ProductiveMath: A Generative-AI-Powered App to Support Productive Failure Teaching

Seyedahmad Rhaimi(✉), Deniz Ercan, Ran Gao, Salah Esmaeiligoujar, Maryam Babaee, Hongming Li, Shan Zhang, Seiyon Lee, Avery Closser, and Anthony Botelho

School of Teaching and Learning, University of Florida, Gainesville, USA
{srahimi,deniz.ercan,gaoran,s.esmaeiligoujar,maryambabaee,hli3, zhangshan,leeseiyon,avery.closser}@ufl.edu, abotelho@coe.ufl.edu

Abstract. Productive Failure (PF) engages students in problem-solving before instruction but designing effective PF problems is challenging. To address this, we developed *ProductiveMath*, an AI-powered tool to support teachers in generating PF problems. Across three studies, we explored five research questions related to problem quality, AI-generation, assessment accuracy, and teacher perceptions. In Study 1, we conducted a literature review to define high-quality PF problems, created a rubric, and used GPT-4o to generate and evaluate 30 problems alongside human raters. Study 2 replicated this process with 60 additional problems, showing strong correlations between AI and human ratings, confirming GPT-4o's capability to produce high-quality problems. In Study 3, seven math teachers evaluated human- and AI-generated problems through surveys and interviews. They rated AI-generated algebra problems as high-quality, reported positive perceptions of *ProductiveMath*'s usability, and expressed intentions to use it. Key teacher feedback included suggestions to adjust problem difficulty, simplify text, enhance visuals, and provide additional PF support.

Keywords: Generative AI · Productive Failure · Algebra · K-12 · ProductiveMath

1 Introduction

In 2023, the average math score for U.S. 13-year-olds dropped 9 points from 2020, one of the largest declines since the 1970s [1]. In 2024, only 28% of eighth graders reached math proficiency, with low-performing students scoring lower than in 2022 [1]. This decline highlights gaps in foundational math skills, especially in algebra, which is crucial for academic and career success [2]. Algebra develops abstract reasoning, variable manipulation, and problem-solving, but algebra's abstract nature poses challenges [3]. Traditional instruction often disconnects algebra from daily life, limiting engagement and learning. Contextualizing algebra with real-world applications enhances its relevance and appeal [4]. Therefore, developing innovative instructional methods is essential to strengthen students' reasoning skills.

A. I. Cristea et al. (Eds.): AIED 2025, CCIS 2591, pp. 344–351, 2025.
https://doi.org/10.1007/978-3-031-99264-3_43

The Productive Failure (PF) teaching method [4] offers a promising solution to these challenges. Failure is often perceived negatively in educational settings, yet it is a necessary component of the learning process [4]. This pedagogical strategy involves guiding students through a phase of planned failure before providing direct instructions which promotes active engagement, collaborative learning, and higher-order thinking skills [5]. Algebra is well-suited for PF due to its complexity and problem-solving opportunities [6, 7].

The PF process consists of two main phases during classroom activities: (1) Response Generation & Exploration, where students engage actively with challenging problems that typically lead to initial failures, promoting problem-solving skills and resilience; and (2) Consolidation & Knowledge Assembly, where teachers guide students toward correct solutions, reinforcing understanding. Additionally, two less-discussed phases occur before and after classroom activities: Preparation, involving teachers creating or selecting algebra problems and forming student groups, and Reflection, in which students reflect on the canonical solutions developed collaboratively with peers and teachers. Rooted in constructivist theories [8], PF posits that engaging with complex tasks before instruction helps learners restructure knowledge and build new schemas [5]. Problem-solving first connects learning to prior experiences [9], normalizes productive struggle, and fosters perseverance [10]. PF typically outperforms traditional instruction, promoting resilience, creativity, deeper learning, and knowledge transfer [7–11].

The success of PF relies on high-quality materials and effective implementation. Despite potential benefits, PF remains underused by teachers, partly because designing suitable PF problems independently is demanding and lacks extensive repositories. Effective PF problems must activate students' prior knowledge, engage them through scenarios or stories, and target a "sweet spot" of difficulty—challenging yet attainable [5]. To address this, we developed *ProductiveMath*, a generative-AI-powered application leveraging Large Language Models (LLMs) to help teachers generate PF problems. Grounded in human-centered AI principles [12, 13], *ProductiveMath* allows teachers to review, refine, and adapt AI-generated problems, supporting teacher agency and teacher-AI collaboration [14, 15]. This partnership reduces teachers' workload while maintaining instructional quality and leveraging their pedagogical expertise [16]. LLMs show strong potential in generating and evaluating educational content, particularly math word problems [17]. By adapting difficulty and context, they enable personalized instruction—crucial for PF approaches [18]. While earlier work focused on simpler problems [19], AI-generated content often requires human refinement [20]. Advanced uses include generating reading items [21], Bloom-aligned physics questions [22], creativity assessments [23], and PIRLS-style passages [24]. The three studies here address the gap in producing complex, PF-aligned math problems.

1.1 *ProductiveMath* Design Based on HCI and UX Principles

Through an iterative, rapid prototyping design process, and based on research in human-computer interaction (HCI) and user experience (UX) design (e.g. [25, 26]), the initial design of *ProductiveMath* was developed using Figma (see Fig. 2). This prototype allowed for early feedback from teachers before further development. *ProductiveMath* allows teachers to create classes and organize students into groups within the "My

Classes" page. The application was designed to gather the critical information needed for generating a PF problem in a step-by-step manner to ensure that teachers could focus on each step carefully before proceeding to the next.

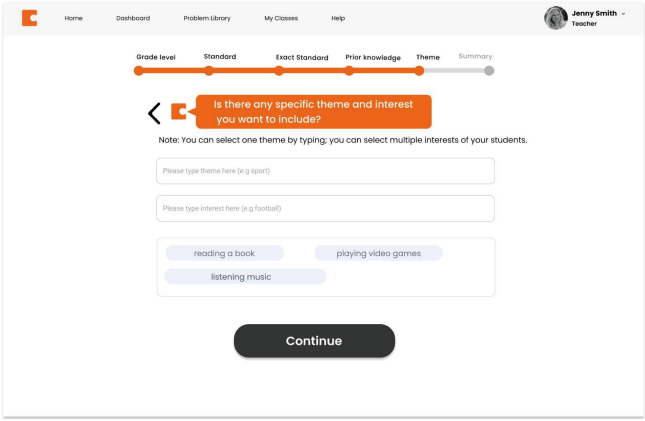

Fig. 2. *ProductiveMath*: Teacher input for AI-generated problems.

2 Study 1: Characterizing Quality for Productive Failure Content

In Study 1, we conducted a systematic literature review of algebra-related PF studies by searching databases (e.g., Google Scholar, ERIC ProQuest, Web of Science) and journals (e.g., STEM Education, Review of Educational Research). From existing PF reviews [7–11], we collected human-generated algebra problems ($n = 44$). Using seminal PF literature [10, 11], we developed a comprehensive rubric (Table 1) to evaluate problem quality, which served to train GPT-4o to generate algebra PF problems. In the second phase of Study 1, we compared zero-shot and few-shot prompt engineering approaches with GPT-4o to generate algebra PF problems. The zero-shot method provided PF guidelines without examples, whereas the few-shot method included exemplar human-generated problems from Phase 1. Iterative testing revealed that detailed contexts, such as character personas, skill descriptions, and clear task guidelines, significantly improved problem quality. The few-shot method proved most effective, generating a set of refined problems ($n = 30$). Two pairs of human raters independently assessed problem quality using the rubric, while GPT-4o simultaneously performed AI-based evaluations. Human and AI ratings were then compared to determine assessment correlation.

2.1 Results of Study 1: Establishing a Rubric for PF Problems' Quality

Our rubric (Table 1), developed from the PF literature, included five criteria: (1) Prior Knowledge Activation (connecting to students' existing knowledge), (2) Sweet Spot

Calibration (challenging yet attainable), (3) Affective Engagement (emotional engagement via scenarios), (4) Multiple Representations and Solution Methods (variety in approaches), and (5) Open-endedness and Ill-structuredness (complexity allowing multiple interpretations). Each criterion was rated on a 1–3 scale, resulting in total scores ranging from 5 to 15. The rubric was validated by a prominent PF researcher.

Human evaluations showed AI-generated problems were high-quality overall (M = 13.6, SD = 1.94), scoring well on Activation of Prior Knowledge (M = 2.88, SD = 0.37), Affective Engagement (M = 2.90, SD = 0.35), Open-endedness (M = 2.40, SD = 0.72), and Multiple Representations (M = 2.60, SD = 0.67). Pearson correlation (r = 0.78) indicated strong alignment between human and GPT-4o assessments. To further evaluate robustness, we expanded the set of PF items across multiple grade levels and standards in Study 3.

Table 1. The rubric for assessing PF problems' quality.

	1	2	3
Prior knowledge activation	Problem does not include any prerequisite concepts	Problem includes only one prerequisite concept	The problem includes 2 and more than 2 prerequisite concepts
Sweet spot calibration	Too easy or too difficult	Moderate challenge, inconsistent difficulty	Well-calibrated challenge, productive engagement
Affective engagement	Not relatable or meaningful	Some relatable elements	Highly engaging, relatable, meaningful scenario
Multiple representations and solution methods	Few or no diverse methods	Some diversity in methods	Encourages wide range of methods
Open-endedness and ill-structuredness	Simple, single solution path	Moderate complexity, few variables	Highly complex, multiple variables

3 Study 2: Exploring Automated Content Generation

In Study 2, we trained GPT-4o, through the same comprehensive prompt engineering process as we did in Study 1, to generate algebra problems suitable for PF. Specifically, in Study 2, we asked GPT-4o to generate 60 problems across 5 algebra Common Core Standards: *CCSS.Math.Content.3.OA.D.9* (n = 10), and *CCSS.Math.Content.4.OA.A.3* (n = 10), *CCSS.Math.Content.5.OA.A.1* (n = 10), *CCSS.Math.Content.7.G.B.5* (n = 15), and *CCSS.Math.Content.8.F.B.5* (n = 15). Two human raters and GPT-4o assessed the quality of those problems using the rubric.

3.1 Result of Study 2: Evaluation of LLM-Generated Content

Human raters evaluated AI-generated problems highly across rubric criteria: Activation of Prior Knowledge ($M = 2.95$, $SD = 0.19$), Sweet Spot Calibration ($M = 2.67$, $SD = 0.40$), Affective Engagement ($M = 2.96$, $SD = 0.12$), Multiple Representations ($M = 2.55$, $SD = 0.49$), and Open-endedness ($M = 2.45$, $SD = 0.52$). The overall problem quality was high ($M = 13.59$, $SD = 1.28$ out of 15), indicating GPT-4o successfully generated PF-aligned problems using iterative prompt engineering. Correlation analysis between human and GPT-4o assessments of original human-generated problems revealed strong convergent validity ($r = .78$), affirming GPT-4o's alignment with human judgment. These findings supported progression to Study 3, involving feedback from seven teachers.

4 Study 3: Teachers' Perceptions of Generated Content

The purpose of Study 3 was to compare teacher perceptions of AI-generated PF math problems to human-generated and simple problems, and evaluate the *ProductiveMath* platform. Using a mixed-methods approach [27], we collected quantitative survey ratings based on our rubric (Study 1), and qualitative data through semi-structured interviews. A participatory design method [28] guided interviews, including think-aloud sessions to gather detailed feedback on *ProductiveMath*'s usability. Seven middle-school math teachers (Grade 6 = 2, Grade 7 = 3, Grade 8 = 2) participated, recruited through district partnerships.

Teachers completed a structured Qualtrics survey evaluating three AI-generated PF problems (from *ProductiveMath* using GPT-4o), one human-generated PF problem, and one simple math problem. They rated problems independently before interviews, unaware of their origins. The survey adapted our rubric from Studies 1 and 2 into five Likert-scale items (1 = Low, 5 = High): prior knowledge activation, affective engagement, sweet spot calibration, open-endedness, and multiple representations. The survey had a maximum total score of 25. We analyzed survey data using descriptive statistics and thematic analysis [29–31].

Following the survey, teachers participated in one-hour semi-structured Zoom interviews with a live demonstration of *ProductiveMath.* Interviews explored experiences with PF, perceptions of *ProductiveMath*, usability, practical classroom applications, and improvement suggestions. Interviews were conducted by one researcher with another taking notes; notes were cross-checked for accuracy [30]. Data underwent inductive thematic analysis [31], including independent coding by researcher pairs, collaborative merging, and a final consensus meeting among five researchers, resulting in a codebook and frequency analysis.

4.1 Results of Study 3: Teacher Surveys and Interviews

Survey results indicated teachers rated AI-generated problems positively, closely matching human-generated problems (AI: $M = 17.19$, $SD = 3.07$; Human: $M = 17.43$, $SD = 3.49$). Simple problems scored lowest ($M = 13.43$, $SD = 4.20$). AI-generated problems

excelled in Sweet Spot Calibration and Affective Engagement, suggesting appropriate challenge and student engagement. However, AI-generated problems scored lower in Curriculum Alignment ($M = 2.57$, $SD = 1.43$) than human-generated problems ($M = 3.29$, $SD = 0.49$), highlighting potential improvements in aligning content with educational standards.

Qualitative data provided insights into teachers' perceptions of human- and AI-generated PF math problems. Teachers noted that human-generated problems often imposed a high cognitive load and were overly complex, especially for low-achieving students. AI-generated problems, while curriculum-aligned, sometimes lacked clarity, contained vague information, were excessively long, argumentative, or required advanced literacy skills. Despite these challenges, teachers praised AI-generated problems for effectively activating prior knowledge, fostering engagement, critical thinking, and creativity through relatable contexts (Table 2).

Table 2. Teachers' evaluation of different problem types based on the rubric. *M* (*SD*).

Problem type	*n*	PKA	SSC	AE	MR	OE	Overall	CA
Simple	7	3.29 (1.25)	2.29 (1.25)	2.43 (1.40)	3.29 (1.38)	2.00 (1.15)	13.29 (4.15)	3.29 (1.25)
Human-generated	7	3.29 (1.25)	3.14 (1.68)	3.43 (1.13)	4.00 (0.82)	3.57 (1.27)	17.43 (3.55)	3.29 (0.49)
AI-generated	21	3.00 (1.14)	3.38 (1.28)	3.71 (1.31)	3.52 (0.98)	3.57 (0.98)	17.19 (2.98)	2.81 (1.33)

Note: PKA = Prior Knowledge Activation, SSC = Sweet Spot Calibration, AE = Affective Engagement, MR = Multiple Representations and Solution Methods, OE = Open-endedness, CA = Curriculum Alignment, Overall = Average of Sum of PKA, SSC, AE, MR, and OE per problem type

Teachers recommended making AI-generated problems shorter and more concise, providing clearer instructions, adding scaffolding support, and incorporating more visual elements. Interview participants suggested improvements including visual enhancements (100%), better formatting (71%), simplified text (57%), and reading-level adjustments (42%). Specifically, teachers recommended positioning visuals clearly at the top, segmenting text with spacing and bolding key points, and aligning text complexity with students' reading abilities.

All teachers had positive initial impressions of the *ProductiveMath* prototype, describing it as "clear," "intuitive," and "user-friendly." Teachers expressed willingness to use the platform, highlighting its potential for warm-ups and small group activities, ease of problem creation and customization, and alignment with best teaching practices. Teachers also offered suggestions for enhancing *ProductiveMath*, including adding a keyword-based search feature, comprehensive support packages (solutions, misconceptions, PF guidance, answer keys), a social feature for collaboration, and integrating CLEVER for seamless account access.

5 Discussion and Conclusion

Through three studies, we evaluated GPT-4o's ability to generate Productive Failure (PF) algebra problems, informing the development of ProductiveMath—an AI-powered tool for teachers. Studies 1 and 2 showed GPT-4o produced PF-aligned problems, with some clarity and curriculum alignment issues. In Study 3, teachers valued the problems for encouraging critical thinking but suggested simplifying text, clarifying instructions, and adding visuals. They also praised ProductiveMath's design for supporting content creation and teacher-AI collaboration [23]. Future work will add student-facing features and continue educator-informed development.

References

1. National Center for Education Statistics. NAEP 2023 long-term trend assessment results. Retrieved from NCES website (2023)
2. Dougherty, S.M., Goodman, J.S., Hill, D.V., Litke, E.G., Page, L.C.: Middle school math acceleration and equitable access to eighth-grade algebra: evidence from the Wake County Public School system. Educ. Eval. Policy Anal. **37**(1), 80S–101S (2015). from https://scholar.harvard.edu
3. Common Core State Standards High School: algebra. Common Core State Standards Initiative (n.d.). https://www.thecorestandards.org/Math/Content/HSA/introduction/
4. Kapur, M.: Productive failure. Cogn. Instr. **26**(3), 379–424 (2008). https://doi.org/10.1080/07370000802212669
5. Piaget, J.: The development of thought: equilibration of cognitive structures (Trans. A. Rosin). Viking (1977)
6. Schwartz, D.L., Martin, T.: Inventing to prepare for future learning: the hidden efficiency of encouraging original student production in statistics instruction. Cogn. Instr. **22**(2), 129–184 (2004). https://doi.org/10.1207/s1532690xci2202_1
7. Sinha, T., Kapur, M.: When problem solving followed by instruction works: evidence for productive failure. Rev. Educ. Res. **91**(5), 761–798 (2021). https://doi.org/10.3102/00346543211019105
8. Darabi, A., Arrington, T.L., Sayilir, E.: Learning from failure: a meta-analysis of the empirical studies. Educ. Technol. Res. Dev. **66**, 1101–1118 (2018). https://psycnet.apa.org/doi/10.1007/s11423-018-9579-9
9. Jackson, A., Godwin, A., Bartholomew, S., Mentzer, N.: Learning from failure: a systematized review. Int. J. Technol. Des. Educ. **32**(3), 1853–1873 (2022). https://link.springer.com/article/10.1007/s10798-021-09661-x
10. Kapur, M.: Examining productive failure, productive success, unproductive failure, and unproductive success in learning. Educ. Psychol. **51**(2), 289–299 (2016). https://doi.org/10.1080/00461520.2016.1155457
11. Kapur, M., Hattie, J., Grossman, I., Sinha, T.: Fail, flip, fix, and feed–rethinking flipped learning: a review of meta-analyses and a subsequent meta-analysis. Front. Educ. **7**, 956416 (2022). https://doi.org/10.3389/feduc.2022.956416
12. Holmes, W., Bialik, M., Fadel, C.: Artificial intelligence in education: promises and implications for teaching and learning. Center for Curriculum Redesign (2019)
13. Qassrawi, R., Al Karasneh, S.M.: Redefinition of human-centric skills in language education in the AI-driven era. Stud. English Lang. Educ. **12**(1), 1894–1912 (2025)

14. Zawacki-Richter, O., Marín, V.I., Bond, M., Gouverneur, F.: Systematic review of research on artificial intelligence applications in higher education: promises and challenges. Int. J. Educ. Technol. High. Educ. **16**(1), 1–27 (2019)
15. Kim, J.: Leading teachers' perspective on teacher-AI collaboration in education. Educ. Inf. Technol. **29**(7), 8693–8724 (2024)
16. Seo, K., Yoo, M., Dodson, S., Jin, S.H.: Augmented teachers: K–12 teachers' needs for artificial intelligence's complementary role in personalized learning. J. Res. Technol. Educ. 1–18 (2024)
17. Mohan, G.B., Prasanna, K., Vishal, K., Keerthinathan, A., Lavanya, G., Meghana, M.K.U.: An analysis of large language models: their impact and potential applications. Knowl. Inf. Syst. **66**, 5047–5070 (2024)
18. Fernández, A.A., López-Torres, M., Fernández, J.J., Vázquez-García, D.: ChatGPT as an instructor's assistant for generating and scoring exams. J. Chem. Educ. **101**(9), 3780–3788 (2024)
19. Sayin, A., Kıyak, Y.S., Kononowicz, A.A.: Using OpenAI GPT to generate reading comprehension items. Educ. Measur. Issues Pract. (2024)
20. Lin, Z., Chen, H.: Investigating the capability of ChatGPT for generating multiple-choice reading comprehension items. System **123**, 103344 (2024). https://doi.org/10.1016/j.system.2024.103344
21. Omopekunola, M.O., Kardanova, E.Y.: Automatic generation of physics items with large language models (LLMs). Res. Eval. Educ. **10**(2), 168–185 (2024). https://doi.org/10.21831/reid.v10i2.76864
22. Laverghetta Jr, A., Luchini, S., Linell, A., Reiter-Palmon, R., Beaty, R.: The creative psychometric item generator: a framework for item generation and validation using large language models. https://arxiv.org/pdf/2409.00202
23. Bezirhan, U., von Davier, M.: Generating Reading Assessment Passages Using a Large Language Model
24. Nielsen, J.: Enhancing the explanatory power of usability heuristics. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 152–155 (1994)
25. Norman, D.: The Design of Everyday Things: Revised and Expanded Edition. Basic Books (2013)
26. Design-Based Research Collective. Design-based research: an emerging paradigm for educational inquiry. Educ. Res. **32**(1), 5–8 (2003). https://doi.org/10.3102/0013189X032001005
27. Creswell, J.W., Creswell, J.D.: Research Design: Qualitative, Quantitative, and Mixed Methods Approaches. Sage Publications (2018)
28. Spinuzzi, C.: The methodology of participatory design. Tech. Commun. **52**(2), 163–174 (2005)
29. Thomas, D.R.: A general inductive approach for analyzing qualitative evaluation data. Am. J. Eval. **27**(2), 237–246 (2006)
30. Lincoln, Y.S., Guba, E.G.: Naturalistic Inquiry. Sage Publications (1985)
31. Braun, V., Clarke, V.: Using thematic analysis in psychology. Qual. Res. Psychol. **3**(2), 77–101 (2006)