# The Impact of AI Assistance on the Quality of Undergraduate Graduation Design: An Empirical Study Based on Counterfactual Inference

Honghe Shi and Bin Duan*

School of Automation and Electronic Information, Xiangtan University, Xiangtan City, Hunan Province, China
Email: 874872786@qq.com (H.S.); db61850@163.com (B.D.)
*Corresponding author
Manuscript received May 30, 2025; accepted August 11, 2025; published December 25, 2025.

***Abstract*—Generative AI is rapidly penetrating higher education, and its net effect on the quality of undergraduate graduation theses remains controversial. This study aims to estimate the net causal effect of AI-assisted writing on the grades of undergraduate graduation theses. The research collected the thesis scoring and questionnaire data of 120 undergraduate students in a certain university. A counterfactual framework with propensity score matching was adopted and supplemented by semi-structured interviews with 10 students. The results showed that the average score of the graduation project of students using AI tools increased by approximately 5.75 points ($p < 0.01$). Interviews show that AI can enhance writing efficiency and standardization, but there is a risk of over-reliance. Colleges and universities should enhance the training of AI literacy and academic integrity to give full play to its value and avoid potential risks.**

***Keywords*—AI-assisted writing, propensity score matching, counterfactual inference, undergraduate graduation thesis**

## I. Introduction

With the rapid development of generative Artificial Intelligence (AI) technology, AI-assisted tools are increasingly being applied in the field of higher education. Especially the rise of large language models like ChatGPT has provided students with assistance in multiple aspects such as immediate feedback, academic writing support, and information retrieval [1–3]. Current research indicates that AI tools have the potential to enhance students' writing skills, learning efficiency and innovation capabilities. However, there are also views suggesting that excessive AI intervention may affect students' independent thinking and cognitive development. Therefore, there is still a lack of consistent conclusions on the actual impact of AI-assisted tools on students' learning outcomes, especially in highly comprehensive educational links such as undergraduate graduation projects. Its mechanism of action and benefits urgently need empirical evaluation [4, 5].

This study focuses on the core issue of "whether AI-assisted tools can significantly improve the quality of undergraduate graduation projects", aiming to systematically evaluate the impact of AI-assisted tools on students' graduation project grades through rigorous quantitative and qualitative methods. The research not only examines the direct effect of AI tools on the scoring of graduation projects, but also focuses on their specific manifestations in improving the standardization of writing, technical details and efficiency. By revealing the mechanism of the effect of AI tools on students' learning outcomes, this paper provides empirical evidence for universities to scientifically introduce AI-assisted tools and improve the quality of talent cultivation.

In terms of the theoretical framework, the research is based on the cognitive load theory and holds that AI-assisted tools can effectively reduce students' external cognitive load by sharing low-level cumbersome tasks (such as grammar proofreading and format adjustment), enabling them to devote more energy to content innovation and logical thinking. Combining the theories of motivation and self-efficacy, this paper further analyzes how AI tools can enhance students' learning confidence and motivation through personalized feedback, while being vigilant about the possible risk of dependence it may cause [6]. Theoretical analysis provides a basis for understanding the positive and negative effects of AI tools and also sets the core variables for subsequent empirical research.

The innovation of this study is mainly reflected in two aspects: First, based on counterfactual inference and propensity score matching, the actual educational effect of AI tools in non-experimental environments is empirically evaluated to enhance the credibility of causal inference; Second, by integrating semi-structured interviews, systematically reveal the specific paths of AI tools in enhancing students' learning quality, influencing cognitive and motivational mechanisms, etc.

## II. Theoretical Basis

### A. The Application of AI in Education

Since large language models came into the public eye at the end of 2022, generative artificial intelligence has rapidly permeated and integrated into core links such as teaching, learning and evaluation in higher education [7]. Its core technical capabilities, including natural language generation, semantic retrieval and multimodal reasoning, have demonstrated significant application advantages in educational scenarios such as academic writing assistance, program code development, literature verification and immediate feedback.

Existing empirical studies generally show that by rationally applying large language models, the external cognitive load of students in basic learning activities such as grammar correction and literature review can be effectively reduced. This reduction in load enables students to invest more cognitive resources in higher-order thinking activities, such as constructing rigorous argumentative structures and conducting innovative designs. Meanwhile, the rapid response ability demonstrated by the model during the

conversational Q&A process can also significantly enhance students' sense of task control and self-efficacy.

Overall, generative artificial intelligence shows considerable potential in enhancing learning efficiency and improving the surface quality of academic outcomes. This technology-enabled educational model has opened up a new path for optimizing the learning process and enhancing the efficiency of knowledge acquisition and application.

### B. Counterfactual Causal Inference Framework

In the field of causal inference, the Rubin Potential Outcome model provides us with the theoretical cornerstone for defining intervention effects [8]. The model assumes that each research subject corresponds to two potential outcomes: denoted as Y(1) when receiving intervention and as Y(0) when not receiving intervention. Since an individual can only be in one processing state at the same time, we can observe at most one of the results, and the other result is in a "counterfactual" state. This gives rise to the fundamental problem that "causal effects cannot be directly observed" – what we truly care about is the average treatment effect:

$$ATE = E[Y(1) - Y(0)] \quad (1)$$

that is, the expectation of the result difference of all objects in the two potential worlds.

Fig. 1 visually presents this logical chain: The leftmost box "Student & Covariates X" represents the observable covariates (such as GPA, Topic Type, Gender, Self-Efficacy, etc.). These covariates simultaneously affect whether students Use AI (above "Treatment Z: AI Use vs No AI Use") and their graduation project scores. The downward arrow leads to "Observed Outcome Y", reminding us that each student can only observe one of Y(1) or Y(0). The "Counterfactual Outcome" box on the right indicates the potential outcomes that can never be directly observed, while the two nodes "Propensity Score e(X)" and "Matching" below correspond to the propensity score estimation and matching steps. It is used to construct a quasi-random experiment and ultimately obtain the ATE.

To enable observational data to support causal identification, we need to meet two key assumptions:

- Conditional independence (no confusion)

$(Y(1),Y(0)) \perp Z \mid X$. Given the covariate X, the processing allocation is approximately random. In Fig. 1, this is reflected through the covariates pointing to the processing and result paths – as long as X is successfully controlled, the "detours" in the figure will be "truncated".

- Mutual support (overlap)

For any value of X, it satisfies $0 < P(Z = 1 \mid X) < 1$. Corresponding to the illustration, only when there is sufficient overlap between the treatment group and the control group in the covariate space can similar control objects be found for each treatment individual in the "Matching" stage.

Under this framework, Propensity Score Matching (PSM) [9–11] is the most commonly used quasi-experimental technique. The propensity score is defined as:

$$e(X) = P(Z = 1|X) \quad (2)$$

As shown in Fig. 1, we first estimate e(X) using covariates. Then, at the "Matching" node, we pair the students in the AI group with those in the non-AI group based on the similarity tendency score, so that the two groups are as balanced as possible on X and simulate the random allocation situation. Subsequently, by comparing the average differences of Y in the paired samples, the causal effect estimation of AI usage can be obtained.

The matching quality strongly depends on the accuracy of the propensity score model and the comprehensiveness of covariates. If the estimation error is large, key covariates are omitted, or a too narrow caliper leads to the discarding of a large number of samples, it may reduce the estimation accuracy and external representativeness.

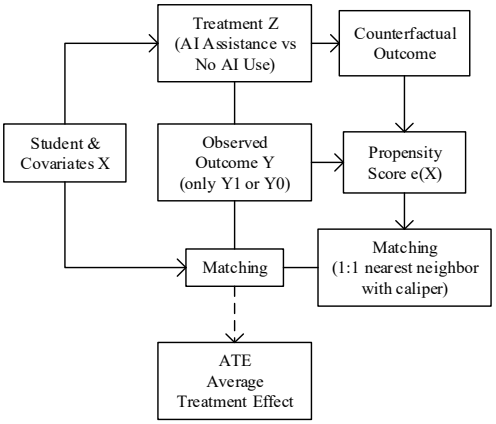


Fig. 1. Schematic diagram of the counterfactual causal inference framework.

## III. Method Design

### A. Data Source

The study selected 120 undergraduate graduation design students from a certain university as samples and adopted multi-source data. Specifically, on the one hand, the academic records of students are obtained through the school's teaching management system, including the semester GPA and the graduation project grades; On the other hand, a questionnaire survey was carried out to collect information such as students' use of AI-assisted tools, self-efficacy and learning autonomy. In addition, semi-structured interviews were conducted with 10 of the students, and the interview records were sorted out. The AI usage questionnaire includes items such as the frequency, duration and main functions of students' use of AI tools; The self-assessment scales adopted the Chinese version of the General Self-Efficacy Scale and the College Students' Autonomous Learning Scale. The former had 10 items (with a 4-point score), while the latter had good structural validity through exploratory factor analysis (KMO = 0.957, $p < 0.001$). All questionnaires and interviews were subject to ethical review, and the anonymity of students and the confidentiality of data were strictly protected during the research process.

### B. Tools and Measurements

The main variables and measurement methods used in the study are as follows: The quality of the graduation project is expressed by the final project grade; The intensity of AI usage is expressed by the score of the questionnaire scale, and a comprehensive index can be constructed based on the Likert score of items such as "weekly usage frequency", "usage duration", and "usage frequency". The topic type is a categorical variable (such as theoretical type = 0, practical

type = 1); Self-efficacy is measured by the total score of GSES. The higher the score, the stronger the student's confidence in completing the learning tasks by themselves. Learning autonomy is measured by the total score of the College Students' Autonomous Learning Scale (Likert score, the higher the total score, the stronger the autonomous learning ability), and this scale has confirmed good structural validity. In addition, the control variables include gender (male/female), major category (liberal arts/science /engineering, etc.) and foundation grades (previous GPA), etc. The specific definitions and quantitative representations of each variable are shown in Table 1 for example.

Table 1. Variable table

| Variable category | Variable name | Variable definition | Measurement method |
|---|---|---|---|
| Result variable | Graduation Project Grade | The final grade of a student's graduation project | Percentage-based grading |
| Handle variables | The usage of AI | Whether to use AI-assisted tools or not | Binary variable (yes = 1, no = 0) |
| Covariates | GPA | Average academic performance | GPA |
| | Project type | Nature of the graduation project topic | Categorical variable (Theoretical type = 0, Practical type = 1) |
| | Professional category | The major category to which the student belongs | Categorical variables (such as liberal arts = 0, science = 1, engineering = 2) |
| | Gender | The gender of the student | Binary variable (male = 0, female = 1) |
| Psychological variables | Self-efficacy | The level of confidence that students have in their ability to complete tasks | General Self-Efficacy Scale (GSES) total score |
| | Learning autonomy | The ability of students to regulate their own learning behaviors | Score of the College Students' Self-Study Scale |

### C. Quantitative Analysis Strategy

To eliminate the selectivity bias caused by students' self-selection of AI tools, the study mainly uses Propensity Score Matching (PSM) for causal inference. Taking the usage status of AI (for example, high-intensity usage = 1, low-intensity or non-usage = 0) as the processing variable, the individual propensity score was estimated through the Logit regression model. The model included covariates such as gender, major, basic academic performance, self-efficacy, and learning autonomy. The nearest neighbor 1:1 matching strategy was adopted and a caliper width of 0.05 was set to ensure a high similarity in propensity scores between the treatment group and the control group. After matching, the standardized mean differences of all covariates were controlled within 10%, indicating that a good balance was achieved. Based on this matching sample, the mean difference between the graduation project scores of the two groups is the estimated value of the Average Treatment Effect (ATE), which is used to quantify the net impact of AI assistance on the quality of graduation projects.

A multiple linear regression model was constructed, the specific form of the regression model is as follows:

$$Y_i = \beta_0 + \beta_{21} AI_i + \beta_2 GPA_i + \beta_3 ProjectType_{1i} + \beta_4 Major_i + \beta_5 Gender_i + \beta_6 SelfEfficacy_i + \beta_7 Autonomy_i + \varepsilon_i \quad (3)$$

$Y_i$: The graduation project grade of the i-th student;
AI usage situation: Indicates whether the i-th student uses the AI tool (using = 1, not using = 0);
GPA, project type, major category, gender, self-efficacy, and learning autonomy: These are the respective control variables;
$\varepsilon_i$: Random error term.

This regression result serves as a supplementary test for the PSM result. Overall, the PSM logic simulates a quasi-experimental environment by "controlling" confounding variables. The effect in the "counterfactual" scenario can be expressed as:

$$ATE = E(Y|AI = 1) - E(Y|AI = 0) \quad (4)$$

where the average result of the matched control group is used as the counterfactual estimate in the absence of AI.

This model can effectively control the potential confounding effect and estimate the net effect of the use of AI tools on the graduation project grade.

## IV. Empirical Research

### A. Quantitative Analysis Strategy

The sample of this study consists of 120 students (with an average age of approximately 21 years old and an equal ratio of male to female students), covering multiple professional directions. Descriptive statistics (see Table 2) show that students using AI-assisted tools account for approximately 50% of the sample. There are differences between the two groups of students in variables such as basic performance, self-efficacy, and autonomy (the standardized mean differences of unmatched samples are mostly above 0.2). The equilibrium of covariates before and after matching is shown in Table 2: Before matching, there were significant differences between the treatment group and the control group in multiple covariates. After matching, the standardized mean differences of all covariates decreased to less than 0.1, which was in line with the equilibrium criteria suggested in the literature, indicating that matching effectively eliminated observational confounding.

The propensity score matching estimation results are shown in Table 3: In the matched samples, the average score of the graduation project of the students using the AI tool was 85.3 points, and that of the control group was 80.1 points. The difference between the two, that is, the average treatment effect ATE, was approximately 5.2 points ($p < 0.05$), indicating that the use of AI-assisted tools has a significant positive impact on the quality of the design.

Table 2. The test results of covariate equilibrium before and after propensity score matching

| Covariate name | Mean of the treatment group: Before | Mean value of the control group: Before | Standardized mean difference: Before | Mean of the treatment group: later | Mean value of the control group: later | Standardized mean difference: later |
|---|---|---|---|---|---|---|
| GPA | 3.42 | 3.17 | 0.32 | 0.36 | 3.34 | 0.04 |
| Project Type (Proportion of Practical Type) | 0.65 | 0.43 | 0.44 | 0.58 | 0.55 | 0.05 |
| Professional category (proportion of science and engineering) | 0.70 | 0.52 | 0.38 | 0.65 | 0.64 | 0.03 |
| Gender (proportion of females) | 0.52 | 0.48 | | 0.08 | 0.50 | 0.00 |
| Self-efficacy score | 31.4 | 29.2 | 0.35 | 30.5 | 30.3 | 0.03 |
| Self-directed Learning Score | 72.8 | 68.3 | 0.40 | 71.4 | 71.0 | 0.05 |

Table 3. The propensity score matches the estimation result

| Analysis group | Average graduation design score | Sample size (n) | Standard Error (SE) | Average Treatment Effect (ATE) | Significance level (p). |
|---|---|---|---|---|---|
| The AI group (processing group) | 85.46 | 46 | 0.29 | — | — |
| The group that did not use AI (the control group) | 79.71 | 46 | 0.30 | 5.75 | < 0.001 |

The corresponding OLS regression results (unmatched samples) showed that after controlling for covariates, the coefficient of AI usage was approximately +4.5 ($p < 0.05$), which was consistent with the direction of the PSM results. For causal inference, according to the counterfactual framework, when the non-confounding assumption holds, ATE simplifies to the difference between two sets of means. Our results indicate that the use of AI tools can significantly improve the grades of graduation projects. To visually demonstrate the matching effect, a propensity score distribution graph (see Fig. 2), a comparison graph of the mean difference of covariates before and after matching (see Fig. 3), and a further ATE result table (Table 4) can be drawn.

All five estimation strategies showed that the ATE was between 5.5 and 5.9 points, with little difference. The directions were consistent and all significant ($p < 0.01$), indicating that the positive effect of AI-assisted tools on the graduation project grades was robust.

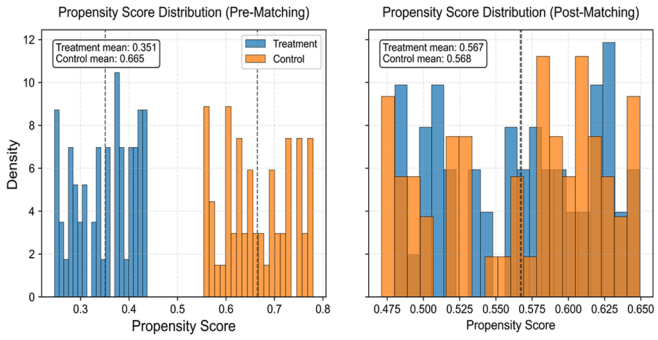


Fig. 2. Propensity score distribution before and after matching.

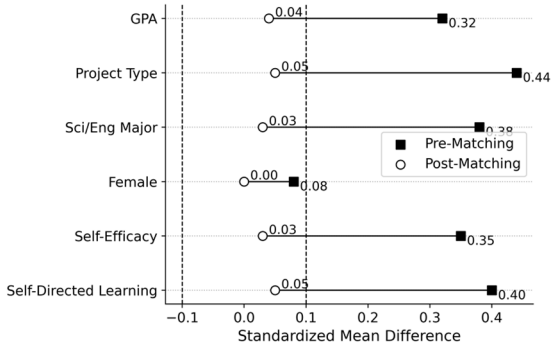


Fig. 3. Comparison of covariate mean differences before and after matching.

Table 4. ATE result table

| Method | Matching/Estimating Details | ATE | Standard Error (SE) | 95% confidence interval | p |
|---|---|---|---|---|---|
| Nearest Neighbor PSM | 1: 1 Nearest Neighbor, Caliper 0.05 | 5.75 | 0.42 | 4.9–6.6 | < 0.01 |
| Kernel PSM | Epanechnikov kernel, bandwidth 0.06 | 5.73 | 0.31 | 5.1–6.3 | < 0.01 |
| IPW | Average treatment effect weight | 5.37 | 0.43 | 4.5–6.2 | < 0.01 |
| OLS Regression | Multiple Linear Regression | 5.84 | 0.54 | 4.8–6.9 | < 0.01 |
| Replacement caliper (0.10) | 1: 1 Nearest Neighbor, Caliper 0.10 | 5.58 | 0.38 | 4.8–6.3 | < 0.01 |

### *B. Sensitivity Test*

To verify the robustness of causal estimation, we conducted multi-angle sensitivity tests, and the results are concentratingly presented in Figs. 4 and 5. For the caliper width setting with propensity score matching, we gradually increased the caliper from 0.01 to 0.20 (see Fig. 4). It can be seen that the Average Treatment Effect (ATE) has always remained between 5.5 and 5.9 points, with extremely small curve fluctuations and significant p values, indicating that the matching results are not sensitive to the selection of calipers. To test the reliability of different estimation strategies, we adopted methods such as many-to-one matching, Kernel matching, Inverse Probability Weighting (IPW), and multiple OLS regression (see Fig. 5 for details). The ATE and its 95% confidence intervals given by each method are highly consistent and none of them cross zero, further proving that the positive effect of AI-assisted tools is methodological robust.

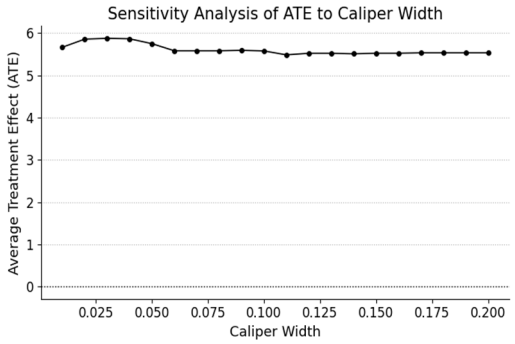


Fig. 4. Caliper width sensitivity curve.

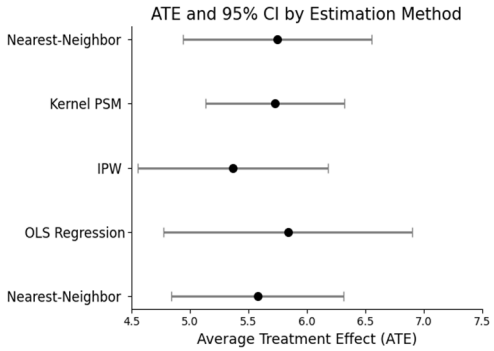


Fig. 5. Robust forest plot of different estimation methods.

The Rosenbaum boundary test shows that if the assumption of unobserved confusion doubles the probability of students using AI ($\Gamma = 2.0$), the ATE remains significant, indicating that the conclusion has strong resistance to potential hidden biases. Finally, after binarizing the dependent variable and adding the covariate interaction term to the regression model, the estimation results were still consistent with the direction of the main analysis, further consolidating the reliability of the study.

### C. Analysis of Semi-structured Interview Content

Semi-structured interviews were conducted with undergraduate students, and the interview contents were encoded using the topic analysis method [12, 13]. Multiple topics were summarized through sentence-by-sentence coding and multiple rounds of discussions. Double-person cross-checking improves the consistency of the coding results. The interview results are highly consistent with the data from the previous questionnaire survey, achieving quantitative and qualitative triangulation verification. The interview content mainly revolves around the following themes:

Efficiency of literature retrieval and data organization: Most respondents believe that AI tools can accelerate literature search and data organization.

Design idea expansion and inspiration acquisition: Many students have indicated that the suggestions provided by AI sometimes bring about new inspirations for ideas.

Technical implementation and programming assistance: Some engineering students mentioned that AI helps generate code snippets or technical solution templates, shortening the implementation time.

Reliance on risk and result reliability: Some students pointed out that AI sometimes makes mistakes and emphasized the need for manual screening of the results output by AI.

These themes were all reflected in the interview and questionnaire data, forming an effective corroborative relationship. For example, the result of "improving efficiency" was leading in the questionnaire, and similar feedback was also obtained in the interview. The interview analysis enriched the understanding of the AI-assisted impact, supplemented and verified the quantitative results, and enhanced the credibility of the research conclusions.

## V. Conclusion

This study introduced counterfactual causal inference into the undergraduate graduation project scenario. After controlling for covariates such as GPA and project type, the average graduation project score of students using generative AI was approximately 5.75 points higher. The permutation test $p < 0.001$, verifying the net effect of AI on learning outcomes and providing quasi-experimental evidence for the educational value of generative AI. Semi-structured interviews show that AI tools can enhance the efficiency of literature retrieval, data organization, design idea expansion and technical implementation, etc. AI reduces cognitive load by sharing low-level tasks, enabling students to focus more on higher-order thinking. However, the interview also pointed out that there is a risk of over-reliance on AI. It is suggested that universities strengthen the training of AI literacy and academic integrity to give full play to the value of AI and avoid potential risks.

## Conflict of Interest

The authors declare no conflict of interest.

## Author Contributions

Honghe Shi collected the data and wrote the paper; Bin Duan provided guidance; both authors had approved the final version.

## References

[1] C. Wang, Z. Li, and C. Bonk, "Understanding self-directed learning in AI-assisted writing: A mixed-methods study of post-secondary learners," *Comput. & Educ.: AI*, vol. 5, art. 100152, 2024.

[2] M. Liu, L.-J. Zhang, and C. Biebricher, "Investigating students' cognitive processes in generative-AI-assisted digital multimodal composing and traditional writing," *Comput. & Educ.*, vol. 201, art. 104738, 2024.

[3] J. Kim, S. Lee, R. Detrick, J. Wang, and N. Li, "Students–generative-AI interaction patterns and their impact on academic writing," *J. Comput. Higher Educ.*, vol. 36, pp. 505–526, 2025.

[4] E. Nurchurifiani, A. Pranowo, and Y. Sasmita, "Leveraging AI-powered tools in academic writing: Insights from English faculty members in Indonesia," in *Proc. 11th Int. Conf. Educ. Psychol. Sci. (ICEPS)*, Tokyo, Japan, 2024, pp. 85–90.

[5] A. M. B. Ugli, N. Khalimova, and F. Turg'unov, "Uzbekistan university EFL teachers' perceptions of ChatGPT: Benefits and ethical challenges," in *Proc. ICEPS 2024*, Tokyo, Japan, 2024, pp. 126–131.

[6] O. Tafhi, H. Boudlal, and A. Qobadi, "Mapping Moroccan physics students' perceptions of ChatGPT," in *Proc. ICEPS 2024*, Tokyo, Japan, 2024, pp. 140–145.

[7] J. Roe and M. Perkins, "Generative AI in self-directed learning: A scoping review," arXiv preprint arXiv:2402.12345, 2024.

[8] D. B. Rubin, "Estimating causal effects of treatments in randomized and non-randomized studies," *J. Educ. Psychol.*, vol. 66, no. 5, pp. 688–701, 1974.

[9] P. R. Rosenbaum and D. B. Rubin, "The central role of the propensity score in observational studies for causal effects," *Biometrika*, vol. 70, no. 1, pp. 41–55, 1983.

[10] S. Becker and A. Ichino, "Estimation of average treatment effects based on propensity scores," *Stata J.*, vol. 2, no. 4, pp. 358–377, 2002.
[11] S. L. Morgan, K. M. Todd, and D. J. Thomson, "Propensity score estimation and causal inference in educational research," *Rev. Res. Educ.*, vol. 34, pp. 1–38, 2010.
[12] C. B. Fontao, M. L. Santos, and A. Lozano, "ChatGPT's role in the education system: Insights from future secondary teachers," in *Proc. ICEPS 2023*, Paris, France, 2023, pp. 210–215.
[13] M. Moundridou and N. Matzakos, "Generative AI in an educational technology course for pre-service mechanical-engineering educators: A case study," in *Proc. ICEPS 2023*, Paris, France, 2023, pp. 165–170.