frontiers in Psychology

TYPE Original Research
PUBLISHED 25 November 2026
DOI 10.3389/fpsyg.2026.2005055

Check for updates

OPEN ACCESS

EDITED BY
Daniel H. Robinson,
The University of Texas at Arlington College of Education, United States

REVIEWED BY
Jesús Catherine Saldaña Bocanegra, Universidad César Vallejo, Peru
Le Thi Kim Hoa, Industrial University of Ho Chi Minh City, Vietnam

*CORRESPONDENCE
Xuan Tang
✉ tangxuan2021@gmail.com

RECEIVED 21 September 2026
REVISED 16 October 2026
ACCEPTED 21 October 2026
PUBLISHED 25 November 2026

CITATION
Tang X, Zeng M, He Z and Zeng W (2026) Acceptance beliefs predict undergraduates' self-assessed research competency but not their instructor-assessed coursework. *Front. Psychol.* 17:2005055. doi: 10.3389/fpsyg.2026.2005055

COPYRIGHT
© 2026 Tang, Zeng, He and Zeng. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Acceptance beliefs predict undergraduates' self-assessed research competency but not their instructor-assessed coursework

Xuan Tang[1]*, Mianmin Zeng[2], Ziyi He[3] and Weimei Zeng[4]

[1]School of Management, Guangzhou Huashang College, Guangzhou, China, [2]School of Public Administration, Nanfang College, Guangzhou, China, [3]School of Political Science and Law, Zhaoqing University, Zhaoqing, China, [4]School of Digital Intelligence and Management, Guangzhou Xinhua University, Guangzhou, China

Introduction: Self-assessment is central to self-regulated learning, yet people's insight into their own ability is moderate at best, and generative artificial intelligence (GenAI) may weaken it further by making work feel fluent whatever the learner has mastered. This study asked how closely undergraduates' ratings of their own research competency correspond to instructors' assessment of their coursework and which beliefs go with each.

Methods: Over one semester, management undergraduates at four institutions in Guangdong, China, completed two research assignments scored by their instructors on a common four-dimension rubric. Intact classes were assigned, without randomisation, to structured GenAI guidance (a guidebook, tutorials, demonstrations, and briefs specifying which steps AI could assist with) or to a comparison condition with the same assignments and conventional guidance; GenAI tools were available to both. Rubric scores covered complete class rosters (n = 284), and the questionnaire measures a linked pre-post panel (n = 169).

Results: The 19-item self-rating was dominated by a general factor, and its correlation with instructor-assessed performance was small (r = 0.04, 95% CI [−0.12, 0.21]; the interval excludes moderate but not small positive correlations), including when content areas were matched. The hypothesised association between change in perceived usefulness and the self-rating did not survive conservative record linkage and is treated as unsupported. In exploratory analyses, post-test perceived usefulness, ease of use, and attitude were positively associated with the self-rating (b = 0.38 to 0.44, p < 0.001) in both conditions and after weighting for attrition, and none showed a significant positive association with instructor-assessed performance. In all three institutions with usable rubric scores, the class given structured guidance scored higher on the second assignment, by 0.57 points on the 16-point rubric when pooled (d = 0.24), most clearly on analytical depth.

Discussion: With six non-randomised classes, non-blind single-rater scoring, and a covariate measured after the guidance was issued, this contrast is descriptive, not causal. The two outcomes are not equivalent instruments: one is a global self-perception, and the other is an evaluation of AI-assisted work. Within those limits, students' self-ratings went with their beliefs about the tools and could not stand in for the instructors' assessment of their coursework.

KEYWORDS
research competency, rubric-based assessment, self-assessment, self-assessment accuracy, self-regulated learning, technology acceptance

1 Introduction

Learners are indifferent judges of their own ability. Synthesising 22 meta-analyses across many domains, Zell and Krizan (2014) concluded that people's insight into how well they perform is real but moderate, and research on self-regulated learning shows that the error is systematic rather than random: the fluency of an experience is read as evidence of learning, so conditions that make study feel easy raise confidence while doing little for retention (Bjork et al., 2013), and those least able to perform a task are often least able to recognise it (Dunning, 2011). The gap is important for instruction and not only for measurement because what students believe they can already do governs what they choose to practise next (Panadero et al., 2016).

Generative artificial intelligence (GenAI) is, from this angle, an unusually strong manipulation of how competent work feels. A model returns organised, confident prose in seconds, and the student who prompted it experiences a hard problem as solved. Its reach is no longer in doubt: 602 million GenAI users in mainland China by the end of 2025 (China Internet Network Information Center, 2026), and 94% of 1,054 full-time United Kingdom undergraduates surveyed in December 2025 using GenAI for assessed work (Stephenson and Armstrong, 2026). The question this study asks is therefore not whether students accept these tools but whether their rating of their own research competency keeps any relation to how an instructor assesses the coursework they produce with those tools to hand.

Two bodies of work speak to this question, and they largely talk past each other. The first is an effects literature of experiments and quasi-experiments whose aggregate verdict is positive but strikingly unstable. Pooling 37 experimental studies, Liu et al. (2025) reported a moderate advantage for ChatGPT-supported learning on academic achievement ($g = 0.577$, 95% CI [0.395, 0.759]), and Deng et al. (2025) reached a broadly similar conclusion from a separate synthesis. Yet a meta-analysis restricted to higher education reported a much larger pooled effect on intellectual outcomes whose interval almost touched zero ($g = 1.096$, 95% CI [0.087, 2.104]), a spread that says more about heterogeneity across tasks, durations, and outcome measures than about any stable average (Liu et al., 2026).

The second is acceptance literature, most of it built on the technology acceptance model (TAM; Davis, 1989). Studies in this tradition report consistently that students find GenAI useful and easy to use, that these perceptions predict favourable attitudes, and that attitudes predict continued use (Dahri et al., 2024; Xu and Zhang, 2026). Reviewing 54 empirical studies of ChatGPT acceptance in higher education, Yang (2026) found the model's core structure reproduced across settings and closed by calling for longitudinal designs and more diverse contexts.

The gap between the two literatures is partly a measurement problem. A large share of acceptance studies, including those that place a competency construct at the end of the model, measure that competency with a self-report scale administered at a single time point to the same respondents who supplied the acceptance ratings. Zhu and Yang (2026), for instance, found that three distinct modes of GenAI use each predict research competence among Chinese graduate students, with critical thinking as a mediator; competence is measured by self-report. When the dependent variable is a student's own estimate of their ability, a positive path is compatible with two quite different stories. The intervention may have built competence. Or it may have built confidence, and confidence is what the scale records.

This ambiguity would be a minor nuisance if there were no reason to expect the two to diverge. It is present in this domain, and the evidence has become specific. Fernandes et al. (2026) gave reasoning problems to samples of 246 and 452 adults and found that participants working with ChatGPT expected to have solved approximately 16.5 of 20 items when they had in fact solved approximately 13; more revealing still, the usual gradient by which stronger performers judge themselves more accurately flattened almost completely under AI assistance so that miscalibration became uniform across skill levels. Lee et al. (2025) documented the behavioural counterpart among knowledge workers, for whom higher confidence in GenAI accompanied reduced critical effort of their own, and Fan et al. (2025) found that GenAI support raised students' essay scores without improving knowledge gain or transfer. Bastani et al. (2024) reported the sharpest version: secondary students given unrestricted access to a GPT-4 assistant improved substantially while it was available and then performed worse than controls on an unassisted examination, an effect that disappeared once the assistant was rebuilt to withhold answers and prompt instead. Zhu et al. (2026) showed why students are poorly placed to notice any of this: dependent and autonomous modes of offloading felt the same in the moment while diverging in their downstream correlates, so self-report cannot separate them.

Taken together, these findings imply that a student's own rating of their research competency is not a safe proxy for how an assessor would judge the work that student hands in. Testing the correspondence requires carrying two outcomes at once through the same students: what the student says they can do and what an instructor, working from a rubric and a written artefact, judges the work to be. Studies that do this are rare. Li et al. (2025) come closest, with a quasi-experiment on 366 Chinese undergraduates showing that ChatGPT-integrated instruction improved research skills, although the mechanisms were traced through self-reported motivation and self-directed learning. What is still missing is a design in which the belief is measured before and after a real semester of use, the work is scored by instructors against a common rubric, and the two sit side by side in the same sample.

The present study was built around that comparison. Over one semester, undergraduates in management programmes at four institutions in Guangdong Province completed two research assignments, an annotated bibliography and a mini-research proposal, both scored by their instructors on the same four-dimension rubric. Intact classes were assigned either to a structured GenAI guidance condition, which received a guidebook, video tutorials, in-class demonstrations, and briefs specifying which steps AI could assist with, or to a comparison condition with the same assignments and conventional research-methods guidance. GenAI tools were available to students in both conditions, so the contrast is one of guidance and not of access. Acceptance beliefs were measured at both waves: AI literacy at the first and self-rated research competency at the second.

The study makes three contributions to work on self-assessment in learning. It measures the correspondence between a self-rated competency scale and instructor-assessed performance in the same students and the same term, where acceptance research and self-assessment research alike usually have only one side. It examines what the self-rating is made of by asking which beliefs predict it and how it behaves as an instrument. It tests those same beliefs against an outcome the student does not supply so that the divergence can be located rather than assumed. A secondary aim, addressed with the coursework data, is whether classes given structured guidance on GenAI use are assessed differently from classes taught as usual and, if so, on which parts of the work.

Three terms are used in fixed senses. Self-assessment is the general process by which learners judge their own ability or work; self-rated research competency is the score on the 19-item scale used here, read as a global self-perception; and instructor-assessed performance is the rubric score given by the course instructor to coursework completed with GenAI tools available. The two outcomes are not alternative measures of one construct, and the second is not a measure of unaided ability.

2 Theoretical background and hypotheses

The hypotheses are presented in the order in which they were specified before estimation, which runs from the acceptance structure and the instructional contrast to the question that motivates the study.

2.1 Perceived usefulness, perceived ease of use, and attitude toward GenAI use

TAM holds that two beliefs govern the uptake of a technology: perceived usefulness, the belief that using it improves performance, and perceived ease of use, the belief that using it demands little effort. Both shape attitude toward use, which, in turn, drives intention and behaviour (Davis, 1989). Dahri et al. (2024) found perceived usefulness to be the strongest predictor of attitude toward ChatGPT among pre-service teachers, with attitude in turn a strong predictor of intention, and Yang's (2026) review of 54 studies reports the same core structure across settings and instruments.

The present setting makes the model worth re-examining. Almost all the available evidence is cross-sectional, often collected soon after students first meet the tool, yet beliefs formed at first contact need not survive a semester of demanding use. The evidence is also drawn overwhelmingly from ChatGPT and comparable Western systems, whereas Chinese undergraduates work with domestic models under different conditions of access, cost, and institutional legitimacy; Xu and Zhang (2026) found that in this setting, habit outweighed the classical determinants, which raises the question of whether the belief structure itself remains stable as use routinises. If the TAM structure reproduces at both waves here, it provides a stable measurement platform for the questions that follow.

> *H1a*. Perceived usefulness is positively associated with attitude toward using domestic GenAI tools for academic research, at both waves.

> *H1b*. Perceived ease of use is positively associated with attitude toward using domestic GenAI tools for academic research, at both waves.

2.2 Structured GenAI guidance and instructor-assessed coursework performance

Research competency is not one skill. Undergraduate research work involves locating and screening sources, appraising them critically, formulating a researchable question and matching a design to it, handling data, and writing the result in disciplinary form (Li et al., 2025). These components differ in how much of the cognitive work a language model can absorb, which is one reason aggregate effect sizes across syntheses range so widely (Deng et al., 2025; Liu et al., 2025; Liu et al., 2026).

The instructional argument for guiding students in their use of GenAI, as opposed to leaving them to it, is that the tool lowers the cost of the iteration on which research training depends. A model that responds in seconds, can be asked to argue against a design choice, and can produce alternative framings makes that iteration available in a way that a lecturer teaching ninety students cannot. Li et al. (2025) observed this pattern, with the caveat that benefit depended on autonomous motivation and self-directed learning; access alone was not sufficient. The moderator profile reported by Liu et al. (2025) points the same way, since the strongest effects appeared in sustained interventions of 5 to 10 weeks rather than in brief exposures.

The counter-evidence reviewed above cuts the other way, and its common thread is substitution: if the model performs the step, the student does not practise the step. The two arguments differ on whether the intervention constrains substitution, and the disappearance of the harm in Bastani et al. (2024) once the assistant was redesigned suggests that this constraint is the operative variable. The intervention evaluated here was built to impose it by requiring verification of every source against a real database, by reserving synthesis, design justification, and limitation identification as the student's own work, and by requiring disclosure of AI use at each stage. Under those conditions, we expect a positive but modest difference in instructor-assessed performance in favour of the guided classes.

H2. Controlling for performance on the first assignment, students in the structured GenAI guidance condition score higher on the second assignment than students in the comparison condition.

2.3 Task type as a condition: where a difference in assessed performance should appear

Salomon et al. (1991) drew a durable distinction. The *effects with* a technology are what a person can accomplish while working in partnership with it; the *effects of* a technology are what remains once the partnership ends. Residual change, on their account, comes from mindful engagement in which the learner keeps directing the work, not from the technology working on the learner's behalf.

Zhu et al. (2026) separate dependent offloading, in which the user delegates the thinking and treats the output as finished, from autonomous offloading, in which the user treats the output as a starting point and retains ownership of the judgement. Across three waves with 589 respondents, the two modes produced comparable immediate gains in efficiency but diverged in their downstream correlates, with dependent offloading associated with transferred cognitive agency and reduced intrinsic motivation. Reviews of over-reliance on AI dialogue systems converge on the same distinction while noting how thin the causal evidence remains (Zhai et al., 2024).

Applied to a student research project, the two most offloadable steps are also the two most visible. Generating search terms and producing polished academic prose can both be handed to a model almost in full, and a student who does so obtains a better-looking bibliography and a better-written paragraph without having exercised judgement; these invite dependent offloading. Deciding what five heterogeneous studies jointly imply, locating the gap a question should address, and justifying why a survey fits that question better than an interview study are steps a model can inform but cannot settle. Here,

the student must keep steering, and here practice and residual change should concentrate.

> *H3*. The effect of the intervention on instructor-assessed performance is concentrated in the higher-order rubric dimensions, analytical depth and research design coherence, rather than in information retrieval and academic writing.

2.4 Acceptance, self-rated competency, and instructor-assessed performance: one belief, two outcomes

Acceptance beliefs and self-rated competency are formed in the same episode of use and from the same evidence. A student who asks a model to compare three sampling strategies and receives a clear, well-organised answer experiences the tool as useful and experiences the problem as solved. Both impressions derive from the fluency of the output, which is the cue that the self-regulated learning literature identifies as the most common source of misplaced confidence (Bjork et al., 2013). Instructor scores are formed outside that episode, from a rubric applied to a written artefact by someone who was not present for the fluent exchange. The two are therefore not two measures of one thing but two judgements built from different evidence, and they do not have the same object either: the self-rating concerns research competency in general, and the rubric is one piece of coursework written with AI assistance available. Their correspondence is an empirical question rather than a psychometric assumption (Panadero et al., 2016).

The evidence set out in Section 1 points the same way: AI assistance inflates self-assessment and flattens its calibration (Fernandes et al., 2026), confidence in the tool travels with reduced critical effort (Lee et al., 2025), and the immediate experience of adaptive and maladaptive use is indistinguishable (Zhu et al., 2026). On this evidence, a student's judgement of their own competency may owe more to the felt quality of the interaction than to anything an assessor would later see in the work.

Movements in acceptance should register on the self-report side and should not, on their own, register on the instructor-assessed side once prior performance is held constant.

> *H4a*. Change in perceived usefulness across the semester is positively associated with self-rated research competency at the end of the semester.
>
> *H4b*. Change in perceived usefulness is not associated with instructor-assessed coursework performance once prior performance and institution are controlled.
>
> *H4c*. Self-rated research competency and instructor-assessed performance are only weakly related.
>
> *H4b* is stated as a null prediction deliberately, and the evidential limits of doing so are addressed in the analytic strategy below.

2.5 AI literacy as a moderator of the difference between conditions

If skill in working with a model governs what a student extracts from it, students should not benefit equally from identical instruction. Ng et al. (2021) conceptualised it as knowing what AI is, using it, evaluating its output, and recognising its ethical implications, and subsequent scale development has operationalised it for non-specialists (Carolus et al., 2023; Laupichler et al., 2023).

Recent evidence is consistent with a conditioning role, although it has been obtained on self-reported outcomes. Wang (2026), surveying 744 Chinese university students, found that GenAI literacy predicted academic achievement through academic self-efficacy, with the effect amplified among students higher in critical thinking. Whether the same conditioning appears on an outcome the student does not report is untested, and it is the version of the claim that is important for instructional design.

> *H5*. Baseline AI literacy moderates the effect of the intervention on (a) instructor-assessed coursework performance and (b) self-rated research competency, such that the effect is stronger at higher levels of AI literacy.

2.6 Conceptual framework

The framework, shown in Figure 1, places two outcomes in parallel rather than in sequence. A training path runs from the assigned condition to instructor-assessed coursework performance and is expected to concentrate in the higher-order rubric dimensions (H2, H3). Within the acceptance block, perceived usefulness and perceived ease of use predict attitude toward use (H1a, H1b); attitude is an endpoint inside that block, not a step on the way to the self-report outcome. A separate belief path runs from acceptance directly to self-rated research competency (H4a). The assigned condition also enters the models of that outcome without a directional hypothesis, and baseline AI literacy is tested as a moderator of the difference between conditions on each outcome (H5a, H5b). The two outcomes are joined only by the predictions that acceptance does not reach the instructor-assessed outcome once prior performance is controlled and that the two outcomes correlate weakly with one another (H4b, H4c).

3 Materials and methods

3.1 Design and setting

The study used a two-wave quasi-experimental design with cluster assignment at the class level and two parallel outcomes, conducted during the second semester of the 2025–2026 academic year at four undergraduate institutions in Guangdong Province, China. Institutions are identified by code throughout. Participants were undergraduates in management programmes, principally human resource management, public administration, and tourism management.

Sampling was criterion-based and purposive, since the design required an instructor at each site willing to deliver a prescribed set of materials and to score two assignments on a common rubric. Sites had to offer an undergraduate management programme with an applied research or research-methods component in the spring semester, institutional permission for anonymous questionnaire data collection, and enrolment sufficient for at least two intact classes.

The four participating institutions span a range of entry standards and programme sizes while holding context constant. All are

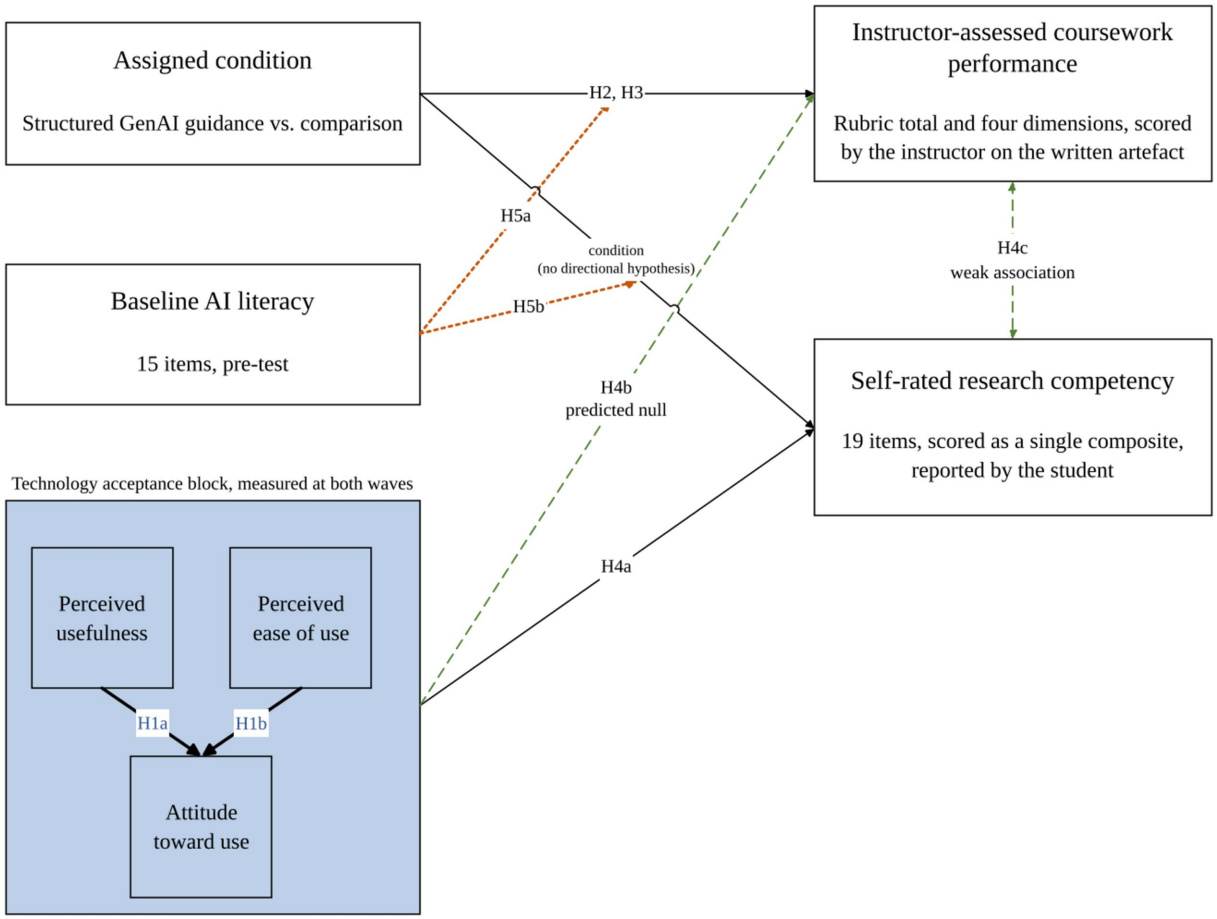

FIGURE 1
Conceptual framework: Two outcomes carried in parallel and the hypotheses linking them. Line types are defined in the note within the figure. Solid: predicted positive associations, apart from the path labelled as carrying no directional hypothesis. Dashed green: cross-path predictions. Dotted orange: moderation tests. H5a and H5b moderate the condition difference on each outcome. H4a runs from the acceptance block as a whole, not from attitude.

teaching-oriented undergraduate institutions, and all operate under the same provincial education authority, academic calendar, language of instruction, and national policy environment for GenAI in higher education. Because sites were not randomly drawn, the sample supports analytic rather than statistical generalisation, and institution fixed effects are carried in every model of instructor-assessed performance.

Within each institution, the collaborating instructor assigned intact classes to the structured GenAI guidance condition (hereafter the structured condition) or to the comparison condition. No student was assigned individually, and no class was split. The assignment was not randomised; the analytic strategy addresses this directly. The condition was recorded through the questionnaire form, with the structured condition completing form A at both waves and the comparison condition form B, so that group membership does not depend on instructor recall. Table 1 summarises the design.

3.2 The two conditions

Both conditions completed the same two assignments, within the same topic domain, under the same word limits, on the same submission platform, and were scored against the same rubric. They differed in one respect: the guidance and materials governing GenAI use, manipulated through four components. Table 2 sets out the contrast. The guidebook issued to the structured condition is provided as Supplementary Data Sheet 2, and the assignment briefs, questionnaires, weekly log template, and rubric descriptor table as Supplementary Data Sheet 1.

The organising principle of the structured condition was that the model assists the work but does not perform it. The assignment briefs named the steps for which AI assistance was recommended: refining a question, sharpening search terms, understanding an English-language source, and checking a proposal for logical gaps. They also named the steps reserved for the student. Every reference had to come from CNKI, Web of Science, or Scopus and be verified by DOI or original link, and fabricated references were explicitly prohibited. Judgements about relevance, about what a body of literature jointly implied, about which method fitted the question, and about the study's real limitations had to be the student's own, and any AI-assisted text had to be rewritten to reflect the student's understanding. Each submission required an AI use statement naming the tools used and the steps they were used for, section by section for the second assignment. The statement was excluded from the word count and the grade, and students were told so. It was handed in with the assignment and was therefore visible to the instructor during scoring (Section 3.3.1). The briefs also framed

10.3389/fpsyg.2026.2005055

TABLE 1 Participating institutions, condition allocation, and contribution to each analytic sample.

Institution	Linked panel (structured/ comparison)	Coursework records with both assignments scored	Included in analyses of instructor-assessed performance
A	61 (30/31)	135	Yes
B	55 (26/29)	64	Yes
C	28 (17/11)	39	No (scores converted from percentage grades)
D	25 (10/15)	85	Yes
Total	169 (83/86)	284 in included institutions	

Each dimension is scored 1–4. Entries are differences between the structured and comparison conditions on the second assignment. All models include institution fixed effects; standard errors are at the student level. The *p* columns are raw *p*-values. Holm-adjusted values across the four dimensions, D1 to D4, are 0.437, 0.036, 0.168, and 0.097 for the unadjusted contrast, 0.645, 0.009, 0.087, and 0.117 for ANCOVA, and 0.901, 0.014, 0.052, and 0.651 for the gain score.

TABLE 2 Components of the structured and comparison conditions.

Component	Structured GenAI guidance condition	Comparison condition	Timing
Written guidance	Purpose-written guidebook organised by research stage: ground rules on hallucination with a worked case of fabricated references; search strategy and Boolean query construction; synthesising abstracts into a review framework; drafting and correcting annotations; narrowing a research question; evaluating a methodological choice; stress-testing a proposal's internal logic. Each chapter supplied editable prompt templates, a compliant-versus-substitution comparison, and at least one worked model error with its correction.	Conventional research-methods guidance of comparable length within the assignment briefs: a staged procedure for defining a question, planning a search, annotating sources and identifying a gap; then deriving a question from that gap, matching method to question and reflecting honestly on limitations.	Released early April, before assignment 1
Video tutorials	Six recorded sessions totalling 6 h 50 min, concentrating on DeepSeek, including its reasoning mode for analytic tasks, and on Doubao's document upload function for working with downloaded articles.	None	Released early April; self-paced
In-class support	Short instructor demonstrations in regular class meetings; one optional online question session.	Regular teaching only	Throughout the semester
Assignment briefs	Identical task, plus a section specifying recommended AI-assisted steps, steps reserved to the student, a source-verification requirement, and a mandatory AI use statement (section by section for assignment 2).	Identical task, with students asked to work independently and to draw on real academic databases.	Mid-to-late April (assignment 1); early May (assignment 2)
Weekly usage log	Kept by students as a weekly record; checked by the instructor, not collected as data	Kept by students as a weekly record; checked by the instructor, not collected as data	Every Friday

assignment 1 as information organisation, with AI as an information assistant, and assignment 2 as independent design, with AI as a thinking partner.

The comparison condition was deliberately a business-as-usual condition rather than an AI-free one. GenAI use was neither taught nor prohibited, since prohibition would have been unenforceable and unlike any real programme. Students in both conditions therefore had the same tools available, and the contrast under test is one of guidance, not of access or exposure. The frequency-of-use item administered at both waves allows that reading to be checked against what students reported (Section 4.2).

3.3 Measures

3.3.1 Instructor-assessed coursework performance

The primary outcome was performance on two written assignments, scored by the course instructor against a common four-dimension rubric. Assignment 1 was a structured literature search and annotated bibliography of approximately 800 to 1,000 Chinese characters, comprising a statement of topic and search strategy, five annotated sources including at least two in English, and an identified research gap. Assignment 2 was a mini-research proposal of approximately 1,200 to 1,500 characters on the same specific topic, comprising a background and research question, a brief literature review, a proposed methodology with justification, and expected contributions together with at least two acknowledged limitations. Both were set within a single topic domain, the impact of digital technology or artificial intelligence on university students' learning and daily life, so that difficulty was comparable across students and classes.

The rubric scored four dimensions on a four-point scale with behaviourally anchored descriptors at each level: D1 information retrieval quality, D2 analytical depth, D3 research design coherence, and D4 academic writing quality, giving totals from 4 to 16. Instructors received the descriptor table and written instructions to score each dimension independently and apply identical criteria to both assignments. Totals were recomputed from the four-dimensional scores,

since the rubric's total column disagreed with their sum in 22 cases at Institution D. Both conditions completed both assignments as ordinary out-of-class coursework with access to GenAI tools, the structured condition having been taught to use them and the comparison condition neither taught nor prohibited. The rubric therefore scores assisted performance on a written artefact. It is not a measure of unaided capability, and no result reported here should be read as one.

Each institution was scored by a single rater, the instructor who taught both of its classes, so the rater is completely confounded with the institution, and all models for this outcome include institution fixed effects. Scoring was not blind. Each instructor knew which class had received the structured guidance, and scripts from that class, unlike those from the comparison class, carried an AI use statement that the instructor could read before assigning rubric scores. No script was scored by a second rater, so inter-rater reliability cannot be estimated. At Institution C, assignment scores were available only as institutional percentage grades converted to the four-point scale, which left the four dimensions identical within each student and produced a ceiling on the second assignment; that institution is excluded from all analyses of instructor-assessed performance and contributes only to the acceptance and self-report analyses.

3.3.2 Self-report measures

All self-report items used five-point Likert scales from 1 (strongly disagree) to 5 (strongly agree) and were administered in Chinese with parallel English wording.

AI literacy was measured at the pre-test with a 15-item subset drawn from the Scale for the Assessment of Non-Experts' AI Literacy (Laupichler et al., 2023), reworded for domestic tools, and translated by forward and back translation. Three five-item subscales were used: technical understanding, critical appraisal of AI-generated content, and practical application.

Perceived usefulness and *perceived ease of use* were measured at both waves with four items each, adapted from Davis (1989) and reworded for domestic GenAI in academic research; at the post-test, the same items were rephrased from anticipatory to experiential wording, with the referent unchanged. *Attitude toward use* was measured at both waves with three items adapted from Dahri et al. (2024); the third of these was replaced at the post-test by a continuance-intention item, so the attitude items are only partly matched across waves. Change scores for the three acceptance constructs were computed as post-test minus pre-test. Because the attitude item set is not fully matched, every attitude result reported below was re-estimated on a two-item composite restricted to the items common to both waves; the substantive conclusions are unchanged, and the comparison is reported in Supplementary Table S3. Item wording changed between waves, so comparability was tested with longitudinal confirmatory models (Supplementary Table S8). Perceived usefulness met equal loadings and intercepts across waves without loss of fit (ΔCFI = 0.000 at each step), which supports, for this sample, the comparability across waves of the change scores used in H4a and H4b. Perceived ease of use did not meet equal loadings ($\Delta\chi^2(3) = 12.65$, $p = 0.005$; ΔCFI = −0.019); freeing one loading brought the model within the ΔCFI criterion but not past the difference test, and its change score is read with that caveat. For attitude, a restricted test on the two common items did not reject equal loadings or intercepts, and the two-item check remains the safeguard.

Self-rated research competency was measured at the post-test with a 19-item scale written for this study and piloted with 30 students. It was not administered at the pre-test to avoid priming students on the competencies against which their assignments would be scored; on the instructor-assessed side, the first assignment's rubric score serves as the measure of prior performance, with the caveat set out in Section 3.6.1 that it was completed after the guidance had been issued. Items were generated to cover five content areas informed by Li et al. (2025): literature searching (4 items), critical appraisal of literature (4), research design (4), data literacy (3), and academic writing (4). The full item list is given in Supplementary Data Sheet 1.

The item set is factorable (Kaiser–Meyer–Olkin = 0.925; Bartlett's test of sphericity $p < 0.001$) and is dominated by one general factor, although it is not strictly unidimensional (Supplementary Tables S2 and S4). The first eigenvalue of the inter-item correlation matrix is 9.83 and accounts for 51.7% of item variance, against 1.26 and 1.12 for the second and third, and it is the only one to exceed its parallel-analysis threshold. A single-factor confirmatory model nevertheless fitted poorly (robust CFI = 0.854, RMSEA = 0.106), and a correlated five-factor model built on the intended content areas fitted better (CFI = 0.929, RMSEA = 0.076; scaled $\Delta\chi^2(10) = 92.9$, $p < 0.001$). Those five factors correlated between 0.69 and 0.91 and loaded between 0.81 and 0.97 on a second-order general factor. Omega total for the composite was 0.95 (95% CI [0.93, 0.96]) and omega hierarchical 0.81 to 0.92 depending on the model, so at least four-fifths of the variance in the total score reflects the general factor. We therefore use the composite as a measure of global self-rated research competency and do not score the content areas separately. Their means are intercorrelated between 0.59 and 0.77, against 0.32 to 0.60 for the four rubric dimensions of the second assignment (Supplementary Table S5), so the self-rating is less differentiated than the instructors' scoring without being undifferentiated.

The critical-appraisal items of the AI literacy scale (pre-test) and of the competency scale (post-test) measure different constructs and are never compared across waves. Frequency of AI use for coursework was recorded at both waves, together with tools used and, at the post-test, purposes.

Weekly usage log. Students in both conditions were asked to keep a short weekly record of whether, with which tools and for what purposes they had used domestic AI for academic work. Instructors checked that the record was being kept, but entries were not collected or coded, so no response rate can be computed and the logs enter no analysis; evidence on how much students used the tools comes only from the frequency item in the questionnaires.

3.4 Procedure

The pre-test was administered around the Qingming holiday, before any AI materials were released, and was completed by most participants between 5 and 9 April 2026, with a small number of late responses through 24 April. The guidebook and video tutorials were then released to the structured condition, followed by a period of self-study. Assignment 1 was posted in mid-to-late April with a 2-week window; assignment 2 was posted in early May, also with a 2-week window. Instructors scored each assignment on the shared rubric after its deadline. The post-test was administered at the end of the semester, between 26 June and 3 July 2026.

Because the questionnaires were anonymous, students supplied a self-generated code at both waves, formed from the initials of their Chinese name. Each questionnaire opened with a statement of

purpose, and completion constituted consent (see the Ethics Statement). Students in the comparison condition were offered the guidebook and video materials after the post-test. The platform's submission timestamp and approximate region were used only for the duplicate-removal and linkage checks in Section 3.5 and were removed before analysis.

3.5 Participants and data preparation

A total of 493 questionnaires were returned across the four instruments (147 and 143 at the pre-test for forms A and B; 102 and 101 at the post-test). Retaining the final submission for each institution-by-code combination within each form removed 27 earlier submissions and left 466 valid records.

Linking the two waves required care, since self-generated codes are not unique identifiers and students do not always reproduce them exactly. Exact matches within institution and form produced 148 pairs. Remaining candidates, arising from truncations, prefixes, and single-letter differences, were resolved by a three-step rule: each pair had to be anchorable to a single named student on the institution's grading roster, with positions already claimed by exact matches removed; candidates were pruned by relationship strength; and any code still admitting more than one counterpart was left unmerged. Of the 57 candidates, 21 were accepted and 36 were rejected, each with its reason recorded. This yielded the linked panel of 169 students shown in Table 1, of whom 159 were linked exactly or by a reliable roster-anchored match. Within that core sample, 27 codes may still not identify a single student and are flagged as unclear. A wrong link joins one student's pre-test to another's post-test, so the change score it yields compares two people: mislinkage adds noise to change scores and, where wrong links are not random, can manufacture an association. One-wave analyses are unaffected. Models that pair a pre-test level with a post-test outcome, such as the AI literacy models, do not carry that artefact, but a wrong link still adds error and, where wrong links are not random, can bias them. The H4 change-score tests and the AI literacy models are therefore repeated on the securely linked subsample (Sections 4.6 and 4.7).

One case required a further correction. A respondent in the comparison arm named an institution that was not a study site; the anchoring rule used for the linkage resolved the case to Institution D, and its coursework record was merged (Supplementary Note S1). Every result in Section 4 is unchanged by this to two decimal places.

The linked panel was 70.4% female. Sixty-seven participants were in their second year, 93 in their third, and nine in their fourth. Every participant had used at least one domestic GenAI tool before the study began, and 88 had used three or more. At the pre-test, 89 of 169 described their use of AI for coursework as frequent or near-constant.

Data preparation was carried out in Python and checked by re-reading all 12 source files independently of the pipeline. The scoring sheet at Institution B was revised by its rater during the semester; the version the rater confirmed as final was used, and the results are reported with and without that institution.

3.6 Analytic strategy

The analysis plan was specified before estimation and is reported in full regardless of outcome, including for H5. Analyses that go beyond the plan are identified as additional where they are reported. Because the two outcomes come from different sources, they are estimated on different samples. Acceptance and self-report models use the linked questionnaire panel, with the core sample ($n = 159$) as the primary analytic sample and the full panel ($n = 169$) as a sensitivity check. Models of instructor-assessed performance do not depend on questionnaire linkage and use the complete rubric rosters of the three institutions with dimension-level scoring ($n = 284$); models that combine an outcome from each source are restricted to panel members who also have dimension-level coursework scores ($n = 136$, core $n = 130$), since Institution C's converted percentage grades are excluded from every analysis of the instructor-assessed outcome. Supplementary Table S1 maps each hypothesis onto its outcome, sample, estimator, and software.

No correction for multiple testing was applied to the pre-specified hypotheses, each of which is one test on its own outcome. The four rubric dimensions under H3 form a family, and Holm-adjusted p-values are given in the note to Table 3. The analyses labelled additional, that is, the post-test belief models, the arm-specific estimates, and the alternative specifications in the supplement, are exploratory: their p-values are unadjusted; they are reported with confidence intervals (Table 4) or standard errors (Supplementary Table S7), and no conclusion rests on any one of them.

3.6.1 Estimation

Three choices in the specification need justification.

The comparison of the two conditions on instructor-assessed performance (H2, H3) is reported in three forms, which estimate different quantities: the unadjusted difference on the second assignment; the difference on the second assignment between students with the same first-assignment score (analysis of covariance [ANCOVA]); and the difference in change from the first assignment to the second (gain-score regression). None is a baseline-adjusted treatment effect in the conventional sense because the design has no measure of performance taken before the guidance was issued. The first assignment was completed after the materials had been released, so it is a post-treatment covariate, and the two assignments are different tasks, so their difference is not a repeated measurement of one performance. Both arms completed both tasks, so the change in task demand is common to them, but the adjustment is not thereby harmless: if the guidance had already raised first-assignment scores,

TABLE 3 Dimension-level differences between conditions on the second assignment ($n = 284$).

Rubric dimension	Unadjusted b (SE)	p	ANCOVA b (SE)	p	Gain score b (SE)	p
D1 Information retrieval quality	+0.059 (0.075)	0.437	+0.032 (0.070)	0.645	−0.010 (0.084)	0.901
D2 Analytical depth	+0.204 (0.077)	0.036	+0.222 (0.072)	0.009	+0.253 (0.086)	0.014
D3 Research design coherence	+0.147 (0.085)	0.168	+0.176 (0.080)	0.087	+0.242 (0.101)	0.052
D4 Academic writing quality	+0.161 (0.075)	0.097	+0.133 (0.070)	0.117	+0.084 (0.085)	0.651

TABLE 4 Acceptance against the two outcomes: the change-score tests specified in H4 and the additional post-test analyses.

Sample	Self-rated competency *b* [95% CI], *p*	*n*	Instructor-assessed score *b* [95% CI], *p*	*n*
Change in perceived usefulness (H4a, H4b as specified)				
Core sample	+0.179 [+0.056, +0.302], 0.005	159	−0.088 [−0.514, +0.338], 0.683	130
Full panel	+0.190 [+0.070, +0.310], 0.002	169	−0.120 [−0.535, +0.295], 0.569	136
Conservative identity matching	+0.065 [−0.069, +0.199], 0.336	132	−0.168 [−0.639, +0.303], 0.481	115
Core sample, weighted for retention	+0.195 [+0.046, +0.344], 0.010	159	−0.019 [−0.480, +0.442], 0.935	130
Post-test perceived usefulness (additional analysis)				
Core sample	+0.383 [+0.240, +0.526], <0.001	159	−0.488 [−1.025, +0.050], 0.075	130
Full panel	+0.391 [+0.249, +0.532], <0.001	169	−0.454 [−0.984, +0.076], 0.092	136
Conservative identity matching	+0.334 [+0.181, +0.488], <0.001	132	−0.579 [−1.162, +0.004], 0.052	115
Institution B excluded	+0.467 [+0.300, +0.634], <0.001	104	−0.737 [−1.534, +0.061], 0.070	76
Core sample, students with both outcomes	+0.261 [+0.097, +0.424], 0.002	130	−0.488 [−1.025, +0.050], 0.075	130
Core sample, weighted for retention	+0.399 [+0.244, +0.555], <0.001	159	−0.555 [−1.214, +0.103], 0.098	130
Core sample, structured arm only	+0.336 [+0.127, +0.546], 0.002	79	+0.035 [−0.721, +0.792], 0.926	63
Core sample, comparison arm only	+0.442 [+0.237, +0.648], <0.001	80	−1.121 [−1.921, −0.320], 0.007	67
Test of the difference between the two paths				
Change in perceived usefulness	difference = −0.283 [−0.606, +0.040]		$p = 0.086$	130
Post-test perceived usefulness	difference = −0.746 [−1.157, −0.335]		$p < 0.001$	130

Unstandardised coefficients with 95% confidence intervals. Models of self-rated competency control for condition and institution; models of the instructor-assessed score add assignment 1. All models of the instructor-assessed outcome exclude Institution C (Section 3.3.1). Weighted rows use retention weights with HC3 standard errors (Section 3.6.2). The difference test stacks both standardised outcomes and clusters by student (260 rows); *p*-values in the weighted rows and in the difference test refer to the normal distribution.

adjusting for them removes part of any difference, and conditioning on a variable the treatment may have moved can bias a comparison in either direction. In a non-randomised design, the two adjusted estimators can also diverge when groups differ on the covariate, and that divergence is itself diagnostic (van Breukelen, 2006). All three forms are therefore reported, and we set an agreement between the two adjusted estimators in advance as a condition for interpreting a difference. H3 and H5 are worded as specified in terms of the effect of the intervention; in reporting them, the word refers to the difference between conditions and carries no causal claim. Institution fixed effects enter every pooled model of this outcome because rater and institution are fully confounded. Classical standard errors are reported with heteroskedasticity-consistent (HC3) errors alongside.

The condition was assigned at the class level, and because each institution contributed one class per condition, the class coincides exactly with the institution-by-condition cell: six such cells in the coursework sample and eight in the questionnaire panel. That is too few for dependable cluster-level inference. With institution fixed effects in the model, the regressors nearly span the six cluster means, so the cluster-robust variance estimator collapses and returns implausibly small standard errors; without those fixed effects, the model is misspecified, since rater is confounded with institution. We therefore report wild cluster bootstrap *p*-values with Webb weights, as recommended for a small number of clusters (MacKinnon et al., 2023), together with the same test estimated without institution fixed effects so that the sensitivity is visible.

Since no formal cluster-level procedure is trustworthy here, the design-level evidence is presented directly and first: each institution contributes one class per condition, so the study contains three within-institution contrasts, and these are reported individually ahead of the pooled estimate. The student-level standard errors that accompany every estimate, HC3 included, quantify sampling variation among students within the six classes observed. They treat those classes as fixed and say nothing about how other classes, instructors, or institutions would have fared. The *p*-values for H2 and H3 are reported on that understanding; they are not tests of whether structured guidance raises assessed performance in classes generally, a question on which the design supplies three contrasts and no more.

Because H4b predicts a null, a non-significant coefficient is not treated as evidence of absence on its own: the estimate is reported with its 95% confidence interval so that the precision of the null can be judged, and the interpretation rests on the contrast between the two paths, not on the absence of one.

3.6.2 Robustness checks

Two features of the sample required checks. The first is attrition, which was substantial and may not have been random; non-retention here combines students who did not complete the post-test with students whose two records could not be linked (Section 3.5). Pre-test scores of students who did and did not enter the linked panel are compared by condition using independent-samples t-tests with Cohen's d, and the H4 analyses are re-estimated in three ways: with each panel member weighted by the inverse of their estimated probability of retention, with pre-test covariates added, and separately within each arm. The retention model, a logistic regression on the pre-test respondents' condition, institution, demographic, and baseline measures, is reported with the resulting weights and balance diagnostics in Supplementary Table S6; weighting corrects for selection on the characteristics in that model and on nothing else. The

second is uncertainty in the records themselves, in the linkage of some questionnaire pairs, and in the scoring sheet revised at Institution B (Section 3.5): every model that uses the panel is re-run on the core sample, on the full panel, and with Institution B excluded, the coursework models are re-run without that institution, and conclusions are reported as robust only where they hold throughout.

3.6.3 Software

The analysis file was constructed using Python 3.11 (pandas, numpy). The analyses specified in the plan were conducted using Stata 19 (StataNow/MP; StataCorp LLC, College Station, TX, United States); the commands used for each hypothesis are listed in Supplementary Table S1. Each of these models in the main tables was then re-estimated in Python (statsmodels) from the same analysis file as an independent check, and only estimates that agreed are reported. The supplementary analyses (unadjusted and within-institution contrasts, confidence intervals for correlations, the retention model, and the weighted, covariate-adjusted, and arm-specific estimates) were estimated using Python and re-estimated using R 4.3.3 and using Stata 19 with identical results; parallel analysis, the confirmatory factor models (robust maximum likelihood), the longitudinal invariance models, and omega were obtained in R with the psych and lavaan packages.

4 Results

4.1 Preliminary analyses

The coursework rosters of the three institutions with dimension-level scoring contain 286 records, 284 with both assignments scored; within the linked panel of 169, 157 students have a coursework score of some kind, and 136 carry dimension-level information. Internal consistency was acceptable to high across every composite and subscale, with Cronbach's alpha between 0.814 and 0.948 (Table 5).

How comparable the two arms are differs between the two analytic samples, and the difference matters for interpretation. On the outcome that carries H2 and H3, the arms were closely matched on the first assignment, which followed the release of the materials and is a covariate rather than a pre-intervention baseline (Section 3.6.1): scores differed by $d = 0.05$ ($p = 0.70$) in the pooled sample and by no more than $d = 0.21$ within any institution, none of these differences approaching significance. In the questionnaire panel, however, the comparison arm reported higher perceived usefulness at the pre-test ($d = 0.37$, $p = 0.018$), while the structured arm reported more frequent prior use ($d = 0.35$, $p = 0.024$). Perceived ease of use, attitude, AI literacy, and the number of tools previously used did not differ (Table 6).

The attrition analysis explains this pattern. Each arm contributed 136 unique pre-test respondents, and 83 and 86, respectively, could be linked to a post-test record, so non-retention was similar in the two arms (39.0% and 36.8%); it was much lower at Institution B (5 of 60) than at the other three (97 of 211; one respondent's institution could not be established). What differed between the arms was its selectivity. In the comparison arm, students who entered the panel had reported significantly higher perceived usefulness at the pre-test than those who did not ($d = 0.59$, $p < 0.01$), with a weaker difference on AI literacy ($d = 0.30$, $p < 0.10$). No such selection appeared in the structured arm (all $|d| \leq 0.27$, ns). Among all 272 pre-test respondents, the two arms did not differ on perceived usefulness (M = 3.90 and 3.98, $d = 0.13$, $p = 0.28$), whereas the difference in prior frequency of use was already present ($d = 0.28$, $p = 0.021$). The imbalance on perceived usefulness in the panel is therefore a product of who remained, not of how classes were assigned (Sections 4.2 and 4.6); the coursework rosters, where retention plays no part, are balanced.

4.2 Changes in acceptance across the semester

Change in acceptance did not differ reliably between the arms once selective retention is taken into account. In the linked panel, perceived usefulness increased in the structured arm (M change = +0.20, $p = 0.019$) and not in the comparison arm (−0.01, $p = 0.82$), a nominally significant difference ($p = 0.040$, $d = 0.32$). That contrast is not robust. The comparison arm entered the panel with higher perceived usefulness because of who was retained (Section 4.1), the two arms finished the semester level (4.08 in each), and the pre-test to post-test correlation of approximately 0.28 leaves ample room for regression to the mean. Adjusted for the

TABLE 5 Descriptive statistics and internal consistency (linked panel, $n = 169$).

Measure	Wave	Items	α	Structured M (SD)	Comparison M (SD)
AI literacy	Pre	15	0.908	3.59 (0.57)	3.73 (0.50)
Perceived usefulness	Pre	4	0.901	3.88 (0.67)	4.10 (0.50)
Perceived ease of use	Pre	4	0.814	3.73 (0.66)	3.81 (0.59)
Attitude toward use	Pre	3	0.873	3.84 (0.65)	3.93 (0.54)
Perceived usefulness	Post	4	0.850	4.08 (0.63)	4.08 (0.48)
Perceived ease of use	Post	4	0.841	3.90 (0.62)	3.85 (0.59)
Attitude toward use	Post	3	0.824	4.00 (0.64)	3.97 (0.46)
Self-rated research competency	Post	19	0.948	3.57 (0.61)	3.57 (0.49)
Frequency of AI use	Pre	1	n.a.	3.65 (0.65)	3.43 (0.60)
Frequency of AI use	Post	1	n.a.	3.88 (0.53)	3.74 (0.69)

Alpha for the three AI literacy subscales was 0.817 (technical understanding), 0.819 (critical appraisal of AI-generated content), and 0.881 (practical application). Coursework totals (4–16): assignment 1 M = 10.72 (SD 1.91), assignment 2 M = 11.91 (SD 2.24), $n = 284$.

pre-test score, the post-test difference between arms is +0.06 ($p = 0.467$); weighted for retention, the difference in change is +0.15 ($p = 0.238$). Neither estimate reproduces the difference in raw change, so it is not one we can attribute to the guidance. Perceived ease of use (+0.17, $p = 0.023$ against +0.05, $p = 0.52$; between-arm $p = 0.226$) and attitude (+0.16, $p = 0.054$ against +0.05, $p = 0.47$; between-arm $p = 0.284$) did not differ between arms on any estimator; the ease-of-use change rests on items that were at best partly invariant across waves (Section 3.3.2).

Frequency of use increased in both arms, and if anything, more in the comparison arm (+0.31, $p < 0.001$) than in the structured arm (+0.23, $p = 0.007$), with no difference between them ($p = 0.470$).

4.3 Technology acceptance structure (H1)

The core structure of TAM reproduced at both waves and in both samples (Table 7). Perceived usefulness and perceived ease of use each predicted attitude toward use at the pre-test ($\beta = 0.458$ and 0.416, both $p < 0.001$, $R^2 = 0.619$) and again at the post-test, after a semester of sustained use ($\beta = 0.412$ and 0.383, both $p < 0.001$, $R^2 = 0.536$). Estimates on the full panel were almost identical. H1a and H1b are supported.

4.4 Differences between conditions in instructor-assessed coursework performance (H2)

Within each of the three institutions, the class given structured guidance scored higher on the second assignment than the comparison class (Table 8). Unadjusted, the differences on the 16-point rubric were +0.39 points at Institution A ($p = 0.270$), +0.78 at Institution B ($p = 0.105$), and +0.71 at Institution D ($p = 0.122$), none of which were individually significant. Adjusted for the first assignment, they were +0.67 ($p = 0.001$), +0.48 ($p = 0.126$), and +0.66 ($p = 0.145$), and as gain scores, +0.70 ($p < 0.001$), +0.44 ($p = 0.158$), and +0.45 ($p = 0.442$). The one contrast that reaches significance at Institution A does so only after adjustment, because the structured class there had scored 0.31 points lower on the first assignment. These three contrasts are the evidence at the level at which the guidance was assigned. All

TABLE 6 Comparability of the two arms and differential attrition.

Measure	Structured	Comparison	d	*p*
First-assignment comparability, coursework rosters ($n = 284$)				
Assignment 1 total	10.77	10.68	0.05	0.70
Institution A	10.75	11.06	0.17	0.34
Institution B	11.16	10.81	0.21	0.41
Institution D	10.39	10.13	0.12	0.59
Baseline equivalence, linked panel ($n = 169$)				
AI literacy	3.59	3.73	0.27	0.081
Perceived usefulness	3.88	4.10	**0.37**	**0.018**
Perceived ease of use	3.73	3.81	0.12	0.43
Attitude toward use	3.84	3.93	0.14	0.36
Frequency of prior use	3.65	3.43	**0.35**	**0.024**
Number of tools used	2.69	2.80	0.11	0.48
Attrition: panel entrants vs non-entrants, by arm				
Perceived usefulness (pre)			0.08 ns/0.59**	
AI literacy (pre)			0.27 ns/0.30†	
Attitude (pre)			0.00 ns/0.24 ns	

Effect sizes are reported as absolute values. On the first assignment, the structured arm scored higher in the pooled sample and at Institutions B and D and lower at Institution A; the direction of the panel differences can be read from the means. Attrition rows report the structured arm first and the comparison arm second. † $p < 0.10$, ** $p < 0.01$. Bold values indicate effect sizes and *p*-values for baseline between-condition comparisons in the linked panel with $p < 0.05$.

TABLE 7 Perceived usefulness and perceived ease of use predicting attitude toward use at both waves.

Predictor / statistic	Pre-test (core, $n = 159$)	Post-test (core, $n = 159$)	Pre-test (full, $n = 169$)	Post-test (full, $n = 169$)
Perceived usefulness	0.451 (0.062)***	0.409 (0.075)***	0.433 (0.062)***	0.404 (0.073)***
Standardised β	0.458	0.412	0.435	0.406
Perceived ease of use	0.392 (0.059)***	0.350 (0.069)***	0.403 (0.060)***	0.356 (0.068)***
Standardised β	0.416	0.383	0.424	0.388
R^2	0.619	0.536	0.599	0.534

Unstandardised coefficients with standard errors in parentheses. *** $p < 0.001$.

10.3389/fpsyg.2026.2005055

TABLE 8 Differences between the structured and comparison conditions on the second assignment (instructor-assessed total, 4–16).

Specification	*n*	*b*	SE	*p*
Within-institution contrasts				
Institution A: unadjusted	135	+0.389	0.351	0.270
Institution A: adjusted for assignment 1 (ANCOVA)	135	+0.668	0.201	0.001
Institution A: gain score	135	+0.696	0.202	<0.001
Institution B: unadjusted	64	+0.781	0.475	0.105
Institution B: adjusted for assignment 1 (ANCOVA)	64	+0.476	0.307	0.126
Institution B: gain score	64	+0.437	0.306	0.158
Institution D: unadjusted	85	+0.709	0.454	0.122
Institution D: adjusted for assignment 1 (ANCOVA)	85	+0.658	0.447	0.145
Institution D: gain score	85	+0.452	0.585	0.442
Pooled, with institution fixed effects				
Unadjusted	284	+0.570	0.240	0.018
Adjusted for assignment 1 (ANCOVA)	284	+0.568	0.191	0.003
ANCOVA, HC3 standard error	284	+0.568	0.182	0.002
Gain score	284	+0.567	0.207	0.007
Gain score, HC3 standard error	284	+0.567	0.195	0.004
ANCOVA, within-institution standardised outcome	284	+0.284	0.093	0.002
ANCOVA, Institution B excluded	220	+0.565	0.229	0.014
Gain score, Institution B excluded	220	+0.606	0.253	0.018
ANCOVA, questionnaire panel only	136	+0.596	0.275	0.032

Positive values favour the structured condition. ANCOVA rows adjust for the first-assignment score, and gain-score rows compare the change from it; the first assignment was itself completed after the guidance had been issued (Section 3.6.1). Standard errors are at student level. Wild cluster bootstrap *p*-values (Webb weights, 9,999 replications, six clusters) are given in the text.

favour the structured condition under every form of the comparison, but three pairs of classes cannot support a formal test: an exact sign test on three contrasts cannot return a two-sided *p*-value below 0.25.

Pooled across institutions with institution fixed effects, the unadjusted difference on the second assignment was +0.570 points (SE 0.240, $p = 0.018$), the difference adjusted for the first assignment was +0.568 points (SE 0.191, $p = 0.003$; HC3 SE 0.182, $p = 0.002$), and the gain-score difference was +0.567 points (SE 0.207, $p = 0.007$; HC3 SE 0.195, $p = 0.004$). The three agree because the arms were level on the first assignment, so the adjustment sharpens the estimate without moving it. The adjusted difference corresponds to d = 0.24 (95% CI [0.08, 0.40]) against the comparison condition's standard deviation and to approximately 5% of the rubric's 12-point range, a small difference by the usual conventions. On analytical depth, the dimension where it is clearest (Section 4.5), the share of students scoring 3 or 4 on the four-point scale, standardised over institutions, was 83% in the structured condition and 75% in the comparison condition.

Excluding Institution B, whose scoring sheet was revised during the semester, or restricting the analysis to students who completed both questionnaires leaves the adjusted estimates essentially unchanged (Table 8). Wild cluster bootstrap *p*-values with Webb weights are 0.002 (ANCOVA) and 0.018 (gain score), but removing institution fixed effects, which misspecifies the model, moves both to approximately 0.37, and we place no weight on either figure.

The data are therefore consistent with H2, but we do not treat the hypothesis as confirmed: the contrast rests on six non-randomised classes, its adjusted forms condition on an assignment completed after the guidance had been issued, and the scores were given by instructors who knew the allocation (Section 3.3.1).

4.5 Where the difference appears (H3)

Under both adjusted estimators, on raw *p*-values, the difference is confined to two of the four rubric dimensions, the two that H3 identified in advance (Table 3, Figure 2): analytical depth (+0.222, $p = 0.002$ under ANCOVA; +0.253, $p = 0.004$ as a gain score) and research design coherence (+0.176, $p = 0.029$; +0.242, $p = 0.017$). Information retrieval shows no significant difference under either, and academic writing does not meet the criterion of agreement between the two adjusted estimators set in Section 3.6.1 (+0.133, $p = 0.058$; +0.084, $p = 0.325$).

The unadjusted contrasts on the second assignment are less tidy: +0.204 for analytical depth ($p = 0.009$), +0.147 for research design coherence ($p = 0.084$), +0.161 for academic writing ($p = 0.032$), and +0.059 for information retrieval ($p = 0.437$). With a Holm correction across the four dimensions, analytical depth remains significant under every estimator (adjusted $p = 0.036$, 0.009, and 0.014) and research design coherence under none (0.168, 0.087, and 0.052); Figure 2 gives the intervals. The separation of research design from writing depends on the adjustment. We read the pattern as consistent with H3 for analytical depth and as suggestive, no more, for research

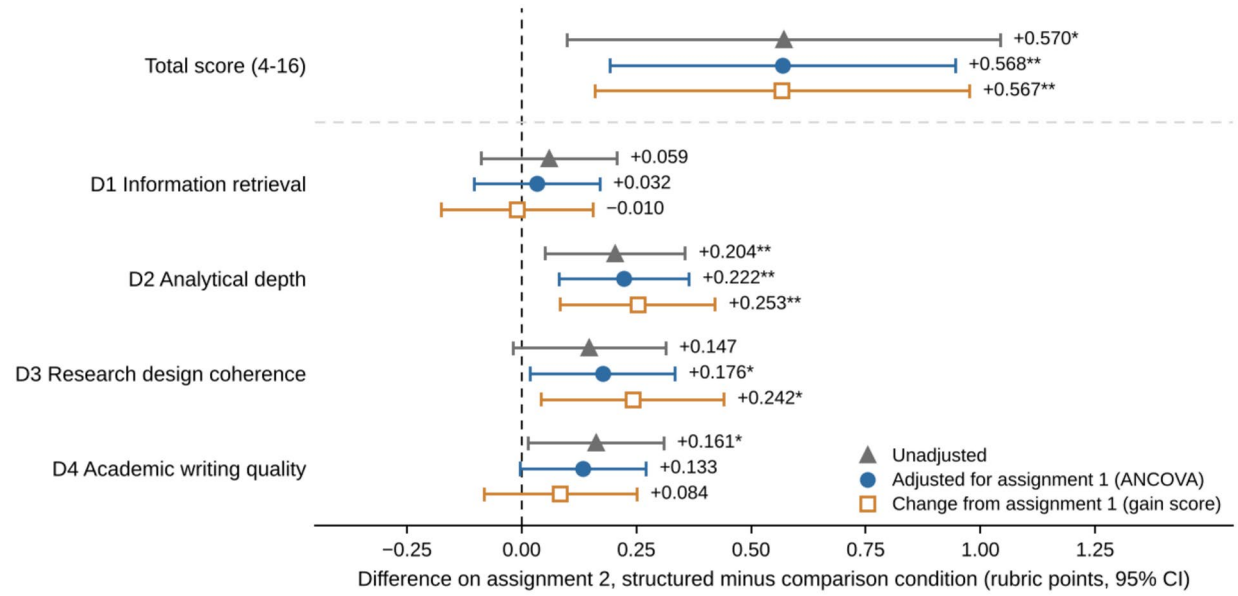

FIGURE 2
Differences between the structured and comparison conditions on the second assignment, for the rubric total and for each rubric dimension: Unadjusted, adjusted for the first assignment (ANCOVA), and as a change from the first assignment (gain score). All models include institution fixed effects (n = 284); Bars are 95% confidence intervals from student-level standard errors. * $p < 0.05$, ** $p < 0.01$, raw p-values; Holm-adjusted values are given in the note to Table 3.

design coherence, and the inferential limits stated for H2 apply here in full.

4.6 Acceptance, self-rated competency, and performance (H4)

H4 was stated in terms of change in perceived usefulness, and its three parts are taken in that form first (Table 4; Figure 3B). H4a is not supported. Change in perceived usefulness is positively associated with self-rated competency in the core sample (b = +0.179, p = 0.005) and on the full panel (+0.190, p = 0.002), but the association is carried by the 27 students whose self-generated codes may not identify a single person: the slope is +0.076 (p = 0.275) for the 132 students whose linkage is secure and +0.496 for the 27 whose linkage is uncertain (interaction b = +0.420, p = 0.003), and on the id-secure subsample alone, it is +0.065 (95% CI [−0.069, +0.199], p = 0.336). Outliers are not the explanation (Supplementary Table S7). A change score means something only if both waves belong to the same person (Section 3.5), and where that is secure, the association is small and not significant. We therefore treat H4a as unsupported, a statement about the evidence and not a demonstration that the association is zero.

H4b is not contradicted, but the data cannot establish that the path is absent. Change in perceived usefulness is not significantly related to second-assignment performance once the first assignment, condition, and institution are held constant (b = −0.088, 95% CI [−0.514, +0.338], p = 0.683, n = 130). The interval admits associations of approximately half a rubric point in either direction per unit of change: an absence of evidence at this sample size, not a demonstration that no association exists.

The results are consistent with H4c. Self-rated competency and instructor-assessed performance are only weakly related: r = 0.044 (95% CI [−0.125, 0.211], p = 0.608, n = 136), Spearman ρ = 0.051 (95% CI [−0.119, 0.217]), and r = −0.034 (95% CI [−0.201, 0.135]) against coursework gain; including Institution C increases the correlation only to 0.129 (95% CI [−0.028, 0.280], n = 157). Correlations of the moderate size usually reported for self-assessment (Zell and Krizan, 2014) lie at or beyond the upper limit of these intervals. The intervals do not exclude small positive correlations, so the result bounds the relation from above rather than showing that none exists. Nor is this simply a matter of setting a general scale against one piece of work: when each self-rated content area is paired with the rubric dimension closest to it, the correlations range from −0.053 to 0.083, and every interval includes zero (Supplementary Table S5). The pairings are imperfect, but they show that mismatched content alone does not produce the result.

The remaining analyses in this section were not specified in H4 and are reported as additional and exploratory, with unadjusted p-values (Section 3.6; Figure 3A). Post-test perceived usefulness, which requires no linkage across waves, is positively associated with self-rated competency (b = +0.383, $p < 0.001$, n = 159) and is not significantly associated with second-assignment performance once the first assignment, condition, and institution are held constant (b = −0.488, 95% CI [−1.025, +0.050], p = 0.075, n = 130). The first association is stable across every restriction (+0.261 to +0.467, $p \leq 0.002$); the second is negative in all four specifications (−0.454 to −0.737) without reaching significance (p between 0.052 and 0.092). The other two beliefs behave in the same way: perceived ease of use gives +0.443 ($p < 0.001$) against −0.244 (p = 0.327), and attitude toward use +0.379 ($p < 0.001$) against −0.208 (p = 0.445).

To test the difference between the two paths directly, both outcomes were standardised, stacked, and regressed on acceptance with an indicator for the outcome, clustering by student. For post-test perceived usefulness, the difference is −0.746 (SE 0.210, $p < 0.001$); for change in perceived usefulness, it is −0.283 (SE 0.165, p = 0.086). What these analyses support is that the two paths differ for the level measure, not that the path to instructor-assessed performance is absent.

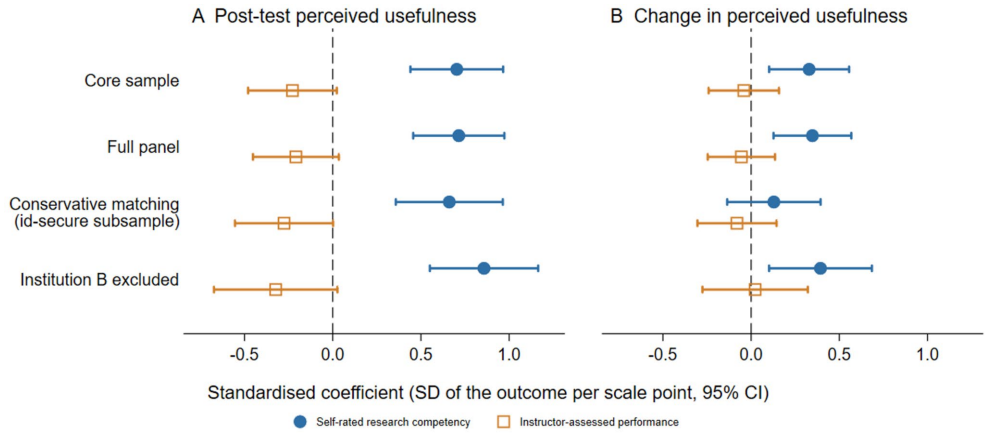

FIGURE 3
The same predictor against two outcomes: **(A)** post-test perceived usefulness, an additional analysis; **(B)** change in perceived usefulness, the predictor specified in H4a and H4b. Models of the instructor-assessed outcome add the first assignment and exclude Institution C; the specifications are given in Section 4.6. Difference between the two paths (n = 130); **A** −0.746 (SE 0.210), $p < 0.001$, **B** −0.283 (SE 0.165), $p = 0.086$.

Because retention in the comparison arm was selective on perceived usefulness (Section 4.1), these associations were re-examined in three ways (Supplementary Table S7). Weighting the panel by the inverse probability of retention leaves the estimates close to their unweighted values: for post-test perceived usefulness, +0.399 ($p < 0.001$) on the self-rating against −0.555 ($p = 0.098$) on the instructor-assessed score, with a path difference of −0.825 ($p < 0.001$) and a correlation between the two outcomes of 0.090 (95% CI [−0.110, 0.281]). Adding pre-test covariates changes nothing material. Estimating the two arms separately is more informative, since selection was detected in only one. The positive association with the self-rating is present in both (structured +0.336, $p = 0.002$; comparison +0.442, $p < 0.001$). The negative coefficient on the instructor-assessed side is not: it is −1.121 in the comparison arm ($p = 0.007$; weighted −1.069, $p = 0.035$) and +0.035 in the structured arm ($p = 0.926$). The difference between the slopes is not robust (interaction $p = 0.058$ unweighted and 0.186 weighted; for attitude, 0.029 and 0.203), so we do not claim that the association depends on condition. What holds in both arms is the weaker statement that acceptance goes with the self-rating and shows no significant positive association with the instructor-assessed score; the negative sign comes from the comparison arm, where retention was selective, and weighting does not remove it.

4.7 AI literacy as a moderator (H5)

H5 is not supported on either outcome. The interaction between baseline AI literacy and condition was not significant for instructor-assessed performance (b = +0.593, SE 0.569, $p = 0.299$, n = 136) or for self-rated competency (b = +0.195, SE 0.166, $p = 0.243$, n = 159).

The simple slopes on the instructor-assessed outcome require care. The estimated difference between conditions was +0.285 at one standard deviation below the mean of AI literacy ($p = 0.495$), +0.604 at the mean ($p = 0.032$) and +0.923 at one standard deviation above ($p = 0.026$). Two are individually significant, and the gradient runs in the predicted direction, but without a significant interaction, the slopes are not evidence of moderation. On the self-rated outcome, none of the three slopes were significant.

AI literacy is nonetheless related to one of the two outcomes, and it divides along the same line as acceptance. Its association with self-rated research competency is positive and significant (b = +0.243, $p = 0.003$). Its association with instructor-assessed performance is not significant and imprecisely estimated (b = +0.075, 95% CI [−0.467, +0.618], $p = 0.784$). On the securely linked subsample, the two associations keep their sign and significance (+0.377 and −0.061). The condition × AI literacy interaction on the self-rating, not significant in the core sample, is significant there (b = +0.409, $p = 0.017$; Supplementary Table S7); because a wrong link does not build the outcome's own respondent into a level predictor (Section 3.5), that subsample is a sensitivity check here rather than the decisive sample it is for H4a, and we do not treat the result as support for H5b.

5 Discussion

5.1 What the study found

The findings fall into four parts, in descending order of how firmly the data support them.

First, self-rated research competency and instructor-assessed coursework performance were only weakly related ($r = 0.04$), including when self-rated content areas were paired with the nearest rubric dimensions; the interval excludes correlations above 0.21 but not small positive ones. Second, in exploratory analyses additional to the hypotheses, all three acceptance beliefs measured at the post-test were positively associated with the self-rating, in both arms and with or without weighting for attrition, and none showed a significant positive association with instructor-assessed performance; for perceived usefulness, the two standardised paths differed significantly in the pooled sample, a divergence carried mainly by the comparison arm. The association hypothesised for change in perceived usefulness (H4a) was not supported. Third, the self-rating was dominated by a general factor, and the rubric scores were more differentiated. Fourth, in each of the three institutions, the class given structured guidance scored higher on the second assignment, by 0.57 points on a 16-point rubric in the pooled comparison and most clearly on analytical depth. That contrast is descriptive: it comes from six non-randomised classes, was scored by instructors who knew the allocation, and in its adjusted versions conditions on an assignment completed after the guidance had been issued.

The core structure of the technology acceptance model was reproduced at both waves; what that structure did not do was show any significant positive association with the instructor-assessed outcome.

5.2 Acceptance goes with the self-rating rather than with the assessed work

The measurement problem set out in Section 1 is more than a methodological nuisance. Had we measured research competency only by self-report, as most acceptance studies in this area do (Yang, 2026), the post-test associations would have invited the conclusion that acceptance goes with greater competency, a conclusion that finds no support in the instructor-assessed outcome. What the self-report tracks is partly legible in its own structure: a composite dominated by one general factor, predicted most strongly by how easy the tool feels to use, and only weakly related to the rubric even where content is matched. That is the profile of a global self-perception, which the self-regulated learning literature associates with judgements made from fluency and not of an inventory of separately judged skills (Bjork et al., 2013; Zell and Krizan, 2014). We did not measure fluency, so that link is an interpretation and not a finding.

This is a boundary condition for TAM in educational settings, not a refutation of it. TAM was built to explain adoption, and the part of it tested here, the path from perceived usefulness and perceived ease of use to attitude, was reproduced at both waves. The error lies in treating the end of the acceptance chain as though it were an outcome of learning. Studies that place a self-reported competency construct at the terminus of an acceptance model, and there are many, measure students' belief that the technology is working for them. Whether their assessed work is any better is a separate question, and it requires a separate measurement.

The sign on the instructor-assessed side deserves care. The pooled point estimates are negative in every specification without reaching significance, and the arm-specific analysis shows where they come from: the association is clearly negative among comparison students and not significant in the structured condition (Section 4.6). Four readings fit that pattern, and the design cannot choose among them; none of the processes they invoke was measured. Students who find the tool more useful may lean on it more substitutively, the mechanism Fan et al. (2025) documented and Zhu et al. (2026) describe as dependent offloading; its absence under guidance designed to constrain substitution would fit that reading, but the difference between the slopes is not robust to weighting. Students who were struggling may have found the tool more valuable because they needed more from it. Fluent AI output may both increase the user's estimate of the tool and reduce the effortful processing that produces a good assignment (Lee et al., 2025). Or the association may be an artefact of retention: comparison students were more likely to stay in the panel if they rated the tools highly, and if those who stayed despite rating them less highly were disproportionately diligent, a negative association would appear among the retained without existing in the class. What the data support is narrower and holds in both arms: acceptance was positively associated with the self-rating and showed no significant positive association with instructor-assessed performance.

The small correlation between the two outcomes makes the same point from another angle: students' self-ratings told us little about their instructors' scores, whether or not content areas were matched, and the interval allows at most a small positive relation. Some of that gap is built in because the scale asks about research skills in general and the rubric scores one piece of AI-assisted coursework, and neither measures unaided ability. They are different instruments with different objects; what the result shows is that the first cannot stand in for the second.

5.3 Reading the dimension-level pattern

On the account of Salomon et al. (1991), residual change requires the learner to keep directing the work. The dimension-level pattern here runs in the direction that argument predicts, most clearly for analytical depth. Analytical engagement and reduced offloading were not measured, so they enter this section as candidate mechanisms and not as demonstrated ones.

Section 2.3 divided the four rubric dimensions on that basis before the data were analysed. Retrieval and writing are steps a model can absorb almost entirely; synthesis and design justification are steps it can inform but cannot settle, and the briefs reserved them for the student. On raw p-values under both adjusted estimators, the difference appeared where the division said it would; with the Holm correction, only analytical depth remains significant, so the evidence is firm there and suggestive for research design coherence; in the unadjusted comparison, academic writing differs as well (Section 4.5).

Even the agreement under adjustment cannot be taken at face value, because scoring was not blind, and rater expectancy is a full alternative explanation of the condition difference. Each instructor scored the classes they taught, knew which had received the guidance, and could read the AI use statement attached to every script from that class (Section 3.3.1). An instructor who had delivered materials aimed at synthesis and design justification might expect, and so find, stronger work on those criteria, and a statement describing careful use of AI might itself incline a rater to credit the student's judgement. Analytical depth and research design coherence are also the dimensions that leave most to the assessor, since retrieval and writing are anchored by features that can be checked on the page. The dimension-level pattern is therefore predicted by expectancy and by the instructional account. One observation bears on the question without settling it: the same instructors scored the first assignment under the same conditions, and on it, the arms did not differ in total ($d = 0.05$; Table 6) or on either of these two dimensions. A general disposition to mark the guided class generously would have shown there too; a bias specific to the second task would not, and only blind scoring by a second rater could exclude it.

If the pattern reflects the guidance at all, the design of the guidance matters more than the presence of the tool. Bastani et al. (2024) found that the harm from unrestricted access to a GPT-4 assistant vanished once the assistant was rebuilt to withhold answers and prompt instead. Our intervention was built to withhold in the same sense (Section 3.2). Read alongside Bastani et al., and subject to the reservations just stated, the present result is consistent with the view that what separates benefit from harm is not how much GenAI is used but which steps it is allowed to perform.

The frequency result belongs here too. Reported use increased in both arms over the semester, slightly more in the comparison arm, so whatever distinguished the structured condition, it was not exposure to the tools. The contrast this study describes is therefore between guided and unguided use, not between use and abstention.

5.4 AI literacy goes with belief more than with performance

AI literacy did not moderate the difference between conditions on either outcome in the specified samples (Section 4.7 notes a subsample exception), which runs against a proposition that is widely assumed

in this literature and rarely tested against anything but self-report. Its direct associations divide along the same line as acceptance: a clear positive association with self-rated competency and no significant association with instructor-assessed performance.

Universities introducing AI literacy instruction often justify it on the grounds that students who understand these systems will use them more productively. Wang (2026) reports the supporting pattern, though on self-reported achievement. In this sample, prior AI literacy predicted how students rated their own research competency and was not significantly related to how their coursework was scored. These skills are plainly worth teaching, but this evidence does not support the expectation that they translate, on their own, into coursework that instructors assess more highly.

5.5 Implications for practice

For institutions, the weak correspondence between the self-rating and the instructor-assessed outcome argues against evaluating GenAI initiatives by student satisfaction or self-reported benefit alone. A serious evaluation needs an outcome the student does not supply, scored by someone who does not know which students took part.

For instructors, the dimension-level pattern suggests a design rule that is worth testing, although this study cannot confirm it. The task steps a model can complete end to end are the steps to reserve; the steps where it can be made to argue, compare, and object are where it earns its place.

For students, the discrepancy between the self-rating and the coursework score is the one worth passing on. We did not measure how fluent the work felt nor what students could do without the tool, so we cannot show that confidence outran capability. What we can say is that a student's rating of their own research competency carried little information about how the work was graded, and that this is consistent with a standing explanation in the self-regulated learning literature: the ease of an experience is a poor guide to what has been learned from it (Bjork et al., 2013).

5.6 Limitations

The assignment was by intact class and was not randomised, and the design contains only six assignment clusters, coinciding with the institution-by-condition cells. No cluster-level procedure is dependable at six clusters, and the student-level standard errors for H2 and H3 describe variation among students within these classes; at the level of assignment, there are three contrasts. The study also has no measure of coursework performance taken before the guidance was issued, so the adjusted and gain-score comparisons condition on a post-treatment variable, and unmeasured differences between classes cannot be ruled out. The condition difference therefore describes what was observed in six classes and is not an estimate of a causal effect.

Scoring was neither blind nor duplicated. Each institution's coursework was scored by the instructor who taught both of its classes, who knew the allocation and could read the AI use statements attached to scripts from the structured class, so the rater is confounded with the institution, and no inter-rater reliability can be computed. Behaviourally anchored descriptors constrain how far a rater's disposition can move a score, but nothing in the design excludes expectancy, and under the adjusted estimators, the difference appears on the dimensions that leave most to the assessor's judgement (Section 5.3). Expectancy is therefore a complete alternative explanation of the H2 and H3 results.

The two outcomes are not equivalent measures of one construct. The instructor-assessed outcome is assisted performance on two specific pieces of coursework: both conditions completed the assignments with GenAI available, and the study contains no delayed or unassisted assessment. The self-rating asks about research skills in general, so the divergence reported here is between a global self-perception and an instructor's evaluation of AI-assisted coursework, not between belief and unaided ability. The self-rating is also a new instrument examined in a single sample: a general factor dominates it, but a single-factor model does not fit well, and its relation to established measures is unknown. The evidence assembled here concerns internal structure and reliability only; convergent, discriminant, and criterion evidence has not been collected, and the scale should be treated as provisional until it has.

Attrition was substantial and, in the comparison arm, selective on perceived usefulness. Linkage relied on self-generated codes, so non-retention mixes dropout with failure to link, and for 27 members of the core sample, the code may not identify a single student, so their change scores may pair two different people (Section 3.5); H4a is judged on the securely linked subsample for that reason. Weighting for retention left the main associations unchanged, but it corrects only for what was measured at the pre-test, and two results proved sensitive: the association hypothesised in H4a is not found among securely linked students, and the apparent difference between arms in the change in perceived usefulness disappears under baseline adjustment and weighting. The negative coefficients on the instructor-assessed side come from the arm in which retention was selective. Self-rated competency was measured only at the post-test, by design (Section 3.3.2), so the two sides of the H4 comparison are not symmetric: one is a terminal level, the other a covariate-adjusted quantity with part of its variance removed. Item wording also changed between waves; invariance held for perceived usefulness but at best partly for perceived ease of use (Section 3.3.2).

Finally, the sample is management undergraduates at four institutions in one Chinese province, working with domestic models over a single semester, and the findings should not be generalised to other disciplines, regions, or education systems without replication. The intervention bundled a guidebook, video tutorials, in-class support, and structured assignment briefs, and the design cannot say which component, if any, did the work. Implementation fidelity was not assessed: no record was kept of how many tutorials each student watched, how closely each instructor followed the demonstration schedule, or how fully the briefs were applied, and delivery may have varied between instructors. Beyond a single frequency item, which did not separate the arms at the post-test, we have no process data on how students used these tools, so the mechanisms proposed in Section 5.2 remain inferences.

5.7 Future research

The most useful next step is process data. If the difference between benefit and harm lies in which steps a model is allowed to perform, then interaction logs, prompt records, and drafts at successive stages would let that claim be tested directly. A second step is a delayed unassisted assessment, which is the only way to separate what students can do with the tool from what remains without it. A third is to measure self-rated competency at both waves in a design where priming can be managed so that the two outcomes can be compared on equal terms. A fourth concerns the design itself: enough classes, randomly

allocated, to support inference at the level of assignment; a performance measure taken before any materials are released; and scripts scored blind by two raters.

5.8 Conclusion

Over one semester, the students who, by its end, regarded these tools as most useful and easiest to use were the ones who rated themselves the most capable researchers, and that rating was at best weakly related to how their instructors scored the coursework they had produced with those tools available. The self-rating was dominated by one general factor, where the instructors' scores were more differentiated. In each of three institutions, the class given structured guidance was scored higher on the second assignment, most clearly on analytical depth; with six non-randomised classes and non-blind scoring, that is a pattern worth testing properly and not an established effect. The two instruments do not measure the same thing, and that is the point: a design that only asks students how capable they have become cannot see how their work is judged.

Data availability statement

The datasets analysed for this study contain individual-level questionnaire responses and coursework scores collected under a consent form that did not cover public deposition, so they are not publicly archived. The Stata, Python and R scripts that reproduce every estimate reported here, including the factor models, the retention weights and the sensitivity analyses, are provided as Supplementary Data Sheet 4 without restriction; the de-identified data are available from the corresponding author on reasonable request.

Ethics statement

The study involving human participants was reviewed and approved by the Academic Affairs Committee of the School of Management, Guangzhou Huashang College (Ref. No.: GZHS-MGY-AAC-2026-002). The study was conducted in accordance with local legislation and institutional requirements. All participants were aged 18 years or older. Before completing the questionnaire, participants were informed of the study purpose, that their responses would be used for research purposes only, and that participation was voluntary. They were also informed that participation would not affect their course grades and that they could withdraw without penalty. Informed consent to participate was indicated by voluntary completion of the questionnaire.

Author contributions

XT: Writing – original draft, Methodology, Visualization, Investigation, Software, Project administration, Data curation, Funding acquisition, Conceptualization, Writing – review & editing, Formal analysis, Supervision. MZ: Data curation, Resources, Validation, Investigation, Writing – review & editing. ZH: Investigation, Validation, Resources, Data curation, Writing – review & editing. WZ: Data curation, Investigation, Writing – review & editing, Resources, Validation.

Funding

The author(s) declared that financial support was received for this work and/or its publication. This work was supported by the Research Talent Programme of Guangzhou Huashang College (Grant No. 2025HSPY12).

Conflict of interest

The author(s) declared that this work was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declared that Generative AI was used in the creation of this manuscript. The author(s) declare that Claude (Anthropic, claude.ai) was used for grammar checking, English language polishing and formatting adjustments to text written by the authors.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2026.2005055/full#supplementary-material

SUPPLEMENTARY DATA SHEET 1
Research instruments and study materials.

SUPPLEMENTARY DATA SHEET 2
Student GenAI guidebook.

SUPPLEMENTARY DATA SHEET 3
Supplementary Tables S1–S8 and Supplementary Tables S1–S8.

SUPPLEMENTARY DATA SHEET 4
Analysis scripts and supporting documentation.

References

Bastani, H., Bastani, O., Sungu, A., Ge, H., Kabakcı, Ö., and Mariman, R. (2024) Generative AI can harm Learning. SSRN Working Paper No. 4895486.

Bjork, R. A., Dunlosky, J., and Kornell, N. (2013). Self-regulated learning: beliefs, techniques, and illusions. *Annu. Rev. Psychol.* 64, 417–444. doi: 10.1146/annurev-psych-113011-143823

Carolus, A., Koch, M. J., Straka, S., Latoschik, M. E., and Wienrich, C. (2023). MAILS – Meta AI literacy scale: development and testing of an AI literacy questionnaire based on well-founded competency models and psychological change- and meta-competencies. *Comput. Hum. Behav. Artif. Hum.* 1:100014. doi: 10.1016/j.chbah.2023.100014

China Internet Network Information Center (2026). *The 57th Statistical Report on China's Internet Development*. Beijing: CNNIC.

Dahri, N. A., Yahaya, N., Al-Rahmi, W. M., Aldraiweesh, A., Alturki, U., Almutairy, S., et al. (2024). Extended TAM based acceptance of AI-powered ChatGPT for supporting metacognitive self-regulated learning in education: a mixed-methods study. *Heliyon* 10:e29317. doi: 10.1016/j.heliyon.2024.e29317

Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Q.* 13, 319–340. doi: 10.2307/249008

Deng, R., Jiang, M., Yu, X., Lu, Y., and Liu, S. (2025). Does ChatGPT enhance student learning? A systematic review and meta-analysis of experimental studies. *Comput. Educ.* 227:105224. doi: 10.1016/j.compedu.2024.105224

Dunning, D. (2011). "The Dunning–Kruger effect: on being ignorant of one's own ignorance," in *Advances in Experimental Social Psychology*, eds. J. M. Olson and M. P. Zanna, vol. 44 (San Diego: Academic Press), 247–296.

Fan, Y., Tang, L., Le, H., Shen, K., Tan, S., Zhao, Y., et al. (2025). Beware of metacognitive laziness: effects of generative artificial intelligence on learning motivation, processes, and performance. *Br. J. Educ. Technol.* 56, 489–530. doi: 10.1111/bjet.13544

Fernandes, D., Villa, S., Nicholls, S., Haavisto, O., Buschek, D., Schmidt, A., et al. (2026). AI makes you smarter but none the wiser: the disconnect between performance and metacognition. *Comput. Hum. Behav.* 175:108779. doi: 10.1016/j.chb.2025.108779

Laupichler, M. C., Aster, A., Haverkamp, N., and Raupach, T. (2023). Development of the "scale for the assessment of non-experts' AI literacy" – an exploratory factor analysis. *Comput. Hum. Behav. Rep.* 12:100338. doi: 10.1016/j.chbr.2023.100338

Lee, H.-P., Sarkar, A., Tankelevitch, L., Drosos, I., Rintel, S., Banks, R., et al. (2025). *The impact of Generative AI on Critical Thinking: Self-Reported Reductions in Cognitive Effort and Confidence Effects from a Survey of Knowledge Workers Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. New York: ACM.

Li, Y., Sadiq, G., Qambar, G., and Zheng, P. (2025). The impact of students' use of ChatGPT on their research skills: the mediating effects of autonomous motivation, engagement, and self-directed learning. *Educ. Inf. Technol.* 30, 4185–4216. doi: 10.1007/s10639-024-12981-9

Liu, C., Xie, L., and Xu, G. (2026). Generative AI in higher education: a meta-analysis of intellectual and social–emotional outcomes. *Front. Psychol.* 17:1848745. doi: 10.3389/fpsyg.2026.1848745

Liu, Z., Zuo, H., and Lu, Y. (2025). The impact of ChatGPT on students' academic achievement: a meta-analysis. *J. Comput. Assist. Learn.* 41:e70096. doi: 10.1111/jcal.70096

MacKinnon, J. G., Nielsen, M. Ø., and Webb, M. D. (2023). Cluster-robust inference: A guide to empirical practice. *J. Econom.* 232, 272–299. doi: 10.1016/j.jeconom.2022.04.001

Ng, D. T. K., Leung, J. K. L., Chu, S. K. W., and Qiao, M. S. (2021). Conceptualizing AI literacy: an exploratory review. *Comput. Educ.: Artif. Intell.* 2:100041. doi: 10.1016/j.caeai.2021.100041

Panadero, E., Brown, G. T. L., and Strijbos, J.-W. (2016). The future of student self-assessment: a review of known unknowns and potential directions. *Educ. Psychol. Rev.* 28, 803–830. doi: 10.1007/s10648-015-9350-2

Salomon, G., Perkins, D. N., and Globerson, T. (1991). Partners in cognition: extending human intelligence with intelligent technologies. *Educ. Res.* 20, 2–9. doi: 10.3102/0013189X020003002

Stephenson, R., and Armstrong, C. (2026). *Student Generative AI Survey 2026 (HEPI Report 199)*. Oxford: Higher Education Policy Institute.

van Breukelen, G. J. P. (2006). ANCOVA versus change from baseline had more power in randomized studies and more bias in nonrandomized studies. *J. Clin. Epidemiol.* 59, 920–925. doi: 10.1016/j.jclinepi.2006.02.007

Wang, Y. (2026). The effect of generative artificial intelligence literacy on academic achievement: the mediating role of academic self-efficacy and the moderating role of critical thinking. *Front. Psychol.* 17:1761562. doi: 10.3389/fpsyg.2026.1761562

Xu, S., and Zhang, G. (2026). Applying UTAUT2 model to investigate acceptance of DeepSeek by Chinese university students for academic use. *Front. Psychol.* 17:1828331. doi: 10.3389/fpsyg.2026.1828331

Yang, F. (2026). ChatGPT acceptance and use in higher education: evidence from a systematic review using the technology acceptance model. *TechTrends* 70, 466–483. doi: 10.1007/s11528-026-01166-5

Zell, E., and Krizan, Z. (2014). Do people have insight into their abilities? A meta-synthesis. *Perspect. Psychol. Sci.* 9, 111–125. doi: 10.1177/1745691613518075

Zhai, C., Wibowo, S., and Li, L. D. (2024). The effects of over-reliance on AI dialogue systems on students' cognitive abilities: a systematic review. *Smart Learn. Environ.* 11:28. doi: 10.1186/s40561-024-00316-7

Zhu, Q., Li, X., Dong, Y., Chang, P., and Fan, M. (2026). Not all cognitive offloading is equal: distinguishing dependent and autonomous offloading to generative AI. *Front. Psychol.* 17:1878629. doi: 10.3389/fpsyg.2026.1878629

Zhu, H., and Yang, S. (2026). The impact of generative AI use on graduate students' research competence: the mediating role of critical thinking and the moderating role of research self-efficacy. *Behav Sci* 16:304. doi: 10.3390/bs16020304