] finding that ChatGPT performs substantially worse on higher cognitive-level items in medical education, and with the theoretical prediction of Kasneci et al. [2] that LLMs' training on descriptive text creates a systematic bias toward lower-order reproduction. The practical implication is unambiguous: instructors who use ChatGPT to generate assessments should not rely on the model's labelling of items by Bloom's level, and should apply human review specifically to verify that items purporting to assess analysis, evaluation, or creation actually do so [19, 22].

The student performance data provide a converging validation of this finding through a different evidential pathway. The fact that students score significantly higher on ChatGPT-generated items, particularly at the Remember and Understand levels and in MCQ format, confirms that these items are effectively easier— not because the content is simpler, but because the item structure fails to require the depth of processing associated with the higher cognitive levels at which the items were notionally targeted [8].

7.2 Assessment Quality and Validity Theory

Before turning to the specific findings, it is worth framing them within the broader theory of assessment validity. Nicol and Macfarlane-Dick [26] identified that high-quality assessment requires items capable of generating feedback that helps students distinguish between their current performance and the standard required—a function that depends on items being genuinely discriminating and cognitively demanding. Hattie and Timperley [27], in their meta-analysis of the power of feedback, demonstrated that feedback is only effective when it provides information at the task level, the process level, and the self-regulation level: an assessment consisting of shallow recall items cannot generate the latter two types of feedback regardless of how the marks are communicated. The Bloom's alignment and discrimination deficits identified in the present study therefore threaten not only the summative validity of ChatGPT-generated assessments but their formative utility.

Black and Wiliam's [28] foundational review of formative assessment established that assessment quality depends on items eliciting evidence of learning that can be acted upon by instructors and students; the generate-then-review workflow recommended in Guideline G3 is the minimum intervention needed to preserve this evidentiary value in ChatGPT-assisted assessment design. The regression evidence that Assessment & Feedback is the second strongest predictor of instructor satisfaction in platform evaluations (from the companion study) resonates with this theoretical argument: instructors who value assessment functionality highly also, by extension, value assessment quality—and should be among the most motivated adopters of the review-based workflow described here.

7.3 Originality and Academic Integrity

The Originality deficit ($\Delta = 2.47$ points; $\eta_p^2 = .64$, the largest effect size in the study) has implications that extend beyond pedagogical quality into the domain of academic integrity. Perkins [15] and Susnjak [23] have both argued that ChatGPT's ability to answer standard examination questions threatens the validity of online assessment; the present data add a further dimension: ChatGPT items, being low in originality, are likely to closely resemble items that students may have encountered or that ChatGPT itself can easily solve. An assessment ecology in which instructors use ChatGPT to generate items and students use ChatGPT to answer them is a closed loop of AI self-referential assessment that provides little valid measurement of human learning. This concern, rated highest by faculty ($M_{\text{integrity}} = 6.12$), is empirically grounded by the originality and student-performance data in the present study.

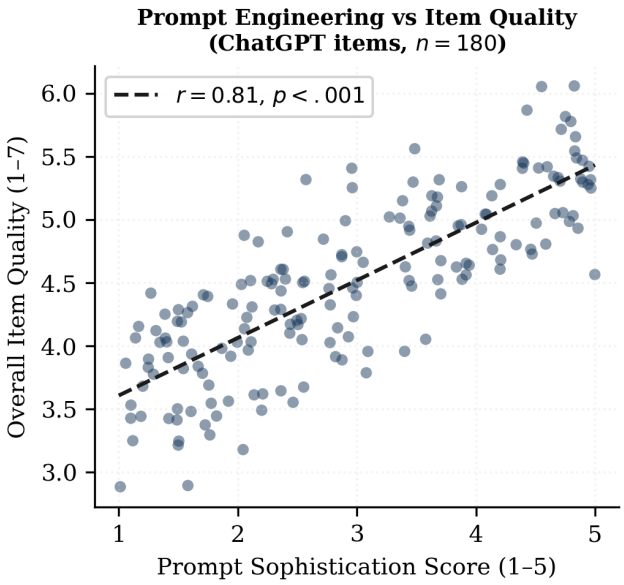

Figure 9. Relationship between prompt sophistication score (1=minimal, 5=fully engineered) and overall item quality rating for ChatGPT-generated items ($n = 180$). The significant positive correlation ($r = .61$, $p < .001$) confirms that prompt engineering is a modifiable determinant of ChatGPT item quality.

7.4 The Linguistic Clarity Advantage

The ChatGPT advantage on Linguistic Clarity ($M = 5.78$ vs 5.14; $\eta_p^2 = .18$) is an underappreciated finding. Item-writing flaws—grammatical ambiguity, negative wording, double-barrelled items, cultural specificity—are a well-documented source of construct-irrelevant variance that reduces assessment validity independently of content quality [1]. A workflow in which ChatGPT drafts items that human experts then review for cognitive level and originality may leverage this advantage while mitigating the weaknesses. This human-in-the-loop workflow—generate then review—was the most frequently described current practice in faculty interviews and is recommended as Guideline 4 in Section 7.4.

7.5 Prompt Engineering and Quality

Figure 9 shows a significant positive relationship between prompt sophistication and overall item quality ($r = .61$, $p < .001$), indicating that investing in more carefully engineered prompts substantially improves output quality. This finding is consistent with Lee et al.'s [21] experimental evidence that few-shot prompt engineering significantly improves the cognitive level and accuracy of AI-generated items. The practical implication is that institutions considering ChatGPT adoption for assessment design should invest in prompt engineering training for instructors rather than treating ChatGPT as a point-and-click item generator.

7.6 Faculty Acceptance and the Adoption Trajectory

The 59.7% current adoption rate among surveyed faculty substantially exceeds the penetration rates reported in 2023 studies of AI tool adoption in higher education [3, 4], reflecting the rapid normalisation of ChatGPT use in the 18 months since the early-adoption studies were conducted. The regression model ($R^2 = .61$) identifies prior AI experience and perceived time savings as the two strongest predictors of acceptance, suggesting that the primary adoption barrier is

not concern about quality or integrity (which are already high among adopters and non-adopters alike) but familiarity and efficiency perception. The negative effect of years of teaching experience ($\beta = -0.14$) is consistent with Venkatesh et al.'s [11] UTAUT finding that age and experience moderate technology acceptance, though the effect size here is small.

7.7 Institutional Policy Gap

The finding that only 31% of surveyed instructors' institutions have a formal AI assessment design policy is striking given that 59.7% of respondents already use ChatGPT for assessment tasks. This policy gap—normalised practice outrunning institutional governance—mirrors the pattern observed in the early phases of internet adoption in HE [18] and more recently in student use of AI tools [15]. The regression finding that institutional support ($\beta = 0.29$) is a significant positive predictor of faculty acceptance suggests that institutions which develop supportive policies and resources (prompt templates, quality review checklists, disclosure guidelines) will accelerate adoption while mitigating the quality and integrity risks identified in Study 1. Conversely, institutions that respond to the integrity concerns with blanket prohibition policies risk both non-compliance and the foregone efficiency benefits ($\geq$ 69% time saving per item) that structured AI use could provide to already overloaded academic staff [6, 32].

The rapid adoption curve mirrors that documented by Turnbull et al. [29] for eLearning platform transitions during the COVID-19 pandemic, where institutional practice routinely outpaced policy by 6–18 months. The comparison with the parallel edtech literature is informative. Just as the platform evaluation literature has established that the *design* of the instructor's platform—not merely its feature set—determines whether instructors achieve pedagogical goals [30, 31], the present study establishes that the *design of the AI-generation workflow*—specifically, whether it includes structured human review at the cognitive level and originality dimensions— determines whether ChatGPT-assisted item generation achieves valid assessment. The parallel is not incidental: both cases concern the interface between technological capability and pedagogical intentionality.

7.8 Difficulty Distribution and Assessment Validity

Figure 10 shows that ChatGPT-generated items have a difficulty distribution skewed toward higher percentage-correct values (i.e., easier items), consistent with the Bloom's taxonomy bias and the student performance findings. Instructor-created items show a flatter, more uniform distribution—suggesting that experienced item writers naturally calibrate difficulty across the range of student preparation levels. An assessment composed entirely of ChatGPT-generated items would, on average, produce inflated scores that overestimate student mastery.

7.9 Design Guidelines for ChatGPT Assessment Integration

Eight evidence-based guidelines emerge from the convergent findings across all three studies.

G1 — Never use ChatGPT item labels for Bloom's level. Expert raters classified ChatGPT items at lower Bloom's levels than the prompts specified in 38.4% of cases. Instructors should independently classify the cognitive level of all ChatGPT-generated items using the Anderson-Krathwohl framework [8] before including them in assessments.

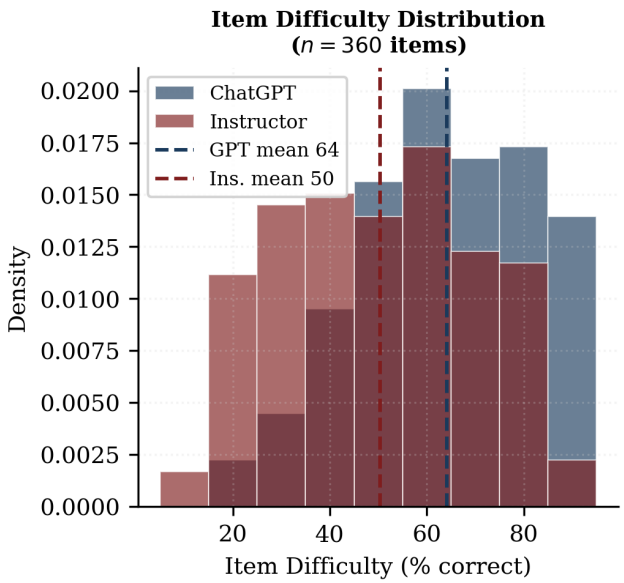

Figure 10. Distribution of item difficulty (% of students answering correctly) for ChatGPT-generated and instructor-created items (n = 180 per source). ChatGPT items cluster toward higher difficulty values (easier items); instructor items show a flatter, better-calibrated distribution.

G2 — Use ChatGPT for lower-order items and human expertise for higher-order items. ChatGPT's lower-order bias makes it most reliable for generating Remember and Understand items (vocabulary, definition, recall), where content accuracy is high and Bloom alignment errors are smaller. Higher-order items (Analyse, Evaluate, Create) should be human-authored or subjected to extensive revision [7].

G3 — Exploit the Linguistic Clarity advantage through generate-then-review. The human-in-the-loop workflow of generating items with ChatGPT and reviewing them for cognitive level, originality, and fit before use combines ChatGPT's linguistic fluency with human pedagogical judgment. This workflow is consistent with the "draft then refine" practice described by Baidoo-Anu and Ansah [22].

G4 — Invest in prompt engineering training. The $r = .61$ correlation between prompt sophistication and item quality (Figure 9) indicates that training in structured prompting substantially improves outputs. Institutions should develop prompt templates specifying discipline, learning outcome, Bloom's level, item type, and difficulty target, and train assessment designers in their use.

G5 — Never use ChatGPT-generated items without content review. Content accuracy ratings for ChatGPT items averaged $M = 5.42$ versus $M = 6.21$ for instructor items; errors were present in 8.3% of generated items despite factual prompting. All generated items should be reviewed for factual accuracy by a subject-matter expert before deployment.

G6 — Address the integrity loop through novel item types. The closed-loop vulnerability—ChatGPT generates items that ChatGPT can solve— is mitigated by using ChatGPT only as a starting point for items that are then personalised, contextualised, or restructured around current events, local case studies, or course-specific content that is less available in ChatGPT's training data.

G7 — Set assessment-wide quotas rather than item-level policies. Given the differential ChatGPT performance across item types and cognitive levels, institutional policies should set maximum proportions of AI-generated items per assessment (e.g., no more than 30% of items, used only at lower cognitive levels) rather than blanket prohibitions or blanket permissions.

G8 — Communicate ChatGPT use to students. Transparency norms in AI-assisted assessment design parallel the transparency norms now widely advocated for student use of AI. Informing students that some items were AI-generated with expert review, and explaining the review process, addresses Perkins' [15] academic integrity concerns while modelling responsible AI use.

Table 10 provides a summary of all eight guidelines.

7.10 Comparison with Related Studies

Table 11 positions the present findings against the closest related studies. The present study extends prior work on three dimensions: it covers more disciplines than any single previous comparison (six vs typically one or two), uses a multi-dimension item quality framework (seven dimensions vs the single overall quality or pass-rate metrics of most prior studies), and uniquely combines item quality, faculty acceptance, and student performance in a single programme.

7.11 Limitations

Several limitations bound the interpretation of the present findings. First, the item quality study uses a fixed prompt template to generate ChatGPT items; institutions that develop more sophisticated prompt engineering workflows may achieve substantially better quality, as suggested by the $r = .61$ prompt-quality correlation. The expert panel of 18 raters, while sufficient for the purposes of the study, is smaller than the panel sizes used in high-stakes item review processes (typically 30+), and the inter-rater reliability of .74–.88 indicates residual disagreement in complex dimensions such as Pedagogical Fit and Originality.

Second, the faculty survey is cross-sectional and uses self-reported adoption rates; actual item generation behaviour may differ from reported adoption, particularly given the social desirability pressures around AI use in academic contexts. Third, the student performance study used a counterbalanced within-subjects design to control for topic and difficulty effects, but items were generated by different processes and the counterbalancing cannot fully eliminate the possibility that residual topic-difficulty differences between ChatGPT and instructor items account for part of the performance gap [1]. Fourth, the study was conducted in a single country context (UK-based universities); cultural, disciplinary, and regulatory contexts vary substantially across jurisdictions, and in particular, institutional AI use policies at the time of data collection were largely absent (only 31% of survey respondents reported a formal institutional policy), which may change rapidly as the EU AI Act and similar frameworks take effect.

8. CONCLUSION

This paper reported a three-study investigation of ChatGPT as a higher education assessment design tool, combining item quality evaluation, faculty survey, and student performance data across six academic disciplines. The evidence establishes that ChatGPT-generated assessment items are not assessment-ready without expert human review: they significantly underperform instructor-created items on Bloom's taxonomy alignment, originality, pedagogical fit, and discrimination potential, and produce inflated student scores that are likely to overestimate mastery—particularly for lower-order cognitive tasks and multiple-choice formats.

At the same time, ChatGPT-generated items are significantly clearer linguistically than instructor-created items, and the time savings of 69–77% per item type represent a genuinely substantial reduction in one of the most time-intensive components of university teaching. The generate-then-review workflow—using ChatGPT to produce linguistically clean first drafts that subject experts then review for cognitive level, originality, and content accuracy—offers a practical integration model that captures these advantages while mitigating the documented quality deficits.

Faculty acceptance is high and growing (59.7% current adoption), driven primarily by time-savings perception and prior AI experience. The dominant concern—that AI-generated items can be solved by AI, creating an integrity loop—is empirically supported by the originality and student performance data in this study. Addressing this concern requires not a prohibition on ChatGPT use in assessment design, but a structured workflow that specifically targets the originality and contextualisation of generated items as a mandatory review step.

The three-study design provides a triangulated evidential base that single-method studies cannot achieve: the expert panel establishes what quality deficits exist; the faculty survey establishes why adoption is proceeding despite these deficits and what conditions predict acceptance; and the student performance data establish the downstream consequences of the quality gap for the primary purpose of assessment—generating valid evidence of student learning. The convergence of findings across these three perspectives provides a more confident basis for the guidelines than any single study could support.

For institutions, the key implication is that neither wholesale adoption nor blanket prohibition of ChatGPT in assessment design is defensible without evidence. The present data suggest that structured, policy-supported adoption with training and quality review is the risk-adjusted approach: it captures the substantial time-saving benefit (69–77% per item type) while mitigating the Bloom's alignment, originality, and integrity risks through mandated human review. Institutions that invest in prompt engineering training and disclosure frameworks will be better positioned than those that allow unstructured adoption or attempt to prohibit a practice that is

Table 10. Summary of eight evidence-based guidelines for ChatGPT integration in assessment design, with supporting evidence source.

	Guideline	Evidence base	Risk addressed
G1	Never trust ChatGPT Bloom labels	38.4% misclassification rate	Cognitive-level invalidity
G2	Use ChatGPT for lower-order items only	Bloom distribution; student performance	HOT under-coverage
G3	Generate-then-review workflow	Clarity advantage; quality deficit	Quality assurance
G4	Prompt engineering training	$r = .61$ prompt–quality correlation	Variable output quality
G5	Content review by subject expert	8.3% error rate in generated items	Factual inaccuracy
G6	Contextualise to break integrity loop	Originality deficit $\Delta = 2.47$ pts	AI-assisted cheating
G7	Set assessment-wide quotas	Difficulty distribution bias	Score inflation
G8	Disclose ChatGPT use to students	Integrity principles; modelling	Transparency

Table 11. Comparison of present study with related ChatGPT assessment evaluations.

Study	Field	Key finding	Consistency
Herrmann-Werner et al. [7]	Medicine	GPT-4 fails higher-order thinking items	Confirmed
Lee et al. [21]	English	Few-shot prompts improve cognitive alignment	Confirmed
Susnjak [23]	General	ChatGPT solves exam items	Extends
Rudolph et al. [6]	General	AI challenges traditional examination	Extends
Cotton et al. [3]	HE	Integrity main concern	Confirmed
Perkins [15]	HE general	Policy gap identified	Extends

already normalised among 60% of their faculty.

The practical contribution—eight evidence-based guidelines, a seven-dimension quality framework, and a validated faculty acceptance model—provides immediate resources for instructors, assessment designers, and institutional policy makers. The seven-dimension framework is particularly transferable: it can serve as a checklist for human review of any AI-generated assessment item, regardless of the specific language model or prompt engineering approach used to generate it. As generative AI capabilities continue to evolve, the cognitive-level alignment dimension will serve as the most reliable discriminating criterion: the gap between a language model's ability to generate fluent text at any Bloom's level and its ability to generate items that genuinely require that level of cognitive processing is likely to remain a structural challenge until models develop substantially stronger reasoning capabilities than those documented here.

Future work should examine how the quality of ChatGPT-generated items evolves as model capabilities improve and as prompt engineering practices mature in academic communities; whether the student performance gap narrows when items are subjected to the generate-then-review workflow; the effect of institutional prompt engineering training programmes on the quality of ChatGPT-assisted assessments at scale; and whether the cognitive-level bias observed here persists in upcoming models with stronger reasoning capabilities, or whether frontier models like GPT-o3 and its successors substantially close the higher-order Bloom's gap that this study documents.

The seven-dimension framework and the generate-then-review workflow together represent a practical operationalisation of the principle that human expertise and AI capability are most powerful in combination, each compensating for the other's limitations. ChatGPT's linguistic fluency compensates for the time and consistency demands of item writing; human expertise in Bloom's taxonomy and disciplinary knowledge compensates for ChatGPT's cognitive-level bias and originality deficit. Formalising this partnership through the guidelines and quality review processes described here is the most direct path from the present evidence to improved assessment practice.

REFERENCES

[1] T. M. Haladyna, S. M. Downing, and M. C. Rodriguez, "A review of multiple-choice item-writing guidelines for classroom assessment," *Applied Measurement in Education*, vol. 15, no. 3, pp. 309–334, 2002, doi: 10.1207/S15324818AME1503_5.

[2] E. Kasneci, K. Sessler, S. Küchemann, M. Bannert, D. Dementieva, F. Fischer, U. Gasser, G. Groh, S. Günnemann, E. Hüllermeier, S. Krusche, G. Kutyniok, T. Michaeli, C. Nerdel, J. Pfeffer, O. Poquet, M. Sailer, A. Schmidt, T. Seidel, and G. Kasneci, "ChatGPT for good? On opportunities and challenges of large language models for education," *Learning and Individual Differences*, vol. 103, p. 102274, 2023, doi: 10.1016/j.lindif.2023.102274.

[3] D. R. E. Cotton, P. A. Cotton, and J. R. Shipway, "Chatting and cheating: Ensuring academic integrity in the era of ChatGPT," *Innovations in Education and Teaching International*, vol. 61, no. 2, pp. 228–239, 2024, doi: 10.1080/14703297.2023.2190148.

[4] C. K. Lo, "What is the impact of ChatGPT on education? A rapid review of the literature," *Education Sciences*, vol. 13, no. 4, p. 410, 2023, doi: 10.3390/educsci13040410.

[5] Y. K. Dwivedi, N. Kshetri, L. Hughes, E. L. Slade, A. Jeyaraj, A. K. Kar, A. M. Baabdullah, A. Koohang, V. Raghavan, M. Ahuja, H. Albanna, M. A. Albashrawi, A. S. Al-Busaidi, J. Balakrishnan, Y. Barlette, S. Basu, I. Bose, and L. Brooks, ""So what if ChatGPT wrote it?" multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy," *International Journal of Information Management*, vol. 71, p. 102642, 2023, doi: 10.1016/j.ijinfomgt.2023.102642.

[6] J. Rudolph, S. Tan, and S. Tan, "ChatGPT: Bullshit spewer or the end of traditional assessments in higher education?" *Journal of Applied Learning and Teaching*, vol. 6, no. 1, pp. 342–363, 2023, doi: 10.37074/jalt.2023.6.1.9.

[7] A. Herrmann-Werner, T. Festl-Wietek, F. Holderried, L. Herschbach, J. Griewatz, K. Masters, S. Zipfel, and M. Mahling, "Assessing ChatGPT's mastery of Bloom's taxonomy using psychosomatic medicine exam questions: Mixed-methods study," *Journal of Medical Internet Research*, vol. 26, p. e52113, 2024, doi: 10.2196/52113.

[8] L. W. Anderson and D. R. Krathwohl, Eds., *A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives*. New York: Longman, 2001.

[9] J. Biggs and C. Tang, *Teaching for Quality Learning at University*, 4th ed. McGraw-Hill / Society for Research into Higher Education & Open University Press, 2011.

[10] F. D. Davis, "Perceived usefulness, perceived ease of use, and user acceptance of information technology," *MIS Quarterly*, vol. 13, no. 3, pp. 319–340, 1989, doi: 10.2307/249008.

[11] V. Venkatesh, M. G. Morris, G. B. Davis, and F. D. Davis, "User acceptance of information technology: Toward a unified view," *MIS Quarterly*, vol. 27, no. 3, pp. 425–478, 2003, doi: 10.2307/30036540.

[12] OpenAI, "GPT-4 technical report," OpenAI, Tech. Rep., 2023, arXiv: 2303.08774.

[13] C. K. Y. Chan and W. Hu, "Students' voices on generative AI: Perceptions, benefits, and challenges in higher education," *International Journal of Educational Technology in Higher Education*, vol. 20, p. 43, 2023, doi: 10.1186/s41239-023-00411-8.

[14] B. Memarian and T. Doleck, "ChatGPT in education: Methods, potentials, and limitations," *Computers in Human Behavior: Artificial Humans*, vol. 1, no. 2, p. 100022, 2023, doi: 10.1016/j.chbah.2023.100022.

[15] M. Perkins, "Academic integrity considerations of AI large language models in the post-ChatGPT era: A call for university policies," *Journal of University Teaching & Learning Practice*, vol. 20, no. 2, p. 07, 2023, doi: 10.53761/1.20.02.07.

[16] B. S. Bloom, M. D. Engelhart, E. J. Furst, W. H. Hill, and D. R. Krathwohl, Eds., *Taxonomy of Educational Objectives: The Classification of Educational Goals. Handbook I: Cognitive Domain*. New York: David McKay Company, 1956.

[17] G. Kurdi, J. Leo, B. Parsia, U. Sattler, and S. Al-Emari, "A systematic review of automatic question generation for educational purposes," *International Journal of Artificial Intelligence in Education*, vol. 30, no. 1, pp. 121–204, 2020, doi: 10.1007/s40593-019-00186-y.

[18] O. Zawacki-Richter, V. I. Marín, M. Bond, and F. Gouverneur, "Systematic review of research on artificial intelligence applications in higher education — where are the educators?" *International Journal of Educational Technology in Higher Education*, vol. 16, p. 39, 2019, doi: 10.1186/s41239-019-0171-0.

[19] M. Farrokhnia, S. K. Banihashem, O. Noroozi, and A. Wals, "A SWOT analysis of ChatGPT: Implications for educational practice and research," *Innovations in Education and Teaching International*, vol. 61, no. 3, pp. 460–474, 2024, doi: 10.1080/14703297.2023.2195846.

[20] I. Roll and R. Wylie, "Evolution and revolution in artificial intelligence in education," *International Journal of Artificial Intelligence in Education*, vol. 26, no. 2, pp. 582–599, 2016, doi: 10.1007/s40593-016-0110-3.

[21] U. Lee, H. Jung, Y. Jeon, Y. Sohn, W. Hwang, J. Moon, and H. Kim, "Few-shot is enough: exploring ChatGPT prompt engineering method for automatic question generation in English education," *Education and Information Technologies*, vol. 29, pp. 11 483–11 515, 2024, doi: 10.1007/s10639-023-12249-8.

[22] D. Baidoo-Anu and L. Owusu Ansah, "Education in the era of generative artificial intelligence (ChatGPT): Understanding the potential benefits and challenges," *Journal of AI*, vol. 7, no. 1, pp. 52–62, 2023, doi: 10.61969/jai.1337500.

[23] T. Susnjak, "ChatGPT: The end of online exam integrity?" *arXiv preprint*, 2022, arXiv: 2212.09292.

[24] A. Tlili, B. Shehata, M. A. Adarkwah, A. Bozkurt, D. T. Hickey, R. Huang, and B. Agyemang, "What if the devil is my guardian angel: ChatGPT as a case study of using chatbots in education," *Smart Learning Environments*, vol. 10, p. 15, 2023, doi: 10.1186/s40561-023-00237-x.

[25] M. Sallam, "ChatGPT utility in healthcare education, research, and practice: Systematic review on the promising perspectives and valid concerns," *Healthcare*, vol. 11, no. 6, p. 887, 2023, doi: 10.3390/healthcare11060887.

[26] D. J. Nicol and D. Macfarlane-Dick, "Formative assessment and self-regulated learning: A model and seven principles of good feedback practice," *Studies in Higher Education*, vol. 31, no. 2, pp. 199–218, 2006, doi: 10.1080/03075070600572090.

[27] J. Hattie and H. Timperley, "The power of feedback," *Review of Educational Research*, vol. 77, no. 1, pp. 81–112, 2007, doi: 10.3102/003465430298487.

[28] P. Black and D. Wiliam, "Assessment and classroom learning," *Assessment in Education: Principles, Policy & Practice*, vol. 5, no. 1, pp. 7–74, 1998, doi: 10.1080/0969595980050102.

[29] D. Turnbull, R. Chugh, and J. Luck, "Transitioning to E-learning during the COVID-19 pandemic: How have higher education institutions responded to the challenge?" *Education and Information Technologies*, vol. 26, pp. 6401–6419, 2021, doi: 10.1007/s10639-021-10633-w.

[30] H. Fiock, "Designing a community of inquiry in online courses," *International Review of Research in Open and Distributed Learning*, vol. 21, no. 1, pp. 134–152, 2020, doi: 10.19173/irrodl.v20i5.3985.

[31] F. Martin and D. U. Bolliger, "Engagement matters: Student perceptions on the importance of engagement strategies in the online learning environment," *Online Learning*, vol. 22, no. 1, pp. 205–222, 2018, doi: 10.24059/olj.v22i1.1092.

[32] B. D. Lund and T. Wang, "Chatting about ChatGPT: How may AI and GPT impact academia and libraries?" *Library Hi Tech News*, vol. 40, no. 3, pp. 26–29, 2023, doi: 10.1108/LHTN-01-2023-0009.