sustainability

MDPI

Article

From Answer-Seeking to Reasoning: A Request Protocol for Sustainable Integration of Generative AI in Engineering Graphics Education

Jue Li [1,2], Hengwu Hu [1,3,*] and Chao Huang [4]

[1] Key Laboratory of Highway Engineering of Ministry of Education, Changsha University of Science & Technology, Changsha 410114, China
[2] College of Traffic & Transportation, Chongqing Jiaotong University, Chongqing 400074, China
[3] Department of Project Management, Hunan Vocational College of Engineering, Changsha 410151, China
[4] School of Civil Engineering, Central South University, Changsha 410004, China
* Correspondence: huhengwu07@163.com

Abstract

Sustainable Development Goal 4 (SDG 4) emphasizes inclusive and equitable quality education and lifelong learning opportunities for all. Within this broader framework, the sustainable integration of generative artificial intelligence (AI) in engineering education raises a specific question: whether students can retain and apply the competencies developed during AI-supported learning when the tool is no longer used. Generative AI can provide learning support beyond limited classroom hours, but it can also perform the cognitive operations that students are expected to master. Using an engineering graphics course, this study compared three strategies of AI use for their effects on how long student competence lasts. The three strategies were no AI, unrestricted AI use, and a request protocol. Under the protocol, students may ask questions about work they have already produced, but they may not ask for the answer. Competence was measured four times under an instruction not to use the tool, and the last measurement fell four weeks after the course ended. The ranking changed after withdrawal. Immediately after instruction, the unrestricted group scored highest before adjustment, but the adjusted between-group differences were not significant. At the delayed test the protocol group held the largest advantage of any contrast and the unrestricted group fell to last. Both adjusted differences were significant at $p < 0.001$. The advantage was largest among students who started lowest and was no longer distinguishable from zero above a baseline score of 64.95. The difference between the two tool conditions sat in classification rather than drafting. Curve-type misjudgment was about six times as frequent in the unrestricted condition as under the protocol, while the drafting conventions did not differ detectably at the end of term. Protocol adherence acted as a graded skill rather than a compliance record and tracked delayed competence ($r = 0.63$, $n = 33$). In this 48 h foundation course, regulating what students asked, not how much they used the tool, was associated with competence that persisted after withdrawal.

Keywords: generative artificial intelligence; engineering graphics education; request protocol; instructional design; sustainable education; educational technology

Check for updates

Academic Editor: Aharon Gero

Received: 28 September 2026
Revised: 30 October 2026
Accepted: 3 November 2026
Published: 5 November 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

The sustainable development agenda asks engineering education to build technical ability that lasts, rather than a skill that expires with each generation of tools [1]. The 2030

Agenda for Sustainable Development defines SDG 4 as ensuring inclusive and equitable quality education and lifelong learning opportunities for all [2]. Within this framework, engineering education is expected to train graduates who can turn technical knowledge into responsible, durably operating solutions [3]. Generative artificial intelligence (AI) has meanwhile entered the learning routines of most students [4]. The tool can supply practice outside class hours and give every student feedback at any time. This shift moved the choice educators face from whether to allow AI to the role AI may play for students.

Engineering graphics shows this change most directly. The course is the first technical foundation course in mechanical, civil and cognate engineering programs [5]. It carries two tasks, training spatial thinking and enforcing the conventions of engineering representation. Its core is not the operation of drawing software but the ability to move in both directions between a solid and its projection [6]. Spatial visualization is trainable rather than a fixed endowment, and the graphics course is one of the main places where it is trained. This two-way transformation is exactly the operation language models replace the best. A course built on that operation has the most to lose from unrestricted use. The practice space of the course is also shrinking. To meet the goals of Chinese New Engineering Disciplines initiative, universities have added applied courses in recent years, and hours for foundation courses keep being squeezed [7]. Some curricula have merged engineering graphics with hands-on computer-aided drawing into a single course of about 48 h. Quantified assessment concentrates weight on a few gradable outputs, which makes process practice harder to protect [8]. Cuts fall on practice first, and the most practice-dependent part, the intersection of a plane with a solid and the intersection of two solids of revolution is cut first [9]. The correctness of these operations can only be derived from the geometry of the two surfaces. Neither memory nor imitation passes the examination. Generative AI looks like a remedy for this shortfall, and courses have good reason to admit it. These operations, however, are also the ones a language model performs on its own, so they are the ones students can delegate most completely once the tool is available.

Research on generative AI in engineering education has followed two lines, application effects and teacher perceptions. In a foundation course for an AI major, GenAI showed direct positive effects on method and ability, with indirect effects on professional literacy [10]. In a design drawing course, integrating generative AI improved learning experience and outcomes, and the weakest entrants gained the most [11]. A survey of 266 pre-service vocational teachers found AI knowledge partially mediating the link between perceived benefit and perceived value [12]. The potential is real and greatest among weakly prepared students. All of these studies, however, rest on data collected while the tool was available. Whether students can still perform weeks after withdrawal remains largely untested. The evidence that does exist points the other way. In a randomized trial of about 1000 high school students, unrestricted GPT-4 use raised performance during assisted practice but scored below a no-AI control on the unaided test after withdrawal. A tutor interface that gave hints without answers avoided this harm [13]. Heavier AI use is also linked to lower critical-thinking scores in survey evidence [14]. The mechanism is cognitive offloading, the delegation of a cognitive operation to an external device [15]. Language models extend the range of delegation from single steps to whole chains of reasoning. A student who delegates the entire chain bypasses the very competence the practice was meant to build. The retrieval interruption framework adds that help given before learners attempt recall or self-explanation can raise immediate performance while weakening delayed recall and transfer [16]. None of this evidence comes from a course whose core operation the model can perform end to end, and none tests a rule that restricts what students ask rather than how much they use the tool.

In the graphics course the default request students make of AI is for the answer, the shape of a section curve, the position of an intersection point, or the finished drawing. Deriving these results is the practice itself. The closest evidence comes from design drawing courses and teacher attitude surveys [12], and neither tests whether ability persists without the tool. Productive failure research shows that fluent success can leave less conceptual knowledge than struggling forward [17]. Students who restudied material were ahead immediately and behind a week later [18]. The assistance dilemma describes help as nonmonotonic. More help removes unproductive floundering but thins processing [19]. The standard response is to fade scaffolding as expertise grows, but that assumes the instructor holds the scaffold and decides when to remove it [20]. Once students hold the model themselves, that control is gone. The only variable a course can still regulate is the kind of request students are allowed to make of the tool.

Generative learning research supplies a way to use that variable. Self-explaining and drawing produce durable understanding, distinct from activities that raise performance in the moment [21]. A meta-analysis of 69 effect sizes puts the mean benefit of self-explaining prompts at $g = 0.55$ [22]. Cognitive load theory adds that only extraneous load, created by the way material is presented, can be reduced without cost. Cutting the intrinsic processing involved in schema construction harms learning [23]. These results separate help into two layers. Help at the layer of expression works from the student's own output, the written description, its completeness, and the reasoning behind it. The transformation stays with the student. Help at the layer of the operation supplies the curve shape, the key-point position or the final answer, and performs the transformation in the student's place. The difference is one of kind, not degree. Constraining requests to the first layer turns the student's behavior from asking for answers into stating grounds and answering questions about their own work. Within the broader objectives of SDG 4, this study focuses on one specific dimension of sustainable engineering education: the development and retention of task-specific competence that students can demonstrate without AI assistance. This focus does not redefine SDG 4; rather, it provides an operational criterion for examining whether AI-supported learning contributes to competence beyond immediate tool-assisted performance. Such an integration is judged by three criteria. The competence must survive withdrawal. The arrangement must require no recurring new contact hours, equipment or procurement. The rule must be carried by the task structure rather than by the teacher's continuing presence.

This study aims to examine how the way students use generative AI in engineering graphics education is associated with the retention of task-specific competence and the reasoning processes involved in learning. Rather than focusing solely on whether AI improves immediate task performance, the study investigates whether different forms of AI assistance are associated with competence that students can demonstrate independently after formal instruction has ended. Two research questions guide the investigation:

(1) How do different AI-use strategies relate to students' retention of task-specific competence, and in which aspects of geometric reasoning and drawing performance are the differences most evident?

(2) How do students' AI-request behaviors differ between unrestricted and protocol-guided use, and how is adherence to the request protocol associated with delayed competence?

To address these questions, a quasi-experimental study was conducted with 100 engineering undergraduates who self-selected into three conditions: no AI, unrestricted AI use, and protocol-guided AI use. Task-specific competence was assessed at four measurement points, including a delayed assessment four weeks after the end of the course. Drawing errors, transfer performance, and submitted AI dialogues were also examined to characterize differences in learning outcomes and request behaviors. This design distinguishes

immediate performance from delayed competence while exploring how the structure of students' requests relates to their independent learning outcomes.

2. Course Context and Students at Entry

The course is a single section of about 100 first-year engineering undergraduates. Before any instruction on projection, the students took the eight-item test used in this study on the course platform, on their own initiative, in their own time and outside class hours. The mean was 48.5 out of 100 with a standard deviation of 30.9. Of the cohort, 31% scored below 30, and 28% scored 70 or above. Entry scores spread this wide are a problem for the weakest students. Many of them held some prior idea of projection, and some of these ideas were wrong but plausible: a drawing can look right while the projection behind it is not. The weakest students are the least able to tell the difference. The practice that would close this gap cannot come from class hours. It has to happen outside class, which is where students in this course already work, and where an unguided tool hands them the answer instead of the practice.

The entrance examination did not account for this spread. The mathematics component of the exam correlated almost zero with the baseline score ($r = -0.02$). Students who scored high on it did no better on the baseline than those who scored low. We therefore used the baseline score, not the entrance examination, as the pre-test control in the analyses.

In the first weeks, students could copy the worked solution and produce a correct drawing. Asked what the configuration was or how they would recognize it again, they could not answer. The module on intersection curves made this weakness visible. A typical case: one student drew a single smooth curve where the correct answer was two straight lines. We asked what would have to be true about the two cylinders for the answer to change. The student could not answer. Two features of this failure set the requirements for the design in Section 3. First, the error does not show in the drawing itself, because the wrong curve looks plausible on the page. Second, the student misclassified the case rather than drawing poorly, so more drawing practice would not fix it. A design that protects learning in this course has to protect the classification step, not the drawing hand.

3. The Request Protocol and Instructional Design

3.1. Design Principle

The protocol regulates two kinds of help, and the difference between them is who performs the transformation. At the expression layer, the tool works from what the student has already produced: it reads the written description of the configuration, checks the description for completeness, and questions the reasoning behind it. The transformation itself still belongs to the student. At the operation layer, the tool performs the transformation for the student by supplying the curve shape, the key-point position, or the final answer. On a task of this kind, such help empties the exercise. The student can finish the session with a correct drawing and never perform the operation the drawing was meant to teach. The protocol therefore governs what the tool may be asked to do, not how often it is used.

Figure 1 shows how the 48 contact hours and the assessment weight are distributed across the curricular objectives. The objective at the center of this study carries 35% of the assessment weight but only 12.5% of the contact hours, six hours in total, and the derivation of the intersection curve takes about one of them. Practice at this scale has to move outside class, where the tool already sits with every student. Three options remain. Giving the transformation to the tool leaves the course nothing to teach. Leaving the student unaided is more than one hour of instruction can support. The third option places the help at the expression layer: the student describes the configuration, the tool questions the description,

and the student performs the transformation. This option adds no contact hours, requires no equipment or procurement, and the operation is still learned.

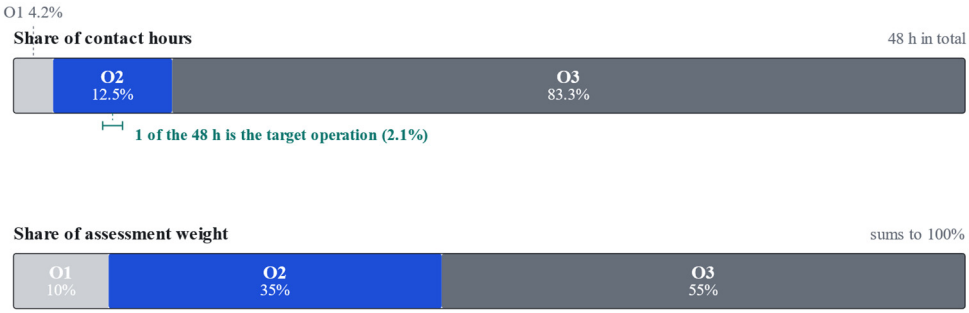


Figure 1. Contact hours and assessment weight by curricular objective.

The protocol holds only if no deliverable can be produced by the tool. That takes two closures, one on what may be asked and one on what may be submitted. The card closes the first, since every route to the answer is prohibited on it. The rule that every drawing be produced without the tool closes the second.

3.2. The Four Permitted Requests

The protocol is handed out as a printed card in the first class, as shown in Figure 2. A request may take one of four forms: the student describes the configuration in their own words and asks the tool to check the description; states how far their reasoning has reached and asks the tool to test it; asks what is missing from the solution; or asks whether the conventions follow the standard. Asking for the answer is not permitted, nor is asking for the shape of a curve, the position of a key point, a classification of the configuration, or a drawing to copy. For each prohibited request the card states what the student must do instead, and for each permitted form it gives a sample request the student can copy.

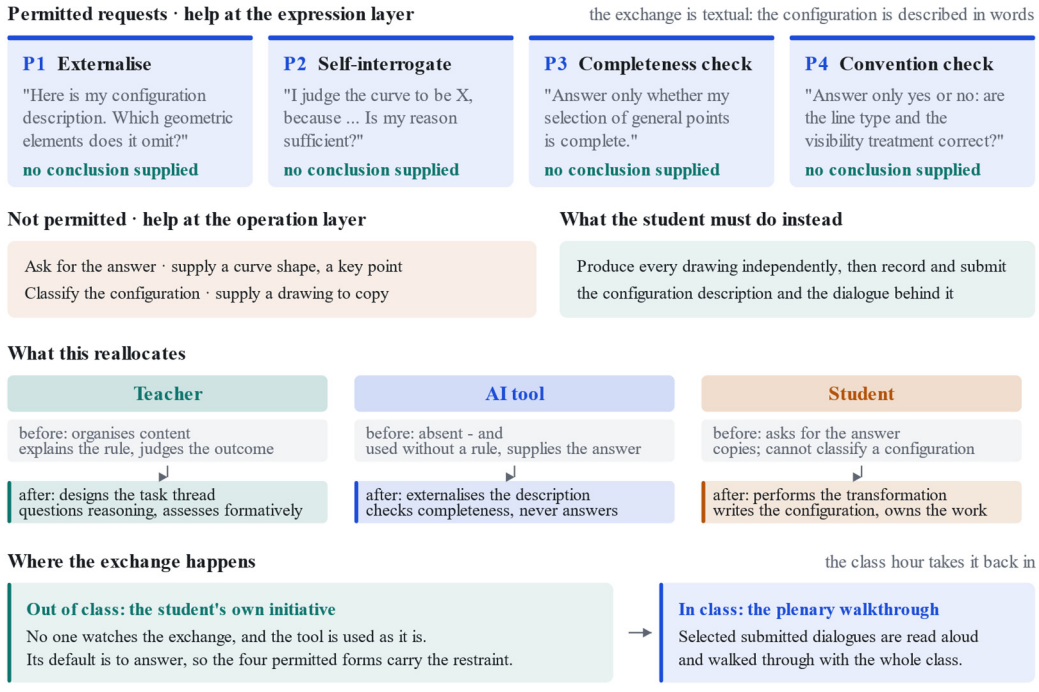


Figure 2. The four permitted request patterns, the prohibited requests, and the roles the protocol assigns to teacher, tool, and student.

Three choices in the design determine how the protocol works in practice. First, the exchange is textual. A figure problem must pass through the student's own written description of the configuration. At the time of the study, the tool handled geometric input unreliably [24]. The stronger reason is that stating a shape precisely enough for someone else to check is itself a skill the course aims to teach, and the textual route trains it. Second, the rule relies on the student's self-discipline. The tool is used without modification and without anyone watching the exchange. A student who asks for the answer receives it. The card therefore gives the student four permitted forms, so that following the rule means asking one of them rather than the answer. The tool also sits outside the course platform. It is a free public service that students reach from their own devices and accounts, and access control was neither possible nor attempted. Both tool conditions used the DeepSeek web interface. In standard chat it served DeepSeek-V3, with V3-0324 in place from late March 2025; with DeepThink enabled it served the DeepSeek-R1 line, with R1-0528 in place from late May 2025. Both modes were reachable through the same interface throughout, and students could use either, with no configuration, no system prompt and no course-specific setup. Third, every drawing is produced without the tool. The card governs what the student may ask, and this rule governs what the student may submit. If the tool can produce the deliverable, using it is the fastest way to a finished drawing. A ban that leaves this shortcut open secures only surface compliance.

3.3. Obligations of Teacher, Model and Student

The protocol assigns a fixed set of duties to each party, and the lower part of Figure 2 summarizes the shift. The teacher designs the task thread and questions reasoning, instead of explaining content and judging each attempt as it is made. The tool restates and checks the student's own description, instead of supplying answers. The student performs the transformation and owns the deliverable, instead of asking for the answer. The exchanges themselves happen outside class, on the student's own initiative, and they return to the classroom in one form: the teacher reads selected submitted dialogues aloud and walks through them with the whole class. This class hour is attended by all three conditions, and every example is anonymized before it is read out. The walkthrough is therefore not exclusive to the protocol group, and it cannot by itself separate the conditions. What the protocol group had that the other two did not was the rule to apply to those examples. This walkthrough brings the reasoning practiced with the machine into the open. It also sustains the rule without any monitoring of the exchanges, which is consistent with the self-discipline on which the protocol rests. In full, the tool may restate a description, question the reasoning behind it, check the description for completeness, and check whether the conventions follow the standard. It may not supply the answer, in any of the forms the card names. The student learns to ask in a way that helps without doing the work.

These role changes are consistent with recent literature. Studies of classroom teaching with machine support describe a shift from a teacher–student structure to a teacher-machine-student structure, with the teacher moving toward designing experience, questioning reasoning, and formative assessment [25]. The same literature warns that students who work with such tools without metacognitive support show declining self-regulation [26]. In this course the tool is used outside class and no one watches the exchanges, so the design has to bring that use back into the classroom in another form. The submission of dialogues and the plenary walkthrough serve this purpose. Student literacy frameworks make the same distinction, between operating a tool and directing it [27], and directing the tool is what the protocol trains. This division of roles is also what keeps the arrangement sustainable in engineering education. The teacher's new duties, designing tasks and questioning reasoning,

remain necessary when the tool is replaced by the next one. Training students to direct the tool gives them the same durability.

3.4. Course Organization Around One Task

The course is organized around one task that runs the whole term, the complete graphical representation of a tubular truss, whose critical detail is the curve where a web member meets a chord. The design follows the whole-task approach to complex learning, in which learners work on simplified but complete versions of the target task instead of accumulating separate parts [28]. In this course the obstacles that scarce hours leave untaught stand in the way of a task the student already wants to finish.

The detail is a suitable bottleneck for two reasons. Its correctness follows from geometry, so an answer can be checked without relying on judgement. It also cannot be copied, because what must be reproduced is the relation between two cylinders, not a line on the sheet. The task is unlocked one obstacle per module. In the first class, students attempt it and stall, because they have not yet met projection. Each later module removes one obstacle, until the module on section and intersection curves reaches the critical detail, where the protocol applies. Six deliverables accumulate over the term, the initial attempt with a written list of what the student cannot yet draw, line conventions, the node detail with the intersection curve, split views and dimensioning, a CAD-digitized drawing, and the complete package. Progress is scored on a six-point ordinal scale. Between deliverables, students work out of class. This arrangement carries the practice that the class hours cannot accommodate, and it adds nothing to the contact hours. Figure 3 shows the arrangement. Twelve modules span the 48 contact hours, the six deliverables are attached to the modules that unlock them, and the out-of-class task thread runs underneath.

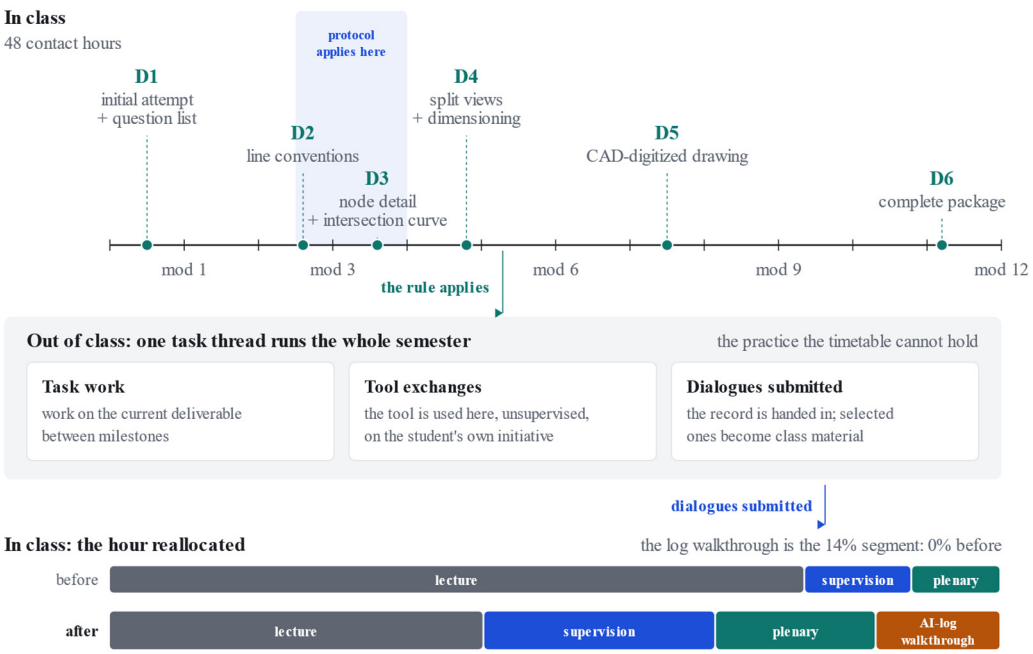


Figure 3. One semester-spanning task, unlocked one obstacle per module.

Two consequences of this organization shaped the design. Students have a reason to care about topics that would otherwise arrive as abstract geometry, because each one is the next obstacle to a finished drawing [29]. Difficulty, instruction and a question worth asking all arrive in the same week, which no timetable achieves on its own. The timing matters for the protocol. The tool is used outside class, and the task makes sure there is something worth asking about at exactly that point.

4. Methods

4.1. Participants and Conditions

The study included 100 undergraduate students enrolled in an engineering graphics course at a university in China. The participants came predominantly from transportation-related engineering programs. According to the administrative enrollment records, 100 students belonged to six transportation-category classes with 2024 entry-year codes. The sample comprised 76 male students (76%) and 24 female students (24%). As documented in the study, all participants were aged 18 years or older; however, individual ages were not available in the demographic records used for this analysis. The participants followed the same 48 h course schedule and completed the same instructional activities and assessments.

The students selected one of three AI-use conditions: no AI (C0, $n = 34$), unrestricted AI use (C1, $n = 33$), or protocol-guided AI use (C2, $n = 33$). All three conditions were compared within the same course cohort. The available administrative records describe the demographic composition of the overall sample but do not establish gender or academic-background distributions within each experimental condition.

Students formed the three groups themselves. After the teaching team explained what each condition required, that is, no AI, unrestricted AI, or the request protocol, each student joined the group that matched their choice. No student was assigned by the teaching team, and no randomization was used at any point. The design is therefore quasi-experimental, and the composition of the three groups is open to self-selection. Whether this selection favored any condition is not known, although students who expected to rely on the tool may have gravitated to the unrestricted condition and students disposed to work carefully may have gravitated to the protocol. If that is what happened, the contrast overstates what the rule contributed, and this design cannot separate the two. We therefore report the between-condition differences as associations rather than as effects of the rule.

Two attainment variables enter the analyses as controls. The baseline score is the eight-item composite measured before any instruction on projection, and it serves as the pre-test. Prior attainment is the score on the mathematics component of the entry examination, taken from institutional records and expressed in cohort standard-deviation units. It serves only as a covariate.

Three elements of the course were unique to the protocol condition. The first was the protocol card, which listed the four permitted request forms, gave a workable sample of each, and specified a substitute action for every prohibited request. The second was the instruction that a prohibited request be replaced by its substitute rather than simply dropped. The third was the rule that every drawing be produced without the tool. The pairing of what may not be asked with what the student does instead is the treatment itself. All three conditions shared the cohort, the timetable, the task thread, the six deliverables, the assessment conditions and the plenary walkthrough of submitted requests. The study does not separate the contribution of the rule from that of these accompanying supports.

A student's condition could not be concealed, because each condition is defined by the rule its members must follow. The protocol card, the four permitted forms and the paired prohibited requests with their substitutes were explained to the protocol group in the first class. The exchanges took place outside class at the student's initiative, and the rule was restated at each module. Because students in one cohort interact freely, contamination between conditions cannot be measured or excluded.

4.2. Measurement Points and Availability of AI

The eight-item test ran as a practice task on the course platform rather than as an in-class examination. A version was released at each measurement point, and students

completed it on their own initiative, outside class hours and without supervision. The item type was held constant, and the geometric parameters were changed at every point, so an answer from an earlier point did not transfer. Each practice was to be completed without the tool. This was an instruction rather than a technical restriction, and we do not claim that every student complied. The conditions differed in the rule, not in access. The tool was available to every student outside class throughout the term. Before the M1 practice, the target module had been taught, and the protocol had been in force for that module only. Between M1 and M2 the tool remained available outside class to everyone. C2 kept to the protocol, C0 was asked to work without it, and C1 was given no rule. Between M2 and M3 the term ended, and the four-week interval fell entirely after it. The task points on the teaching platform were closed, and no instruction or task work was set in any condition. Whether students studied the topic on their own during the interval was not recorded, so contact with the topic outside the course cannot be excluded. Withdrawal, as the term is used here, refers to the end of course work on the topic. It does not record whether students used the tool during the interval. The delayed practice was answered through the course platform under the study code and carried no name field. The C2 protocol was therefore a written rule with no scheduled work to apply it to. The course ran from March to May 2025, and every exchange submitted for the study fell inside that period. The baseline practice followed the first class, the end-of-term practice fell in the final week, and the delayed practice took place four weeks after the term ended. Table 1 summarizes the timing of each point, the item family used there, and the rule in force at the time.

Table 1. Measurement schedule and the availability of AI at each point.

Point	Timing	Content	AI During Measurement	Rule in Force
M0	After the first class, on the platform, before instruction	8 items, family A	No	None; protocol not yet in force
M1	Immediately after the target module	8 items, family B	No	Protocol, target module only (C2)
M2	End of term	8 items, family C	No	Protocol for the whole term (C2); none for C0, C1
M3	Four weeks after the end of term	8 items, family D; near transfer	No	No teaching or task work in any condition

Note: The entry "No" under AI during measurement records the instruction given to students, not a condition verified by supervision or by a technical block.

The eight-item assessment was developed specifically to measure the target geometric transformation taught in the course, rather than general spatial ability or overall engineering competence. Its content was aligned with the curricular objective of interpreting and constructing section and intersection representations of solids of revolution. All items belonged to the same parameterized problem family, with geometric parameters varied across measurement points while the underlying task structure and scoring criteria remained unchanged. Correct responses were determined using analytically generated geometric standards, allowing the assessment to evaluate the specified operation against a consistent reference.

4.3. Measures and Scoring

Table 2 lists the variables, the instrument for each, and its role in the analysis. The primary outcomes are competence at the target transformation and error-type incidence. Drawing quality at the end-of-term measurement and near transfer at the delayed measure-

ment are secondary. Two process variables, generative processing and cognitive ease, record how students worked through the practice. Three role variables describe how students used the tool. These are protocol adherence, question quality, and autonomy and role perception.

Table 2. Variables, instruments and role in the analysis.

Variable	Instrument	Role
Competence at the target transformation	Parameterized item family, each item scored 0–2 against an analytically generated standard, with 10% human audit	Primary
Error-type incidence	E1–E5 classification coding of drawings, coded blind to condition	Primary
Drawing quality	Four-dimension rubric (projection relations, line conventions, completeness, curve accuracy), 1–5, coded blind to condition; no agreement coefficient estimated	Secondary
Near transfer	Same structure, a different component family, at the delayed measurement	Secondary
Generative processing	Coded from process records, 0–3	Process
Cognitive ease	Practice-stage self-report, 1–9	Process
Protocol adherence	Coded from submitted dialogue records, 0–1	Role
Question quality	Coded from submitted dialogue records, 1–5, coded blind to condition; no agreement coefficient estimated	Role
Autonomy and role perception	1–5, pre and post	Role
Out-of-class study time	Self-reported log, hours	Covariate

The primary outcome is competence at the target operation. The operation moves in both directions between a solid of revolution and its two-dimensional section or intersection representation. The three-dimensional intersection of the two surfaces and its appearance in a given projection are not the same object. The assessment scores the projection, and the error typology classifies the configuration behind it. Throughout this paper, competence means this narrow, task-specific ability, not general engineering competence or design capability. The construct is deliberately narrower than spatial ability, a relatively stable individual difference that is not expected to move within one term, and this narrowness is what makes the delayed transfer test informative. The measure has eight parameterized items of the same type, each scored against a standard computed analytically from the item's parameters. The item type and the scoring rule were fixed at all four points, so the composite is comparable across them. Every condition answered the same version at each point, so a difference in form difficulty would move all three conditions together and cannot by itself produce a between-condition difference. The case used in instruction is a chord of Ø219 mm with a branch of Ø114 mm meeting it at 90° with intersecting axes, a diameter ratio of 0.52. Each item carries a score of 0, 1 or 2 read off that standard, so the composite records how much of the operation a student can complete, not only whether the final answer is right. The eight item scores are summed and expressed on a 0 to 100 scale as 100 times the sum divided by 16. Figure 4 shows one item of the family with its standard and its three orthographic views at a single scale.

The error typology serves as a measurement instrument and is reported as a distribution rather than as a score. E1 locates only extreme points. E2 assigns visibility segments incorrectly. E3 ends a contour line away from the intersection point. E4 misjudges the type of the section or intersection curve. E5 replaces the curve with a circular arc. E4 is the most informative of the five because the geometry of the two solids settles it completely, and it is the error the protocol exists to prevent. It is therefore the error type tracked in the

drawing analysis. E4 records whether the family of the section or intersection curve was identified correctly, and it is coded dichotomously because that identification is either right or wrong. The error code is read from the drawing that each practice requires, one drawing per student per point, so the composite and the error code describe the same response rather than two instruments. The quality of execution once the family is right is graded separately by the curve accuracy dimension of the rubric (1–5). E4 detects a wrong type decision, and curve accuracy grades the execution of a right one.

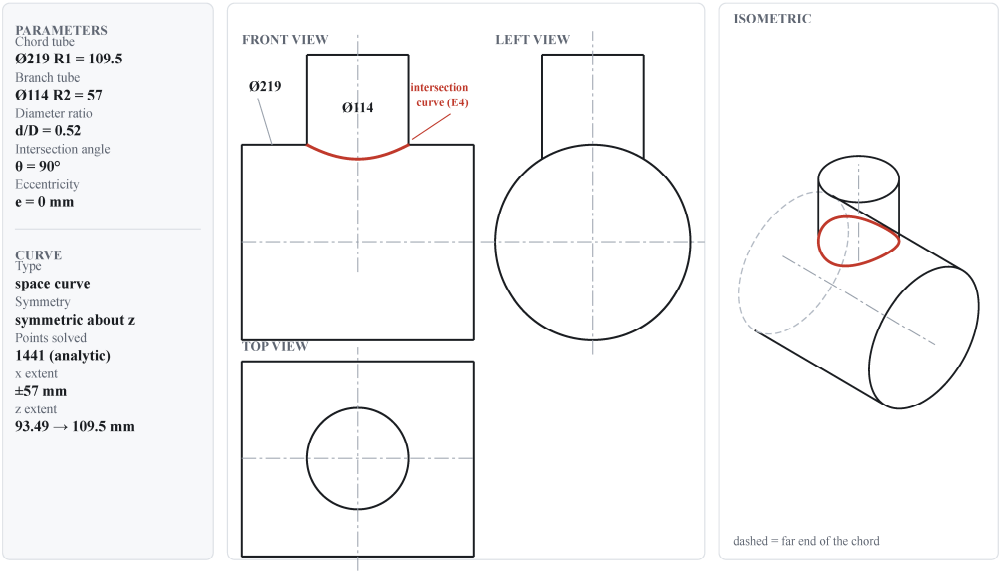


Figure 4. One item of the semester task with the standard it is scored against, shown in three orthographic views at a single scale.

The two primary contrasts are C2 versus C0 and C2 versus C1 at the delayed measurement. Drawing quality on the four rubric dimensions and near transfer are secondary. Everything else is exploratory, including the immediate and end-of-term contrasts, the five error types taken individually apart from E4, the two process variables, the role analyses within C2 and the individual rubric dimensions. No multiplicity correction is applied to the exploratory set, because that set was never intended as a test and no pre-specified family exists. The exploratory comparisons are labelled as such in the text and in the table notes, and the Conclusions rest only on the two primary contrasts. Effect estimates and confidence intervals accompany the *p*-values for the contrasts that carry the argument.

Two models carry the main analyses, both estimated by ordinary least squares. The first takes the delayed score as the outcome with three covariates. The second is a difference-in-differences model estimated on the pooled student-by-point data. No student fixed effects are included, and standard errors are clustered by student (100 clusters). Johnson–Neyman boundaries are obtained from the estimated coefficient covariance matrix of the interaction model. Full coefficient tables, standard errors and residual degrees of freedom are given in Appendix A.

4.4. Coding of Protocol Adherence

Protocol adherence was measured from the submitted dialogue records rather than from self-report, and was coded on three dimensions combined into one score. Structure conformity was scored from 0 to 3, with 3 meaning that all requests conformed to the protocol and at least two of the four forms were used, and 0 meaning that the rule was not applied at all, either because the student asked for the final answer or because the student submitted no exchange in which the protocol could be applied. The two routes to 0

are distinct behaviors and are scored alike because both record a failure to apply the rule. Conclusion-seeking frequency was the proportion of requests asking for the final answer, the shape, a key-point position, or the classification of the configuration. Self-attempt priority was coded as a shortfall from 0 to 3, where 0 means the student had carried the attempt as far as the item allowed before asking and 3 means the request came before any attempt. It was read from the student's own statement of how far they had got, checked against the first turn of the dialogue, and the dialogue took precedence where the two disagreed. The composite was $A_i = 0.5 \cdot (s_i/3) + 0.3 \cdot (1 - p_i) + 0.2 \cdot (1 - r_i/3)$, where s_i is structure conformity, p_i is conclusion-seeking frequency and r_i is the self-attempt shortfall, so each term scores higher as the student departs less from the rule. The weights were fixed from the design rationale before the data were seen. The composite is used inferentially only in the within-condition analysis of the protocol group, whereas the corresponding value for C1 is reported for descriptive purposes only. A complete record is one submitted window containing both sides of every exchange for the module, in sequence.

The weights order the three dimensions by how central each one is to the rule itself, with structure conformity first, conclusion seeking second and self-attempt priority third. Structure conformity carries the largest weight because it is conformity to the rule itself, and it is the only dimension that can be scored for every request in a record. Conclusion-seeking frequency enters with its sign reversed because it is the prohibited behavior. It is weighted below conformity because a student who asks for the answer once in an otherwise conforming record is not equivalent to a student who never conformed. Self-attempt priority carries the smallest weight. It is a precondition for the request rather than a property of it, it is the least reliably coded of the three, and it can be checked only against the first turn of the dialogue. The ordering $0.5 > 0.3 > 0.2$ follows from these roles. The particular values are a design choice.

Because the individual dimension scores were not retained per student, the robustness of the composite to the choice of weights could not be examined. The composite is not a calibrated scale. It is used only within the protocol condition, where it describes how closely the rule was followed. The scoring rules of all three dimensions are stated above in enough detail for another coder to reproduce the composite.

4.5. Reliability, Ethics, and Blinding

Internal consistency of the eight-item composite was high at every measurement point, as shown in Table 3. Cronbach's α was 0.893 at M0, 0.911 at M1, 0.892 at M2 and 0.893 at M3, and the mean inter-item correlation at the delayed measurement was 0.512. Because each item carries partial credit, part of this consistency reflects the rubric itself, not only the items.

Table 3. Reliability of the eight-item competence composite.

Index	Value
Cronbach's α, M0	0.893
Cronbach's α, M1	0.911
Cronbach's α, M2	0.892
Cronbach's α, M3	0.893
Mean inter-item correlation, M3	0.512

The coded variables are human classifications, and their reliability is reported separately. Twenty-one students (seven per condition, 21% of the sample) had their end-of-term drawings double-coded, and the fourteen of them in the two tool conditions also had their dialogues double-coded; C0 produced none. For the categorical and ordinal dimensions we report Krippendorff's α [30], which applies across numbers of coders, levels of measure-

ment and missing data. For the continuous dimension, conclusion-seeking frequency, we report the two-way random-effects intraclass correlation for a single rating under absolute agreement, ICC(2,1), which treats coders as a sample from a population of coders and gives reliability for an individual judgement rather than for a mean [31]. Agreement was $\alpha = 0.86$ for error-type classification, ICC(2,1) = 0.87 for conclusion-seeking frequency, $\alpha = 0.76$ for structure conformity and $\alpha = 0.75$ for self-attempt priority. No dimension fell below the pre-set threshold of 0.70. These four variables are the ones that were double-coded. The drawing-quality dimensions, question quality and generative processing were each coded once, the first two blind to condition, and no agreement coefficient is claimed for any of them. For the error typology, agreement is reported for the classification as a whole across the five categories, and no separate estimate is given for E4, which the coding rules define as a dichotomous family decision. All coefficients are estimated on the double-coded subsample, 21 drawings and 14 dialogues. Coders worked blind to condition, with the condition table stored separately, the drawings numbered at random, and the condition field hidden in the coding interface. Hiding the label does not remove every cue, since a dialogue can reveal whether a student worked under the protocol. The blind therefore applies to the label and not to the content of the record.

Scoring was carried out by the teaching team after submission, and no score was returned to students. The four practices served as research measurements only, not formative assessment, so no commentary from them reached the student and a practice score could not feed into later performance. The six deliverables of the task thread were treated in the same way, and nothing further was asked of the students after the final measurement.

The study is part of a curriculum teaching reform and collected no sensitive personal information beyond what relates to course learning. Students were informed of the measurement arrangement before the first measurement. Participation was voluntary and did not affect course grades, and no incentive or compensation was offered to any group. Responses were recorded under a study code; no name and no student number appears in the research dataset, and the code was assigned when the response was entered rather than taken from any institutional record. The correspondence between codes and students was held by the data custodian outside the dataset and was used only to attach the entry-examination record, the sole source of prior attainment. The four measurements are linked to one another by the code alone. The records are therefore pseudonymous rather than anonymous. All participants were first-year undergraduates aged 18 or older. Dialogue records were submitted by the students themselves as part of their course deliverables, and only the request–response text of the submitted window was collected. No account credentials or personal information were collected. After the delayed measurement, students in the no-AI condition received the same protocol card and the same guidance on tool use as the other two groups. Automated scoring was audited by hand on a 10% random sample of item responses, re-scored against the analytic standard and compared with the automated score. Each discrepancy was resolved by re-reading the response against that standard. The sample was drawn as a check on the scoring script and not as an agreement study, so no disagreement count was recorded for it and no agreement coefficient is claimed.

5. Results

5.1. The Trajectory

All 100 students completed all four practices, with no attrition and no missing item response (3200 item responses in total). Completion was voluntary, and none of the four practices entered the course mark. During the term each practice coincided with the task thread the students were already working through, and at the delayed point, four weeks after the term, all 100 still completed it. Completion therefore did not vary across conditions

and entered no analysis. An attempt made with little effort still counts as complete, so the records do not certify the effort behind them. Baseline competence did not differ across conditions ($F(2, 97) = 0.04$, $p = 0.965$; means 47.61, 49.62 and 48.30). Prior attainment did not differ either ($F(2, 97) = 0.34$, $p = 0.712$; means +0.11, +0.31 and +0.22 in cohort standard-deviation units), and it did not predict the delayed outcome ($\gamma_2 = -2.15$, $p = 0.289$). At the delayed measurement, scores still ranged widely across the cohort (SD = 31.23), almost unchanged from baseline. Figure 5 plots the trajectories by condition with 95% confidence intervals, and Table A4 gives the means and standard deviations behind them.

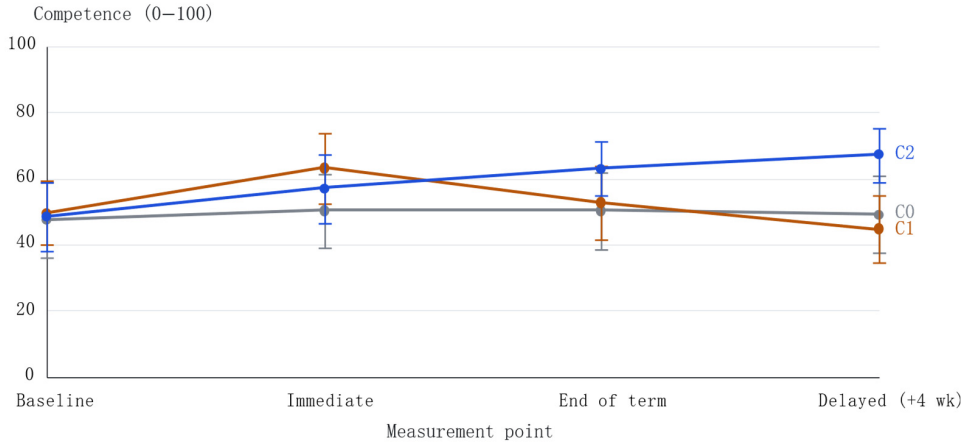


Figure 5. Competence by condition across the four measurement points, with 95% confidence intervals.

The confidence intervals in Figure 5 overlap at the first three measurement points and separate only at the delayed one. The separation that survives is between the protocol condition and the unrestricted condition, with no overlap at the delayed measurement, while the no-tool interval still overlaps both. In raw terms the no-tool condition sits closer to the unrestricted one, which is why the analysis adjusts for baseline and prior attainment rather than relying on the trajectory alone.

The order of the three conditions changed across the four points. In the week the module was taught, the unrestricted condition was ahead on unadjusted means, by 6.06 points over the protocol group. The no-tool condition barely moved across the four points. By the delayed measurement the protocol group was the strongest of the three and the unrestricted group was the weakest. Of the protocol group, 51.5% scored at or above 70, more than double the share in the unrestricted condition (24.2%) and well above the no-tool condition (32.4%). The value of 70 was not the pass mark of the course and was not fixed before the study. It was chosen after data collection as a round cut-point for describing where delayed scores fell. No practice score counted towards the course mark, so the cut-point carries no grading consequence and enters no inferential test.

5.2. Delayed Competence

Controlling baseline competence, prior attainment and out-of-class study time, C2 exceeded C0 by 18.22 points and C1 by 22.79 points at the delayed measurement. Both adjusted contrasts come from one model of the delayed score, shown in Equation (1).

$$Y_{i,M3} = \beta_0 + \beta_1 C1_i + \beta_2 C2_i + \gamma_1 Y_{i,M0} + \gamma_2 P_i + \gamma_3 T_i + \varepsilon_i, \quad (1)$$

where $Y_{i,M0}$ is the baseline score; P_i is prior attainment and T_i is out-of-class study time.

The condition main effect was significant ($F(2, 94) = 12.06$, $p < 0.001$), and the model as a whole accounts for $R^2 = 0.612$. In addition, C1 did not differ from C0 (-4.57, $p = 0.371$). Study time did not predict the delayed outcome ($\gamma_3 = 0.046$, $p = 0.530$). It is in the model

because out-of-class practice is the mechanism the design relies on, and reported hours do differ across conditions: the no-tool group reported the most out-of-class hours, 127.9 against 116.9 under the protocol and 110.2 under unrestricted use ($F(2, 97) = 3.47$, $p = 0.035$), and still scored lowest. The coefficient alone would not show that hours played no part. Removing study time from the model leaves both contrasts essentially unchanged, at 17.70 for C2 − C0 and 23.12 for C2 − C1. At the immediate measurement, after adjustment, the contrasts were small and not significant. Withholding the answer had no detectable effect in the week the module was taught. The complete coefficient table for this model, including the covariate block and the model R2, is given in Appendix A, Table A1. Table 4 separates the two estimation bases, unadjusted contrasts from independent-samples comparisons and adjusted contrasts from the model above.

Table 4. Contrasts at each measurement point, unadjusted and covariate-adjusted.

Point	Contrast	Unadjusted Means	ANCOVA-Adjusted [95% CI]
M1	C2 − C1	−6.06 (d = −0.20)	−4.64 [−14.49, 5.22], $p = 0.352$
M1	C2 − C0	+6.82 (d = 0.21)	+5.66 [−4.21, 15.53], $p = 0.258$
M2	C2 − C1	+10.23 (d = 0.36)	+11.17 [1.49, 20.85], $p = 0.024$
M2	C2 − C0	+12.69 (d = 0.43)	+12.34 [2.65, 22.04], $p = 0.013$
M3	C2 − C1	+22.35 (d = 0.83)	+22.79 [12.97, 32.61], $p < 0.001$
M3	C2 − C0	+17.96 (d = 0.60)	+18.22 [8.38, 28.06], $p < 0.001$

Note: M3 contrasts (C2 − C0, C2 − C1) are the two primary contrasts; the M1 and M2 contrasts are reported for the trajectory and are exploratory. Estimates are in points on the 0–100 composite; adjusted contrasts come from Equation (1), with baseline composite, prior attainment and out-of-class study time as covariates. The adjusted contrasts come from a model that assumes a common slope across conditions. The joint test of the condition × baseline interaction rejects that assumption ($F(2, 92) = 5.17$, $p = 0.007$). Estimated from the interaction model over the observed baseline distribution, the average contrasts are 17.75 (C2 − C0) and 23.07 (C2 − C1).

Adding condition × covariate interactions rejects that assumption for the baseline ($F(2, 92) = 5.17$, $p = 0.007$): the adjusted advantage of the protocol group is 0.43 points smaller for every point of baseline score above the mean. The form of the interaction is reported in Appendix A, Table A3, and the contrast is plotted against the baseline in Figure A1. For orientation, the contrast against C0 is +30.88 points at one SD below the baseline mean, +17.75 at the mean and +4.63 at one SD above it, with intervals in Table A3. The boundary of significance falls at a baseline of 64.95 on the 0–100 composite, so the advantage is no longer distinguishable from zero above that score. The contrasts in Table 4 are averages over the baseline distribution of this cohort, not uniform effects. Estimated from the interaction model over the same distribution, the average contrast is 17.75 points for C2 − C0 and 23.07 for C2 − C1, within half a point of the common-slope estimates, so the reduced model does not create the average it reports.

A difference-in-differences model on the pooled student-by-point data, with standard errors clustered by student (100 clusters), traces each tool condition against the no-tool reference over time. The unrestricted group's advantage was concentrated in the teaching week (C1 × M1 = 10.87, $p = 0.024$), had disappeared by the end of term (C1 × M2 = 0.46, $p = 0.918$), and had turned negative four weeks after withdrawal (C1 × M3 = −6.40, $p = 0.167$). The protocol group's contrast was small and not significant at the immediate measurement (C2 × M1 = 6.14, $p = 0.264$) and was significant at both later points (C2 × M2 = 12.01, $p = 0.026$; C2 × M3 = 17.28, SE 5.86, $p = 0.004$). The C2 × M3 coefficient equals the change in C2 minus the change in C0, $(67.05 - 48.30) - (49.08 - 47.61) = 17.28$. The full interaction set is reported in Appendix A, Table A2. These time-course contrasts are exploratory, and the two primary contrasts remain the delayed ones in Table 4.

At M2, the end-of-term measurement, the protocol group was already ahead on adjusted means (+11.17 against C1, +12.34 against C0, both $p < 0.05$), while the unrestricted group had fallen back to the level of the no-tool group. The gap did not appear only after the break. It had opened while the tool was still in use (C2 × M2 = 12.01, $p = 0.026$), and the change between M2 and M3 is not significant on its own (C2 × [M3 − M2] = 5.27, SE 5.19, $p = 0.310$).

5.3. Adherence and Role Perception

Table 5 reports protocol use, blind-coded question quality, and the changes students reported about their own working. In each tool condition 31 of the 33 students submitted at least one exchange (62 of 66 in total, 93.9%), and four students, two in each condition, submitted none. The figures for request content in Table 5 are computed based on the 62 students with records.

Table 5. Protocol use, question quality and learner-reported change.

Index	C1	C2
Protocol adherence (0–1)	0.37	0.77
Requests asking for the conclusion	63.30%	11.80%
Question quality (1–5), blind-coded	2.18	3.51 (d = 1.70, $p < 0.001$)
Autonomy, pre → post	3.05 → 2.52 (−0.53, $p = 0.004$)	3.16 → 3.69 (+0.53, $p = 0.005$)
Role perception, pre → post	2.87 → 2.20 (−0.67, $p < 0.001$)	3.29 → 3.63 (+0.34, $p = 0.098$)
Autonomy and role perception, C0	Autonomy −0.03 ($p = 0.897$); role perception +0.16 ($p = 0.294$)	

Note: Percentages for request content are computed based on the students with submitted records (31 of 33 in each tool condition). Adherence is used inferentially only within the protocol condition; the value for C1 is descriptive. The role analyses are exploratory.

Within C2, adherence correlated with delayed competence (r = 0.63, 95% CI [0.37, 0.80], $p < 0.001$; Figure 6). Adherence ran from 0.17 to 0.97 across the 31 students who submitted records, the two who submitted none taking the floor by the coding rule. The difference is in what students asked, not only in what they were permitted to ask. How closely a student kept to the rule tracked how much the same student retained four weeks later. Adherence varied continuously across students rather than splitting them into compliant and non-compliant, which is why we describe it as a graded performance and not a compliance record. That estimate has limits. The interval is wide, it rests on a single condition of 33 students, and the correlation cannot separate adherence as a cause of retention from adherence as a marker of students already disposed to work in this way.

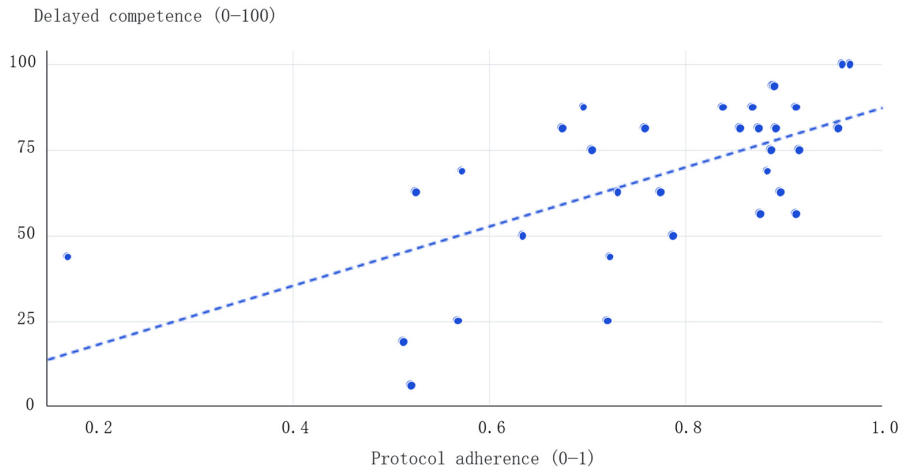


Figure 6. Protocol adherence and delayed competence within C2 ($n = 33$).

The self-reports point in opposite directions in the two tool conditions and are the weaker part of the evidence in this table. Under unrestricted use, students reported attempting problems themselves less often (-0.53, $p = 0.004$) and described the teacher's role as less about designing tasks and questioning reasoning (-0.67, $p < 0.001$). Under the protocol, autonomy rose by the same margin in the other direction ($+0.53$, $p = 0.005$), and role perception rose by $+0.34$, not significant ($p = 0.098$). The no-tool condition did not move on either measure (-0.03, $p = 0.897$; $+0.16$, $p = 0.294$).

5.4. The Process Record Behind the Advantage

Two process variables were measured alongside the outcome and describe how the conditions differed. They were measured at the same time as the outcome, so they are not a decomposition of the effect. Generative processing, coded from the logs, was highest under the protocol at the delayed measurement (C2 = 1.36 against C0 = 0.71 and C1 = 0.65) and correlated with delayed competence across all participants (r = 0.50, $p < 0.001$). Cognitive ease, the practice-stage self-report, was lowest under the protocol (C2 = 4.91 against C0 = 5.77 and C1 = 7.17) and correlated negatively with delayed competence (r = -0.21, $p = 0.033$). A mediation test would require the process variables to be measured before the outcome, so no indirect effect is reported. Table 6 reports both comparisons. Both patterns run in the direction the design predicted, in that the protocol group worked harder on the material and found it less easy than the unrestricted group, which found it easiest of the three.

Table 6. Process variables at the delayed measurement, by condition.

Quantity	Value
Generative processing, C2/C0/C1	1.36/0.71/0.65
Generative processing and delayed competence	r = 0.50, $p < 0.001$
Cognitive ease, C2/C0/C1	4.91/5.77/7.17
Cognitive ease and delayed competence	r = -0.21, $p = 0.033$

5.5. Where the Difference Sits

The aggregate score shows only that unrestricted use is the weakest condition four weeks later. Unrestricted use did not differ from no AI on the delayed composite (-4.57, $p = 0.371$) or on near transfer (d = -0.37, $p = 0.130$), so the study does not establish that it lowered competence relative to no AI. What it establishes is a difference between the two tool conditions and between the protocol condition and no AI. Two further analyses show whether the difference sits on drafting or on reasoning. Table 7 gives the incidence of each error type by condition at M1 and M3. Percentages are of students, since each measurement produced one drawing per student and the denominators are the condition samples.

At the immediate measurement the two tool conditions did not differ on the diagnostic error type (C1 = 36.4%, C2 = 48.5%, $\chi^2(1) = 0.99$, $p = 0.319$), and each of the two tool conditions was below the no-tool condition on all five types; for E3 the values were 27.3% and 18.2% against 38.2%. Four weeks later, the unrestricted condition had the highest incidence on all five, including the three that concern the structure of the curve. The sharpest indicator is E4, misjudgment of curve type. At the delayed measurement the two tool conditions differed by a factor of six (C1 = 54.5%, C2 = 9.1%, $\chi^2(1) = 15.71$, $p < 0.001$), with a risk ratio of 6.00 (95% CI 1.95 to 18.44) and a risk difference of 45.5 percentage points (95% CI 25.8 to 65.1). C1 was about one and a half times as likely as the no-tool condition to make the error (54.5% against 35.3%), a difference that is not significant on its own ($\chi^2(1) = 2.51$, $p = 0.113$). Four weeks later, students who could obtain the answer still could not classify the configuration.

Table 7. Percentage of drawings exhibiting each error type.

Condition	Time	E1 Missing Points	E2 Visibility	E3 Contour Termination	E4 Type Misjudgment	E5 Arc Substitution
C0	M1	35.3	35.3	38.2	50.0	58.8
C0	M3	47.1	38.2	32.4	35.3	20.6
C1	M1	27.3	18.2	27.3	36.4	12.1
C1	M3	48.5	48.5	33.3	54.5	24.2
C2	M1	27.3	33.3	18.2	48.5	30.3
C2	M3	33.3	18.2	9.1	9.1	12.1

Note: Entries are percentages of students in the condition, not of drawings: each measurement produced one drawing per student, so the denominator is the condition sample at that point (C0 = 34, C1 = 33, C2 = 33). Counts for the diagnostic error type E4 at the delayed measurement are 12 of 34 (C0), 18 of 33 (C1) and 3 of 33 (C2); E1–E5 are defined in the text.

Figure 7 shows the two configurations side by side, in front view at a single scale. They differ only in the diameter ratio. At d/D = 1.00 each branch of the intersection is a straight line and the two branches cross on the chord axis, while at d/D = 0.52 the intersection is a space curve that leaves the chord outline and sags 16 mm toward the axis.

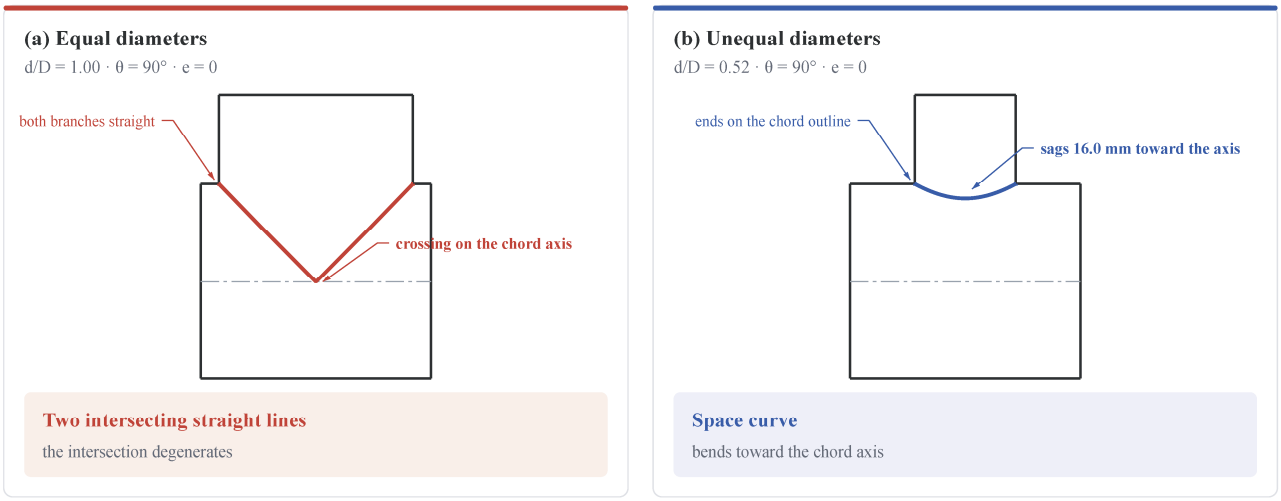


Figure 7. The two configurations behind E4. At d/D = 1.00 the intersection degenerates into two straight lines; at d/D = 0.52 it is a space curve. Both panels are computed from the same analytic solutions used to score the responses.

Table 8 separates drafting quality from the configural transformation and adds the transfer test. The argument rests on two results. Curve accuracy is where a conceptual deficit should show, and among the four rubric dimensions it carries the widest gap to C1 (d = 0.77, $p = 0.003$) and to C0 (d = 1.16, $p < 0.001$). Completeness moved the same way but less strongly (d = 0.66 and 0.74), and near transfer, reported below, separates the two tool conditions more widely still (d = 1.23). Line conventions are where copying a correct answer would appear, and the three conditions did not differ detectably (C0 = 3.63, C1 = 3.38, C2 = 3.54, every pairwise $p \geq 0.17$). On transfer to a different component family, C1 fell below C0 (2.64 against 3.01, d = −0.37) without reaching significance, while C2 scored above both (3.67). The difference between the two tool conditions therefore falls on deciding what to draw, not on drawing it once the decision is made.

Projection relations is the layout dimension, that is, whether the views are placed in the correct relationship to each other. Both tool conditions drew throughout the term, and since a generated figure can be copied, the unrestricted group is unlikely to have drawn less than the protocol group. Volume of drawing is what maintains layout practice, which is a plausible reason the unrestricted group did not fall below the no-tool group on this

dimension (3.00 against 3.16, d = −0.16, $p = 0.506$). The protocol group's advantage over it (3.52 against 3.00, d = 0.55, $p = 0.029$) then reflects more disciplined work rather than more practice. Line conventions behave differently. It is a rule-based dimension that all three conditions rehearsed directly in the same course, and it is where every group converged.

Table 8. Drawing quality at end of term and near transfer at the delayed measurement (1–5).

Dimension	C0	C1	C2	C2 vs. C1	C2 vs. C0	C1 vs. C0
Projection relations	3.16	3.00	3.52	d = 0.55, $p = 0.029$	d = 0.35, $p = 0.161$	d = −0.16, $p = 0.506$
Line conventions	3.63	3.38	3.54	d = 0.23, $p = 0.345$	d = −0.12, $p = 0.619$	d = −0.34, $p = 0.170$
Completeness	2.93	3.02	3.56	d = 0.66, $p = 0.010$	d = 0.74, $p = 0.004$	d = 0.10, $p = 0.671$
Curve accuracy	2.74	3.05	3.79	d = 0.77, $p = 0.003$	d = 1.16, $p < 0.001$	d = 0.35, $p = 0.156$
Near transfer	3.01	2.64	3.67	d = 1.23, $p < 0.001$	d = 0.71, $p = 0.005$	d = −0.37, $p = 0.130$

Note: Entries are condition means on a 1–5 scale; d is Cohen's d and p is uncorrected. Curve accuracy is the dimension on which a conceptual deficit must show, and line conventions the one on which copying a correct answer would show. The three pairwise comparisons within each row were not part of the primary plan and are reported without multiplicity correction. They are exploratory.

6. Discussion

AI will sit in every workplace our graduates enter, and a course that bans it trains students for a world that does not exist. Embracing the tool is therefore not optional, but it is not neutral either. In this course the model can already perform the core operation, the projection of a solid onto a plane. Once the tool reaches into the middle of what a course teaches, the teacher–student relationship and the teaching method are themselves challenged: who does the explaining, who checks the reasoning, and who owns the mistake. Sustainability, in the sense used in this paper, asks that the ability outlasts the tool. This study tested one way to meet that standard in one first-year graphics course, on one operation whose correctness is settled by geometry.

Instructional theory asks how much help to give, and the usual advice is to fade scaffolding as expertise grows [19,20]. That advice assumes an instructor holding the scaffold. When students control the tool themselves, the amount of help is no longer ours to set. What remains ours is what the tool may be asked to do. The results are consistent with treating that variable as the operative one. Unrestricted use showed a gain concentrated in the teaching week that did not survive withdrawal, whereas under the protocol the teaching-week gain was small and not significant, and the delayed gain was the widest in the study. The protocol group also received the card and submitted dialogues that were read out in class, so the two tool conditions differed in more than the rule. In this cohort, protocol-guided AI use was associated with stronger retention of task-specific competence at the four-week delayed assessment than unrestricted AI use.

The observed differences between immediate and delayed performance are consistent with, but do not establish the same causal relationship as, previous research on AI-assisted learning. Bastani et al. [13] reported that unrestricted access to generative AI could improve performance during assisted mathematical practice while undermining subsequent unaided performance, whereas a tutor interface providing guidance without direct answers mitigated this disadvantage. Although their randomized study involved secondary-school mathematics and the present study examined engineering graphics through a self-selected quasi-experimental comparison, both findings highlight the importance of distinguishing immediate task completion from independently demonstrated competence. In the present study, the unrestricted AI condition showed the highest unadjusted performance immediately after instruction, but this advantage did not persist at the delayed measurement. Conversely, the protocol-guided condition showed the strongest delayed performance. This

pattern is also compatible with the cognitive-offloading perspective [15], which suggests that external assistance may reduce the cognitive processing undertaken by learners themselves. The present data, however, do not directly measure cognitive offloading or establish it as the mechanism responsible for the differences.

The findings can also be considered in relation to previous research on generative AI in design-oriented education. Liu et al. [11] reported positive learning outcomes from GenAI integration in a design drawing course, particularly among students with weaker initial preparation. The present study extends this discussion by distinguishing the type of AI assistance from its general availability and by examining competence at a delayed measurement. The larger protocol-associated advantage among students with lower baseline competence suggests that structured assistance may be particularly relevant where foundational geometric understanding is limited. This interpretation is consistent with research on self-explanation [21,22], productive failure [17], and the assistance dilemma [19], all of which emphasize the importance of preserving learners' active cognitive engagement. Nevertheless, the baseline-dependent findings were exploratory, and the present design cannot determine whether differences resulted from the request restriction itself, students' initial learning dispositions, or other accompanying instructional supports.

At the delayed measurement, the unrestricted group had fallen to the last position and the protocol group held the widest advantage of any contrast in the study. The drawings show where the difference sits. Line conventions did not separate the three groups at the end of term, but curve-type misjudgment did, by a factor of six between the two tool conditions. Students who could obtain the answer still could not decide what to draw. The dialogue records point the same way. Under the protocol, most requests questioned the student's own description instead of asking for the conclusion. We did not change the tool, the hours or the assessment. We changed what students were allowed to ask of it, and that is the lever a course still holds once the tool sits in the student's own hands. For evaluation the lesson is direct. What AI brings to a course can only be measured after the tool is gone, because a test taken with the tool open counts the tool's performance as the students'. Evidence for sustainability has to survive the tool.

A second point concerns supervision. Nearly all AI use happens out of sight, and it did here. The tool sat with every student all term, and the protocol was a printed rule with no monitoring. Adherence tracked delayed competence, at $r = 0.63$ across 33 students. Guidance cannot rest on an appeal to self-restraint. It has to be built into what students are asked to do. In this course the task did that work, and the teacher's hours moved from grading drawings to reading the class's dialogues aloud. The arrangement added no recurring contact hours to the timetable, required no equipment and involved no procurement. The tool is a free public service the students already used, the exchanges took place outside class, and the task thread ran inside the existing 48 h. We did not measure teaching hours before and after the change, so the claim covers scheduled hours and purchases only, not total staff time. Courses whose core work is open-ended design will need their own boundary. Ours holds because geometry settles the answer.

The evidence in this study comes from one class in one course at one institution. The three conditions were formed by the students' own choice and not by randomization, and the study was not registered in advance. Baseline balance across the three groups was checked, but a nonsignificant baseline difference does not make the groups equivalent. The study was designed for the contrast between conditions at the delayed measurement and was not powered for interaction, subgroup or process analysis, so those analyses are exploratory and cannot test a mechanism. The difference between the two tool conditions cannot be separated from the card, the dialogue submissions and the plenary walkthrough that accompanied it. Contamination between students in one cohort cannot be measured. The

unaided practices were completed under an instruction that was not technically enforced, since the tool sits outside the course platform. Study during the four-week interval was not recorded. The outcome is narrow task competence on one target operation, not general engineering competence. Four weeks bounds the claim about durability. Tool use came from student submissions, so the unrestricted group's record may be incomplete. Future studies should verify use with independent records, replicate across courses and institutions with larger protocol groups, pre-register the design, and follow students beyond four weeks with a validated measure of transfer.

7. Conclusions

Engineering education is expected to develop sustainably while AI tools enter every part of engineering practice. We carried out a study in one first-year engineering graphics course. Students used generative AI outside class under three different strategies, no AI, unrestricted use, and a request protocol that governs what may be asked. Students formed the three groups themselves rather than being assigned, so the comparison is quasi-experimental. The comparison ran for one term and ended four weeks after the tool was withdrawn, so that what each strategy leaves behind could be examined. The findings below hold for tasks whose correctness is settled by geometry rather than by convention. The findings are as follows.

(1) The delayed measurement changed the order of the three conditions. Four weeks after withdrawal the protocol group scored 67.05 against 44.70 under unrestricted use and 49.08 without AI, and 51.5% of the protocol group scored at or above 70, against 24.2% and 32.4% in the other two conditions.

(2) In this cohort the two tool conditions diverged on classification, not on drafting. Curve-type misjudgment appeared in 9.1% of the protocol group's delayed drawings against 54.5% under unrestricted use ($p < 0.001$), while line conventions did not separate the conditions at the end of term. The protocol group's advantage extended to a component family it had not trained on.

(3) Adherence to the protocol acted as a graded skill rather than a compliance record. It ranged from 0.17 to 0.97 across the 31 students who submitted records, with the two who submitted none taking the floor, and it was associated with delayed competence at r = 0.63 (95% CI 0.37 to 0.80). Under the protocol, 11.8% of the recorded requests asked for the conclusion, against 63.3% under unrestricted use, and blind-coded question quality rose by more than a standard deviation.

Four weeks is the interval this study measured, and it bounds what can be claimed. The difference between the conditions survived withdrawal for that period, in one class, on one target operation. Two constraints of the present test point to what further research can improve when conditions permit. The outcome rests on one operation whose correctness is settled by geometry, and the follow-up stops at four weeks. A stronger test would cover open-ended design tasks where no geometry settles the answer, track students beyond four weeks, verify tool use with independent records, and run the protocol group at a larger scale.

Author Contributions: Conceptualization, J.L.; methodology, J.L.; software, J.L.; validation, H.H.; formal analysis, C.H.; investigation, J.L., H.H. and C.H.; resources, J.L.; data curation, C.H.; writing—original draft preparation, J.L.; writing—review and editing, H.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Higher Education Scientific Research Project of the Chongqing Higher Education Association (2023–2024), grant number cqgj23060C.

Institutional Review Board Statement: This research is based on the existing practice of curriculum teaching reform; all the data collected are for evaluating the teaching effect; apart from information

related to course learning, no sensitive personal information or data involving privacy was collected. Therefore, it is exempt from ethical review and approval.

Informed Consent Statement: Written informed consent was obtained from all student participants prior to data collection. Participants were clearly informed that participation was completely voluntary; they could withdraw from the evaluation at any time without penalty to their course grades. All academic records, questionnaire responses and examination data were anonymized; personal identifiable information was stripped before analysis to guarantee participant confidentiality.

Data Availability Statement: The individual records are course records held under an institutional ethics approval that does not extend to public deposit, and the item-generation, scoring and analysis code forms part of a teaching-project deliverable that the authors are not authorized to distribute. Neither the data nor the code can therefore be released with the manuscript. The estimation procedures, software, model specifications and clustering are set out in Appendix A, so that the estimation is fully specified and could be reproduced on equivalent records.

Acknowledgments: During the preparation of this manuscript, the authors used DeepSeek-V4.1-Flash for the purposes of language polishing and editing of the text. The authors have reviewed and edited the output and take full responsibility for the content of this publication.

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

Appendix A

All item generation, scoring and model estimation were carried out in Python 3.13, using the least-squares and cluster-robust routines of the Python scientific stack. No spreadsheet or point-and-click package was used at any stage.

Table A1. Delayed competence regressed on condition and covariates [Equation (1)].

Term	Coef.	SE	t	p
Intercept	9.25	10.17	0.91	0.366
C1 vs. C0	−4.57	5.08	−0.90	0.371
C2 vs. C0	18.22	4.95	3.68	< 0.001
Baseline composite (M0)	0.718	0.065	11.00	< 0.001
Prior attainment	−2.15	2.01	−1.07	0.289
Out-of-class study time (M3)	0.046	0.073	0.63	0.530

Note: N = 100, $R^2 = 0.612$, adjusted $R^2 = 0.592$, residual df = 94. The derived contrast C2 vs. C1 is 22.79 points (SE 4.95, t = 4.61, $p < 0.001$).

Table A2. Difference-in-differences model on the pooled data (complete interaction set).

Term	Coefficient	Cluster-Robust SE	t	p
C2 × M1	6.14	5.47	1.12	0.264
C2 × M2	12.01	5.31	2.26	0.026
C2 × M3	17.28	5.86	2.95	0.004
C1 × M1	10.87	4.73	2.30	0.024
C1 × M2	0.46	4.44	0.10	0.918
C1 × M3	−6.40	4.59	−1.39	0.167

Note: Estimated on the pooled student-by-point data without student fixed effects, with standard errors clustered by student (100 clusters). Coefficients are in points on the 0–100 composite. Main effects and the intercept are omitted. With M2 as the reference point, C2 × (M3 − M2) = 5.27 (SE 5.19, $p = 0.310$) and C1 × (M3 − M2) = −6.85 (SE 4.54, $p = 0.132$). A model with student fixed effects returns C2 × M3 = 17.28 (SE 6.76, $p = 0.011$).

Table A3. Baseline × condition interaction: joint test, simple slopes and region of significance.

Term	Coef.	SE	t	*p*
(Intercept)	47.89	9.62	4.98	<0.001
C1 vs. C0	−5.32	4.88	−1.09	0.278
C2 vs. C0 (at baseline mean)	17.75	4.75	3.74	<0.001
Baseline (centered)	0.850	0.100	8.52	<0.001
C1 × baseline	0.019	0.154	0.13	0.900
C2 × baseline	−0.425	0.150	−2.84	0.006
Prior attainment	−2.41	1.94	−1.24	0.217

Note: N = 100, R^2 = 0.651, adjusted R^2 = 0.625, residual df = 92. Joint test of the two interaction terms: $F(2, 92) = 5.17$, $p = 0.007$. Baseline SD = 30.9, centered at its mean of 48.50. Averaged over the observed baseline distribution, the contrast is 17.75 points for C2 − C0 (SE 6.61) and 23.07 for C2 − C1 (SE 6.86).

Simple slopes of the baseline within each condition: 0.850 (SE 0.100) under C0, 0.870 (SE 0.118) under C1 and 0.425 (SE 0.110) under C2. Under the protocol the delayed score is therefore markedly less dependent on where the student started than under the other two conditions.

Region of significance (Johnson–Neyman): the protocol contrast against C0 is +30.88 points (95% CI 17.93 to 43.82) at one SD below the baseline mean, +17.75 (95% CI 8.32 to 27.19) at the mean, and +4.63 (95% CI −8.74 to 18.00) at one SD above it. The boundary of significance falls at a baseline of 64.95 on the 0–100 composite, above which the difference is no longer distinguishable from zero. The other root (197.4) lies outside the scale.

Table A4. Means and standard deviations of the composite by condition and measurement point.

Point	C0 (*n* = 34)	C1 (*n* = 33)	C2 (*n* = 33)
M0	47.61 (33.68)	49.62 (28.89)	48.30 (30.75)
M1	50.18 (33.20)	63.07 (31.34)	57.01 (30.37)
M2	50.18 (34.52)	52.65 (32.25)	62.88 (23.74)
M3	49.08 (34.96)	44.70 (29.98)	67.05 (23.86)

Note: Values are means with standard deviations in parentheses, on the 0–100 composite. N = 100, and all 100 students completed all four points.

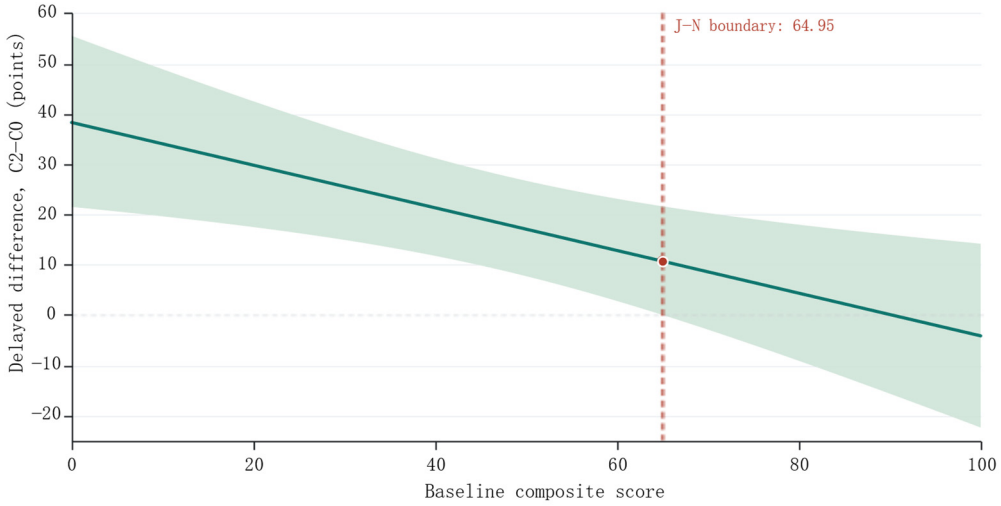


Figure A1. Adjusted difference in delayed competence (C2 − C0) as a function of the baseline score. Solid line, estimated difference from Table A3. Shaded band, 95% confidence interval. Dashed line, Johnson–Neyman boundary.

References

1. Mangalore, P.; Gamit, J.S.; Kanchan, M. Developing a framework for sustainable engineering education through transformative learning principles. *Discov. Sustain.* **2025**, *6*, 1473. [CrossRef]
2. United Nations. *Transforming Our World: The 2030 Agenda for Sustainable Development*; A/RES/70/1; United Nations General Assembly: New York, NY, USA, 2015.
3. Huang, T.; Wang, T.; Zhang, F.; Hao, Y.E.; Liang, L.E.; Wang, X.; Yao, M.; Yuan, C. Integrating sustainability and ethical responsibility into building water supply and drainage engineering education: A CDIO-based curriculum reform. *Sustainability* **2026**, *18*, 1933.
4. Matey-Sanz, M.; Granell, C.; Mollineda, R.A. Responsible Integration of Generative AI in Software Engineering Education. *IEEE Rev. Iberoam. Tecnol. Aprendiz.* **2026**, *21*, 375–384. [CrossRef]
5. Kadam, K.; Mishra, S.; Deep, A.; Iyer, S. Enhancing engineering drawing skills via fostering mental rotation processes. *Eur. J. Eng. Educ.* **2021**, *46*, 796–812. [CrossRef]
6. Potter, C.; Van Der Merwe, E.; Kaufman, W.; Delacour, J. A longitudinal evaluative study of student difficulties with engineering graphics. *Eur. J. Eng. Educ.* **2006**, *31*, 201–214. [CrossRef]
7. Zheng, L.; Ge, D. Cultivating key disciplinary competencies among university students against the background of China's new engineering education initiative. *PLoS ONE* **2025**, *20*, e0318561. [CrossRef]
8. Corsiuc, G.; Graur, A.-M.; Mârza, C. The Influence of the Pandemic in the Teaching of Graphic Subjects. *J. Ind. Des. Eng. Graph.* **2024**, *19*, 247–250.
9. Tsutsumi, E. A mental cutting test using drawings of intersections. *J. Geom. Graph.* **2004**, *8*, 117–126.
10. Xie, H.; Zheng, H.; Huang, Y.; Wu, G. An Empirical Study on the Influencing Factors of GenAI on Learning Achievement in Basic Courses of Artificial Intelligence Major. In Proceedings of the 2026 15th International Conference on Educational and Information Technology (ICEIT), Xi'an, China, 27–29 March 2026; pp. 639–645.
11. Liu, X.-K.; Chiang, F.-K.; Tan, Y.-Q.; Wu, X.-Y.; Cha, Y.-Y.; Lim, C.-P. Leveraging GenAI tool in design drawing: A case study of an architecture course in a Chinese university. *Int. J. Artif. Intell. Educ.* **2025**, *35*, 3529–3558. [CrossRef]
12. Setiyawan, A.; Soeharto, S.; Wijaya, T.T.; Lavicza, Z. Understanding Pre-Service Vocational Teachers' Perceptions of GenAI Using ChatGPT Through the Lens of Perceived Benefit. *Technol. Knowl. Learn.* **2026**, *31*, 2237–2266. [CrossRef]
13. Bastani, H.; Bastani, O.; Sungu, A.; Ge, H.; Kabakcı, Ö.; Mariman, R. Generative AI without guardrails can harm learning: Evidence from high school mathematics. *Proc. Natl. Acad. Sci. USA* **2025**, *122*, e2422633122. [CrossRef] [PubMed]
14. Gerlich, M. AI tools in society: Impacts on cognitive offloading and the future of critical thinking. *Societies* **2025**, *15*, 6, Correction in *Societies* **2025**, *15*, 252. [CrossRef]
15. Risko, E.F.; Gilbert, S.J. Cognitive offloading. *Trends Cogn. Sci.* **2016**, *20*, 676–688. [CrossRef] [PubMed]
16. Morales, C. Retrieval Interruption Framework: AI-Assisted Cognition and Retrieval-Dependent Learning in Higher Education. *Educ. Sci.* **2026**, *16*, 1179. [CrossRef]
17. Kapur, M. Productive failure. *Cogn. Instr.* **2008**, *26*, 379–424. [CrossRef]
18. Roediger, H.L., III; Karpicke, J.D. Test-enhanced learning: Taking memory tests improves long-term retention. *Psychol. Sci.* **2006**, *17*, 249–255.
19. Koedinger, K.R.; Aleven, V. Exploring the assistance dilemma in experiments with cognitive tutors. *Educ. Psychol. Rev.* **2007**, *19*, 239–264. [CrossRef]
20. Van Merriënboer, J.J.; Kirschner, P.A.; Kester, L. Taking the load off a learner's mind: Instructional design for complex learning. *Educ. Psychol.* **2003**, *38*, 5–13. [CrossRef]
21. Fiorella, L.; Mayer, R.E. Eight ways to promote generative learning. *Educ. Psychol. Rev.* **2016**, *28*, 717–741. [CrossRef]
22. Bisra, K.; Liu, Q.; Nesbit, J.C.; Salimi, F.; Winne, P.H. Inducing self-explanation: A meta-analysis. *Educ. Psychol. Rev.* **2018**, *30*, 703–725. [CrossRef]
23. Sweller, J.; Van Merriënboer, J.J.; Paas, F. Cognitive architecture and instructional design: 20 years later. *Educ. Psychol. Rev.* **2019**, *31*, 261–292. [CrossRef]
24. Zhang, Y.; Su, R.K.L. Evaluating performance of large language models on fundamental structural knowledge. *Proc. Inst. Civ. Eng.-Struct. Build.* **2025**, *178*, 1009–1017. [CrossRef]
25. Kahn, P.; Carrigan, M.; Smith, P.; Murtagh, L.; Liu, R.; Song, F. Teacher agency and generative artificial intelligence: Teaching in higher education as a responsive, cultural activity. *Learn. Media Technol.* **2025**, 1–12. [CrossRef]
26. Xu, X.; Qiao, L.; Cheng, N.; Liu, H.; Zhao, W. Enhancing self-regulated learning and learning experience in generative AI environments: The critical role of metacognitive support. *Br. J. Educ. Technol.* **2025**, *56*, 1842–1863. [CrossRef]
27. Chee, H.; Ahn, S.; Lee, J. A competency framework for AI literacy: Variations by different learner groups and an implied learning pathway. *Br. J. Educ. Technol.* **2025**, *56*, 2146–2182. [CrossRef]
28. Van Merriënboer, J.J.; Clark, R.E.; De Croock, M.B. Blueprints for complex learning: The 4C/ID-model. *Educ. Technol. Res. Dev.* **2002**, *50*, 39–61. [CrossRef]

29. Shute, V.J. Focus on formative feedback. *Rev. Educ. Res.* **2008**, *78*, 153–189. [CrossRef]
30. Hayes, A.F.; Krippendorff, K. Answering the call for a standard reliability measure for coding data. *Commun. Methods Meas.* **2007**, *1*, 77–89. [CrossRef]
31. Shrout, P.E.; Fleiss, J.L. Intraclass correlations: Uses in assessing rater reliability. *Psychol. Bull.* **1979**, *86*, 420–428. [CrossRef] [PubMed]

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