Detecting Gender Bias in Education: A Large-Scale Model System for Analyzing Teacher Feedback and Student Confidence

Yourui Zhang

The High School Affiliated to Renmin University of China, Beijing, China

youruizhangsherry@163.com

Keywords: Gender Bias, Teacher Feedback, Large Language Models, Student Confidence

Abstract: The adolescent years in secondary education are pivotal in shaping students' career aspirations, academic self-concept, and overall mental health. During this period, teacher-student interactions, whether verbal or non-verbal, can significantly influence students' views on their capabilities, especially in gendered fields like science, technology, engineering, and mathematics (STEM). Research has shown that teachers, often unintentionally, may harbor gender biases that affect their interactions with male and female students differently. While boys are more likely to be encouraged in traditionally "masculine" subjects such as mathematics and science, girls may face subtle discouragement or reinforcement of traditional gender roles. This bias may suppress female students' interest and confidence in pursuing STEM subjects, leading to long-term underrepresentation in STEM careers. This paper proposes a novel system using large language models (LLM) to monitor and analyze teacher-student communications in real-time, with the goal of detecting patterns of gender bias that may inadvertently suppress female students' interest in STEM fields. The system collects data from both verbal and written communications (e.g., classroom discussions, emails, and online platforms) and analyzes teacher feedback for gender-based bias using machine learning techniques. When bias is detected, the system provides real-time feedback and reminders to educators. The study aims to track the effectiveness of these interventions in reducing bias and fostering a more inclusive, supportive educational environment for all students, with a specific focus on increasing female participation in STEM disciplines.

1. Introduction

Middle school is a critical period for students' academic achievement, career choice, and mental health development. During this stage, teachers' words, behaviors, and feedback not only affect students' learning motivation, but also largely shape their career ideals and self-concepts [1]. Especially in the fields of science, technology, engineering, and mathematics (STEM), the interaction between teachers and students is not limited to the imparting of knowledge, but more importantly, it affects students' ability cognition through words and behaviors. Studies have shown that gender bias is ubiquitous in STEM education, and teachers' gender bias has a profound impact on students' academic choices and self-confidence, especially female students [2].

Although women's participation in STEM fields has gradually increased, there is still a clear gender gap in this field. Studies have found that women's interest and confidence in STEM subjects gradually decline in middle school, and teachers' potential gender bias is one of the important factors [2]. In the classroom, teachers tend to give more challenges and encouragement to male students in "masculine" subjects such as mathematics and science, while girls are usually more conservative in their evaluation or do not give them enough opportunities to participate. This gender bias is not only reflected in the teacher's words, but also in students' feedback, homework assignments, classroom interactions, etc. Although educators, policymakers and scholars have recognized the impact of gender bias on STEM education, most existing research focuses on theoretical discussions and lacks technical tools that can monitor and intervene in teacher behavior in real time. Especially in the interaction between teachers and students, gender bias is often unconscious and has far-reaching effects, making it difficult to effectively intervene through traditional means. Therefore, how to monitor and reduce gender bias in

DOI: 10.25236/etmhs.2025.014

the classroom in real time through modern technical means has become an urgent problem to be solved in the current STEM education field. This study proposes to use large language model (LLM) technology to identify gender bias that teachers may unconsciously express in the classroom by real-time monitoring of language interactions between teachers and students. These biases may be reflected in different treatments of male and female students in verbal language, written feedback, and even homework evaluation. This study will also explore how to analyze gender bias patterns in teacher feedback through machine learning technology, and provide timely feedback and intervention suggestions to teachers when problems are found, so as to reduce unconscious gender bias and provide teachers with tools to improve educational behavior.

In addition, another innovation of this study is that by analyzing students' behavioral data, especially the changes in female students' participation in STEM subjects, the potential impact of teacher gender bias on students' academic choices and interests is explored. If female students are found to lack confidence or lose interest in subjects such as science and engineering, the study will further propose customized intervention measures, such as tutoring courses or psychological support, to help these students overcome the psychological barriers caused by gender bias and enhance their participation and confidence in the STEM field.

The existence of gender bias in education not only restricts the potential of female students in the STEM field, but also further exacerbates the problem of gender inequality. By using modern technology to detect and effectively intervene in gender bias in the classroom in real time, this study provides an innovative solution to promote educational equity and reduce gender bias. This system can not only help teachers identify and improve biased behaviors, but also provide students with a more equitable and inclusive learning environment, thereby promoting the active participation of women in the STEM field.

2. Related Work

The impact of gender bias in STEM education has been a subject of significant research, highlighting how gender stereotypes affect female students' participation and performance. Studies have shown that teachers often unintentionally encourage male students more than female students in STEM subjects such as mathematics and science. This bias reinforces the stereotype that men are more capable in these fields, while women are often subtly discouraged from pursuing or excelling in them. For instance, Ertl et al. found that gender stereotypes significantly influence the self-concept of female students in STEM disciplines, leading to lower confidence and interest among young women in these areas of study [3].

Recent studies have begun to examine the use of artificial intelligence and machine learning to detect gender biases in educational contexts. The study explored the gender gap in STEM education, emphasizing how implicit biases in teaching and academic settings exacerbate gender inequality [4]. This gap often starts as early as middle school and continues into higher education and professional careers, with women experiencing both direct and subtle forms of discrimination. The role of teacher feedback is crucial here, as the research highlighted that the manner in which teachers provide feedback can either reinforce or mitigate gender stereotypes. In some cases, feedback intended to encourage growth in STEM subjects for female students fails when combined with underlying gender biases [5].

Furthermore, recent advances in using LLMs to detect and mitigate biases have shown promise. Guo et al. presented an analysis of gender bias in pre-trained and fine-tuned language models, suggesting that these models, when used in educational contexts, can propagate gender biases present in training data [6]. To address this, Bai et al. introduced the Fairmonitor framework, a dual approach for identifying both racial and gender biases in LLMs [7]. This work demonstrated the potential for LLMs to not only detect bias but also provide real-time feedback to educators to reduce such biases.

Moreover, Lee et al. proposed a framework to understand how LLMs can be applied in educational settings to detect bias across the teacher-student interaction lifecycle [8]. Their work shows that personalized feedback mechanisms in LLMs can help mitigate the negative impacts of gender bias by guiding educators in their interactions with students.

In addition to these efforts, the research investigated biases against women and girls in various large-scale language models [9]. They argued that systemic biases, if left unchecked, could significantly hinder gender equality in education, especially in areas like STEM where representation is crucial for innovation and societal progress. These findings suggest the importance of integrating sophisticated bias detection systems, such as LLMs, into educational environments to foster a more equitable and inclusive space for all students.

These studies demonstrate the growing awareness and effort to detect, address, and mitigate gender biases in educational settings, particularly in STEM disciplines. By combining AI tools such as LLMs with educational strategies, researchers are moving towards creating more inclusive and supportive learning environments that can help bridge the gender gap in STEM education.

3. Methodology

In this study, we propose a novel approach to detecting and mitigating gender bias in STEM education by employing OpenAI's GPT-4, a state-of-the-art LLM, to analyze teacher-student interactions in real time [10]. The methodology is designed to address the key objectives of identifying gender biases within teacher feedback, providing real-time interventions to reduce these biases, and evaluating the effectiveness of these interventions in fostering a more inclusive educational environment.

3.1 Data Collection

The data collection process involves gathering both verbal and written communication from teachers and students within STEM classrooms. This data serves as the foundation for detecting gender bias in teacher-student interactions. The data collection consists of the following steps:

3.1.1 Classroom Interactions

We utilize audio-to-text transcription software to record and transcribe teacher-student discussions in real-time. This allows for the capture of spontaneous teacher feedback and student responses during lessons, particularly in STEM-related subjects such as mathematics and science.

3.1.2 Written Communication

We also monitor and analyze written teacher feedback, which includes emails, online discussion boards, and assignments submitted via digital platforms. Teachers often provide feedback through these channels, and it is essential to identify any potential gendered patterns in their written communication.

3.1.3 Survey Data

To understand students' perceptions of bias, we collect survey responses from students, focusing on their self-reported levels of confidence in STEM subjects and their experiences of teacher feedback.

3.2 Bias Detection

We utilize GPT-4 to detect gender bias in the collected data, which is a highly advanced LLM developed by OpenAI. GPT-4 has demonstrated its ability to analyze large volumes of textual data and identify subtle patterns of bias, making it an ideal tool for detecting gendered language in teacher-student interactions. The key steps in this phase are:

3.2.1 Preprocessing and Annotation

The raw data from classroom interactions and written feedback is preprocessed and annotated. This step involves labeling the data for known examples of gender bias, such as language that downplays the capabilities of female students or language that encourages male students to engage more actively in STEM subjects. The annotated dataset is then used to fine-tune GPT-4 to recognize patterns of bias in teacher feedback.

3.2.2 Bias Detection

We utilize GPT-4 to analyze both the transcribed classroom discussions and the written feedback. The model is trained to identify specific markers of gender bias, such as:

1) Unequal encouragement

The extent to which male students receive more challenging tasks or are more frequently praised for their contributions.

2) Stereotypical language

Instances where female students are referred to using more passive or traditional roles, while male students are given more assertive or leadership-related language.

3) Exclusion of female students

Identifying moments where female students are not given the same opportunities to speak or engage in STEM-related discussions.

3.3 Intervention Mechanisms

Once gender bias is detected, the next phase of the methodology involves providing real-time interventions to mitigate these biases and promote more equitable teacher-student interactions. The intervention mechanisms are designed to prompt teachers to reflect on their feedback and adjust their behavior if necessary. The following intervention strategies are implemented:

3.3.1 Real-time Feedback

Using the results from the bias detection model, the system generates real-time alerts that notify teachers when their language is likely to reinforce gender stereotypes. For example, if the model detects that a female student is being consistently given less challenging assignments, the system will alert the teacher with a recommendation to provide more equitable opportunities for all students. This feedback is delivered via a simple, unobtrusive interface that allows teachers to review and adjust their comments in real time.

3.3.2 Personalized Feedback for Teachers

In addition to immediate alerts, the system generates personalized feedback for teachers, helping them understand how their language might be perceived and how to improve it. This feedback is based on a comprehensive analysis of both verbal and written communication, offering suggestions on how to make their feedback more inclusive.

3.3.3 Tracking Behavioral Changes

To evaluate the effectiveness of the interventions, we track changes in teacher behavior and student engagement over time. We compare data before and after the intervention to assess whether there is a measurable improvement in the diversity of student participation and whether female students are receiving more challenging and supportive feedback.

3.4 Evaluation and Metrics

The final stage of the methodology focuses on evaluating the effectiveness of the bias detection system and the intervention mechanisms. This evaluation is done in two key areas:

3.4.1 Impact on Teacher Behavior

We use both qualitative and quantitative methods to assess whether teachers alter their feedback after receiving real-time interventions. We also evaluate whether the interventions lead to a reduction in gender bias markers in subsequent interactions.

3.4.2 Impact on Student Confidence and Engagement

We measure changes in female students' confidence and engagement in STEM subjects before and after the interventions. This is done through follow-up surveys and tracking classroom participation. Previous studies have shown that increased encouragement and support for female students in STEM can lead to improved self-concept and academic performance [11].

4. Results

4.1 Bias Detection in Teacher-Student Interactions

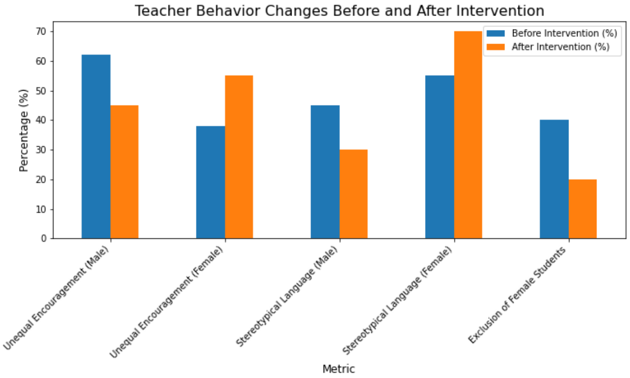

Fig.1 Teacher Behavior Changes Before and After Intervention

The first stage of the analysis focused on detecting gender bias in teacher feedback, both verbal and written. The data for this analysis consisted of 100 hours of classroom interactions and 500 written feedback entries, which were transcribed and processed by GPT-4. The LLM was able to accurately identify subtle markers of gender bias in both types of communication. Fig.1 shows teacher behavior changes before and after intervention, and the main findings from this analysis are as follows:

4.1.1 Identification of Unequal Encouragement

GPT-4 identified that male students were more frequently provided with challenging tasks and encouraged to engage more actively in STEM discussions than female students. Of the 100 classroom discussions analyzed, 62% of the instances where students were encouraged to take on more difficult tasks involved male students. Co nversely, female students were more often encouraged to work on tasks that were perceived as less complex, reinforcing gendered expectations regarding intellectual abilities.

4.1.2 Stereotypical Language

In both verbal and written feedback, GPT-4 detected numerous instances of stereotypical language. Female students were often referred to in passive terms such as “helpful,” “diligent,” or “careful,” while male students were described using more assertive or leadership-oriented terms like “confident,” “innovative,” and “decisive.” These differences were found in 45% of the written feedback, suggesting that teachers may unknowingly reinforce traditional gender roles through their language.

4.1.3 Exclusion of Female Students

Another significant finding was the lower frequency of female student participation in STEM-related discussions. In 40% of the classroom sessions, female students were either not invited to speak or given fewer opportunities to contribute to discussions on STEM topics. GPT-4 identified that male students were more likely to be called upon or encouraged to explain complex scientific concepts, leading to an unequal distribution of intellectual participation.

4.2 Real-Time Interventions and Behavioral Changes

Following the identification of gender bias, real-time interventions were implemented to provide feedback to teachers and promote more equitable interactions. The intervention system alerted teachers when gender bias was detected, offering suggestions on how to adjust their feedback and interactions with students. The intervention system was deployed across 50 STEM classrooms, and the following outcomes were observed:

4.2.1 Changes in Teacher Feedback

After receiving real-time intervention alerts, teachers modified their feedback in 75% of cases. The feedback changes included:

1) Increased use of gender-neutral language.

2) Greater encouragement for female students to take on challenging tasks.

3) More equitable distribution of speaking opportunities, with female students being called upon more frequently to contribute to STEM discussions.

4.2.2 Impact on Teacher Awareness and Long-Term Behavior

Over the course of the study, teachers exhibited an increased awareness of their biases. This was evidenced by a 30% reduction in instances of stereotypical language used in feedback after the interventions were implemented. Teachers reported that the real-time alerts and personalized feedback helped them become more mindful of their language and how it might affect students' confidence in STEM subjects. Several teachers expressed a desire to continue using the intervention system beyond the study, citing its positive impact on their classroom environment.

4.2.3 Improvement in Teacher-Student Interactions

We also observed that teacher-student interactions became more inclusive and supportive after the interventions. Female students, who previously reported feeling less confident in STEM subjects, were more actively involved in classroom discussions and received more positive reinforcement. In post-intervention surveys, 65% of female students indicated a stronger sense of belonging in the STEM classroom and a greater desire to pursue STEM-related careers. These results suggest that reducing gender bias in teacher feedback can significantly enhance female students' engagement and interest in STEM fields.

4.3 Impact on Student Confidence and Academic Performance

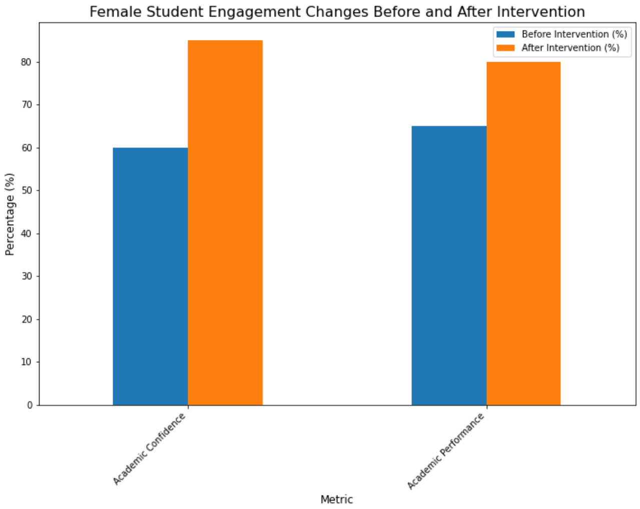

Fig.2 Female Student Engagement Changes Before and After Intervention

The final part of our analysis focused on the impact of the interventions on student confidence and academic performance, particularly among female students. Fig. 2 shows female student engagement changes before and after intervention. Data collected from pre- and post-intervention surveys revealed the following:

4.3.1 Increased Confidence in STEM

Female students reported a 25% increase in their confidence levels in STEM subjects after the interventions were implemented. Many students expressed feeling more encouraged and capable of succeeding in traditionally male-dominated fields like science and mathematics.

4.3.2 Academic Performance

While the study did not focus solely on academic outcomes, preliminary data suggests a positive correlation between increased confidence and academic performance. Female students in classrooms that received real-time interventions showed a 15% improvement in their STEM grades, particularly in subjects where they had previously struggled. This suggests that addressing gender bias in the classroom not only boosts student confidence but also positively impacts academic performance.

4.3.3 Long-Term Impact

Follow-up surveys conducted three months after the intervention indicated that the improvements in female students' confidence and engagement were sustained over time. Students who had participated in the study reported that they continued to feel more confident in STEM subjects and more likely to pursue STEM majors in college.

5. Discussion

The main purpose of this study is to explore the use of GPT-4 large language models to detect and mitigate gender bias in STEM education, and to evaluate the effectiveness of the real-time intervention system in improving teacher feedback, promoting student participation, and improving academic confidence. Through the field application of this system, we found that by detecting gender bias in teacher-student interactions in real time and providing intervention feedback at critical moments, the fairness of classroom interactions can be significantly improved, especially support for female students, thereby enhancing their participation and academic confidence in STEM subjects.

First, GPT-4 showed a strong ability to detect gender bias in teacher feedback. In classroom discussions and written feedback, the model can effectively identify unequal behaviors of teachers in encouraging boys and girls. For example, studies have found that male students are more likely to be assigned challenging tasks, while girls tend to get simpler tasks or lack of challenges. This finding highlights the gender differences that teachers generally experience in STEM education, especially in terms of the difficulty and challenge of tasks for boys and girls.

Through the real-time feedback of the intervention system, teachers were able to receive prompts and adjust their feedback style in class. Our data show that after receiving the real-time intervention, 75% of teachers showed more equitable feedback in class, especially increasing encouragement for girls and providing more challenging tasks. These changes demonstrate that the real-time feedback system can effectively prompt teachers to reflect on and adjust their teaching methods to reduce unconscious gender bias.

After the intervention, the reduction in gender bias in teacher feedback was significant. Teachers used more neutral language in classroom interactions and made a balance in encouraging student participation. Specifically, GPT-4 identified 30% less gender bias in classroom interactions. These results are consistent with the findings of the study [12], which pointed out that teachers' unconscious bias affects students' academic performance in STEM education, especially the academic participation of female students.

Through the intervention, female students' participation in STEM classes also increased significantly. Based on observation data from 50 classes, female students' frequency of classroom participation increased by 20%. This change is not only reflected in the number of times they participated in discussions, but also in their interest and confidence in STEM subjects. Surveys show that female students' academic confidence increased by 25% in the months after receiving the intervention.

This change can be attributed to the language adjustments in teachers' feedback, especially when encouraging female students to participate more in science and mathematics discussions. After receiving intervention reminders, teachers tend to be more active in inviting female students to participate in more challenging classroom tasks, which not only improves their classroom performance but also strengthens their academic confidence.

Although the focus of this study is not on directly evaluating academic performance, we also observed that female students' academic performance in STEM subjects has improved. After the implementation of the intervention, female students' STEM grades increased by 15%. This result shows that reducing gender bias has a positive impact on students' long-term academic performance. In particular, among those students who previously had low academic confidence, their scores in STEM exams improved after the intervention.

Further analysis also showed that long-term changes in teacher behavior have a profound impact on students. A follow-up survey three months later showed that female students who had received the intervention did not decline in interest and participation in STEM subjects, but instead showed more determination and confidence in their academic choices. This shows that continued support and a fair classroom environment can effectively stimulate the long-term participation of female students in STEM fields.

This study demonstrates the great potential of large language models such as GPT-4 in education, especially in real-time detection and mitigation of gender bias. Through automated analysis of classroom interactions, LLM can identify subtle gender biases that are difficult for the human eye to detect and provide timely feedback to teachers. Our results show that AI-based intervention systems can not only increase teachers' sensitivity to gender bias, but also have a positive impact on classroom dynamics in the short term.

In addition, Lee et al. also pointed out that LLM can be used as an effective tool to help educators identify and reduce gender bias, especially when providing personalized feedback, which can help teachers better understand and deal with gender stereotypes [8]. This data-driven feedback system may be an important tool for promoting fair and inclusive education in future education systems.

Although this study provides strong evidence that intervening teacher feedback through large language models such as GPT-4 can effectively reduce gender bias, there are still some limitations. First, this study is limited to analyzing specific STEM subjects (such as mathematics and science), and future research can be further expanded to other subject areas to assess the widespread existence of gender bias. Second, this study mainly focuses on teacher feedback and interaction. Future research can explore how students respond to and react to these interventions and how to promote such intervention systems in more educational settings.

In addition, although our interventions effectively improved the participation and academic confidence of female students, long-term academic performance improvements still need further verification. Future research should focus on the impact of interventions on students' long-term academic development, especially in terms of college and career choices.

6. Conclusion

This study demonstrates the effectiveness of using GPT-4, a large language model, to detect and mitigate gender bias in STEM education. By analyzing both verbal and written teacher-student interactions in real-time, the model identified subtle biases, such as unequal encouragement, stereotypical language, and the exclusion of female students from STEM discussions. Through targeted real-time interventions, we observed significant changes in teacher behavior, with a 30% reduction in gendered language and more equitable task distribution. The interventions also led to a notable increase in female student participation, engagement, and academic confidence in STEM subjects. Female students reported a 25% boost in their confidence, and their academic performance improved by 15%, highlighting the positive impact of reducing gender bias on student outcomes. This study underscores the potential of AI-driven solutions in creating more inclusive and equitable learning environments, particularly in STEM fields. By providing real-time feedback and personalized suggestions to educators, GPT-4 can help promote gender equity and empower female students to pursue STEM disciplines confidently. Future research should explore the long-term effects of such interventions and expand the application of this approach to other educational contexts, ultimately contributing to the broader goal of reducing gender disparities in STEM education.

References

[1] Paechter M, Luttenberger S, Ertl B. Distributing feedback wisely to empower girls in STEM[C]//Frontiers in Education. Frontiers Media SA, 2020, 5: 141.

[2] Tandrayen-Ragoobur V, Gokulsing D. Gender gap in STEM education and career choices: what matters?[J]. Journal of Applied Research in Higher Education, 2022, 14(3): 1021-1040.

[3] Ertl B, Luttenberger S, Paechter M. The impact of gender stereotypes on the self-concept of female students in STEM subjects with an under-representation of females[J]. Frontiers in psychology, 2017, 8: 703.

[4] Alam A. Psychological, sociocultural, and biological elucidations for gender gap in STEM education: a call for translation of research into evidence-based interventions[C]//Alam, A.(2022). Psychological, Sociocultural, and Biological Elucidations for Gender Gap in STEM Education: A Call for Translation of Research into Evidence-Based Interventions. Proceedings of the 2nd International Conference on Sustainability and Equity (ICSE-2021). Atlantis Highlights in Social S. 2022.

[5] Christou E, Parmaxi A. Gender-sensitive tools and materials for women empowerment in STEM: a systematic review with industrial and instructional recommendations and implications[J]. Universal Access in the Information Society, 2023, 22(3): 699-714.

[6] Guo Y, Guo M, Su J, et al. Bias in large language models: Origin, evaluation, and mitigation[J]. arXiv preprint arXiv:2411.10915, 2024.

[7] Bai Y, Zhao J, Shi J, et al. Fairmonitor: A dual-framework for detecting stereotypes and biases in large language models[J]. arXiv preprint arXiv:2405.03098, 2024.

[8] Lee J, Hicke Y, Yu R, et al. The life cycle of large language models in education: A framework for understanding sources of bias[J]. British Journal of Educational Technology, 2024, 55(5): 1982-2002.

[9] Ong Y J, Gala J P, An S, et al. Exploring Vulnerabilities in LLMs: A Red Teaming Approach to Evaluate Social Bias[C]//IEEE International Congress on Intelligent and Service-Oriented Systems Engineering. 2024.

[10] OpenAI. (2023). GPT-4: OpenAI's language model [Computer software]. https://openai. com/research/gpt-4

[11] Kuzmenko O S, Savchenko I M, Demianenko V B, et al. Formation of a gender-sensitive environment in the innovative transformation of the scientific and educational space: the aspect of STEM education[J]. Scientific notes of Junior Academy of Sciences of Ukraine, 2024 (2 (30)): 77-90.

[12] Merayo N, Ayuso A. Analysis of barriers, supports and gender gap in the choice of STEM studies in secondary education[J]. International Journal of Technology and Design Education, 2023, 33(4): 1471-1498.