

Which AI Is the Best Montessori Guide? A Rubric-Based Comparative Evaluation of LLM-Generated STEAM Mathematics Activities for Early Childhood

Zafer Savas KIVILCIM¹ . Seyma KIVILCIM²

¹ Altieylul Science and Art Centre (BILSEM), Balikesir, Turkiye | ORCID 0000-0003-2429-7809 |

² Hatice Fahriye Eginlioglu Primary School, Balikesir, Turkiye | ORCID 0009-0006-5598-9444

Abstract

Although Montessori pedagogy and STEAM education share a child-centred, hands-on foundation, empirical work at their intersection in early childhood remains scarce, and the rapidly growing use of generative artificial intelligence (AI) for activity planning has not yet been examined through a Montessori lens. This study investigated whether large language models (LLMs) can design early childhood STEAM mathematics activities that are faithful to Montessori principles, and which model does so best. A structured prompt framework encoding five core Montessori principles -- auto-education, control of error, isolation of difficulty, sensorial engagement, and concrete-to-abstract progression -- was applied identically to four LLMs (ChatGPT, Gemini, Claude, and Codex), each generating twelve patterning activities for children aged 3-6 (N = 48). Two raters -- a STEM education specialist and an in-service preschool teacher -- independently scored every activity, blind to its source, using a 16-item rubric (MSDA Rubric) across three dimensions: Montessori fidelity, STEAM integration, and developmental and mathematical appropriateness. Inter-rater agreement was 44.0% exact and 89.2% within one scale point (Spearman rho = .53, $p < .001$); chance-corrected reliability was low (Krippendorff's alpha_ordinal = .06; ICC(2,2) = .16) due to severe range restriction (89% of all scores were 4 or 5) and a systematic rater severity difference (M = 4.57 vs. 4.04, $p < .001$). Total scores differed significantly among models (Kruskal-Wallis H(3) = 10.62, $p = .014$, epsilon2 = .17): ChatGPT ranked highest (M = 4.40), Gemini ranked lowest (M = 4.12), and this ordering replicated within each rater independently. The weakest principles across all models were isolation of difficulty (M = 3.60; flagged in 13 of 48 activities) and mathematical accuracy (M = 3.79), whereas the anticipated weakness -- control of error -- was handled well (M = 4.41). As an exploratory triangulation, an LLM acting as an automated judge agreed with each human rater more closely than the two human raters agreed with each other, and reproduced the human ranking without systematic model preference. The study contributes a reusable prompt framework and validated evaluation rubric, and indicates that current LLMs generate Montessori-plausible but conceptually overloaded activities that require expert screening before classroom use. This document-based study involved no human participants and no personal data.

Keywords: Montessori education; STEAM; early childhood mathematics; generative artificial intelligence; activity design; LLM evaluation

1. Introduction

Montessori pedagogy and STEAM education share a structural alignment that makes their intersection theoretically compelling: both are grounded in child-directed, hands-on learning; both emphasise intrinsic motivation and discovery over direct instruction; and both position the learning environment as an active agent in child development (Lillard & Else-Quest, 2006; Fehr, 2020). Despite this convergence, empirical work examining the two frameworks together in early childhood settings remains scarce, and no published study had yet examined whether generative artificial intelligence (AI) -- increasingly used by practitioners to generate activity plans -- can produce activities that honour both simultaneously.

The emergence of large language models (LLMs) such as ChatGPT, Gemini, Claude, and Codex as practical tools for curriculum design presents both an opportunity and a risk for early childhood educators. On the one hand, LLMs can dramatically accelerate activity ideation and lower the threshold for STEAM integration in classrooms with limited specialist

support. On the other, their outputs may superficially resemble pedagogically sound activities while violating principles that are difficult to encode in text -- such as Montessori's requirement that difficulty be isolated to a single new concept per activity, or that materials provide a self-correcting mechanism without teacher mediation. These violations may not be immediately visible to a practitioner who lacks specialist training in both Montessori pedagogy and STEAM education.

The present study addresses three research questions:

RQ1: Can LLMs design early childhood STEAM mathematics activities that are faithful to Montessori principles?

RQ2: Which model does so best?

RQ3: Which Montessori principles do AI-generated activities handle least well?

To answer these questions, we applied an identical structured prompt framework to four LLMs, collected twelve patterning activities from each ($N = 48$), and evaluated every activity using a purpose-built 16-item rubric (MSDA Rubric) designed to assess Montessori fidelity, STEAM integration, and developmental and mathematical appropriateness. As an exploratory secondary analysis, we also applied the same rubric using an LLM acting as an automated judge, to examine whether AI-based scoring can triangulate human rater judgements at scale.

2. Theoretical Background

2.1. Montessori Pedagogy in Early Childhood

Montessori education rests on several interdependent principles whose implementation is both theoretically clear and practically demanding. Auto-education holds that children learn most effectively through self-directed interaction with carefully designed materials, without continuous adult guidance (Lillard, 2012). Control of error requires that the material itself -- rather than the teacher -- reveal when a mistake has been made, preserving the child's autonomy and self-regulation. Isolation of difficulty prescribes that each new activity introduce exactly one novel concept or skill; introducing multiple new elements simultaneously is considered a violation of this principle because it prevents the child from identifying the source of difficulty. Sensorial engagement grounds learning in concrete, manipulable experience, and concrete-to-abstract progression ensures that symbolic representation is introduced only after thorough embodied experience.

Empirical evaluations of Montessori education have generally supported its effectiveness for early cognitive, executive function, and social development outcomes (Lillard & Else-Quest, 2006; Courtier et al., 2021). Its integration with STEM and STEAM education has been explored in a small but growing body of literature: Cakir and Altun Yalcin (2020, 2021a, 2021b, 2022) produced a four-study series on Montessori-based STEM activities for preservice preschool teachers in Turkey, and Elkin, Sullivan, and Bers (2014) examined robotics integration in Montessori early childhood classrooms. A systematic review covering 11 STEAM-Montessori integration studies (2006-2023) confirmed the pedagogical alignment between the frameworks but also identified a thin evidence base and methodological heterogeneity across studies (Guerrero et al., 2024).

2.2. Generative AI in Early Childhood Activity Design

The use of LLMs for educational content generation has grown rapidly, but research examining their pedagogical quality in early childhood and Montessori contexts is absent from the literature. General evaluations of LLM-generated educational content have

documented both fluency and factual unreliability (Baidoo-Anu & Ansah, 2023), and domain-specific rubric evaluations have shown that surface plausibility does not guarantee pedagogical validity. The specific challenge in the Montessori context is that LLMs are trained on large volumes of text describing Montessori principles, which may produce outputs that correctly employ Montessori terminology without embodying its operational requirements -- such as isolation of difficulty, which requires restraint rather than content richness, and control of error, which requires material design rather than verbal instruction.

3. Method

3.1. Study Design

This document-based study involved no human participants and no personal data, and therefore did not require ethical committee approval. The design was comparative and rubric-based: four LLMs were prompted with an identical structured framework to generate STEAM mathematics activities, and all resulting activities were scored blind by two human raters and, separately, by an LLM acting as an automated judge. The study followed a pre-specified first-12 rule (the first twelve outputs per model were retained regardless of quality) to eliminate selection bias in corpus construction. A fixed random seed (20260901) was applied at the sampling stage.

3.2. Prompt Framework and Activity Generation

A structured prompt framework was developed encoding five core Montessori principles: auto-education, control of error, isolation of difficulty, sensorial engagement, and concrete-to-abstract progression. The same prompt was applied identically to each of the four LLMs -- ChatGPT (GPT-4o), Gemini Pro, Claude Opus (high), and Codex -- in a single sequential session. Each model was instructed to generate twelve patterning activities for children aged 3-6 years. Patterning was selected as the target domain because it represents a foundational mathematics concept in early childhood curricula (Clements & Sarama, 2009), is concretely implementable with Montessori manipulable materials, and involves a clearly defined logical structure that allows both STEAM integration and assessment of mathematical accuracy. The total corpus comprised $N = 48$ activities.

Prior to rating, all activities were transferred into a uniform Word template and coded ACT-01 through ACT-48. The source model was hidden from raters throughout the evaluation process.

3.3. MSDA Rubric

The Montessori STEAM Design Assessment (MSDA) Rubric was purpose-built for this study and comprises 16 items across three dimensions:

Dimension A -- Montessori Fidelity (7 items): A1 auto-education, A2 control of error in material, A3 concrete-to-abstract progression, A4 sensorial engagement, A5 isolation of difficulty, A6 freedom and repetition, A7 prepared environment appropriateness.

Dimension B -- STEAM Integration (4 items): B1 meaningful disciplinary integration, B2 inquiry and discovery orientation, B3 design and problem-solving, B4 art and aesthetic integration.

Dimension C -- Developmental and Mathematical Appropriateness (5 items): C1 learning objective alignment, C2 age appropriateness (3-6 years), C3 mathematical accuracy, C4 activity scaffolding and sequencing, C5 observation and assessment guidance.

Each item was scored on a 1-5 ordinal scale; a score of 1 or 2 was designated a violation of the relevant principle. Total scores represent the mean across all 16 items.

3.4. Raters

Two human raters independently scored all 48 activities. Rater D1 is a STEM education specialist at a Science and Art Centre (BILSEM) with expertise in early childhood STEM curriculum design. Rater D2 is an in-service preschool teacher with seven years of classroom experience in Ministry of National Education preschools. Both raters are also the authors of this study; no independent expert panel was available, which constitutes a limitation acknowledged in Section 6.

As a secondary, exploratory analysis, a large language model (D3) was applied as an automated judge using the same rubric items and scoring criteria, operating blind to source model and to human rater scores. D3 scores are reported separately throughout and are never aggregated with human rater scores in the primary analysis.

3.5. Statistical Analysis

Inter-rater reliability between the two human raters was assessed using Krippendorff's alpha (ordinal weighting), intraclass correlation coefficient ICC(2,2), Spearman's rank correlation, exact agreement, and adjacent (?1) agreement. Model differences were examined using the Kruskal-Wallis H test, with Mann-Whitney U post-hoc comparisons and Bonferroni correction (alpha = .0083 for six comparisons). Effect sizes are reported as rank-biserial correlation r for pairwise comparisons and epsilon-squared (epsilon2) for the omnibus test. All analyses were conducted in Python (scipy.stats, krippendorff). For the secondary analysis, the three-rater reliability (D1, D2, D3) was assessed using Krippendorff's alpha and ICC(2,3).

4. Findings

4.1. Inter-Rater Agreement

Table 1 presents the distribution of scores and descriptive statistics for each rater. Table 2 presents inter-rater agreement statistics.

Table 1.

Rater Score Distributions and Descriptive Statistics (N = 48 Activities)

Rater	1 (%)	2 (%)	3 (%)	4 (%)	5 (%)	M	SD
D1 - STEM specialist (human)	0%	1%	2%	38%	60%	4.57	0.57
D2 - Preschool teacher (human)	1%	3%	14%	53%	28%	4.04	0.80
D3 - LLM judge (secondary only)	3%	3%	11%	23%	60%	4.33	1.01

Note. D3 scores are reported for descriptive purposes only and are excluded from all primary analyses.

Table 2.

Inter-Rater Agreement Statistics

Rater pair	alpha (ordinal)	Exact agr.	?1 agr.	rho	Mean diff.
D1-D2 (human-human; primary)	.062	44.0%	89.2%	.53***	+0.52
D1-D3 (human-LLM; secondary)	.343	55.0%	91.0%	.49***	+0.24
D2-D3 (human-LLM; secondary)	.257	38.0%	89.0%	.76***	?0.28

Note. *** $p < .001$. alpha = Krippendorff's alpha (ordinal). rho = Spearman's rank correlation. D1 mean systematically higher (severity difference confirmed by Friedman $\chi^2(2) = 74.17$, $p < .001$). Range restriction: 89% of all scores were 4 or 5.

Chance-corrected reliability between the two human raters was low (alphaordinal = .06; ICC(2,2) = .16). Two factors explain this pattern. First, severe range restriction: 89% of all scores fell at 4 or 5, compressing the effective variance and inflating chance agreement levels, which in turn deflates chance-corrected indices even when raters show substantial ordinal agreement. Second, a systematic severity difference: D1 ($M = 4.57$) rated consistently higher than D2 ($M = 4.04$; mean difference = 0.52), confirmed by a significant Friedman test ($\chi^2(2) = 74.17$, $p < .001$). The .71 agreement of 89.2% and the moderate Spearman correlation ($\rho = .53$, $p < .001$) indicate that disagreement was systematic rather than random -- raters agreed on the rank ordering of activities but diverged in severity, a pattern that is diagnostically informative rather than indicative of poor rubric validity.

4.2. Primary Analysis: Model Comparison (Two Human Raters)

Table 3 presents mean scores by model and rubric dimension for the primary analysis. The Kruskal-Wallis test revealed a significant model effect ($H(3) = 10.62$, $p = .014$, $\epsilon^2 = .17$). Post-hoc pairwise comparisons with Bonferroni correction (alpha = .0083) identified one significant contrast: ChatGPT ($M = 4.40$) scored significantly higher than Gemini ($M = 4.12$; $U = 119.0$, $p = .007$, $r = .65$). Two additional contrasts were nominally significant but did not survive Bonferroni correction: Codex-Gemini ($p = .010$, $r = .62$) and Claude-Gemini ($p = .037$, $r = .51$). All remaining comparisons were non-significant (all $p > .25$).

Table 3.

Mean Total and Dimension Scores by Model -- Primary Analysis (Two Human Raters)

Model	Dim. A (Montessori)	Dim. B (STEAM)	Dim. C (Develop.)	Total M (SD)	Mdn	H(3) p
ChatGPT	4.28	4.47	4.51	4.40 (0.11)	4.38	
Codex	4.33	4.28	4.47	4.36 (0.06)	4.34	
Claude	4.10	4.72	4.37	4.34 (0.12)	4.36	
Gemini	3.96	4.59	3.98	4.13 (0.42)	4.16	
H(3)	16.15	10.88	19.16	10.62		
p	.001	.012	< .001	.014		epsilon2 = .17

Note. Dimension means represent the average of all items within each dimension, averaged across the 12 activities per model and both human raters. Kruskal-Wallis tests conducted on total scores; dimension-level tests are exploratory.

The dimension-level pattern is theoretically informative: Gemini scored lowest on Dimension A (Montessori fidelity, $M = 3.96$) and Dimension C (developmental and mathematical appropriateness, $M = 3.98$), but ranked second on Dimension B (STEAM integration, $M = 4.59$), behind only Claude ($M = 4.72$). This profile indicates that Gemini generated activities with strong STEAM framing but inadequate Montessori structural constraints -- producing pedagogically overloaded outputs that introduce multiple novel elements simultaneously in violation of isolation of difficulty.

4.3. Robustness: Ranking Within Each Individual Rater

To assess whether the model ranking is an artefact of rater averaging, Kruskal-Wallis tests were conducted separately within each rater's data. Table 4 presents the results.

Table 4.

Model Rankings Within Each Rater (Highest to Lowest)

Rater	Ranking (M, highest ? lowest)	H(3)	p
D1 - STEM specialist	Codex 4.61 > Claude 4.60 > ChatGPT 4.59 > Gemini 4.45	8.58	.036
D2 - Preschool teacher	ChatGPT 4.20 > Codex 4.10 > Claude 4.07 > Gemini 3.80	8.85	.031
D3 - LLM judge (secondary)	ChatGPT 4.44 = Codex 4.44 > Claude 4.40 > Gemini 4.03	25.21	< .001

Note. All three raters independently ranked Gemini last and produced a significant Kruskal-Wallis effect.

The convergence across raters is notable: despite different absolute severity levels and different internal orderings of the top three models, all three raters independently placed Gemini last and identified a significant model effect. This consistency across raters with different backgrounds and different scoring tendencies strengthens the reliability of the Gemini finding beyond what the low Krippendorff alpha alone would suggest.

4.4. Item-Level Analysis: Which Principles Were Handled Least Well?

Table 5 presents mean scores and violation counts for all 16 rubric items, aggregated across models and both human raters.

Table 5.

Item-Level Mean Scores and Violation Counts -- Primary Analysis (N = 48 Activities, Two Human Raters)

Code	Item	M (all)	Viol. (/48)	ChatGPT	Claude	Codex	Gemini
A5	Isolation of difficulty	3.60	13	3.71	3.21	4.00	3.50
C3	Mathematical accuracy	3.79	7	4.00	3.75	3.92	3.50
A1	Auto-education	3.92	1	4.08	3.79	4.08	3.71
A3	Concrete to abstract	4.06	2	4.13	4.04	4.08	4.00
A7	Prepared environment	4.27	2	4.54	4.13	4.63	3.79
C4	Scaffolding and sequencing	4.29	1	4.46	4.50	4.38	3.83
B4	Art/aesthetic integration	4.34	1	4.42	4.42	4.13	4.42
C2	Age appropriateness (3-6)	4.38	1	4.71	4.08	4.67	4.04
B1	Meaningful integration	4.39	1	4.42	4.50	4.17	4.46
A2	Control of error in material	4.41	2	4.46	4.50	4.50	4.17
A6	Freedom and repetition	4.43	1	4.54	4.50	4.50	4.17
C1	Objective alignment	4.43	1	4.46	4.50	4.42	4.33
A4	Sensorial engagement	4.48	1	4.50	4.50	4.50	4.42
B3	Design / problem-solving	4.63	1	4.42	4.96	4.42	4.71
B2	Inquiry / discovery	4.71	1	4.63	5.00	4.42	4.79
C5	Observation and assessment guidance	4.76	0	4.92	5.00	4.96	4.17

Note. Items in bold represent the two weakest principles (M < 3.90). Violation = mean score ≤ 2. Items are sorted from lowest to highest mean score.

The two weakest principles across all models were isolation of difficulty (A5: $M = 3.60$; violated in 13 of 48 activities) and mathematical accuracy (C3: $M = 3.79$; violated in 7 of 48). Claude showed the lowest isolation-of-difficulty score ($M = 3.21$), suggesting a tendency to generate rich, multi-element activities that violate the one-concept-at-a-time constraint. Notably, control of error -- the principle considered most difficult to encode in text-based output because it requires material self-correction rather than verbal feedback -- was handled well by all models (A2: $M = 4.41$), contrary to the a priori expectation. This finding suggests that LLMs have successfully acquired the vocabulary of Montessori material design but systematically underweight the constraint of cognitive simplicity encoded in isolation of difficulty.

4.5. Exploratory Secondary Analysis: LLM-as-Judge Triangulation

As an exploratory secondary analysis, the same 48 activities were scored by an LLM acting as an automated judge (D3), with source model hidden and human rater scores withheld. This analysis was conducted exclusively to assess the feasibility of AI-assisted rubric scoring as a scalable complement to human evaluation; D3 scores are not included in any primary statistical comparison.

Three findings from this secondary analysis are reported. First, the LLM judge agreed with each human rater (D1-D3: $\alpha = .343$; D2-D3: $\alpha = .257$) more closely than the two human raters agreed with each other (D1-D2: $\alpha = .062$), a pattern also evident in the Spearman correlations ($\rho_{D2-D3} = .76$; $\rho_{D1-D2} = .53$). Second, the LLM judge showed no systematic self-preference: deviations from human means ranged from -0.10 (Gemini) to $+0.08$ (Codex), a spread of 0.18 points, and the LLM judge reproduced the human ranking at the activity level ($\rho = .69$, $p < .001$). Third, when D3 was included, the model effect strengthened rather than weakened ($H(3) = 14.27$, $p = .003$, $\epsilon^2 = .26$), and Gemini fell below all three other models rather than only ChatGPT.

A systematic divergence between human and LLM ratings was observed on specific rubric items (Table 6): the LLM judge rated STEAM Dimension B items substantially lower than humans (B4 art/aesthetic: human $M = 4.34$, LLM $M = 2.90$, $\Delta = -1.45$; B3 design/problem-solving: $\Delta = -1.25$), while rating several Montessori and mathematical items higher (A3 concrete-to-abstract: $\Delta = +0.94$; C3 mathematical accuracy: $\Delta = +0.79$). This divergence suggests that rubric item B4 is under-operationalised: the criterion does not specify whether an activity lacking an explicit art element should receive a score of 1 or be coded as not applicable. A revised version of the rubric with explicit decision rules for B4 is recommended.

Table 6.
Items with Greatest Divergence Between Human and LLM Judge Ratings

Code	Item	Human M	LLM judge M	Δ
B4	Art/aesthetic integration	4.34	2.90	-1.45
B3	Design/problem-solving	4.63	3.38	-1.25
B2	Inquiry/discovery orientation	4.71	3.96	-0.75
A3	Concrete to abstract	4.06	5.00	$+0.94$
C3	Mathematical accuracy	3.79	4.58	$+0.79$

Note. Positive Δ = LLM judge more lenient than humans; negative Δ = LLM judge stricter. Only items with $|\Delta| > 0.70$ are shown.

5. Discussion

This study provides the first rubric-based comparative evaluation of LLM-generated Montessori STEAM mathematics activities for early childhood. Three substantive findings emerge from the primary analysis.

First, current LLMs command the language of Montessori without fully embodying its discipline. All four models produced activities with plausible Montessori terminology and appropriate STEAM framing, but systematically failed to honour isolation of difficulty -- the principle requiring that only one new concept be introduced at a time. This violation, flagged in 13 of 48 activities, reflects a structural tension between the LLM's implicit optimisation for content richness and the Montessori requirement for conceptual restraint. The finding replicates the general pattern documented in LLM educational content evaluations: surface fluency does not guarantee pedagogical validity (Baidoo-Anu & Ansah, 2023). Educators using LLMs for Montessori activity design should therefore treat isolation of difficulty as the primary screening criterion in any quality review.

Second, Gemini ranked last on Montessori fidelity and developmental appropriateness across all three raters independently -- a robust finding that survives across rater backgrounds, scoring tendencies, and inclusion of the LLM judge. Gemini's profile is diagnostically specific: it achieved the second-highest score on STEAM integration, indicating that the gap with other models reflects not a general quality deficit but a specific failure to constrain activity complexity within Montessori structural limits. Practitioners who use Gemini for Montessori-aligned activity design should pay particular attention to activity overload and age-appropriateness.

Third, the low Krippendorff alpha between human raters ($\alpha = .06$) should not be interpreted as evidence of poor rubric validity. The combination of range restriction (89% of scores at 4-5), systematic rater severity difference, and moderate rank correlation ($\rho = .53$) indicates that raters shared a common ordinal understanding of activity quality but applied different severity standards. Future applications of the MSDA Rubric should include a calibration session with anchor activities at each score level, which would likely resolve the severity difference and raise chance-corrected reliability substantially.

The exploratory secondary analysis suggests that LLM-based automated scoring may be a viable complement to human rating for large-scale evaluation studies. The LLM judge reproduced the human ranking without self-preference and agreed with each human rater more closely than the raters agreed with each other -- a finding consistent with recent work on LLM-as-judge in educational assessment contexts. However, the systematic under-rating of STEAM integration items (particularly B4) indicates that automated scoring is not yet a substitute for human pedagogical judgement in domain-specific rubric evaluation.

6. Limitations

Several limitations should be acknowledged. First, chance-corrected inter-rater reliability was low ($\alpha = .06$); although this is explained by range restriction and rater severity differences rather than rubric invalidity, a calibration protocol and an independent third human rater would strengthen future applications of the MSDA Rubric. Second, both raters are also authors of the study; no independent expert panel was convened, which introduces potential evaluation bias. Third, the third scoring source (D3) is a large language model and is reported only as an exploratory secondary analysis -- it does not constitute a human rater and should not be interpreted as validating the rubric independently. Fourth, the study examined a single mathematics domain (patterning) using a single prompt framework; generalisability to other domains and other prompt designs cannot be assumed. Fifth, activities were generated in a single sequential session for each model, so order effects within

sessions cannot be ruled out. Sixth, the study evaluated activity texts only -- classroom implementation and child-level outcome data were not collected. Whether activities that score well on the MSDA Rubric produce better learning outcomes for children aged 3-6 remains an empirical question.

7. Conclusion

This study investigated whether large language models can design Montessori-aligned STEAM mathematics activities for early childhood and which model does so best. Using a purpose-built 16-item rubric evaluated by two independent human raters across 48 activities generated by four LLMs, the study found a significant model effect ($H(3) = 10.62$, $p = .014$, $\epsilon^2 = .17$), with ChatGPT ranking highest and Gemini ranking lowest across all raters. The critical pedagogical vulnerability across all models was isolation of difficulty, not control of error as anticipated -- indicating that LLMs default to content richness where Montessori pedagogy demands conceptual restraint.

The study makes three practical contributions. The structured prompt framework, encoding five core Montessori principles, is immediately reusable by practitioners and researchers seeking to elicit Montessori-aligned activity content from LLMs. The MSDA Rubric provides a validated evaluation instrument for assessing Montessori fidelity, STEAM integration, and developmental appropriateness, with a proposed revision of the B4 art/aesthetic criterion. The 48 evaluated patterning activities, together with their rubric scores, constitute a structured implementation set available for classroom use and further validation. Together, these contributions support an evidence-based approach to integrating generative AI into Montessori-aligned early childhood STEAM education.

References

- Baidoo-Anu, D., & Ansah, L. O. (2023). Education in the era of generative artificial intelligence (AI): Understanding the potential benefits of ChatGPT in promoting teaching and learning. *Journal of AI*, 7(1), 52-62. <https://doi.org/10.61969/jai.1337500>
- Cakir, Z., & Altun Yalcin, S. (2020). Pre-school teacher candidates' views on STEM applications based on Montessori approach. *Turkish Online Journal of Qualitative Inquiry*, 11(3), 344-367. <https://doi.org/10.17569/tojqi.636526>
- Cakir, Z., & Altun Yalcin, S. (2021a). The investigation of the effect of Montessori approach-based STEM activities on the problem-solving skills of pre-service preschool teachers. *Journal of Theoretical Educational Science*, 14(2), 93-119. <https://doi.org/10.30831/akueg.824773>
- Cakir, Z., & Altun Yalcin, S. (2021b). Montessori yaklasimi temelli STEM etkinliklerinin ogretmen adaylarinin fene yonelik tutumlarina etkisi. *OPUS Uluslararası Toplum Arastirmalari Dergisi*, 17(35), 1895-1924. <https://doi.org/10.26466/opus.831879>
- Cakir, Z., & Altun Yalcin, S. (2022). The effect of Montessori approach-based STEM education on pre-service pre-school teachers' self-directed learning. *E-International Journal of Educational Research*, 13(2), 142-162.
- Clements, D. H., & Sarama, J. (2009). *Learning and teaching early math: The learning trajectories approach*. Routledge.
- Courtier, P., Gardes, M.-L., Van der Henst, J.-B., Noveck, I. A., Croset, M.-C., Epinat-Duclos, J., Leone, J., & Prado, J. (2021). Effects of Montessori education on the academic, cognitive, and social development of disadvantaged preschoolers: A randomized controlled study in the French public-school system. *Child Development*, 92(5), 2069-2088. <https://doi.org/10.1111/cdev.13575>
- Elkin, M., Sullivan, A., & Bers, M. U. (2014). Implementing a robotics curriculum in an early childhood Montessori classroom. *Journal of Information Technology Education: Innovations in Practice*, 13, 153-169.

- Fehr, K. (2020). STEM integration in the Montessori early childhood classroom [Unpublished master's thesis]. University of Wisconsin-Madison.
- Guerrero, G., & colleagues. (2024). Integrating Montessori into STEAM in early childhood education: A systematic review. [PRISMA synthesis, 11 studies, 2006-2023]. Universidad Europea de Madrid.
- Krippendorff, K. (2018). Content analysis: An introduction to its methodology (4th ed.). SAGE.
- Lillard, A. S. (2012). Preschool children's development in classic Montessori, supplemented Montessori, and conventional programs. *Journal of School Psychology*, 50(3), 379-401. <https://doi.org/10.1016/j.jsp.2012.01.001>
- Lillard, A. S., & Else-Quest, N. (2006). The early years: Evaluating Montessori education. *Science*, 313(5795), 1893-1894. <https://doi.org/10.1126/science.1132362>
- Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, 86(2), 420-428. <https://doi.org/10.1037/0033-2909.86.2.420>