**Advances in Education** 教育进展, 2026, 16(9), 833-844
Published Online September 2026 in Hans. https://www.hanspub.org/journal/ae
https://doi.org/10.12677/ae.2026.1691968

Hans汉斯

# 生成式人工智能融入工科数学课堂的效果与边界

## ——基于《数值计算方法》课程两轮准实验的实证分析

**张立溥[1]，刘馨莹[1]，徐映红[2]**

[1]浙江传媒学院媒体工程学院，浙江 杭州
[2]浙江理工大学理学院，浙江 杭州

收稿日期：2026年8月5日；录用日期：2026年9月4日；发布日期：2026年9月11日

## 摘 要

**《数值计算方法》的教与学长期卡在一个地方：学生在公式层面“看得懂”，却说不清公式的适用条件，写不出对应的程序，也解释不了误差从何而来。生成式人工智能进入课堂以后，教师可以在讲解现场生成代码、绘制图形、构造反例，学生也能围绕同一算法反复追问。但它究竟只是提高了课堂的新鲜感，还是真的帮助学生跨过了“会套公式”到“能解释算法过程”的那道坎，需要放在一门真实课程、两届自然班和可量化的学习结果里检验。本文以《数值计算方法》连续两年的教学改革为现场，比较两种AI课堂组织方式与传统教学的成绩差异。2024~2025学年，23级一个自然班($n = 24$)采用教师AI演示，同年级三个平行班为对照($n = 97$)；2025~2026学年，24级一个自然班($n = 21$)改为学生AI对抗式学习，同年级三个平行班为对照($n = 84$)。两年均采用覆盖10个知识点的期末分项考核。结果显示，两轮实验班总分均高于对照班：23级高14.55分($p < 0.001$, Hedges $g = 0.78$)，24级高12.00分($p = 0.009$, Hedges $g = 0.62$)，及格率与优秀率同步提升。差异中的差异分析表明，从教师演示转向学生对抗式学习并未带来显著的总分额外增益($-2.55$分，$p = 0.645$)。逐知识点分析揭示明显异质性：曲线拟合、代数精度、数值积分、数值微分受益稳定，Newton法和部分复杂迭代内容提升有限。研究说明，生成式AI能够改善工科数学课程的学习表现，但其效果取决于知识点性质、教师支架与课堂任务结构，不能简单理解为学生使用AI越多越好。**

## 关键词

**生成式人工智能，《数值计算方法》，工科数学，准实验研究，知识点异质性，教学改革**

# Effects and Limits of Integrating Generative AI into Engineering Mathematics Classrooms

## —An Empirical Analysis Based on Two Rounds of Quasi-Experiments in a *Numerical Methods* Course

文章引用: 张立溥, 刘馨莹, 徐映红. 生成式人工智能融入工科数学课堂的效果与边界[J]. 教育进展, 2026, 16(9): 833-844. DOI: 10.12677/ae.2026.1691968

**Lipu Zhang[1], Xinying Liu[1], Yinghong Xu[2]**

[1]School of Media Engineering, Communication University of Zhejiang, Hangzhou Zhejiang
[2]School of Science, Zhejiang Sci-Tech University, Hangzhou Zhejiang

Received: August 5, 2026; accepted: September 4, 2026; published: September 11, 2026

## Abstract

**Students in *Numerical Methods* courses routinely reproduce formulas on paper without understanding when those formulas apply, how to implement them in code, or where errors originate. With generative AI, instructors can now generate code, plot functions, and construct counterexamples in real time, while students can probe the same algorithm from multiple angles. Whether these tools actually deepen understanding, rather than adding surface novelty, can only be judged through controlled trials with intact cohorts and measurable outcomes. This study reports two years of such trials in a *Numerical Methods* course. In 2024~2025, one intact class from the 2023 cohort (n = 24) received instructor-led AI demonstrations, with three parallel classes serving as controls (n = 97). In 2025~2026, one intact class from the 2024 cohort (n = 21) shifted to student-driven adversarial AI learning, again with three parallel controls (n = 84). Both years concluded with a final itemized assessment covering 10 knowledge points. The experimental classes outperformed their controls in both rounds, by 14.55 points for the 2023 cohort ($p < 0.001$, Hedges' $g = 0.78$) and by 12.00 points for the 2024 cohort ($p = 0.009$, Hedges' $g = 0.62$), and both pass rates and excellence rates improved. A difference-in-differences analysis showed that shifting from instructor demonstration to student adversarial learning yielded no significant additional gain in total scores (−2.55 points, $p = 0.645$). Item-level analysis also showed uneven gains across topics: curve fitting, algebraic precision, numerical integration, and numerical differentiation improved consistently, whereas Newton's method and portions of the complex iterative content showed only limited benefits. Generative AI can improve learning outcomes in engineering mathematics, but its value depends on what is being taught, how instructors scaffold the work, and how classroom tasks are designed. The shift from instructor-led demonstration to student-driven adversarial learning brought no extra benefit, and the gains varied sharply by topic. These patterns suggest that effectiveness is not a simple function of how much students interact with AI.**

## Keywords

**Generative Artificial Intelligence, *Numerical Methods*, Engineering Mathematics, Quasi-Experimental Study, Knowledge-Point Heterogeneity, Teaching Reform**

Copyright © 2026 by author(s) and Hans Publishers Inc.
This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).
http://creativecommons.org/licenses/by/4.0/
Open Access

## 1. 引言

《数值计算方法》承担着把数学公式转化为算法、程序和工程判断的任务。学生在学习插值、拟合、数值积分、常微分方程数值解、非线性方程迭代和线性方程组迭代时，遇到的障碍往往不是看不懂公式，而是说不清公式的适用条件，写不出对应的程序过程，也解释不了误差和收敛现象。课堂里一个常见的画面是：学生能照着例题算出一步，却答不上“为什么这一步成立”“误差从哪里来”“换一个初值结

果还稳不稳”。这种“会套公式”与“能解释过程”之间的断裂，正是工科数学教学长期难以通过增加练习量来解决的问题。

生成式人工智能进入高等教育场景后，这类断裂有了被重新组织的可能。UNESCO 在 2023 年发布的生成式 AI 教育与研究指南强调，高校需要在学习支持、教师角色、学术规范和风险治理之间建立新的平衡[1]。对教师而言，AI 的价值不只是替代板书，而是在讲解现场即时生成代码、绘制运行图形、构造反例，把原本停留在想象中的算法过程显性化；对学生而言，AI 意味着可以就同一个算法反复追问、验证和比较。但工具进入课堂，学习并不自动发生。本文关注的正是这个交叉点：在一门真实工科数学课程、两届自然班和可量化的学习结果中，观察生成式 AI 能够改善什么、不能改善什么，以及这种改善依赖怎样的课堂组织。

已有研究为这一问题提供了基础，但留下的空白同样明确。早期高等教育 AI 研究多集中在学习者画像与预测、评价与测评、自适应系统和智能导师等应用，Zawacki-Richter 等指出，该领域长期存在教师角色和教学设计讨论不足的问题[2]；Ouyang 等对 2011~2020 年在线高等教育 AI 实证研究的综述进一步表明，AI 虽能支持学习状态预测、资源推荐、自动评价和学习体验改进，但教育理论与过程数据仍需更紧密地进入设计环节[3]。Crompton 和 Burke 将 2016~2022 年高等教育 AI 研究归纳为评价、预测、AI 助手、智能导师和学习管理等类型，并指出 ChatGPT 类工具正在改变研究议程[4]。生成式 AI 出现后，Kasneci 等从学习者和教师两个视角讨论了大语言模型用于解释、反馈、内容生成和学习支持的机会，同时提醒其可靠性、依赖性和事实核查问题[5]；Lo 对 ChatGPT 教育文献的快速综述发现，ChatGPT 在不同学科中的表现并不均衡，尤其在数学任务中可能出现不稳定解答，需要教师培训和评价调整来回应[6]；Wardat 等围绕 ChatGPT 数学教学案例指出，其即时解释和互动潜力与数学误解纠正能力并存，复杂问题仍需谨慎使用[7]。在国内，朱永新和杨帆、陈玉琨分别从教育创新和教学变革角度提醒高校把生成式 AI 纳入课程重构，而不是停留在工具展示层面[8] [9]。上述文献共同说明了一个事实：生成式 AI 进入教育的可能价值和风险已被充分讨论，但针对工科数学课程、真实自然班、连续两年改革以及知识点层面异质性的准实验证据仍然不足，多数结论停留在“可以使用”的层面，缺少“怎样使用才有效”的细化答案。

本文的动机直接来自笔者连续两年的《数值计算方法》教学改革实践，也来自上述文献未能回答的具体问题。2024~2025 学年，同年级三个平行自然班采用常规教学，一个自然班在课堂中引入教师 AI 演示：教师讲完算法原理后，现场调用 AI 生成 Python 程序、绘制运行图形、解释误差变化，并引导学生把代码结果和数学公式对应起来。2025~2026 学年，同年级三个平行自然班继续采用常规教学，一个自然班进一步调整为学生 AI 对抗式学习：学生围绕当堂知识点提出问题，另一名学生借助 AI 尝试作答或验证，随后由提问者、作答者和教师共同判断答案质量。按 Ouyang 和 Jiao 提出的 AI 教育三种范式来看，第一轮更接近“AI 支持下的学习者协作”，第二轮则把学生推向更主动的 AI 赋能学习[10]。两次改革让本文得以同时回答三个此前研究很少放在一起检验的问题：第一，AI 辅助课堂是否在总分、及格率和优秀率上稳定优于常规课堂；第二，从教师 AI 演示转向学生 AI 对抗式学习，是否带来额外的总分增益；第三，不同知识点的受益是否一致，哪些内容更适合借助生成式 AI 展开，哪些内容仍需要教师强支架。

下文的结构安排如下：第二节说明研究对象、两轮教学改革实施、数据处理原则和统计方法，并在表 1 中给出两轮样本的描述性统计；第三节报告总体成绩差异，结合表 2 和图 1 展示 AI 辅助教学对总分分布的影响；第四节分析及格率和优秀率变化，结果见表 3、图 2 与图 3；第五节使用差异中的差异模型比较两轮改革模式的相对变化，见表 4 和图 4；第六节讨论十个知识点的异质性收益，见表 5、图 5 以及题目诊断表 6；第七节进一步呈现分数段分布与性别亚组的结果，见图 3、图 6 与表 7；最后在统计结果与课堂实践之间作解释，提出生成式 AI 辅助工科数学教学的适用边界与改进方向。

## 2. 研究设计

### (一) 研究对象

研究对象为某高校工科相关专业两届学生。2024~2025 学年纳入 23 级学生 121 人，其中一个自然班 24 人为实验班，三个平行自然班共 97 人为对照班；2025~2026 学年纳入 24 级学生 105 人，其中一个自然班 21 人为实验班，三个平行自然班共 84 人为对照班。两年合计 226 人，实验班 45 人，对照班 181 人。

本研究采用自然班级条件下的准实验设计。由于班级并非随机分配，实验班与对照班在生源基础、学习习惯和专业方向等方面可能存在差异，因此本文将结果解释为课堂改革的准实验证据，不作超出数据范围的强因果断言。这一设计牺牲了随机化的内部效度，换取了在真实教学现场观察两种 AI 组织方式的可行性与外部效度。

### (二) 教学改革实施

第一轮改革面向 23 级实验班，采用教师 AI 演示模式。课堂仍以算法原理讲解为主，AI 出现在讲解之后的验证环节：教师现场生成 Python 程序，结合运行结果、图形输出和误差变化解释算法过程，同时负责筛选 AI 生成内容，指出代码与数学解释中可能存在的问题，引导学生把代码结果与算法原理对应起来。该模式的重点不是让 AI 替代讲解，而是把抽象算法的运行过程变得可观察、可追问。

第二轮改革面向 24 级实验班，采用学生 AI 对抗式学习模式。课堂活动围绕“提问－作答－验证－反思”展开：学生先根据当前知识点提出问题，另一名学生借助 AI 尝试作答或验证，随后由提问者、作答者和教师共同判断答案质量，对低分或错误较多的学生，要求其补充错因说明与修正过程。与第一轮相比，第二轮的主要变化是 AI 使用主体由教师转向学生，课堂重心由教师演示转向学生生成问题与相互检验。

### (三) 考核与数据处理

两年均采用同类分项考核方式，总分 100 分，覆盖 10 个知识点，每个知识点满分 10 分，包括 Lagrange 插值、Newton 插值、曲线拟合、数值积分、代数精度、数值微分、Runge-Kutta 法、Newton 法、Gauss-Seidel 迭代和 LU 分解。考核以闭卷笔试为主，重点评价算法步骤、计算过程和结果解释。

数据来自两份课程考核 Excel 表，原始数据、清洗后的中间结果与分析脚本随论文源文件一并保存。清洗过程严格对应论文声明：剔除 23 级数据中的“满分”说明行；剔除 25~26 学年表中不属于 24 级的旧年级记录；统一处理班级名称中的空格与制表符；根据教学改革方案将实施 AI 辅助教学的自然班标记为实验班，其余同年级平行自然班标记为对照班。经核对，逐题成绩中的空白单元格代表该题 0 分而非数据缺失，因此在逐题分析中按下分处理；总分数据无缺失。全部统计量与图表均由脚本 analysis_repro.py 以固定随机种子重新生成，可完整复现本文表 1~7 与图 1~6。

### (四) 统计方法

连续变量以均值 ± 标准差表示。总分组间比较采用 Welch t 检验，并以 Mann-Whitney U 检验作为稳健性检验。考虑实验班样本量较小，效应量采用 Hedges g [11]，均值差与效应量的置信区间采用非参数 Bootstrap 方法估计(5000 次重抽样) [12]。及格率和优秀率采用 $\chi^2$ 检验和 Fisher 精确检验，并报告比值比(odds ratio, OR)及其近似置信区间。逐知识点检验采用 Benjamini-Hochberg 方法控制错误发现率(FDR) [13]。为检验第二轮教学模式是否产生额外总分增益，建立如下差异中的差异模型：

$$Score = \beta^0 + \beta^1 Treatment + \beta^2 Year2026 + \beta^3 (Treatment \times Year2026) + \varepsilon \quad (1)$$

其中 $\beta^3$ 表示学生 AI 对抗式学习相对于教师 AI 演示的额外变化，回归标准误采用 HC3 稳健估计[14]。需要预先说明的是，实验班每年仅一个自然班(23 级 24 人、24 级 21 人)，效应估计的把握度有限，因此下文对不显著的结果一律作保守解释，不把“未检出差异”读作“没有差异”。

## 3. 结果

### (一) 总分分布与组间差异

表 1 显示，两年实验班均值均高于对应对照班。23 级实验班总分为 78.17 ± 12.37，对照班为 63.62 ± 19.78；24 级实验班总分为 76.95 ± 17.43，对照班为 64.95 ± 19.52。从标准差看，实验班内部差异仍然明显，但整体位置已向高分段移动。

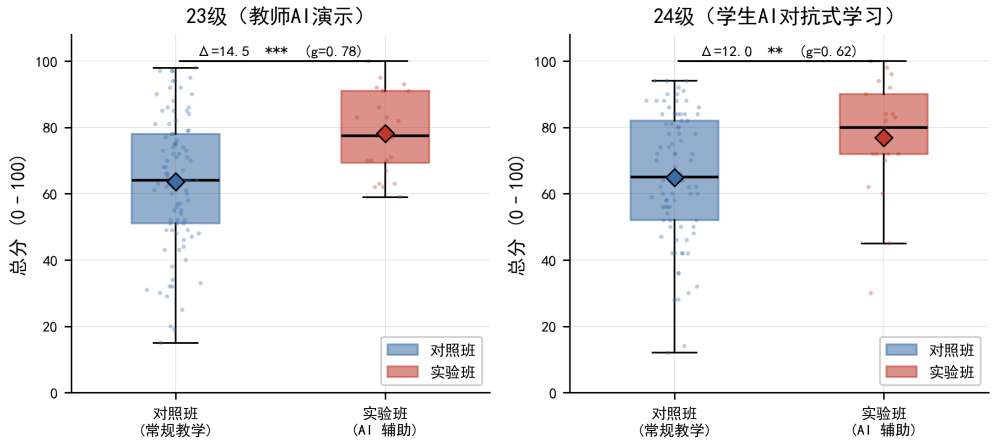
**Table 1.** Descriptive statistics of total scores in two years
**表 1.** 两年总分描述统计

| 年级 | 组别 | n | 均值 ± 标准差 | 中位数 | 最小值 | 最大值 |
|---|---|---|---|---|---|---|
| 23 级 | 对照班 | 97 | 63.62 ± 19.78 | 64.0 | 15 | 98 |
| 23 级 | 实验班 | 24 | 78.17 ± 12.37 | 77.5 | 59 | 100 |
| 24 级 | 对照班 | 84 | 64.95 ± 19.52 | 65.0 | 12 | 94 |
| 24 级 | 实验班 | 21 | 76.95 ± 17.43 | 80.0 | 30 | 100 |

组间检验进一步支持这一差异。表 2 显示，23 级实验班较对照班高 14.55 分，Bootstrap 95%CI 为 [8.27, 20.77]，Welch t = 4.51，p < 0.001，Hedges g = 0.78；24 级实验班较对照班高 12.00 分，Bootstrap 95%CI 为[3.38, 20.17]，Welch t = 2.75，p = 0.009，Hedges g = 0.62。Mann-Whitney U 检验方向一致，说明结果并非只由均值检验的假设推动。图 1 以箱线图叠加抖动散点展示了两年成绩分布，实验班低分段学生减少、高分段更集中，菱形标记为各组均值。

**Table 2.** Comparison of total scores between groups
**表 2.** 总分组间比较

| 年级 | 对照班均值 | 实验班均值 | 均值差(95%CI) | Welch t | p | Mann-Whitney p | Hedges g |
|---|---|---|---|---|---|---|---|
| 23 级 | 63.62 | 78.17 | 14.55 [8.27, 20.77] | 4.51 | < 0.001 | 0.001 | 0.78 |
| 24 级 | 64.95 | 76.95 | 12.00 [3.38, 20.17] | 2.75 | 0.009 | 0.011 | 0.62 |




菱形为各组均值，括号内给出均值差、显著性标记与 Hedges g。

**Figure 1.** Effect of AI-assisted teaching on total score distribution and group mean differences
**图 1.** AI 辅助教学对总分分布和组间均值差的影响

### (二) 及格率、优秀率与分数段迁移

把成绩转化为更常用的及格率和优秀率后，实验班的优势更为直观。23 级实验班及格率为 95.8%，对照班为 62.9%；24 级实验班及格率为 90.5%，对照班为 61.9%。优秀率方面，23 级实验班为 29.2%，对照班为 11.3%；24 级实验班为 28.6%，对照班为 8.3%。表 3 列出分类结果检验，由于实验班样本量较小，同时报告 Fisher 精确检验。

**Table 3.** Comparison of pass rates and excellence rates
**表 3.** 及格率和优秀率比较

| 年级 | 指标 | 实验班 | 对照班 | 风险差(百分点) | OR (95%CI) | Fisher p |
|---|---|---|---|---|---|---|
| 23 级 | 及格 | 23/24 (95.8%) | 61/97 (62.9%) | 32.9 | 13.57 [1.76, 104.82] | 0.001 |
| 23 级 | 优秀 | 7/24 (29.2%) | 11/97 (11.3%) | 17.8 | 3.22 [1.09, 9.49] | 0.049 |
| 24 级 | 及格 | 19/21 (90.5%) | 52/84 (61.9%) | 28.6 | 5.85 [1.28, 26.79] | 0.017 |
| 24 级 | 优秀 | 6/21 (28.6%) | 7/84 (8.3%) | 20.2 | 4.40 [1.30, 14.94] | 0.021 |

注：风险差为实验班比例减去对照班比例；OR 为实验班相对于对照班的比值比。

图 2 给出两年及格率与优秀率的分组对比。更值得追问的是这些提升来自哪个分数段。图 3 将每个学生归入优秀、良好、中等、及格、不及格五个分数段并作百分比堆叠，可以清楚看到：实验班相对对照班的变化主要体现为“不及格”一档的收缩与“及格–良好”一档的扩张，而非高端优秀率的普遍抬升。AI 辅助教学对中下游学生的托底作用比对尖子生的拔高作用更明显，这正是后续讨论中“降低算法实现门槛”这一机制的分数段证据。

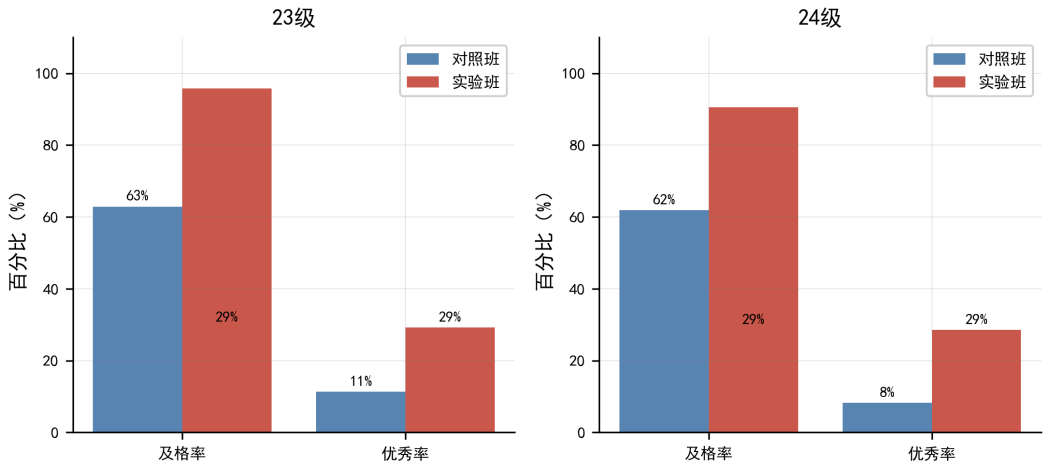



**Figure 2.** Effect of AI-assisted teaching on pass rates and excellence rates
**图 2.** AI 辅助教学对及格率和优秀率的影响

### (三) 模式升级的额外增益

两年实验班都高于对照班，但这并不等同于第二轮“学生 AI 对抗式学习”一定优于第一轮“教师 AI 演示”。更稳妥的比较是看两年实验班相对于各自对照班的差异是否继续扩大：23 级实验班相对对照班高 14.55 分，24 级实验班相对对照班高 12.00 分，差异中的差异为−2.55 分。表 4 显示，交互项不显著($p = 0.645$)。图 4 以 2 × 2 交互图呈现两组均值及其 95%置信区间，两条连线近似平行，直观地说明第二轮并未在总分层面拉开与第一轮的相对差距。

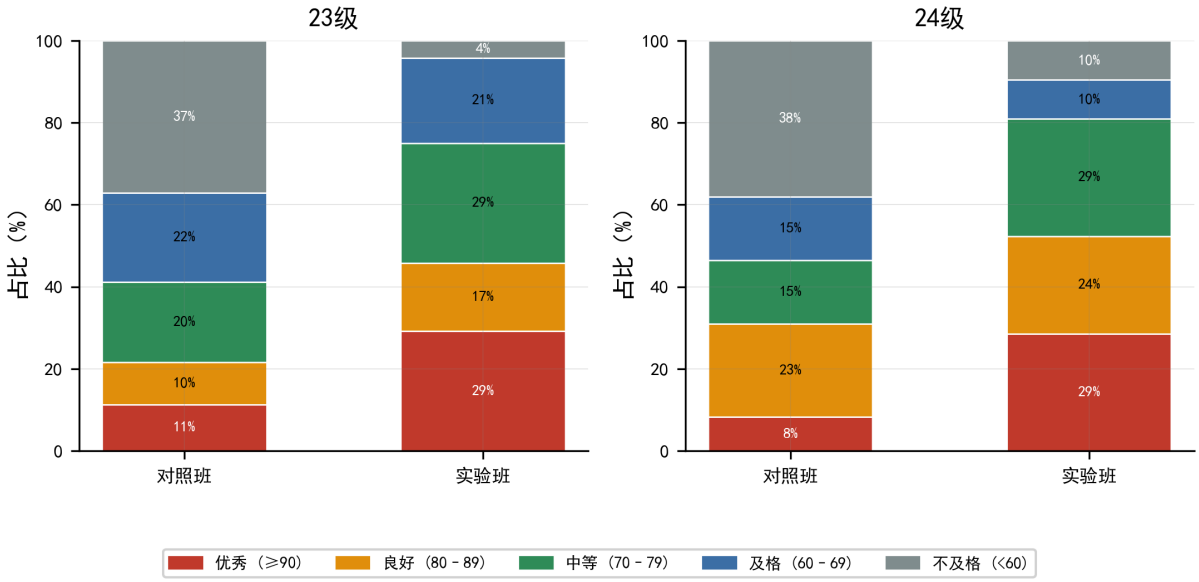



**Figure 3.** Score band composition of experimental and control classes in two years
**图 3.** 两年实验班与对照班的分数段构成。AI 辅助教学的增益主要来自不及格学生向及格－良好段的迁移

**Table 4.** Difference-in-differences regression results
**表 4.** 差异中的差异回归结果

| 变量 | 系数 | HC3 稳健标准误 | t | p |
|---|---|---|---|---|
| 截距 | 63.62 | 2.02 | 31.51 | <0.001 |
| 实验班 | 14.55 | 3.27 | 4.44 | <0.001 |
| 24 级 | 1.33 | 2.94 | 0.45 | 0.651 |
| 实验班 × 24 级 | −2.55 | 5.52 | −0.46 | 0.645 |

注：参照组为 23 级对照班；标准误采用 HC3 稳健估计。

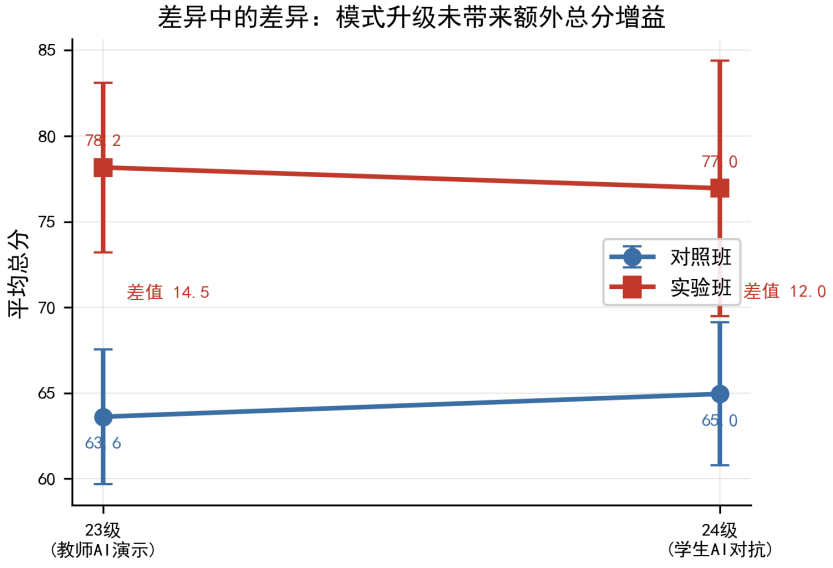



**Figure 4.** Difference-in-differences analysis of two rounds of quasi-experiments
**图 4.** 两轮准实验的差异中的差异分析。误差棒为均值 95%置信区间，两条连线近似平行说明模式升级未带来额外总分增益

这一结果提示，学生 AI 对抗式学习虽然提高了课堂活动的参与性，但并未在总分层面产生显著额外收益。这说明，AI 辅助教学的关键不在于学生使用 AI 的时间多少，而在于其使用是否与知识点性质和课堂任务结构相匹配。

**(四) 知识点层面的异质性**

逐知识点结果揭示了更细的差异。表 5 和图 5 显示，23 级中曲线拟合、代数精度、Runge-Kutta 法、Gauss-Seidel 迭代和 LU 分解在 FDR 校正后仍达到显著水平；24 级中曲线拟合、数值积分、代数精度和数值微分在 FDR 校正后仍显著。相比之下，Newton 法在两年中均未表现出正向优势；Gauss-Seidel 迭代在 23 级教师演示模式下效果较大，但在 24 级学生对抗式学习中优势消失。图 5 中实心点表示 FDR 校正后 $q < 0.05$，空心点表示未达该标准，可见显著效应集中在少数可计算、可视化的知识点。

**Table 5.** Item-level effect sizes and FDR correction
**表 5.** 逐知识点效应量与 FDR 校正

| 年级 | 知识点 | 实验班均值 | 对照班均值 | 均值差 | Hedges g(95%CI) | p | FDR q |
|---|---|---|---|---|---|---|---|
| 23 级 | Lagrange 插值 | 8.08 | 7.60 | 0.49 | 0.23 [−0.17, 0.62] | 0.268 | 0.334 |
| 23 级 | Newton 插值 | 8.21 | 7.39 | 0.82 | 0.31 [−0.11, 0.71] | 0.149 | 0.248 |
| 23 级 | 曲线拟合 | 8.58 | 4.89 | 3.70 | 0.93 [0.54, 1.34] | <0.001 | <0.001 |
| 23 级 | 数值积分 | 7.71 | 7.01 | 0.70 | 0.25 [−0.18, 0.66] | 0.262 | 0.334 |
| 23 级 | 代数精度 | 9.12 | 7.59 | 1.54 | 0.47 [0.12, 0.75] | 0.008 | 0.021 |
| 23 级 | 数值微分 | 7.33 | 6.64 | 0.69 | 0.19 [−0.20, 0.56] | 0.336 | 0.374 |
| 23 级 | Runge-Kutta 法 | 8.12 | 5.62 | 2.51 | 0.68 [0.39, 0.98] | <0.001 | <0.001 |
| 23 级 | Newton 法 | 6.04 | 6.35 | −0.31 | −0.09 [−0.57, 0.37] | 0.695 | 0.695 |
| 23 级 | Gauss-Seidel 迭代 | 7.29 | 4.31 | 2.98 | 0.92 [0.56, 1.32] | <0.001 | <0.001 |
| 23 级 | LU 分解 | 7.67 | 6.23 | 1.44 | 0.45 [0.09, 0.80] | 0.017 | 0.033 |
| 24 级 | Lagrange 插值 | 9.14 | 8.55 | 0.60 | 0.30 [−0.08, 0.63] | 0.119 | 0.169 |
| 24 级 | Newton 插值 | 9.05 | 7.96 | 1.08 | 0.36 [−0.07, 0.75] | 0.099 | 0.166 |
| 24 级 | 曲线拟合 | 9.57 | 7.76 | 1.81 | 0.57 [0.32, 0.78] | <0.001 | 0.001 |
| 24 级 | 数值积分 | 8.57 | 6.90 | 1.67 | 0.55 [0.10, 0.97] | 0.017 | 0.042 |
| 24 级 | 代数精度 | 8.67 | 5.76 | 2.90 | 0.72 [0.34, 1.09] | <0.001 | 0.002 |
| 24 级 | 数值微分 | 8.48 | 6.38 | 2.10 | 0.54 [0.13, 0.93] | 0.012 | 0.040 |
| 24 级 | Runge-Kutta 法 | 5.14 | 4.06 | 1.08 | 0.30 [−0.20, 0.82] | 0.253 | 0.316 |
| 24 级 | Newton 法 | 6.95 | 7.79 | −0.83 | −0.24 [−0.79, 0.27] | 0.372 | 0.414 |
| 24 级 | Gauss-Seidel 迭代 | 3.52 | 3.57 | −0.05 | −0.02 [−0.54, 0.52] | 0.953 | 0.953 |
| 24 级 | LU 分解 | 7.86 | 6.21 | 1.64 | 0.46 [0.03, 0.89] | 0.042 | 0.084 |

注：逐题空白分数按下分处理；q 为 Benjamini-Hochberg FDR 校正后结果。

表 6 从题目难度与区分度的角度补充了上述异质性的来源。以 23 级为例，受益最大的曲线拟合与 Gauss-Seidel 迭代在对照班的平均得分率本就偏低(分别为 0.49 与 0.43)，说明其瓶颈主要在计算执行与可视化；而 Newton 法对照班得分率并不低(0.64)却未获增益，提示其瓶颈在于收敛性判断与理论约束，并非 AI 擅长的执行环节。可见，同样是“难”的知识点，难点性质不同，AI 能起的作用也不同。

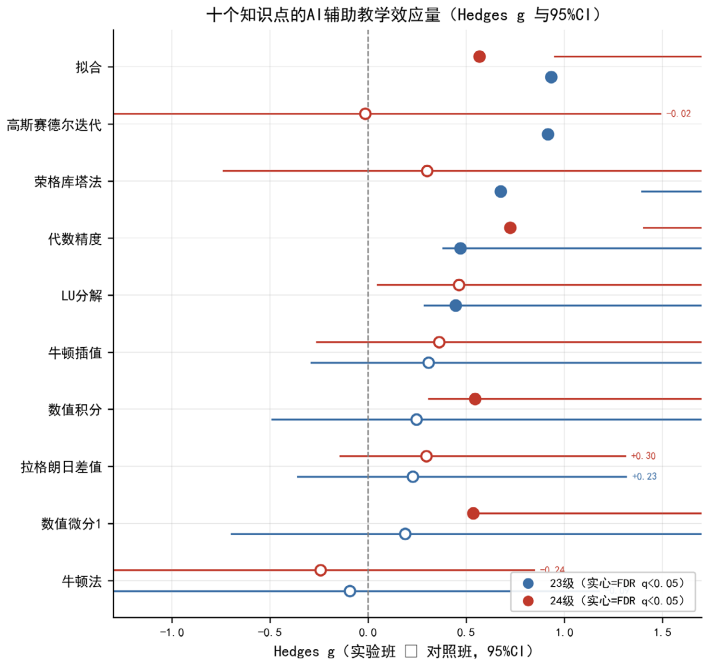



**Figure 5.** Effect sizes of AI-assisted teaching by knowledge point (Hedges g and 95% CI)
**图 5.** 不同知识点的 AI 辅助教学效应量(Hedges g 及 95%CI)。实心点表示 FDR 校正后 q < 0.05

**Table 6.** Item difficulty and discrimination of representative knowledge points (excerpt)
**表 6.** 代表性知识点的题目难度与区分度(节选)

| 年级 | 知识点 | 组别 | 难度系数* | 区分度† |
|---|---|---|---|---|
| 23 级 | 曲线拟合 | 对照班 | 0.49 | 0.41 |
| 23 级 | 曲线拟合 | 实验班 | 0.86 | 0.22 |
| 23 级 | Gauss-Seidel 迭代 | 对照班 | 0.43 | 0.38 |
| 23 级 | Gauss-Seidel 迭代 | 实验班 | 0.73 | 0.31 |
| 23 级 | Newton 法 | 对照班 | 0.64 | 0.33 |
| 23 级 | Newton 法 | 实验班 | 0.60 | 0.35 |
| 24 级 | 代数精度 | 对照班 | 0.58 | 0.36 |
| 24 级 | 代数精度 | 实验班 | 0.87 | 0.19 |
| 24 级 | Gauss-Seidel 迭代 | 对照班 | 0.36 | 0.40 |
| 24 级 | Gauss-Seidel 迭代 | 实验班 | 0.35 | 0.41 |

*难度系数 = 该知识点平均得分/满分；†区分度 = 高分组(前 27%)与低分组(后 27%)得分率之差。完整结果见随文数据表 table_item_diagnostics.csv。

### (五) 分数段迁移与性别亚组

除总体与知识点层面外，两类补充结果对理解“谁受益、从哪受益”有帮助。图 3 已经显示增益主要来自不及格学生向及格–良好段的迁移；图 6 与表 7 进一步给出性别亚组的总分均值。24 级实验班女生均值为 85.50 分(n = 14)，而同班男生仅 59.86 分(n = 7)，甚至低于该年级对照班男生(58.42 分)。23 级

则呈相反方向，实验班男生(80.29 分)高于女生(77.29 分)。性别效应在两个 cohort 中方向不一致，且每一年实验班男生样本仅 7 人，因此这里只能作为假设生成的观察，不能得出“AI 对某一性别更有效”的结论；但它提示后续研究需要系统记录学生使用 AI 的过程数据，才可能解释这种波动。

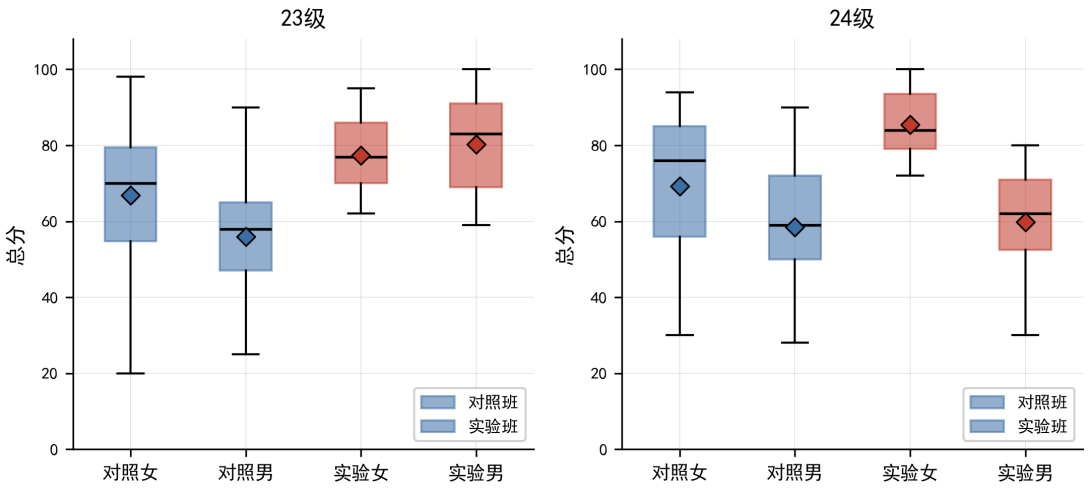



**Figure 6.** Gender subgroup total score distribution of experimental and control classes
**图 6.** 两年实验班与对照班的性别亚组总分分布。实验班男生样本量小(n = 7)，亚组结果仅作提示性参考

**Table 7.** Descriptive statistics of gender subgroups
**表 7.** 性别亚组描述统计

| 年级 | 组别 | 性别 | n | 均值 ± 标准差 |
|---|---|---|---|---|
| 23 级 | 对照班 | 女 | 68 | 66.90 ± 19.65 |
| 23 级 | 对照班 | 男 | 29 | 55.93 ± 18.17 |
| 23 级 | 实验班 | 女 | 17 | 77.29 ± 11.54 |
| 23 级 | 实验班 | 男 | 7 | 80.29 ± 14.95 |
| 24 级 | 对照班 | 女 | 51 | 69.18 ± 20.04 |
| 24 级 | 对照班 | 男 | 33 | 58.42 ± 16.98 |
| 24 级 | 实验班 | 女 | 14 | 85.50 ± 9.76 |
| 24 级 | 实验班 | 男 | 7 | 59.86 ± 17.19 |

## 4. 讨论

### (一) AI 的主要作用是把算法过程显性化

两年数据共同指向同一种解释：AI 辅助教学之所以有效，并不是因为它替代了教师讲解，而是因为它把许多原本停留在板书和想象中的算法过程显性化了。学生能够在课堂上看到代码如何生成、参数如何变化、曲线如何拟合、误差如何累积。从多媒体学习与认知负荷的角度看，恰当的图形、代码和语言解释可以帮助学生把分散的信息组织为可理解的过程，但前提是教师需要控制信息密度，避免把 AI 输出变成新的负担[15] [16]。这也与图 3 的分数段证据一致：显性化最直观地帮助了卡在技术执行环节的中下游学生越过及格线，而不是把已经达标的学生再往上推。

### (二) 学生自主使用 AI 需要明确的任务约束

第二轮改革将 AI 使用主体从教师转向学生，但差异中的差异分析并未发现总分层面的额外增益。这

一结果并不说明学生使用 AI 没有价值，而是揭示了一个常被忽视的约束：AI 进入课堂后，“使用得更多”并不等于“学得更深”。如果学生提出的问题停留在快速求解和答案验证层面，AI 可能只是提高了课堂活动的热度；只有当问题指向概念边界、算法条件、误差判断和方法比较时，AI 才可能转化为深层学习工具。

这与技术整合研究的基本判断一致：技术、学科内容和教学法需要共同设计，而不是简单叠加[17]。以 Gauss-Seidel 迭代为例，当它在 24 级从“教师演示受益最大”变成“学生对抗式学习下无增益”(表 5)，恰好说明同一知识点在不同任务结构下结果可以完全反转。教师如果只把 AI 交给学生自行探索，收敛条件、初值敏感性这类理论约束很容易被跳过。

**(三) 不同知识点不宜采用同一种 AI 策略**

逐知识点结果表明，AI 辅助教学并不适合以同一种方式覆盖所有内容。曲线拟合、代数精度、数值积分和数值微分等内容较容易受益，因为它们本就需要在公式、表格、代码和图形之间建立联系，AI 可以帮助学生较快看到计算结果与可视化效果，使课堂讨论更容易围绕算法意义展开。

Newton 法和复杂迭代类内容则呈现不同情形。这类知识点不仅涉及计算步骤，还涉及初值选择、收敛性判断、误差控制和方法适用条件。表 6 中 Newton 法对照班得分率并不低却未获增益，说明其瓶颈不在“算不出”，而在“说不清为什么收敛、为什么不收敛”。对这类内容，AI 更适合作为辅助解释、反例生成和错误诊断工具，而不是主要的解题工具。

**(四) 异质性的来源：难度、区分度与任务结构**

把表 5 的效应量与表 6 的难度、区分度放在一起看，可以得出更具体的判断。AI 增益集中的知识点(曲线拟合、Gauss-Seidel 迭代、代数精度)在对照班普遍呈现“低得分率、高区分度”，即多数学生本就在这些地方失分，且失分能拉开学生差距；AI 通过把计算执行与可视化过程显性化，压缩了这部分本可由训练量弥补的差距。相反，Newton 法在两年中效应量均为负，却并非因为“太简单”——它的难点是收敛性分析，属于 AI 不容易替学生完成的理论判断。因此，判断一个知识点是否适合 AI 辅助，不能只看“难不难”，而要看“难点在执行还是判断”。这一区分比笼统地说“可视化内容更适合 AI”更接近教学现场的真实机制。

**(五) 研究局限**

本文的局限主要来自研究现场本身，也来自上述亚组结果暴露出的盲区。第一，实验班与对照班来自自然班级，未能随机分组，班级基础差异可能影响结果解释；第二，实验班样本量较小，部分知识点效应量置信区间较宽，且性别亚组每年仅 7 名男生，结论的把握度有限；第三，评价材料主要来自闭卷分项考核，尚未充分覆盖编程实践能力、长期保持效果和迁移能力；第四，学生使用 AI 的过程数据(提问质量、提示词特征、追问次数、课堂互动轨迹)没有被系统记录，这正是性别波动等现象无法被解释的原因。后续研究若能结合课堂日志、学习过程记录和访谈材料，将更有助于回答“AI 究竟通过哪些环节发挥作用”。

## 5. 结论与建议

基于两年 226 名学生的准实验数据，本文发现生成式 AI 辅助教学能够改善《数值计算方法》课程的学习表现，并明显提高及格率和优秀率。两轮改革均显示稳定正向效果，但从教师 AI 演示转向学生 AI 对抗式学习，并未在总分层面产生显著额外增益。AI 的教学价值更确切地体现在知识点层面的结构性改善：计算实现型、可视化程度较高的内容受益更稳定，理论推导与复杂迭代内容则需要更强的教师支架。

由此提出三点具体建议。第一，AI 应优先服务于课程难点中的“执行缺口”，而不是作为课堂装饰；对插值、拟合、积分、微分等可计算、可可视化的内容，可以更多使用 AI 生成代码与图形，把学生的认

知资源从机械计算转移到算法理解。第二，学生自主使用 AI 时必须配套问题规范和评价标准，把“能算出答案”提升为“能解释条件、比较方法、诊断错误”，否则对抗式学习容易退化为热闹的问答。第三，对 Newton 法、Gauss-Seidel 迭代等需要理论判断和收敛性分析的内容，教师仍应占据支架位置，要求学生说明收敛条件、比较方法误差、解释初值影响，AI 只作为辅助解释与反例生成工具。只有把 AI 的使用方式、知识点性质和课堂任务结构对齐，生成式人工智能才能真正从“课堂里的新东西”变成“工程教育中可依赖的理解工具”。

## 基金项目

浙江省“十四五”第二批本科省级教学改革备案项目(JGBA2024362)，浙江省“十四五”第二批研究生省级教学改革常规项目的(JGCG2024162)资助。

## 参考文献

[1] Miao, F. and Holmes, W. (2023) Guidance for Generative AI in Education and Research. UNESCO.

[2] Zawacki-Richter, O., Marín, V.I., Bond, M. and Gouverneur, F. (2019) Systematic Review of Research on Artificial Intelligence Applications in Higher Education—Where Are the Educators? *International Journal of Educational Technology in Higher Education*, **16**, Article No. 39. https://doi.org/10.1186/s41239-019-0171-0

[3] Ouyang, F., Zheng, L. and Jiao, P. (2022) Artificial Intelligence in Online Higher Education: A Systematic Review of Empirical Research from 2011 to 2020. *Education and Information Technologies*, **27**, 7893-7925. https://doi.org/10.1007/s10639-022-10925-9

[4] Crompton, H. and Burke, D. (2023) Artificial Intelligence in Higher Education: The State of the Field. *International Journal of Educational Technology in Higher Education*, **20**, Article No. 22. https://doi.org/10.1186/s41239-023-00392-8

[5] Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., *et al*. (2023) ChatGPT for Good? On Opportunities and Challenges of Large Language Models for Education. *Learning and Individual Differences*, **103**, Article 102274. https://doi.org/10.1016/j.lindif.2023.102274

[6] Lo, C.K. (2023) What Is the Impact of ChatGPT on Education? A Rapid Review of the Literature. *Education Sciences*, **13**, Article 410. https://doi.org/10.3390/educsci13040410

[7] Wardat, Y., Tashtoush, M.A., AlAli, R. and Jarrah, A.M. (2023) ChatGPT: A Revolutionary Tool for Teaching and Learning Mathematics. *Eurasia Journal of Mathematics*, *Science and Technology Education*, **19**, em2286. https://doi.org/10.29333/ejmste/13272

[8] 朱永新，杨帆. ChatGPT/生成式人工智能与教育创新：机遇、挑战以及未来[J]. 华东师范大学学报(教育科学版), 2023, 41(7): 1-14.

[9] 陈玉琨. ChatGPT/生成式人工智能时代的教育变革[J]. 华东师范大学学报(教育科学版), 2023, 41(7): 103-116.

[10] Ouyang, F. and Jiao, P. (2021) Artificial Intelligence in Education: The Three Paradigms. *Computers and Education*: *Artificial Intelligence*, **2**, Article 100020. https://doi.org/10.1016/j.caeai.2021.100020

[11] Hedges, L.V. (1981) Distribution Theory for Glass's Estimator of Effect Size and Related Estimators. *Journal of Educational Statistics*, **6**, 107-128. https://doi.org/10.3102/10769986006002107

[12] Efron, B. and Tibshirani, R.J. (1994) An Introduction to the Bootstrap. Chapman and Hall/CRC.

[13] Benjamini, Y. and Hochberg, Y. (1995) Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society Series B*: *Statistical Methodology*, **57**, 289-300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

[14] MacKinnon, J.G. and White, H. (1985) Some Heteroskedasticity-Consistent Covariance Matrix Estimators with Improved Finite Sample Properties. *Journal of Econometrics*, **29**, 305-325. https://doi.org/10.1016/0304-4076(85)90158-7

[15] Sweller, J. (1988) Cognitive Load during Problem Solving: Effects on Learning. *Cognitive Science*, **12**, 257-285. https://doi.org/10.1207/s15516709cog1202_4

[16] Mayer, R.E. (2020) Multimedia Learning. 3rd Edition, Cambridge University Press.

[17] Mishra, P. and Koehler, M.J. (2006) Technological Pedagogical Content Knowledge: A Framework for Teacher Knowledge. *Teachers College Record*: *The Voice of Scholarship in Education*, **108**, 1017-1054. https://doi.org/10.1111/j.1467-9620.2006.00684.x