Experimental evaluation of Large Language Models for in-class learning experience customization

Daniel Moreno[a], Victor Guerra[b], and Antonio G. Ravelo-García[a]

[a]Instituto para el Desarrollo Tecnológico y la Innovación en Comunicaciones, Universidad de Las Palmas de Gran Canaria, Las Palmas, Spain
[b]Pi Lighting Sarl, Sion, Switzerland

ABSTRACT

This paper explores the utilization of Large Language Models (LLMs) for personalized education in Secondary Education, focusing on motivation and personalization. It uses the GPT 3.5 model by OpenAI to generate tailored exercises and examines their impact on student motivation and academic performance. The study highlights the positive correlation between motivation and performance, emphasizing the need to consider classroom dynamics and teacher-student relationships. While LLMs enhance educational content, they should complement, not replace, the teacher's role. The research calls for further investigation into personalization's impact on education, considering study duration and student samples for a more robust understanding.

Keywords: Large Language Model, Motivation, Personalized Learning, Learning Experience

1. INTRODUCTION

Artificial Intelligence (AI) and Large Language Models (LLMs) have advanced significantly in recent years. LLMs are natural language processing models that utilize machine learning to generate coherent and meaningful text in various languages. They are trained on extensive textual data to understand linguistic patterns and predict subsequent words or phrases within text.[1]

Furthermore, LLMs have gained popularity due to their proficiency in tasks such as translation, text generation, question answering, and text classification. They find applications in chatbots, voice assistants, and personalized recommendation systems,[2,3] although they have also raised ethical and privacy concerns related to bias and the potential for producing misleading text.[4]

In education, LLMs offer opportunities to enhance learning processes and resources. Some researchers propose using these models to generate content and tailor learning experiences for students.[5]

Moreover, motivation in teaching significantly affects student engagement and performance. The Self-Determination Theory emphasizes that meeting psychological needs for competence, autonomy, and social relatedness enhances motivation and engagement.[6] Conversely, unmet needs can lead to disengagement and potential dropouts.

In this context, LLMs can improve instruction quality and personalization, since they enable the creation of personalized exercises, potentially boosting student motivation and engagement.

This work explores the benefits of LLMs in education, focusing on personalization and motivation. The main objectives include using an LLM, specifically the GPT 3.5 model by OpenAI,[7] to generate personalized exercises for Secondary Education students. The expectation is that exercises generated by the LLM will capture student interest, fostering heightened motivation and engagement in the learning process. This innovative approach aims to analyze the motivational impact of personalized learning, examining its correlation with academic performance.

The remainder of this paper is structured as follows: Section 2 establishes a theoretical framework, providing the foundation for the research's concepts and theories. Section 3 outlines the methodology, including research design, data collection, and analysis techniques. Section 4 presents a comprehensive analysis of the study's findings. Finally, in Section 5 the paper concludes by summarizing key insights, discussing their implications, and suggesting potential directions for future research.

Further author information: (Send correspondence to D.M.)
D.M.: E-mail: daniel.moreno@ulpgc.es, Telephone: +34928459966

ISBN 978-84-09-51520-2

2. THEORETICAL FRAMEWORK

This section delves into three primary topics: the application of LLMs in education, student motivation, and the personalized learning experience. Here, key aspects of each topic are presented.

2.1 Educational Application of LLMs

LLMs, exemplified by the ChatGPT model in this study, have gained prominence in recent years for their natural language processing and text generation capabilities.[8] These models excel in understanding intricate linguistic patterns and generating coherent, contextually relevant content, sparking significant research interest in education.[9]

Students can harness these tools to augment their learning experiences, as they can clarify complex concepts, receive examples, and practice content through exercises.[10] Moreover, some scholars advocate using LLMs as virtual personal tutors to address student queries.[11]

For educators, LLMs offer a versatile tool to create supplementary content, tailor activities to individual student needs, and design evaluation materials.[10] Additionally, some studies suggest that AI has the potential to reduce educators' workloads by automating tasks.[12]

However, LLMs also present challenges in academic settings, particularly regarding potential misinformation or inaccuracies. In this regard, three scenarios emerge: banning LLMs, returning to written assessments, or integrating LLMs into teaching-learning processes to address these concerns.[13]

2.2 Student Motivation

Student motivation is a crucial factor in education, influencing engagement, commitment, and academic success. Intrinsic motivation, driven by personal interest and curiosity, is especially important for effective learning according to the self-determination theory.[14] Hence, fostering intrinsic motivation is a key goal for educators.

Motivation's importance spans all educational levels, with its determinants varying based on educational stages and individual circumstances.[15] For secondary education, various studies have explored motivation's impact across different domains and contexts.

For instance, a study found that grouping students by proficiency levels in English classes can enhance learning strategies and motivation compared to mixed-level classes.[16] In Peru, research on secondary students' social skills revealed that a significant portion lacks adequate social skills, potentially affecting their motivation and academic performance.[17]

2.3 Customized Learning Experience

Personalized learning customizes the teaching process to individual student attributes like skills, prior knowledge, and interests, which has been shown to enhance academic performance and student satisfaction.[18,19]

In secondary education, Information and Communication Technologies (ICT) can facilitate personalized learning by tailoring content and activities to students' learning style preferences, using various online tools. This approach can significantly improve student engagement and motivation by aligning tasks with their interests and abilities, ultimately increasing attendance and reducing truancy.[20]

Virtual Learning Environments (VLE) can also strengthen reading skills in basic and secondary education, involving educators from diverse fields.[21] LLMs play a key role in implementing personalized learning by generating relevant content for exercises and instructional materials, meeting individual student needs and interests, due to their ability to create coherent and contextually relevant content.

3. METHODOLOGY

The primary goal of this study, as outlined in Section 1, is to utilize the LLM ChatGPT to create customized tasks for enhancing the motivation of Secondary Education students. The experiments involved five groups, all in the 2nd year of Compulsory Secondary Education (ESO). Among them, three groups (T1, T2, and T3) served as test groups, completing customized exercises, while two groups (C1 and C2) acted as control groups, working on the same exercises but without customization.

ISBN 978-84-09-51520-2

3.1 Experimental setup

The study was conducted at the IES Siete Palmas educational center in Las Palmas de Gran Canaria, Spain, from mid-March to mid-May 2023. This timeline allowed for obtaining real results from 2nd-year ESO students. To ensure the consistency of the study, students with specific educational support needs that necessitated curriculum adaptation were excluded. As a result, there were a total of 62 students in the test groups and 37 students in the control groups, as indicated in Table 1.

Table 1: Number of students per class

Group	T1	T2	T3	C1	C2
Cardinality	20	20	22	19	18

The study focused on a part of the curriculum aligned with evaluation criterion 5 (STEE02C05) in Learning Block IV, Structures and Mechanisms: Machines and Systems, according to the Organic Law for the Improvement of Educational Quality (LOMCE). This criterion involved understanding the mechanical components responsible for transforming and transmitting movements in machines and systems within a structure, including their functionality, movement transformation or transmission, and the relationships between machine elements. In the educational institution where the research took place, this criterion encompassed three learning situations related to levers, pulleys, and gears. However, during the experimental period, only the learning situation concerning gears was covered, as the others had already been addressed.

3.2 Procedures

In Figure 1, the flowchart for the test and control groups is depicted. It consists of four phases: Phase I deals with the motivational assessment of the students, Phase II involves the generation of problem groups, Phase III encompasses classroom intervention, and Phase IV focuses on the assessment of the experience and the examination to measure the outcomes.

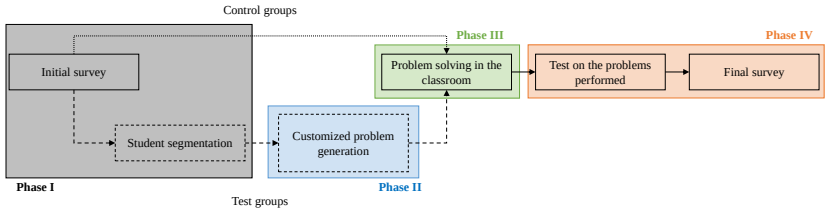


Figure 1: Flow diagram depicting the followed processes during the experimentation. You can grasp the caption to make some short clarifications on the figure.

3.2.1 Phase I: Motivational driver assessment

Both groups initially completed a survey that assessed their motivation level toward the subject, hobbies, interests, preferences for working in groups, willingness to connect hobbies with subject tasks, and preferred learning techniques (theoretical explanation, problem-solving, or construction projects).

After analyzing the survey results and considering students' indicated hobbies, the test group students were grouped in pairs or groups of three. Grouping was based on identifying common or similar interests to select thematic problems, and students with closely aligned preferences were placed together. In cases where students had more restrictive preferences in terms of cardinality, they were initially paired to address the challenge of finding matching interests.

3.2.2 Phase II: Problem batch generation

After segmenting the students, the process of creating custom problems began using ChatGPT. This involved analyzing the curriculum content, specifically related to gear systems, and using sample exercises to define a prompt. In the context of LLMs, a prompt is a text input or instruction given to the model to generate a

ISBN 978-84-09-51520-2

coherent response. Prompts are vital for interacting with LLMs as they provide initial guidance and context for generating relevant responses. Configuring and fine-tuning prompts is essential to influence the model's output, including its style, tone, level of detail, and thematic coherence in the responses.

In this study, after several iterations to refine the quality of the generated responses, the prompts used followed the following steps: definition of the chatbot's role as a gear problem generator for 2nd-year ESO students, description of the information to be provided and the expected outcome, example of a problem, data for each of the problems to be obtained, final instructions and the theme of the problems.

The instructions were written in English because, for some topics, queries in Spanish did not yield satisfactory results. Additionally, after the initial response from the chatbot, an additional request was always made to refine the problems by adding more context, since in most cases the problems lacked sufficient detail to make them engaging.

After receiving the problems generated by ChatGPT, a review process was initiated to ensure the problem statements aligned with the chosen theme, had the required level of detail, matched the provided prompt data, and were free of general errors. Occasionally, certain problems had to be regenerated to include more context or distinguish them from previously generated ones. In some cases, additional information related to the topic was supplied to the chatbot for incorporation into the problem statements. Finally, when preparing the problem sheet to be provided to the students, the necessary modifications were made to enhance and tailor each problem to the context of the theme.

3.2.3 Phase III: In-class intervention

The next phase involved implementing the problems in both the test and control groups. Students engaged in two 55-minute sessions, each focused on solving problems. In the first session, they worked on a set of five problems, serving as their initial exposure to the curriculum exercises. The second session involved eight problems to reinforce their learning. In both sessions, students in the test groups were paired or grouped in threes based on their thematic interests, while the control groups were grouped at the teacher's discretion without considering preferences. Afterward, the problems were collectively corrected on the board for the entire group.

3.2.4 Phase IV: Post-experience survey and exam

Students were presented with a competency test consisting of theoretical questions (40%) and problems (60%) similar to those encountered in the previous sessions. This assessment method evaluated both their grasp of theoretical concepts and their problem-solving skills, offering a comprehensive measure of their learning progress. After the exam, students completed a survey that included a motivation rating aimed to compare their initial and final motivation levels.

3.3 Metrics

To determine if there was a significant difference in motivation before and after the problems of the learning situation covered in this work, both for the groups to whom the problems were personalized and those to whom they were not, the McNemar test was used. This method is widely used to analyze paired or related data, as in this case, where the responses of the same students were compared before and after the intervention.

First, the study collected students' motivation levels through surveys conducted before and after covering the subject matter. The 75th percentile separated highly motivated students from the rest. Thus, four groups emerged: those not highly motivated before or after, those not highly motivated before but highly motivated after (b), those highly motivated before but not after (c), and those highly motivated both before and after.

A contingency table was constructed from this data to calculate the McNemar test statistic using Equation 1. This statistic allows for evaluating the discrepancy between cases where changes in motivation were observed after the intervention.[22] The result, denoted as χ^2, follows a chi-square distribution with one degree of freedom.

$$\chi^2 = \frac{(b-c)^2}{b+c} \tag{1}$$

ISBN 978-84-09-51520-2

To determine whether the obtained result is significant, it is compared with the ρ-value (significance level) with a predefined critical value, in this case, $\rho = 0.05$. If the calculated ρ-value is less than the significance level, the null hypothesis is rejected, and the alternative hypothesis is accepted. In this work, the following hypotheses have been formulated:

- Null Hypothesis (H0): Performing personalized problems based on student interests does not have a significant impact on motivation.
- Alternative Hypothesis (H1): Performing personalized problems based on student interests does have a significant impact on motivation.

The McNemar test and the contingency table were employed for a quantitative analysis of the relationship between motivation before and after the teaching situation, which helps determine significant differences in motivation between the control and test groups, providing statistical support for research conclusions.

Additionally, the Wilcoxon test was used to assess the intervention's impact on student motivation, especially when the data does not follow a normal distribution.[23] The ρ-value obtained was compared to the predefined critical value of $\rho = 0.05$ to ascertain the statistical significance of the differences, offering evidence of the intervention's effectiveness.

On the other hand, the Pearson correlation was applied to investigate the relationship between motivation levels and grades obtained in the part of the syllabus with personalized exercises. It measures the strength and direction of the linear relationship between motivation (measured on a scale from 0 to 10) and corresponding grades. This analysis helps assess the correlation between motivation and academic performance.

$$r_{xy} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_j (x_j - \bar{x})^2 (y_j - \bar{y})^2}} \tag{2}$$

Where x and y represent the variables of interest (motivation levels and grades), and $\bar{x}$ and $\bar{y}$ are the means of these variables.

The Pearson correlation assesses the linear relationship between variables but does not imply direct causality. It is important to recognize that academic performance can be influenced by various factors like skill level, study dedication, and contextual aspects. However, the Pearson correlation is a valuable tool for examining the connection between quantitative variables and can provide insights into how motivation may be linked to academic performance in this particular context.

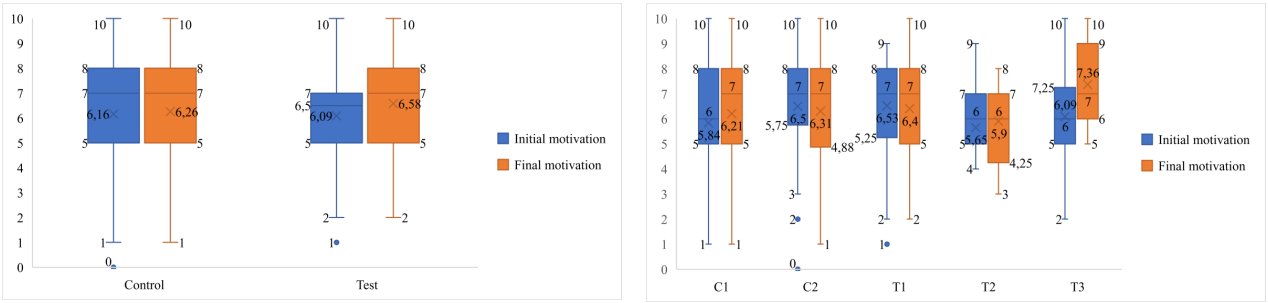
4. RESULTS

The results obtained from the surveys conducted by the students, along with the statistical significance analysis of the implementation of personalized problems in the test and control groups are described in this section. Besides, the relationship between the scores on the curriculum evaluation test where the intervention was carried out and student motivation is analyzed.

Firstly, some boxplots are depicted in Figure 2a showing the initial and final motivations of students in both the control and test groups.

The data analysis reveals that the control groups maintained relatively consistent motivation levels from the start to the end of the study (averaging 6.16 to 6.26). In contrast, the test groups had slightly lower initial motivation than the control groups but showed a slight increase in final motivation (averaging 6.09 to 6.58). Notably, the upper quartile, median, and mean of final motivation in the test groups all increased compared to the initial motivation. This suggests that, on the whole, students experienced a boost in motivation after engaging with personalized problems. Specifically, the higher upper quartile value indicates more students achieved higher motivation levels, the increased median suggests half of the students improved their final motivation, and the higher mean shows an overall rise in student motivation after participating in personalized problems.

ISBN 978-84-09-51520-2



(a) (b)

Figure 2: Box-and-whisker plot on: (a) the motivation of control and test groups. (b) the motivation of individualized control and test groups.

Thus, the increase in the upper quartile, median, and mean of final motivation in the test groups suggests that personalized problems based on topics of interest to students had a positive impact on their motivation. These results support the effectiveness of personalization in improving motivation and highlight the relevance of using educational approaches that take into account individual student interests.

Moreover, Figure 2b shows dis-aggregated boxplots representing initial and final motivation for each group. In C1, most students started with a motivation of 5 to 8 and ended up slightly more motivated (from an average level of 5.84 to 6.21). In C2, there are at least two students with low motivation at the beginning, and some show a decrease in motivation after the intervention. In the test groups, T1 shows an initial motivation similar to the final one, and in T2, the initial motivation is similar to the final, but decreases slightly. In T3, all students show a significant improvement in motivation after the intervention.

On the other hand, Table 2 shows the Wilcoxon test statistics (W^+) and the ρ-values. As can be seen, these values confirm the results analyzed from the graphs in Figure 2b, since the T3 group is the only one with a ρ-value less than the critical value $\rho = 0.05$, which rejects the null hypothesis.

Table 2: Wilcoxon test statistic and ρ-value for control and test groups.

	C1	**C2**	**T1**	**T2**	**T3**
W^+	20.5	38.5	63	67	9
ρ	0.2555	0.6185	0.7924	0.6422	0.0035

The significant improvement in the T3 group suggests that personalized intervention can increase student motivation. However, it is crucial to consider other factors that may influence motivation, such as incidents in the T2 group that generated tension and demotivation. These contextual factors should be taken into account when interpreting the results. Furthermore, more data and additional samples from different contexts are needed for a more complete and generalizable assessment of the effects of personalized intervention on student motivation.

As mentioned above, additional statistical analysis was performed using McNemar's test to assess significant differences in motivation between the control and test groups. A motivation threshold equal to or greater than 8 was established to identify highly motivated students, and these categories were used in the contingency tables for analysis.

Table 3 shows the McNemar test results for the control and test groups. In the control groups, there were 14 highly motivated students before the learning situation, but after completion, it decreased to 13. As for the students without high motivation, there were 23 at the beginning and increased to 24 at the end. In the test groups, before the intervention, there were 13 highly motivated students and 49 without high motivation. After the intervention, the number of highly motivated students increased to 22, and those without high motivation decreased to 40.

After calculating McNemar's statistic to analyze whether the performance of personalized problems based on student interests has a significant impact on motivation, the results shown in Table 4 were obtained.

ISBN 978-84-09-51520-2

Table 3: McNemar's test for control and test groups (control/test).

Motivation		After		
		Demotivated	**Motivated**	**Total**
Before	**Demotivated**	21/34	2/15	23/49
	Motivated	3/6	11/7	14/13
	Total	24/40	13/22	37/62

Table 4: χ^2 and ρ-values of McNemar's statistical test for control and test groups.

	Control groups	**Test groups**
χ^2	0.20	3.86
ρ	0.6547	0.0495

The results were compared with a critical value of $\rho = 0.05$ to determine statistical significance. In the control group, the χ^2 statistic was 0.20 with a ρ-value of 0.6547, indicating no significant difference in motivation before and after the learning situation in this group. In the test group, the χ^2 statistic was 3.86 with a ρ-value of 0.0495, suggesting significant differences in motivation before and after the personalized problem intervention. This indicates that the intervention had a positive and significant impact on student motivation in the test groups compared to the control groups.

Finally, Pearson's correlation was used as a metric to analyze the relationship between student motivation and the grades obtained in the learning situation of the intervention. Table 5 shows the values of the Pearson correlation coefficients for each of the groups.

Table 5: Pearson's correlation coefficients for control and test groups.

	C1	**C2**	**T1**	**T2**	**T3**
Pearson correlation coefficient (r)	0.12	0.33	0.47	0.13	0.55

In group C1, the correlation is positive but weak (0.12). In group C2, it is positive and moderate (0.33). Among the test groups, the T1 group exhibits a stronger positive correlation (0.47), whereas the T2 group shows a weak correlation (0.14). The T3 group has a positive and strong correlation (0.55). These findings indicate that in most test groups where personalized problems based on student interests were utilized, a positive and significant relationship between motivation and grades exists, suggesting that higher motivation is associated with better performance. However, it is crucial to consider other factors that might influence this relationship and conduct further analysis.

5. CONCLUSIONS

The article focused on the use of LLMs, specifically the ChatGPT model, to generate personalized exercises and improve the motivation of Secondary Education students. The results showed that content personalization had a positive impact on student motivation and a positive correlation was observed between motivation and academic performance. The influence of other factors, such as classroom environment and teacher-student dynamics, on the results was recognized. The study has limitations in terms of duration and student sample, therefore, future research considering these aspects is suggested for a more robust understanding of the impact of personalization in education.

This study has significant implications for teaching. Leveraging LLMs for educational content creation offers teachers flexibility and resources to cater to individual student needs, ultimately improving the learning experience and academic achievements. Personalized exercises aligned with students' interests can boost motivation and engagement, accommodating diverse learning styles. It is important to note that while LLMs can enhance education, they should not replace the teacher's role entirely. Teachers remain essential for guiding, motivating, and evaluating the content generated by LLMs, which may contain errors. LLMs are a supplementary tool that enriches the teacher's work, rather than a substitute for human interaction and personalized tutoring.

ISBN 978-84-09-51520-2

REFERENCES

[1] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al., "Language models are few-shot learners," *Advances in neural information processing systems* **33**, 1877–1901 (2020).

[2] Harrington, S. A., "The ultimate study partner: Using a custom chatbot to optimize student studying during law school," *Available at SSRN 4457287* (2023).

[3] Carvalho, I. and Ivanov, S., "Chatgpt for tourism: applications, benefits and risks," *Tourism Review* (2023).

[4] Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P.-S., Cheng, M., Glaese, M., Balle, B., Kasirzadeh, A., et al., "Ethical and social risks of harm from language models," *arXiv preprint arXiv:2112.04359* (2021).

[5] Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., et al., "Chatgpt for good? on opportunities and challenges of large language models for education," *Learning and individual differences* **103**, 102274 (2023).

[6] Reeve, J., "Why teachers adopt a controlling motivating style toward students and how they can become more autonomy supportive," *Educational psychologist* **44**(3), 159–175 (2009).

[7] OpenAI, "GPT-4: OpenAI's Generative Pretrained Transformer 4." https://openai.com/gpt-4 (2023). Accessed: 2nd September 2023.

[8] Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., et al., "Improving language understanding by generative pre-training," (2018).

[9] Lo, C. K., "What is the impact of chatgpt on education? a rapid review of the literature," *Education Sciences* **13**(4), 410 (2023).

[10] Rahman, M. M. and Watanobe, Y., "Chatgpt for education and research: Opportunities, threats, and strategies," *Applied Sciences* **13**(9), 5783 (2023).

[11] Sok, S. and Heng, K., "Chatgpt for education and research: A review of benefits and risks," *Available at SSRN 4378735* (2023).

[12] Opara, E., Theresa, A. M.-E., and Aduke, T. C., "Chatgpt for teaching, learning and research: Prospects and challenges," (3 2023).

[13] Milano, S., McGrane, J. A., and Leonelli, S., "Large language models challenge the future of higher education," *Nature Machine Intelligence* **5**(4), 333–334 (2023).

[14] Ryan, R. M. and Deci, E. L., "Intrinsic and extrinsic motivations: Classic definitions and new directions," *Contemporary educational psychology* **25**(1), 54–67 (2000).

[15] Puentes, A. E., Guerrero Cruz, E., et al., "Factores que intervienen en la motivación durante la adolescencia y su influencia en el ámbito escolar," (2019).

[16] Balasso, C., *Distribución de niveles en las clases de inglés, y su relación con la motivación y estrategias de aprendizaje en la educación secundaria.*, PhD thesis, Universidad de Almería (2018).

[17] Sacaca, L. and Pilco, R., "Habilidades sociales en estudiantes de educación secundaria," *Revista Estudios Psicológicos* **2**(4), 109–120 (2022).

[18] Pane, J. F., Steiner, E. D., Baird, M. D., and Hamilton, L. S., "Continued progress: Promising evidence on personalized learning.," *Rand Corporation* (2015).

[19] Hattie, J., [*Visible learning: A synthesis of over 800 meta-analyses relating to achievement*], routledge (2008).

[20] Macías Sánchez, R., "Metodologías activas de aprendizaje para matemáticas en educación secundaria," Ice (2019).

[21] Ruiz, I. R. B., "Revisión documental sobre el fortalecimiento de la competencia lectora mediante el uso de un ambiente virtual de aprendizaje (ava) en estudiantes de educación básica y secundaria," *Ciencia Latina Revista Científica Multidisciplinar* **6**(5), 2970–2998 (2022).

[22] McNemar, Q., "Note on the sampling error of the difference between correlated proportions or percentages," *Psychometrika* **12**(2), 153–157 (1947).

[23] Wilcoxon, F., "Individual comparisons by ranking methods," in [*Breakthroughs in Statistics: Methodology and Distribution*], 196–202, Springer (1992).

ISBN 978-84-09-51520-2