IRSTI 14.25.09 https://doi.org/10.26577/JES202584312

G. Totikova[1], A. Yessaliyev[1]*, N. Medetbekova[1], A. Bitemir[2]

[1]M. Auezov South Kazakhstan University, Shymkent, Kazakhstan
[2]O. Zhanibekov South Kazakhstan Pedagogical University, Shymkent, Kazakhstan
*e-mail: aidar.esali@mail.ru

ANALYZING STEAM APPROACH EFFECTIVENESS IN PRIMARY SCHOOL: G, R, CLES METRICS

This study evaluates the educational effectiveness of short-cycle 2D/3D modeling in elementary school as a tool for developing mechanical, technical, and spatial thinking. By integrating statistical rigor with practical interpretability, we bridge the gap between research findings and classroom applicability: beyond p-values, we report effect sizes (Hedges' g, r) and the probability of superiority (CLES) to quantify pedagogical impact.

Scientific novelty lies in the systematic application of the g–r–CLES triad for educational data interpretation, emphasizing probabilistic insights into learning gains. Practical significance is demonstrated through a compact, technology-accessible intervention format designed for seamless integration into standard curricula.

Methodology: A parallel-group experiment (Experimental/Control; grades 1–4, $N = 172$) employed a pre-post design using the Bennett test. Statistical analysis (Statistica 10) included: Wilcoxon tests for within-group shifts; Mann–Whitney tests for between-group differences in gains (Δ = Post − Pre); Reporting of p, Z, $r = Z/\sqrt{N}$, Hedges' g, and CLES (derived from U and d under normal approximation).

Results: Experimental groups showed large pre-post improvements ($p < .001$; $r \approx 0.85$–0.88; $g \approx 1.9$–2.4; CLES ≈ 0.91–0.95); Control vs. experimental comparisons consistently favored the intervention (CLES ≈ 0.76–0.86), with an aggregate effect of $Z = 2.83$ ($p = .004$), $r = 0.42$, $g \approx 0.96$, and CLES ≈ 0.74–0.75.

Practical interpretation: In ~75% of cases, a student in the intervention group outperformed a control peer in learning gains.

Contributions: Evidence that brief 2D/3D design cycles yield statistically robust and pedagogically meaningful effects; A replicable analysis framework combining classical effect sizes with probabilistic benchmarks; Implementation guidelines for classroom modules, including growth-threshold monitoring.

Keywords: STEAM, elementary school, 2D/3D modeling, Bennett test, Hedges g, CLES.

Г.А. Тотикова[1], А.А. Есалиев[1]*, Н.Н. Медетбекова[1], А.А. Битемір[2]

[1]М. Әуезов атындағы Оңтүстік Қазақстан университеті, Шымкент, Қазақстан
[2]О. Жәнібеков атындағы Оңтүстік Қазақстан педагогикалық университеті, Шымкент, Қазақстан
*e-mail: aidar.esali@mail.ru

Бастауыш мектепте STEAM-тәсілінің тиімділігін талдау: G, R, CLES метрикалары

Бұл зерттеу бастауыш сыныптағы механикалық, техникалық және кеңістіктік ойлауды дамыту құралы ретінде 2D/3D модельдеудің қысқа мерзімді білім беру тиімділігін бағалайды. Статистикалық қатаңдықты тәжірибелік түсініктілікпен үйлестіре отырып, біз зерттеу нәтижелері мен сыныптағы қолдану арасындағы алшақтықты толтырамыз: p-мәндерден тыс, біз педагогикалық әсерді сандық түрде бағалау үшін эффект өлшемдерін (Hedges' g, r) және үстемдік ықтималдығын (CLES) есептейміз.

Ғылыми жаңалық білім беру деректерін түсінуде g–r–CLES үштігінің жүйелі қолданылуына және оқу жетістіктерінің ықтималдықтық бағалауына негізделген. Тәжірибелік маңыздылық стандартты оқу бағдарламаларына оңай енгізуге арналған ықшам, технологиялық тұрғыдан қолжетімді интервенция форматы арқылы көрсетілген.

Әдіснама: Параллельді топтардың эксперименті (Эксперименттік/Бақылау; 1-4 сыныптар, $N = 172$) Беннет тесті арқылы pre-post дизайнын қолданды. Статистикалық талдау (Statistica 10) мынаны қамтыды: Топ ішіндегі өзгерістер үшін Уилкоксон критерийі; Жетістіктердегі топаралық айырмашылықтар үшін Манн-Уитни критерийі (Δ = Post − Pre); p, Z, $r = Z/\sqrt{N}$, Hedges' g және CLES (қалыпты жуықтауда U және d арқылы есептелген) туралы есеп.

Нәтижелер: Эксперименттік топтар pre-post жетістіктерде айтарлықтай өсу көрсетті ($p < .001$; $r \approx 0.85$–0.88; $g \approx 1.9$–2.4; CLES ≈ 0.91–0.95). Бақылау және эксперимент топтарын

салыстыру интервенцияның артықшылығын әділ көрсетті (CLES ≈ 0.76–0.86) жалпы эффект Z = 2.83 (p = .004), r = 0.42, g ≈ 0.96 және CLES ≈ 0.74–0.75 болды.

Тәжірибелік құндылығы: Шамамен 75% жағдайда эксперимент тобындағы оқушы бақылау тобындағы құрдасына қарағанда жоғары жетістікке қол жеткізді.

Зерттеудің үлесі: 2D/3D модельдеудің қысқа циклдары статистикалық дәлелденген және педагогикалық маңызы бар нәтиже береді; Классикалық эффект өлшемдерін ықтималдықтық критерийлермен біріктіретін қайталанатын талдау схемасы; Оқу модульдерін енгізу және белгіленген өсу шегін асқан оқушылар үлесін бақылау бойынша әдістемелік ұсынымдар.

Түйін сөздер: STEAM, бастауыш мектеп, 2D/3D модельдеу, Беннетт сынағы, Hedges g, CLES.

Г.А. Тотикова[1], А.А. Есалиев[1]*, Н.Н. Медетбекова[1], А.А. Битемир[2]

[1]Южно-Казахстанский университет им. М. Ауэзова, Шымкент, Казахстан
[2]Южно-Казахстанский педагогический университет им. У. Жанибекова, Шымкент, Казахстан
*e-mail: aidar.esali@mail.ru

Анализ эффективности STEAM-подхода в начальной школе: метрики G, R, CLES

В исследовании оценивается образовательная эффективность краткосрочного 2D/3D моделирования в начальной школе как инструмента развития механического, технического и пространственного мышления. Сочетая статистическую строгость с практической интерпретируемостью, мы преодолеваем разрыв между научными результатами и их применимостью в классе: помимо p-значений, мы сообщаем размеры эффекта (Hedges' G, R) и вероятность превосходства (CLES) для количественной оценки педагогического воздействия.

Научная новизна заключается в систематическом применении триады G–R–CLES для интерпретации образовательных данных с акцентом на вероятностную оценку учебных достижений. Практическая значимость демонстрируется через компактный, технологически доступный формат вмешательства, разработанный для плавной интеграции в стандартные учебные программы.

Методология: эксперимент с параллельными группами (экспериментальная/контрольная; 1-4 классы, N = 172), использовался pre-post дизайн с тестом Беннета. Статистический анализ (Statistica 10) включал: критерий Уилкоксона для внутригрупповых изменений; критерий Манна-Уитни для межгрупповых различий в приросте (Δ = Post − Pre); отчетность по p, Z, r = Z/√N, Hedges' g и CLES (рассчитанным из U и d при нормальном приближении).

Результаты исследования: экспериментальные группы показали значительный pre-post прогресс (p < .001; r ≈ 0.85–0.88; g ≈ 1.9–2.4; CLES ≈ 0.91–0.95). Сравнение контрольной и экспериментальной групп последовательно демонстрировали преимущество вмешательства (CLES ≈ 0.76–0.86) с совокупным эффектом Z = 2.83 (p = .004), r = 0.42, g ≈ 0.96 и CLES ≈ 0.74–0.75.

Практическая значимость: примерно в 75% случаев учащиеся экспериментальной группы демонстрировал больший прогресс, чем их сверстники из контрольной группы.

Практическая значимость исследования: доказательство того, что краткие циклы 2D/3D моделирования дают статистически значимый и педагогически ценный эффект; воспроизводимая схема анализа, сочетающая классические размеры эффекта с вероятностными критериями; практические рекомендации по внедрению модулей в учебный процесс и мониторингу доли учащихся, преодолевших заданный порог прогресса.

Ключевые слова: STEAM, начальная школа, 2D/3D моделирование, тест Беннета, Hedges G, CLES.

Introduction

Today, the STEAM approach in elementary schools is considered a tool for the early integration of mathematical and natural science concepts with project activities and visual-spatial thinking. Recent reviews show that the effect of STEAM modules is manifested in educational achievements and meta-subject skills, but depends on the quality of didactic design, the role of the teacher, and the "material anchor" – tools that translate abstractions into substantive actions (prototyping, modeling, 3D printing) (Yim, I.H.Y., et al., 2024; Amanova, A.K., et al., 2025). In the primary school segment, positive results were also recorded in special reviews on 3D modeling/printing: from increased subject understanding to increased engagement and spatial visualization (Fokides, E., Lagopati, G. 2024). These effects correlate with the tasks of developing engineering and technological literacy "from below," when working with 2D/3D representations becomes a routine practice rather than an elective.

The topic has additional social significance for Kazakhstan. According to PISA-2022, the proportion of high–achieving students (levels 5-6) in mathematics is approximately 2% (the OECD average is 9%), indicating the need for approaches that can simultaneously strengthen basic literacy and push the "top tail" of the distribution (OECD, 2023).

Pedagogical publications often focus on the "is/is not" effect (p-levels), while it remains unclear how much we "see" the effect in the classroom and what is the probability that a student will outperform a comparable student from the control in the intervention. The methodological agenda of the last decade calls for accompanying hypothesis testing with a report on the size of the effect and providing an interpretation that is understandable to practice (Lakens, 2013). For nonparametric checks, the transformation of Z-statistics into r (as a measure of strength) is consistently applied, and for standardized differences, the unbiased Hedges correction g (includPedagogical publications often focus on the "is/is ot" effect (p-levels), while it remains unclear how much we "ee" the effect in the classroom and what is the probability that a student will outperform a comparable student from the control in the intervention. The methodological agenda of the last decade calls for accompanying hypothesis testing with a report on the size of the effect and providing an interpretation that is understandable to practice (Lakens, D., 2013). For nonparametric checks, the transformation of Z-statistics into r (as a measure of strength) was consistently applied, and for standardized differences, the unbiased Hedges correction g (including for repeated measurements according to the schemes (Morris. S.B., DeShon, R.P., 2002). However, coefficients r and g do not provide an answer to the question of the probabilistic advantage of intervention. To solve this problem, special probabilistic indicators are used, including the Common Language Effect Size (CLES).

The classical formulation of CLES by McGraw, K.O., & Wong, S.P. (1992): the probability that a randomly selected observation from one group is superior to an observation from another. Vargha, A., & Delaney, H. (D. (2000) proposed a nonparametric generalization, A statistic of "probability of superiority" that is rThe classical formulation of CLES by McGraw, K.O., & Wong, S.P. (1992): the probability that a randomly selected observation from one group is superior to an observation from another. Vargha, A., & Delaney, H.D. (2000) proposed a nonparametric generalization, A statistic of "probability of superiority" that is resistant to deviations from normality; the line of work of Ruscio et al. developed procedures for estimating and confidence intervals for the probability of superiority. For normal assumptions, the analytical bridge is useful: CLES = $\Phi(d/\sqrt{2})$, which makes it possible to interpret standardized differences in probabilistic terms. For ordinal/biased distributions, the A-score was applied directly (Vargha–Delaney A). Collectively, the g+ r + CLES bundle provides a triple perspective: metrics, correlations, and probabilities.

Purpose, novelty and contribution of the research

The study focused on STEAM interventions in elementary schools.

This study examines the relationship between the statistical and practical significance of the results according to the methodology for evaluating design and technical thinking (the Bennett test).

The aim of this study was to evaluate the didactic effectiveness of the 2D/3D module through a comprehensive analysis of statistical indicators, including the p-significance level, Hedges coefficient g, r correlation measure (based on Z-statistics of nonparametric criteria), and Common Language Effect Size (CLES) metric.

The scientific novelty of this study is as follows:

1. Introduction of the probabilistic approach (CLES) for interpreting educational effectiveness, along with traditional effect size indicators.

2. Application of correction methods for repeated measurements (Morris & DeShon) and comparative analysis of CLES calculated based on U-statistics and the d-criterion.

3. The adaptation of the study to the conditions of primary education in the Republic of Kazakhstan, considering the requirements for basic literacy and the development of higher cognitive skills (in the context of PISA-2022), as well as improving the professional readiness of teachers to implement STEAM modules.

A practical contribution is the development of aA practical contribution is the development of a unified system for evaluating the effectiveness of educational interventions for use at the school and municipal levels, which makes it possible to make informed decisions about scaling successful practices.

Research hypotheses

H1. A statistically significant shift (Z-criterion) is predicted with high r/g values and CLES > 0.70 due to the use of material manipulative means and cyclical prototyping in elementary schools (con-

firmed by reviews on 3D modeling and STEAM pedagogy).

H2. The CLES is expected to be in the range of 0.65–0.75, with a minimum average effect level (g), which corresponds to data from meta-analyses on STEAM interventions in school education.

H3. When statistical significance is reached ($p<0.05$), the consistency of metrics is assumed: CLES is significantly higher than 0.50, g is medium/high, and r is moderately high. Discrepancies are allowed (for example, $p<0.05$ with CLES≈0.55–0.60), which should be inter

Literature review

Over the past decade, primary school research has documented a steady, albeit variable, positive effect of STEAM approaches: the benefits depend on the specific didactic design, the role of the teacher, and the presence of "material anchors" (projects, prototyping, visual and spatial work).

A systematic review of recent research (Yim et al., 2024) confirms the ability of STEAM interventions to enhance academic achievement and develop interdisciplinary competencies but notes a significant lack of standardized criteria for evaluating their effectiveness.

The introduction of 2D/3D modeling and 3D printing in elementary schools is associated with increased spatial thinking, the concretization of abstract concepts, and increased academic engagement. A scoping review of the use of 3D printers in elementary school students, as well as empirical work on 3D modeling, showed improvements in meaningful understanding and spatial visualization; the effect was observed even with short cycles of project activity. (Fokides & Lagopati, G. (2024); Toptaş et al., 2012).

The relevance of STEAM approaches in the Kazakh educational system is emphasized by the PISA-2022 data (OECD, 2023, Volume I, II), which revealed the need to develop mathematical literacy and introduce practice-oriented methods in primary schools, as well as a growing number of local research initiatives in this area (Amanova et al., 2025).

In the broader profile of PISA-2022, the link between creative thinking and academic performance is also noteworthy, which is an important argument in favor of design practices (OECD, 2023). Simultaneously, mapping the STEM/STEAM landscape in the Republic of Kazakhstan shows a rapid increase in publications and initiatives since 2019, but also records gaps in methodological support, teacher training, and the operationalization of "practical benefits" in the school classroom (Abdrakhmanova et al., 2025; Zhumabay et al., 2024). In parallel, local cases of integrating AR/3D tools into primary and secondary school courses are accumulating (Beisenbayeva et al., 2024; Tazabekova et al., 2024). This makes relevant research that can link statistical evidence with didactic interpretation.

The methodological basis for assessing the practical significance of the results is a comprehensive analysis using probabilistic metrics, in particular, the Common Language Effect Size (CLES), the concept of which was originally developed by McGraw and Wong (1992). This indicator, interpreted as the probability of superiority of a randomly selected result of an experimental group, has two fundamental advantages: firstly, it provides an intuitive interpretation for practicing teachers, and secondly, it demonstrates resistance to violations of parametric assumptions due to its nonparametric analogue, the Varga-Delaney statistics (Ruscio, J., & Mullen, T. (2012).

In the framework of this study, a set of complementary metrics was used, including a standardized mean difference with a Hedges correction (g) for intergroup comparisons, a correlation measure r calculated using Z-statistics of nonparametric criteria, and probabilistic indicators CLES/A.

Special attention was paid to the methodological aspects of the analysis of repeated measurements using the Morris and DeShon correction (2002), which ensured the comparability of the results with international studies (Liu et al., 2019; Dunlap, P. 1999; Ortelli, O.A. 2018).

The choice of nonparametric analysis methods was determined by the peculiarities of the data distribution of the Bennett test, including its discrete nature and potential ceiling effect. The regional specifics of the study were considered through an analysis of Kazakhstani publications on STEAM education, which revealed the need for standardized reporting schemes on the effectiveness of interventions, which determined the methodological contribution of this work through the development of an integrated assessment system (g-r-CLES/A).

Methods

The study was performed as a experiment with parallel groups (Experimental and Control) and two time measurements (Pre and Post). Students from grades 1-4 of the same school participated, and paired observations were available for each child,

which made it possible to assess individual dynamics. The intervention consisted of 2D/3D modeling cycles embedded in the structure of the educational process as an extracurricular activity (problem statement, flat prototyping, basic 3D design, and solution reflection). This design makes it possible to evaluate intra−group shifts (Pre→Post) and the intergroup difference in increases Δ = Post – Pre with a comparable calendar window and study load.

The study included students in grades 1-4: in the control group, there were 21/21/23/22 students (grades 1-4), and in the experimental group, there were 23/20/21/21 (total N = 172). (Table 1). The inclusion criteria were participation in all mandatory intervention classes and the presence of paired data. The exclusion criteria were incomplete data and individual curricula that were incompatible with the module. At the pre-stage, the groups were comparable in terms of design and technical thinki

The measuring instrument was the Bennett Mechanical Comprehension Test in adapted forms appropriate to the age and curriculum of grades 1-4, which made it possible to assess the understanding of basic mechanical principles and spatial and technical reasoning. For intergroup comparisons, the Δ = Post −Pre increments were used, since this approach removes the problems of direct equivalence of "raw" scores in different classes and focuses interpretation on the learning effect.

The Pre-measurement was carried out procedurally at the beginning of the half-year, the intervention was carried out within the framework of regular lessons lasting 30 minutes with a frequency of once or twice a week, and the post-measurement was carried out at the end of the half-year using the same forms. At the data preprocessing level, the completeness of paired observations was checked, fields were typed (Group/Stage/Class/Score), records were deleted without an identifier or with both measurements missing, and individual Δ\DeltaΔ was calculated for each student. For the descriptive part, averages, medians, quartiles, standard deviations, and numbers were additionally calculated by Class × Group × Stage; aggregation "as a whole" was applied cautiously and only where procedures and scales were uniform. For illustrative purposes, the proportions of "zero shifts" (ties) and the distribution of signs of differences in paired comparisons were additionally evaluated; this is important for the interpretation of nonparametric criteria.

Statistical analysis was based on two lines. For intragroup shifts (Pre→Post), the Wilcoxon criterion was used for the paired samples. We reported the strength of the effect as $r = Z/\sqrt{N}$ (where N is the number of valid pairs), as well as a standardized mean difference for repeated dav measurements with a small Hedges correction (g) using the Morris and DeShon procedures, which prevents overestimation of the effect on the background of intrasubject correlation.

The Mann–Whitney criterion was used for intergroup differences in Δ increments; in addition to the p-level and Z-statistics, $r = Z/\sqrt{N}$ (where N is the total number of observations), Hedges g based on a standardized Δ difference, and probabilistic interpretation through the Common Language Effect Size (CLES) were reported. The latter was estimated in two equivalent ways: (i) directly from U as the probability of stochastic superiority (A is the Vargha–Delaney estimate) and (ii) through a normal approximation from the standardized difference d using the formula $CLES = \Phi(d/\sqrt{2})$.

For Wilcoxon, the standard rules for handling zero differences (ties) and checking the "zero-method" options were used, for Mann–Whitney, the correct accounting of unequal group volumes and possible rank matches was used. In all checks, the significance was considered two-way; reporting was unified: Z, p, r, Hedges g, and CLES were supplemented with 95% confidence intervals for g and CLES (bootstrap or delta method), and the decimal separator in all numbers was a dot. The choice of this particular triad of metrics – g, r, and CLES – is not accidental: g ensures comparability with international educational literature, r makes nonparametric results readable for practitioners, and CLES "translates" the effect into a probabilistic language more convenient for pedagogical decision-making.

The analysis was performed using Statistica 10 with data in long format (ID, Class, Group, Stage [Pre/Post], Test, Score). The increments were calculated as Δ = Post − Pre (Data → Compute Variables); descriptive statistics were obtained through Statistics → Basic Statistics/Tables → Descriptive Statistics over the Class × Group × Stage section.

Intragroup shifts (Pre→Post) were evaluated using Wilcoxon matched pairs (Two Related Samples) with frequencies of W, Z, and p and an effect strength of $r = Z/\sqrt{N}$ (where N is the number of pairs).

The intergroup differences in gains were analyzed using the Mann–Whitney U test (Two Independent Samples) with U, Z, and p reporting; the probability of stochastic superiority was calculated as $CLES\_U = U/(n_1 \cdot n_2)$.

The standardized increment difference is represented by Cohen's d with a small Hedges correction

g. For interpretation "in the language of probability," the normal approximation CLES = $\Phi(d/\sqrt{2})$ is additionally used.

Derived indicators (Δ, r, d, g, CLES) were calculated using spreadsheet tools, and key tables and logs were exported to RTF/XLSX; the source data package and workbook were saved for reproducibility.

Ethical aspects were observed within the school context: parental/legal representatives provided informed consent, personal data were depersonalized, and measurements were carried out in regular educational conditions without interventions that could harm participants. This approach makes the intervention educational in nature and safe in form, and the results obtained are comparable and suitable for replication in other schools.

Results

The study included 172 students in grades 1-4, with similar subsample sizes in the control and experimental groups for each parallel. At the initial measurement (Pre), the distributions according to the Bennett test in the groups overlapped: the median was usually five points, the quartile intervals were similar, and the average differences were small (for example, in the 2nd grade, 5.24 in the control and 5.65 in the experiment; in the 3rd, 5.13 and 5.14, respectively), which made it possible to interpret further shifts as a result of educational impact, rather than initial incompatibility. Cm. The summary characteristics of the sample and descriptive indicators are shown in Tables 1-2, where all values are given with a decimal point and unified notation.

Table 1 – Number of students in classes and groups

Class	Group	Number of students
1	Control	21
	Experimental	23
2	Control	21
	Experimental	20
3	Control	23
	Experimental	21
4	Control	22
	Experimental	21
Altogether		172

The descriptive indicators for each class × group ×stage (pre/post) combination are presented in Table 2. At the level of "raw" distributions, there is a systematic shift to the right from Pre to Post in the experimental group (an increase in averages and/or medians, often with a moderate narrowing of IQR), whereas in the control group, the dynamics are either minimal or heterogeneous. For example, in the 2nd grade, the average score increased from 5.65 (pre) to 8.15 (post) in the experiment, with an almost unchanged control profile (5.24 → 5.38). These patterns were visually consistent with the subsequent nonparametric tests.

Table 2 – Descriptive statistics of scores on the Bennett test (pre/post) in the control and experimental groups, grades 1-4.

Class	Group	Stage	n	Mean	SD	Median	Q1	Q3	Min	Max
1	Control	Pre	21	5,05	1,56	5	4	6	3	8
		Post	21	5,19	1,12	5	5	6	3	8
	Experimental	Pre	23	5,17	0,94	5	5	6	3	7
		Post	23	7,39	1,23	7	6,5	8	5	10
2	Control	Pre	21	5,24	1,22	5	4	6	3	7
		Post	21	5,38	1,02	5	5	6	4	7
	Experimental	Pre	20	5,65	1,5	6	4,75	7	3	8
		Post	20	8,15	1,04	8	7,75	9	6	10
3	Control	Pre	23	5,13	1,1	5	4,5	6	3	7
		Post	23	5,3	1,43	5	4	6	3	8
	Experimental	Pre	21	5,14	1,06	5	4	6	3	7
		Post	21	7,43	1,16	7	7	8	5	10

Continuation of the table

Class	Group	Stage	n	Mean	SD	Median	Q1	Q3	Min	Max
4	Control	Pre	22	5,09	0,97	5	4,25	6	3	7
		Post	22	5,23	1,15	5	4	6	4	8
	Experimental	Pre	21	5,05	1,16	5	4	6	3	7
		Post	21	7,86	1,11	8	7	8	6	10

Paired pre-post comparisons within the Wilcoxon groups confirmed a pronounced positive shift in the experimental classes in the absence of statistically and didactically significant changes in the control. In the experiment, Z lies approximately in the range of 3.72–3.82 in all parallels ($p < 0.001$), with a large r (approximately 0.85–0.88) and standardized Hedges g differences of the order of 1.86–2.38; the probabilistic interpretation reaches CLES_{paired} ≈ 0.91–0.95, that is, in nine out of ten cases, the Post score exceeds its own Pre from a randomly selected student. In the control classes, Z fluctuated around zero, $p > 0.60$, r and g were small, and CLES_{paired} ≈ 0.52–0.59, which corresponded to the absence of a meaningful shift. All the indicators are listed in Table 3. In the final version, technical artifacts were eliminated (repetitions of characters, "rrr/ggg," comma and period confusion), and the Z, p, r, g, and CLES reporting format was unified.

Table 3 – Intragroup effects (Wilcoxon, r, d av, g, CLES)

Class	Group	N	W plus	Z	p value	r from Z	n pos	n neg	n ties	CLES paired	d av	g Hedges
1	Control	21	73	0,259	0,793	0,065	9	7	5	0,548	0,105	0,101
	Experimental	23	190	3,823	0	0,877	19	0	4	0,913	2,025	1,955
2	Control	21	64	0,227	0,815	0,059	8	7	6	0,524	0,127	0,122
	Experimental	20	171	3,724	0	0,878	18	0	2	0,950	1,940	1,862
3	Control	23	75,5	0,388	0,695	0,097	9	7	7	0,543	0,136	0,132
	Experimental	21	206,5	3,789	0	0,847	19	1	1	0,929	2,050	1,972
4	Control	22	54	0,094	0,924	0,025	9	5	8	0,591	0,128	0,123
	Experimental	21	187,5	3,722	0	0,854	18	1	2	0,905	2,475	2,381
All	Control	23	98,5	0,566	0,566	0,133	11	7	5	0,587	0,192	0,186
	Experimental	23	190	3,823	0	0,877	19	0	4	0,913	2,025	1,955

An intergroup comparison of the increase Δ = Post − Pre (Mann–Whitney) revealed a stable advantage for the experimental group in all parallels. The Z values are approximately in the range of 2.93–3.94 ($p < 0.003$), and the U-based CLES shows a probability of superiority of 0.76–0.86 in classes, which is interpreted as a "visible" effect for practice. The standardized differences in Δ with a small Hedges correction g lie in a large zone: approximately ≈1.00 (1 class), 1.46 (2 cl.), 1.15 (3 cl.), 1.51 (4 cl.). In the recalculated summary row "as a whole," $Z = 2.83$, $p = 0.004$, $r = 0.42$, $g = 0.96$ were obtained for Δ, with CLES ≈ 0.74–0.75 (from U and consistently from d for CLES = $\Phi(d/\sqrt{2})$). This alignment of p, r, g, and CLES reduces the risk of overinterpretation and strengthens the conclusions at the aggregate level.

Graphic materials support the numerical outputs. The forest plot of intergroup g by class (Figure 1) demonstrates a unilaterally positive and modulo large effect with a reasonable 95% CI width in accordance with the size of the subsamples.

The probability of superiority diagram (Figure 2) compactly displays CLES in parallel and "in general": the values by class are mainly in the range of 0.79–0.86, with a combined score of approximately 0.74. For a reader without statistical training, this provides an intuitive reading of the result: with a probability of approximately three out of four, the increase in the "experimental" student is higher than the increase in the "control" student.

Table 4 – Intergroup effects by Δ\DeltaΔ (Mann–Whitney, r, g, CLES from U and d)

Class	N {exp}	N {ctrl}	U	Z	p	r	CLES U	Mean Δ (exp)	SD Δ (exp)	Mean Δ (ctrl)	SD Δ (ctrl)	d	g_ Hedges
1	23	21	366,0	2,925	0,003	0,441	0,758	2,217	1,731	0,143	2,330	1,018	1,00
2	20	21	361,0	3,938	0	0,615	0,860	2,500	1,670	0,143	1,493	1,490	1,46
3	21	23	383,5	3,337	0,001	0,503	0,794	2,286	1,586	0,174	1,992	1,167	1,15
4	21	22	391,0	3,887	0	0,593	0,846	2,810	1,861	0,136	1,612	1,538	1,51
All	23	23	393,5	2,834	0,004	0,418	0,744	2,217	1,731	0,261	2,261	0,972	0,96

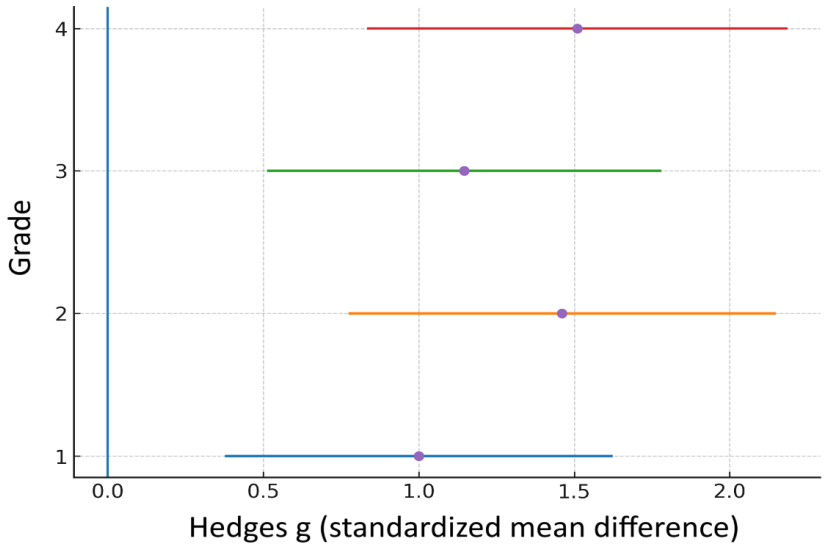

Figure 1 – Between-group effect sizes (Hedges g) by grade (95% CI)

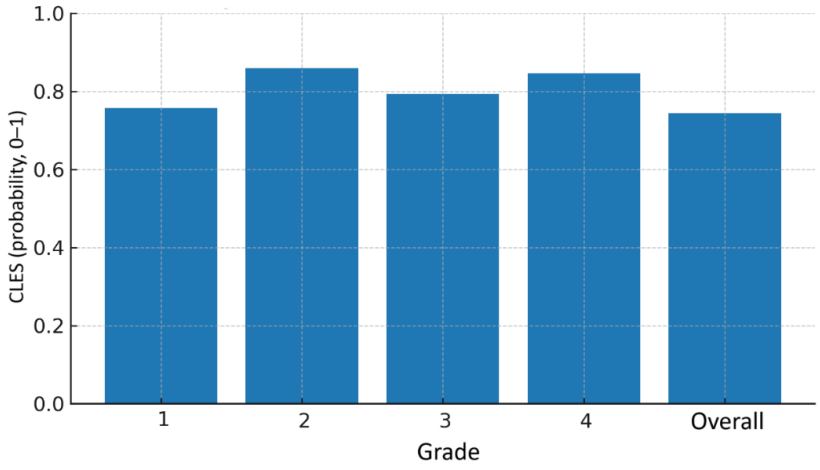

Figure 2 – Probability of superiority (CLES) by grade and overall

Stability checks confirmed the robustness of the conclusions. The normalization of Δ by the maximum of the scale (if the shapes differed by the upper bound) preserves the direction and order of magnitude of the effects; the differences between the "raw" and normalized estimates fall within the expected limits of the scale change. Additionally, the standard rules for working with ties were applied in paired tests, and in intergroup comparisons, the unequal group volumes were correctly accounted for. Aggregation was performed "as a whole" for all observations, without using "All" strings as duplicates of a separate class. Collectively, the results demonstrate a major intra-group shift in the experiment and sustained intergroup superiority in terms of gains, with the effect being "visible" both at the class level and in the summary assessment.

Discussion

The data obtained demonstrate a consistent pattern: in the experimental classes, the distributions according to the Bennett test systematically shifted to the right from Pre to Post, while in the control classes, the dynamics were minimal. Paired comparisons in experimental groups produced large and stable effects (Wilcoxon: $Z \approx 3.7–3.8$; $p < 0.001$; $r \approx 0.85–0.88$; Hedges $g \approx 1.9–2.4$; CLES_paired $\approx 0.91–0.95$), the intergroup difference in increases $\Delta = \text{Post} - \text{Pre}$ was stable in favor of intervention along all parallels ($Z \approx 2.9–3.9$; $p \leq 0.003$; CLES $\approx 0.76–0.86$), and the summary estimate "as a whole" remains significant and didactically "visible" ($Z = 2.83$; $p = 0.004$; $r = 0.42$; $g \approx 0.96$; CLES $\approx 0.74–0.75$). In terms of probabilities, this means that in about three cases out of four, the gain of an "experimental" student is higher than that of a "control" student, and for the same child, the Post is usually higher than his own Pre. These findings are based on unified p, r, g, and CLES reporting and mutually supported by descriptive class statistics.

The reason for this effect is due to the fact that during the intervention, the basic components of the Bennett test were worked out: processing text and graphic information, integrating individual parts into a single structure, converting planar images into three-dimensional models, and identifying mechanical dependencies. Short completed cycles of "sketch → prototype → test → short reflection" create dense feedback, translating mistakes into learning events; regularity (30 minutes 1-2 times a week) provides a "dose" of repetitions without overload. The noted heterogeneity in parallels is natural: we record the greatest intergroup effects in grades 2 and 4 ($g \approx 1.46$ and $g \approx 1.51$), which can be interpreted as a coincidence with the "sensitivity windows" to basic spatial transformations (2nd grade) and with the stage where mechanical ideas receive more points. applications in the educational material (4th grade). Small positive shifts in control are explained by repeated testing and general academic progress and do not change the basic picture of the benefits of the intervention.

A comparison with previous studies on initial STEAM training shows agreement both in the direction and in the scale of the effect: the greatest gains are achieved where the core is material artifacts, the design cycle, and reflection, rather than episodic "paper" projects. Our results add a methodological argument in favor of reporting classical effect sizes (g, r) combined with probabilistic interpretation (CLES), which "translates" statistics into the language of pedagogical decisions without reducing the rigor of the analysis. The data clarify the mechanism: the development of technical and spatial thinking in younger schoolchildren is supported by regular "translation" between external representations (2D or 3D) and meaningful reflection of the result; the active ingredient is a short, repetitive design cycle.

The practical significance lies in the feasibility of the intervention within the framework of regular lessons and a clear metric of the effect for monitoring. The module can be implemented as a series of short, completed tasks related to current topics in technology and mathematics, with explicit quality criteria (assembly, stability, accuracy) and mandatory micro-reflections. At the management control level, in addition to averages and g's, it is advisable to track the proportion of students who have overcome the practical threshold (for example, $\Delta \geq 2$ points to their own Pre according to Bennett) and to keep a checklist of typical errors (supports, leverage, friction) to link numerical shifts with the observed behavior and adjust assignments point by point. The CLES profile of $\approx 0.74–0.86$ indicates that a noticeable increase in the proportion of those who "crossed the threshold" is a realistic and verifiable target.

The limitations of this study set the framework for generalizations. One tool (Bennett) was used, measuring primarily the mechanical and technical components; shifting to creativity, metacognition, and mathematical reasoning requires an expanded

dashboard. The design is experimental, which leaves residual risks of confusion (differences in teachers' stylistic practices, classroom dynamics), although the growth analysis eliminates some of these risks. The age forms of the tests were assumed to be comparable. The sensitivity to normalization has shown the stability of the main conclusions, but formal verification of scale invariance remains a task for future work. Parallel stratification expands the confidence intervals, limiting "subtle" comparisons between classes; it is more correct to focus on the overall effect profile. Finally, the absence of a delayed posttest does not allow us to judge the long-term sustainability of these gains.

The prospects for further research follow directly from these limitations of the study. Methodologically, it is advisable to strengthen the design to cluster randomization by class/school, add delayed post-measurement (after 1-3 months), take into account the moderators of the effect (initial level of spatial skills, educational motivation, "fidelity" to teacher implementation, dosage), explore the "dose curve" and saturation threshold, at which further increase requires increased reflexive parts of the cycle. It is didactically useful to assemble a "minimal package" of scaling: a bank of micro-cases with increasing complexity and explicit mechanical goals, rubrics for product evaluation and reflection, a map of typical errors with brief correction scenarios, and a short teacher training (6-8 hours) in key practices: reading a drawing, translating a view into an assembly, planning operations, and analyzing unsuccessful prototypes.

Conclusion

The 2D/3D intervention integrated into elementary school lessons provides a statistically reliable and pedagogically "visible" effect: large paired shifts in experimental classes are combined with sustained intergroup superiority in gains, and the overall score remains significant ($Z = 2.83$; $p = 0.004$; $r = 0.42$; $g \approx 0.96$; CLES ≈ 0.74–0.75). In practice, this is "three cases out of four" in favor of the student who completed the module. The work contributes to the theory of design-oriented learning in primary school age, clarifying the role of the "short cycle" and translations between 2D and 3D as a mechanism for the development of technical and spatial thinking, and offers a reproducible reporting scheme (p, r, g, CLES), convenient for pedagogical solutions. Considering these limitations, the module can be recommended for scaling in school practice; further research will clarify the stability of the effect and the limits of portability to related academic areas.

Gratitude

The authors express their gratitude to the Committee of Science of the Ministry of Science and Higher Education of the Republic of Kazakhstan, the research teams of the M. Auezov South Kazakhstan State University for their support in collecting and analyzing data.

This research was funded by the Committee of Science of the Ministry of Science and Higher Education of the Republic of Kazakhstan (grant no. AP19678173)

References

1. Abdrakhmanova, K., Kadirbayeva, R., Kudaibergenova, K., Zharmukhanbetov, S., Nurmukhanbetova, G. «Formation of STEM Competencies of Future Teachers: Kazakhstani Experience» Open Education Studies, vol. 7, no. 1, 2025, pp. 20240058. https://do

2. Amanova, A. K., et al. (2025). *A systematic review of the implementation of STEAM education in schools*. **EJMSTE..** https://doi.org/10.29333/ejmste/15894

3. Amir, M.F., Fediyanto, T., Rudyanto, H.E., Afifah, N., Tortop, H.S. (2020). Elementary students' perceptions of 3Dmetric: A cross-sectional study. Heliyon, Volume 6, Issue 6, https://doi.org/10.1016/j.heliyon.2020.e04052.

4. Beisenbayeva, G. K., Mubarakov, A. M., Seylova, Z. T., Zhadrayeva, L. U., & Artymbayeva, B. N. (2024). Evaluating the Impact of an Augmented Reality App on Geometry Learning in Kazakh Secondary Schools. Journal of Information Technology Education: Research, 23, Article 22. https://doi.org/10.28945/5355

5. Dunlap, W.P. (1999). A program to compute McGraw and Wong's common language effect size indicator. Behavior Research Methods, Instruments, & Computers 31, 706–709. https://doi.org/10.3758/BF03200750

6. Fokides, E., Lagopati, G. (2024). The utilization of 3D printers by elementary school-aged learners: A scoping review. Journal of Information Technology Education: Innovations in Practice, 23, Article 6. https://doi.org/10.28945/5288

7. Lakens, D. (2013) Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. Front. Psychol. 4:863. doi: 10.3389/fpsyg.2013.00863

8. Liu, X., Carlson, R., Kelley, K. (2019). Common language effect size for correlations. The Journal of General Psychology. 146. 325-338. 10.1080/00221309.2019.1585321.

9. Morris. S.B., DeShon, R.P. Combining effect size estimates in meta-analysis with repeated measures and independent-groups designs. Psychol Methods. 2002 Mar;7(1):105-25. doi: 10.1037/1082-989x.7.1.105. PMID: 11928886.

10. OECD (2023), PISA 2022 Results (Volume I): The State of Learning and Equity in Education, PISA, OECD Publishing, Paris, https://doi.org/10.1787/53f23881-en

11. OECD (2023), PISA 2022 Results (Volume II): Learning During – and From – Disruption, PISA, OECD Publishing, Paris, https://doi.org/10.1787/a97db61c-en

12. Ortelli, O.A. (2018). RProbSup: Calculates Probability of Superiority. https://doi.org/10.32614/cran.package.rprobsup

13. Ruscio, J., & Mullen, T. (2012). Confidence Intervals for the Probability of Superiority Effect Size Measure and Area Under a Receiver Operating Characteristic Curve. Multivariate Behavioral Research, 47(2), 201–223. https://doi.org/10.1080/00273171.2012.658329

14. Tazabekova, P., Nurbekov, A., Nurbekova, Zh., Sembayev, T. (2024). The use of generated 3D models in teaching the development of AR/VR applications. World Transactions on Engineering and Technology Education. WIETE, Vol.22, No.4, pp. 241-247

15. Toptaş, Veli, Celik, Serkan, and Karaca, Elif. (2012). Improving 8 th grade spatial thinking abilities through A 3D modeling program. Turk. Online Journal Educ. Technol. 11. 128-134.

16. Vargha, A., & Delaney, H. D. (2000). A Critique and Improvement of the CL Common Language Effect Size Statistics of McGraw and Wong. Journal of Educational and Behavioral Statistics, 25(2), 101-132. https://doi.org/10.3102/10769986025002101

17. Yim, I.H.Y., et al. (2024). *STEAM in practice and research in primary schools*. Research in Science & Technological Education. https://doi.org/10.1080/02635143.2024.2440424

18. Zhumabay, N., Varis, S., Abylkassymova, A., Balta, N., Bakytkazy, T., & Bowen, G. M. (2024). Mapping the Kazakhstani STEM Education Landscape: A Review of National Research. European Journal of STEM Education, 9(1), 16. https://doi.org/10.20897/ejsteme/15576

Авторлар туралы мәлімет:

Тотикова Гулдана Арыновна – PhD доктор, Мектепке дейінгі және бастауыш оқыту теориясы мен әдістемесі кафедрасының аға оқытушысы, М.Әуезов атындағы Оңтүстік Қазақстан мемлекеттік университеті (Шымкент, Қазақстан, e-mail: sultik.askarbek@mail.ru)

Есалиев Айдарбек Аскарбекович – медицина ғылымдарының докторы, дене тәрбиесі және бастапқы әскери дайындық теориясы мен әдістемесі кафедрасының профессоры, М.Әуезов атындағы Оңтүстік Қазақстан мемлекеттік университеті (Шымкент, Қазақстан, e-mail: aidar.esali@mail.ru)

Медетбекова Нүргул Ниязбекқызы – педагогика ғылымдарының кандидаты, мектепке дейінгі және бастауыш оқыту теориясы мен әдістемесі кафедрасының доценті, М. Әуезов атындағы Оңтүстік Қазақстан университеті (Шымкент, Қазақстан, e-mail: totikovag@gmail.com)

Битемір Айдана Алиакбарқызы – Өзбекәлі Жәнібеков атындағы Оңтүстік Қазақстан педагогикалық университетінің PhD докторанты (Шымкент, Қазақстан, e-mail: aidar.esali@mail.ru)

Сведения об авторах:

Тотикова Гулдана Арыновна – PhD, старший преподаватель кафедры теории и методики дошкольного и начального обучения, Южно-Казахстанский университет им. М. Ауэзова (Шымкент, Казахстан, e-mail: sultik.askarbek@mail.ru);

Есалиев Айдарбек Аскарбекович – доктор медицинских наук, профессор кафедры теории и методики физической культуры и начальной военной подготовки, Южно-Казахстанский университет им. М. Ауэзова (Шымкент, Казахстан, e-mail: aidar.esali@mail.ru);

Медетбекова Нүргул Ниязбекқызы – кандидат педагогических наук, доцент кафедры теории и методики дошкольного и начального обучения, Южно-Казахстанский университет им. М. Ауэзова (Шымкент, Казахстан, e-mail: totikovag@gmail.com);

Битемір Айдана Алиакбарқызы – PhD докторант Южно-Казахстанского педагогического университета имени У. Жанибекова (Шымкент, Казахстан, e-mail: aidar.esali@mail.ru).

Information about authors:

Totikova Guldana Arynovna – PhD, Senior Lecturer of the department of Theory and Methodology of Preschool and Primary Education of M. Auezov South Kazakhstan University (Shymkent, Kazakhstan, e-mail: sultik.askarbek@mail.ru)

*Yessaliyev Aidarbek Askarbekovich – Doctor of Medical Sciences, Professor of the Department of theory and methods of Physical Education and primary military training of the M. Auezov South Kazakhstan State University (Shymkent, Kazakhstan, *e-mail: aidar.esali@mail.ru)*

Medetbekova Nurgul Niyazbekkyzy – Candidate of Pedagogical Sciences, Associate Professor of department of Theory and Methodology of Preschool and Primary Education of M. Auezov South Kazakhstan University (Shymkent, Kazakhstan, e-mail: totikovag@gmail.com)

Bitemir Aidana Aliakbarkyzy – PhD student of South Kazakhstan Pedagogical University named after O. Zhanibekov (Shymkent, Kazakhstan, e-mail: aidar.esali@mail.ru)

Received 17.08.2025
Accepted 20.09.2025