LONDON JOURNAL OF SOCIAL SCIENCES

eISSN 2754-7671

What Readers Want: Five Gaps in AI-STEM Education Literature and a Roadmap for Multilingual Learners

Yernat Mnuar*
Yetkin Yildirim**
Gulsara Bokembay***

Abstract

Generative artificial intelligence (genAI) is entering STEM education faster than the evidence base can guide it. Public debate is polarized between narratives of transformation and warnings about inequity, algorithmic bias, and overdependence. These tensions matter most for multilingual learners (MLs), for whom genAI can support diverse linguistic practices while also reinforcing dominant language norms. This study uses critical reception analysis to ask what practitioner-readers say is missing from influential books on STEM education, mathematics and science teaching, educational equity, and AI in education. Guided by reception theory, we treat critical and mixed reader reviews of 20 books as evidence of gaps between academic discourse and classroom realities. Reviews were collected manually from publicly available platforms between over summer months 2026 and analyzed through reflexive thematic analysis (N = 138 reviews; 20 books; Cohen's kappa = 0.89). Five recurring gaps emerged: (1) concerns about evidence quality, credibility, and commercial influence; (2) limited practical guidance for educators facing real-world constraints; (3) polarization between AI promotion and resistance without balanced implementation frameworks; (4) insufficient attention to multilingual and multicultural learners; and (5) universal claims that overlook context and boundary conditions. Building on these gaps and the Subjective Intelligence framework (Perez et al., 2025), we propose a research roadmap for equitable AI-supported STEM education. We argue that effective integration requires moving beyond technological optimism or pessimism toward evidence-based, culturally responsive, and learner-centered approaches, so that AI reduces rather than reproduces inequities for multilingual learners.

Keywords: AI in education; STEM education; multilingual learners; equity; reception analysis; generative AI; culturally responsive pedagogy

https://doi.org/10.31039/ljss.2026.12.422

*Department of Curriculum and Instruction, University of Houston, Houston, Texas, USA

** Rice University, Houston, Texas, USA, yetkin@rice.edu

*** Independent Educational Researcher, Houston, Texas, USA

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

Introduction

GenAI has moved into STEM classrooms, tutoring tools, and teacher workflows within a few years, and the literature describing it has grown just as quickly. Much of that literature takes one of two stances. Optimistic accounts present AI as a means of personalizing instruction, automating feedback, and widening access (Holmes et al., 2019; Kasneci et al., 2023). Critical accounts document algorithmic bias, privacy risks, commercial capture of public education, and learner overdependence (Baker and Hawn, 2022; Selwyn, 2019; Williamson and Eynon, 2020). Both stances are well supported, yet they are rarely brought into one implementable account of what a teacher should do on Monday morning. The gap is sharpest for multilingual learners. Large language models (LLMs) perform unevenly across languages and dialects (Hofmann et al., 2024), and common AI-detection tools have been shown to misclassify the writing of non-native English speakers at rates as high as 61% (Liang et al., 2023). At the same time, multilingual learners can use AI to draw on their full linguistic repertoires, in line with translanguaging pedagogy (Garcia and Li Wei, 2014). Research on translanguaging in STEM remains concentrated in science and mathematics and says little about technology and engineering education (Cheung and Ng, 2026). The Subjective Intelligence framework offers one response, asking that genI in STEM be understood through learners' identities and linguistic resources (Perez et al., 2025). It is a position paper, however, and it has not yet been tested against how practitioners actually receive the wider literature. This study takes a different vantage point. Instead of surveying what scholars say, it asks what readers say back. Reviews of influential books are a naturally occurring record of how teachers, school leaders, and other practitioners judge academic and popular writing against the conditions of their work. Critical reviews are especially informative because they concentrate disappointment: a reader who takes the time to write a critical review is usually saying that the book failed to deliver something they needed.

Purpose and research questions. We analyze critical and mixed reviews of 20 influential books on STEM education, mathematics and science teaching, educational equity, and AI in education to identify what practitioner-readers find lacking, and we use those gaps to build a research roadmap. The study asks:

1. **RQ1:** What recurring gaps do critical readers identify in influential books on STEM education, equity, and AI in education?
2. **RQ2:**How do these gaps vary across the four book domains?
3. **RQ3:** What do the gaps imply for research on equitable, culturally responsive AI-supported STEM education for multilingual learners, when read through the Subjective Intelligence framework?

Contribution. The paper offers (a) an empirical map of five reader-identified gaps, (b) a transparent and replicable method for using public review data in education research, and (c) a prioritized roadmap that connects those gaps to multilingual learners' needs.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

BACKGROUND

AI in STEM education: promise and critique

Systematic reviews show that research on AI in education has been driven largely by computer science and technology perspectives, with limited engagement with pedagogical theory and little critical reflection on ethics and risks (Zawacki-Richter et al., 2019). This disciplinary imbalance matters for STEM education in particular, because claims about AI's benefits are often made by the same actors who build and sell the tools, and are rarely tested against what teachers actually need in classrooms. A recent scoping review of 183 publications on genAI in secondary STEM education found that ethical aspects are sidelined and that the research focuses on AI tool performance rather than human-centered application (Fock and Siller, 2025). The result is a literature that is rich in technical possibility and thin in pedagogical evidence. Critical scholars argue that AI in education carries commercial and political interests that are seldom acknowledged in the research that promotes it (Selwyn, 2019; Williamson and Eynon, 2020). The most visible AI-in-education books illustrate the problem. Salman Khan's Brave New Words (2024) is, by reviewers' own accounts, "basically just a really long ad for Khanmigo," and the tool it promotes costs $44 per student per year plus district fees, undercutting the book's claim to widen access. Bowen and Watson's Teaching with AI (2024) deploys what reviewers call the "Silicon Valley inevitability argument" - the claim that educators have no choice but to adopt AI as a substitute for evidence. One of its authors holds a position as a Vice President for Digital Innovation, a role whose job is technology adoption. These are the most widely read books in the field, and their commercial entanglements shape how practitioners understand what AI in STEM education is for. Work on bias in educational algorithms shows that learner populations are unevenly represented in the data behind AI tools and that performance can differ across demographic groups (Baker and Hawn, 2022). LLMs perform unevenly across languages and dialects, and research has documented that they generate covertly racist decisions about people based on their dialect (Hofmann et al., 2024). AI-text detectors compound the problem: they misclassify the writing of non-native English speakers at rates as high as 61%, compared with significantly lower rates for native writers (Liang et al., 2023). For multilingual learners in STEM, this means the tools that promise to support their learning are also the tools most likely to misjudge their work. The bias is built into the data and the design. Technical critiques of LLMs add concerns about scale, opacity, and the reproduction of dominant language patterns (Bender et al., 2021). These concerns are amplified in STEM education, where disciplinary language is already dense and where mathematical notation and scientific registers are not evenly represented in training corpora. A language model that performs well on English-language word problems may perform poorly on the same problems in Spanish, Arabic, or Vietnamese, and it may not recognize the validity of code-switched or dialectal STEM talk at all. The technical literature on LLMs rarely addresses these disciplinary and linguistic dimensions, which is why the critique must come from education research as well as from computer science. The practical consequence of this literature is a field that is polarized between promotion and resistance. On one side, books like Khan's and Bowen and Watson's present AI as an inevitable good and offer little evidence beyond the authors' enthusiasm. On the other, critical accounts such as Reich's Failure to Disrupt (2020) and Watters's Teaching Machines (2021) document a century of ed-tech hype cycles and warn that

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

technology alone cannot transform education, but they offer few proposals for what teachers should do instead. Our analysis of reader reviews found this polarization to be the defining feature of the AI-in-education domain: fourteen reviews across four books raised it, and readers described the literature as "propaganda," "snake oil," and "critique without proposals."

Multilingual learners and genAI in STEM

Multilingual learners bring linguistic resources that are assets for disciplinary sense-making. Translanguaging scholarship treats the use of a learner's whole repertoire as legitimate and productive, not as a deficit (Garcia and Wei, 2014). This perspective matters for STEM because disciplinary meaning-making is not confined to one language: students reason about mathematical relationships, scientific mechanisms, and engineering design in whatever linguistic resources they have available, and the strategic use of multiple languages can support rather than hinder conceptual understanding. A classroom that treats multilingualism as a resource invites students to draw on their full repertoire; a classroom that treats it as a barrier asks them to set aside the very tools that could help them think. A recent systematic review of 109 studies from 2011 to 2024 found that translanguaging research in STEM is concentrated in science and mathematics, relies mostly on naturalistic methods, and often measures outcomes that overlook disciplinary meaning-making (Cheung and Ng, 2026). The review also found that research is concentrated in the Global North and that technology and engineering education remain largely unstudied. This is a significant gap for two reasons. First, engineering and technology are precisely the STEM disciplines where design talk, collaboration, and multimodal representation are central, and where multilingual learners' resources could be most visible. Second, the absence of STEM-specific theoretical frameworks means that translanguaging research has not yet developed the conceptual tools to ask how disciplinary language and multilingual language use interact. That lens has become more urgent as AI enters STEM classrooms. The Subjective Intelligence framework extends it by asking that genAI in STEM be understood through learners' identities and linguistic resources, including gender and cultural identities (Perez et al., 2025). The framework's third tenet (multilingual and multidialectal support in design) is directly relevant to the questions this paper raises. But the framework is a position paper, and it has not yet been tested against how practitioners actually receive the literature. Our review corpus suggests that practitioners rarely raise language at all: only few reviews in 138 explicitly named English language learners.

The research-practice gap

Education research has long struggled with the distance between what is published and what teachers can use. Research-practice partnership scholarship attributes this partly to research agendas that are set without practitioners and to findings that are not designed for use in particular settings (Coburn and Penuel, 2016). The gap is in the questions asked, the settings studied, and the forms of evidence produced. Education technology has its own history of promises that did not survive contact with classrooms, and that history suggests the gap widens when new tools arrive with more enthusiasm than evidence. If this gap exists for AI-

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

STEM research, practitioners should be voicing it somewhere. Public book reviews are one place where they do.

Online reviews as evidence of reception

Reader reviews have been used to study how large groups of readers respond to works at a scale that interviews cannot reach. Holur et al. (2021) analyzed Goodreads reviews of five novels and argued that such reviews are a window onto what readers find important and where their readings diverge, because they are unguided and written for the reader's own purposes. They also note limits like noisy text, reviews written as school assignments, and self-selection. These strengths and limits carry over to our corpus and are addressed in the Methods and Limitations sections. The argument for treating reviews as evidence rests on a simple premise that readers write reviews when a text has failed to meet an expectation they brought to it. A positive review confirms that the text matched the reader's horizon; a critical review records a mismatch. This makes critical reviews unusually informative, because they name the specific expectation that was not met. Reviews, though, are not a measure of book quality but a record of unmet demand, and this is precisely what a field seeking to improve its literature needs to understand.

Gaps in the literature

Three gaps follow. First, we know little about how practitioners receive the AI-in-STEM literature when the field has studied what AI does to learn, but almost never what teachers say back about the books that tell them how to use it. Second, the Subjective Intelligence framework has not been used as a lens for synthesizing practitioner critique, which means its three tenets remain untested against the actual demands of classroom readers. Third, reception analysis has rarely been applied to education scholarship, and almost never to questions of multilingual equity, leaving the field without a method for hearing the readers whose students are most affected by AI's linguistic biases.

THEORETICAL FRAMEWORK

Reception theory

Reception theory emerged from the University of Konstanz in the 1970s as a challenge to the assumption that a text carries a single, author-determined meaning. Hans Robert Jauss argued that a text's meaning is not fixed by its author but is completed by readers, who bring their own "horizons of expectation" that shaped by prior literary experience, social norms, and the questions they bring to the work (Jauss, 1982). Wolfgang Iser added that texts contain *gaps* or *blanks* that invite the reader to fill them, making reading an active process of meaning-constitution rather than passive reception (Iser, 1978). Stanley Fish then located the reader within *interpretive communities*, arguing that shared assumptions shape what counts as a good reading (Fish, 1980). Recent work has extended this framework in two directions relevant here. First, scholars have applied reception theory to AI-generated text, arguing that when no human author exists, the reader is forced to derive meaning entirely from the text and their own interpretive lenses by making the reader's role more visible than ever. Second, research on online reviews has used reception theory to treat public reader responses as evidence of

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

how texts are actually received, rather than how scholars assume they are. For this study, the key move is to treat practitioners as an interpretive community whose horizon is set by classroom constraints such as time, class size, language diversity, assessment pressure, and limited technology. A critical rating then records a mismatch between the horizon the book assumed and the horizon the reader actually occupies.

Subjective Intelligence

Subjective Intelligence is a conceptual framework for genAI in STEM education that reads AI use through students' identities and subjectivities, including linguistic resources, gender, and cultural identities (Perez et al., 2025). The framework emerged from a position paper published in Education Sciences, drawing on the authors' first-hand experiences across an engineering undergraduate course, a graduate STEM seminar, and an engineering design task. It has three tenets: (1) cognitive and moral development toward ethical engagement with data practices; (2) identifying and examining human biases; and (3) multilingual and multidialectal support in design. Recent work has begun to cite the framework, including a study validating a scale for pre-service chemistry teachers' acceptance and use of genAI, which suggests the framework is gaining traction in STEM teacher education research. However, the framework remains a conceptual proposal rather than an empirically tested model. Its authors present it as a lens for inclusive and reflective teaching and they acknowledge that it surfaces "new tensions and possibilities" around GenAI's limitations rather than resolving them. We use the framework as an interpretive lens in the discussion and roadmap. Reviewers are forcing their words into its categories would distort them. Instead, we ask afterwards which tenets the reader-identified gaps speak to, and where the framework itself has blind spots that practitioners point out.

Reception theory supplies the empirical stance, and Subjective Intelligence supplies the normative stance (equitable AI-supported STEM requires attention to who the learner is). Together they let us move from *what readers found missing* to *what research should now provide*.

METHODOLOGY

Design

We conducted a critical reception analysis, a qualitative design in which publicly posted reader responses are treated as data about how a body of work is received. The approach treats reception as a social practice rather than an individual opinion where reviews are written for an audience, shaped by platform norms, and read by other practitioners deciding what to buy or assign. The analysis was reflexive and hybrid with five provisional gaps from an initial reading guided attention, while the codebook was open to revision and to new themes (Braun and Clarke, 2006). This hybrid logic allowed us to begin with a framework drawn from the literature while remaining responsive to what reviewers actually wrote.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

Data sources and collection

We collected reviews manually from publicly available platforms (Amazon and Goodreads) over the summer months of 2026. Manual collection was necessary because platform terms prohibit automated scraping, and because we needed to read each review to code it. Reviews were pasted into an Excel spreadsheet stored on a secure, access-controlled university cloud drive. No user names or identifying information were recorded. Because the data are publicly available and no human subjects were involved, the study did not require institutional review board approval.

Star ratings were available for every review we collected. We nonetheless treat the corpus as a purposive sample of critical and mixed reception rather than a filtered set of low-star reviews, and we chose not to sort or select by rating number for three reasons. First, low-star ratings are unreliable signals of content as reviewers routinely assign one, two, or three stars because the book arrived damaged, was late, or never arrived at all, and these ratings record a logistics failure rather than a judgment about the text. Second, some of the most substantive and detailed criticisms appeared in reviews attached to three- and four-star ratings, which described what the book promised against what it delivered while still acknowledging its strengths. Third, the complexity of the comments did not map neatly onto rating numbers where a one-star review might be a single sentence of frustration, while a three-star review might contain several paragraphs of careful critique. We therefore coded on the content of the review, and we recorded star ratings where they help interpret a review's stance.

Five reviews were written in languages other than English (Persian, Portuguese, Spanish, and Indonesian). These were translated using Google Translate. Key passages were checked against the original by the first author where possible.

Corpus construction

We selected 20 books, five in each of four domains: STEM education; mathematics and science teaching; educational equity; and AI in education. Two entries pair related titles (Inferior, 2017, and Superior, 2019; The Alignment Problem, 2020, and Unmasking AI, 2023), so the 20 entries represent 22 titles. Inclusion criteria were: (a) published between 1999 and 2024; (b) addressed to educators, school leaders, or education researchers; (c) widely read or assigned in teacher education or professional development; and (d) having publicly available reader reviews. *Influential* was operationalized as wide readership, presence on professional reading lists, and prominence in public debate. The full list is in Appendix A.

Within each book we kept items that discussed content, including critical and mixed reviews and, for balance, positive reviews that named specific strengths. We removed items about shipping, formatting, or price, and we excluded three items that were not reader reviews. The final corpus is 138 items across 20 books, of which 135 are genuine reader reviews. The number of reviews per book ranged from 1 (Book 10, The Brilliance of Black Children in Mathematics) to 26 (Make It Stick). Books 2, 3, 4, 10, and 14 are flagged as thin-evidence cases, with one or fewer genuine reviews each. Five reviews were written in languages other

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

than English (one Persian, two Portuguese, one Spanish, one Indonesian) and were translated using Google Translate.

Analysis

We followed the six phases of reflexive thematic analysis: familiarization, initial coding, generating themes, reviewing themes, defining themes, and writing up (Braun and Clarke, 2006).

1. Two authors independently read the full corpus and drafted open codes.
2. The team merged codes into a codebook with definitions and exclusion rules (Appendix B).
3. Two authors coded the full corpus. Agreement was checked with Cohen's kappa and disagreements were resolved by discussion. Kappa was 0.89.
4. Themes were built across codes, then compared by book domain.
5. Prevalence is reported as the number of reviews and books that carry each theme. Counts describe this corpus, not the population of readers.
6. Only after the themes were fixed did we read them against the three tenets of Subjective Intelligence.

Trustworthiness

We applied credibility and confirmability checks from Lincoln and Guba (1985): a coding audit trail, negative-case analysis (reviews that praise the features others criticize), peer debriefing among authors, and a sensitivity check that drops the single most reviewed book to show that no one book drives a theme. The negative cases are reported in Section 5.6.

Ethics and positionality

The data are public and used for secondary analysis, but public does not mean consequence-free. We followed the Association of Internet Researchers ethical guidelines (franzke et al., 2020). We did not use user names, we paraphrase or lightly redact quotations that could be traced by search, and we do not redistribute raw review text. The authors are STEM education researchers with experience in multilingual classrooms and an interest in equity. We acknowledge that this interest shaped what we noticed, and we used negative-case analysis to guard against it.

Findings

The corpus comprised 138 reviews of 20 books. Five gaps recurred across domains. Table 1 shows how widely each appeared.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

Table 1. Prevalence of the five gaps

Gap	Reviews (n)	Books (n of 20)	Strongest domain
1. Evidence quality, credibility, commercial influence	30	10	Mathematics education; Equity
2. Limited practical guidance under real constraints	29	11	General STEM education
3. Polarization without balanced implementation frameworks	14	4	AI in education
4. Insufficient attention to multilingual and multicultural learners	12	7	Equity
5. Universal claims that ignore context and boundary conditions	12	6	Mathematics and science education

Gap 1: Evidence quality, credibility, and commercial influence

Reviewers question whether claims rest on sound evidence, whether the author is credible, or whether the book serves a commercial or ideological interest. Mentions of anecdote, cherry-picked studies, missing citations, outdated sources, marketing, sales pitch, author or vendor conflicts of interest.

Thirty reviews across ten books raised credibility concerns. The pattern was strongest in mathematics education and equity. Mathematical Mindsets drew detailed attacks: “incorrectly-summarized research; cherry-picked anecdotal and qualitative accounts”; “rampant ignorance of basic neuroscience”; “most of the references in the book are to her own work.” One reviewer with neuroscience expertise listed specific falsehoods, including the claim that “every time a student makes a mistake in math, they grow a synapse.” Building Thinking Classrooms drew the complaint that “there is no data proving that it increases learning” and that the research is “based primarily on four papers of his own.” Hammond's Culturally Responsive Teaching and the Brain was faulted for relying on “MacLean's triune brain theory... from the 1960s” and for making “bold claims about neuroscience... but not cited.” Whistling Vivaldi was challenged on sample size: “he gave the test to only 20 students of each race.” Brave New Words was called “snake oil” and accused of “misrepresented studies, out-of-context quotes, and some studies that seem to be fake/fabricated.”

Gap 2: Limited practical guidance for educators facing real-world constraints

The book is judged too abstract or theoretical to use, or it ignores time, class size, resources, curriculum mandates, or school policy: “Not practical,” “no examples,” “too theoretical,” “can’t use this in my class,” requests for lesson-level or tool-level steps.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

Twenty-nine reviews across eleven books raised implementation concerns. This gap appeared in every domain but was strongest in general STEM education. Invent to Learn drew the most: “most of the book is fluff”; “lots and lots of quotes from luminaries... with little clear guidance on how to actually do this”; “I never felt like the author connected her views to finer aspects of actual classroom teaching.” STEM Lesson Essentials was faulted for offering “no practical way to implement STEM lessons in your classroom”; one reviewer advised, “Save your money and go to the Internet blogs.” Successful K-12 STEM Education drew the blunt verdict: “If you are a teacher, don’t bother.” Make It Stick was told to “just read chapters 1 and 8.” Mathematical Mindsets was called “completely devoid of any meaningful solution.” The pattern is consistent: practitioners want steps, examples, and adaptation guidance, not frameworks alone.

Gap 3: Polarization between promotion and resistance

The book is faulted for either uncritical enthusiasm or blanket rejection of technology, with no balanced framework for deciding when and how to implement: “Hype,” “doom,” “one-sided,” “propaganda,” “no middle ground,” absence of decision criteria or implementation stages.

Fourteen reviews across four books raised polarization concerns, all in the AI domain. Brave New Words drew eight of them: “this entire book is basically just a really long ad for Khanmigo”; “pro-AI propaganda”; “nuance-less propaganda”; “utopian googlygook”; “snake oil.” Teaching with AI was called “an ad for AI use in the classroom, not a reasoned and thoughtful guide,” and reviewers explicitly named the “Silicon Valley inevitability argument” as a marketing tactic. Failure to Disrupt drew the opposite complaint: “critique without proposals”; “it is more a review of why... it doesn't get to the revolution.” Teaching Machines was praised for its history but faulted for having “just one short chapter” on the present. The AI shelf, in short, is boosters with products to sell versus critics with no solutions.

Gap 4: Insufficient attention to multilingual and multicultural learners

Reviewers say the book neglects English learners, multilingual classrooms, cultural difference, or the situations of students outside the author's assumed context. Mentions of ELLs/ELs, bilingual, language, culture, immigrant or international students, non-US contexts, “only works for” a particular population.

Twelve reviews across seven books raised equity heterogeneity concerns, strongest in the equity domain. Hammond drew four: “White students are presented as culturally homogeneous”; “Black and Hispanic students are presented as being generally incapable of success without being specially catered to”; “Asian students... are never acknowledged”; “POC Culture is presented in a one-dimensional and idealistic way.” The Brilliance of Black Children in Mathematics was faulted for treating Black students as “more monolithic than they are.” Make It Stick was criticized for examples “written only for folks who like sports and the military” and for “gender dynamics... from the 1950s.” The most consequential review for this paper came from Rough Draft Math: a teacher of “100% English language learners in New York City” reported “some sort of disconnection and disparity between the teaching scenarios presented in the book and my own teaching reality.” Even a book about linguistic participation in mathematics failed the multilingual classroom test. Only one review

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

in the entire corpus explicitly named English language learners. This low prevalence is itself a finding. Most readers did not look for multilingual learners, which suggests that literature has not taught them to see this population. The silence is not evidence that the gap does not matter; it is evidence that the gap has been normalized.

Gap 5: Universal claims that overlook context and boundary conditions

The book generalizes from one setting, subject, or population without stating where its claims stop applying: "Doesn't apply to," "works in theory," "sample of one school," "not for [age, subject, country]," "one size fits all."

Twelve reviews across six books raised boundary-condition concerns, strongest in mathematics and science education. Building Thinking Classrooms drew four: "Based on elementary classrooms, and math and reading. Not based on science classrooms"; "not compatible with my group of students in middle school science"; "published in 2020... a lot of these methods do not work in engaging the current crop of students"; "I question the diversity of the research population." Ambitious Science Teaching drew the complaint that "the method only works when engaging phenomena can be found to anchor instruction" and that it "falls apart quickly" for dense content such as DNA replication. Mathematical Mindsets was faulted for "not enough high school applications." The Teaching Gap was faulted because "the focus is entirely on the teaching of eighth grade mathematics." Invent to Learn was faulted for offering "little practical advice about how to integrate these ideas into a Social Studies or Literature class." Reviewers punish universal claims.

Cross-cutting patterns and negative cases

Implementation (Gap 2) co-occurred with boundary conditions (Gap 5) in reviews of Building Thinking Classrooms, Ambitious Science Teaching, and Invent to Learn. Credibility (Gap 1) co-occurred with equity heterogeneity (Gap 4) in reviews of Hammond and Make It Stick. This suggests that when readers distrust the evidence, they also notice who is missing from it.

Every major book had at least one reviewer praising what others criticized. Invent to Learn was called "a nicely done practical guide" by one reader and "fluff" by another. Make It Stick was praised for "sound, research-based information" and faulted for citing "research that has been disproven." Building Thinking Classrooms drew "Fantastic book overall!" alongside "no data proving it increases learning." Brave New Words drew "every educator should read this book" alongside "snake oil." Hammond drew "worse than useless" from one reader and "pleasantly surprised" from a self-described conservative teacher who found common ground. These contrasts are not contradictions to be resolved; they are evidence that reception is community-dependent, which is exactly what reception theory predicts.

A reviewer of Building Thinking Classrooms observed that "English learners who were shy in my class... were engrossed in English-language conversation about how to solve problems" during randomized group work at vertical whiteboards. This is a concrete, classroom-level mechanism by which a pedagogy choice affects multilingual learners. It is the kind of evidence the literature needs more of.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

DISCUSSION

Practitioners ask for usable, bounded, balanced evidence

The first, second, third, and fifth gaps share a root that readers want claims they can trust, apply, and limit. Credibility (Gap 1) and implementation (Gap 2) together account for 59 of 138 coded reviews, roughly 43% of the corpus. This pattern echoes long-standing concerns that research and commercial narratives about education technology outrun classroom evidence (Selwyn, 2019; Williamson and Eynon, 2020) and that research agendas need to be shaped with practitioners (Coburn and Penuel, 2016). The polarization gap (Gap 3) is confined to AI books, which suggests that the AI literature has a distinct reception problem to read as advocacy rather than as evidence.

Multilingual learners are largely absent from the conversation

The near-invisibility of multilingual learners in the reviews is a finding about the literature, not about the reviewers. Only one review explicitly named English language learners. If practitioners do not raise language, it is because the literature has not made language visible as a dimension of STEM teaching. This matches evidence that AI tools perform unevenly across languages and dialects (Hofmann et al., 2024; Liang et al., 2023) and that translanguaging research in STEM is thin outside science and mathematics (Cheung and Ng, 2026). The silence is the gap.

Reading the gaps through Subjective Intelligence

The framework's three tenets map onto the gaps unevenly, and the mapping shows where the framework needs development.

Subjective Intelligence tenet	Gaps it speaks to	What reviewers add
Cognitive and moral development toward ethical engagement with data practices	1, 3	Readers ask for evidence standards and for criteria that make ethical use workable, not only stated
Identifying and examining human biases	1, 5	Readers point to unexamined author and vendor assumptions as well as algorithmic bias
Multilingual and multidialectal support in design	4, 5	Readers rarely raise language; the framework must make it visible to non-specialists
(Not a tenet) Practical implementation under constraints	2, 3	Readers' strongest demand has no home in the framework; add it

The framework is strong on the normative question (what equitable AI should look like) and weak on the empirical question (what practitioners actually experience). The reviewers supply

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

the empirical corrective: they want evidence, guidance, and boundaries, and they want language to be visible.

Implications for authors, publishers, and researchers

Books on AI in STEM education can respond directly to state the evidence base and conflicts of interest, show worked examples under constraint, give decision criteria in place of advocacy, name the learners the claims were tested with, and state where the claims stop. Publishers can support this by requiring evidence statements and conflict-of-interest disclosures. Researchers can support it by studying reception and by reporting by language background in every AI-in-education study.

Limitations

This study has several limitations. First, reviewers are not representative of practitioners. People who post reviews are self-selected, and some are not educators. Second, reviews can be incentivized or coordinated; our manual collection reduces but cannot remove this. Third, the corpus includes only English-language platforms, and five non-English reviews were translated using Google Translate, which may under-represent multilingual readers. Fourth, reviews are short and rarely state the reviewer's role or setting. Fifth, five books are thin-evidence cases (Books 2, 3, 4, 10, and 14), with two or fewer genuine reviews each; Book 10 contributed only one platform review and is treated as a case study of scholarly critique. Sixth, three reviews attributed to Book 1 carry uncertain attribution, and the three reviews in Book 18 cannot be confidently assigned to one of its two paired titles; both issues are flagged in Appendix A. Eighth, thematic analysis involves judgment; we addressed this with a codebook, double coding (Cohen's kappa = 0.89), and negative-case analysis, but another team might draw the boundaries differently.

A Research Roadmap for Equitable AI-Supported STEM Education

The roadmap converts each reader-identified gap into a research priority. It is a proposal for the field.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

Gap	Research priority	Example question	Suggested designs	Tenet
1. Evidence and credibility	Independent, preregistered evaluations with declared funding and vendor ties	Do AI tutors improve disciplinary reasoning for multilingual learners, measured independently of the developer?	Preregistered trials; replication; open materials	Data ethics
2. Practical guidance	Constraint-aware implementation research with teachers	What is the smallest viable AI-supported routine for a 45-minute class with 30 students and mixed English proficiency?	Design-based research; research–practice partnerships	Data ethics
3. Beyond polarization	Staged, criteria-based implementation frameworks	Under what conditions should a teacher adopt, adapt, delay, or refuse a given AI tool?	Delphi studies with teachers; decision-tool development and testing	Bias examination
4. Multilingual and multicultural learners	Translanguaging-informed design and evaluation across languages and dialects	How do AI tools respond to code-switched and dialectal STEM talk, and how do learners use them to make meaning?	Classroom discourse analysis; audits of model performance by language	Multilingual support
5. Context and boundary conditions	Reporting standards for population, setting, and limits of claims	Which findings hold across grade bands, disciplines, resource levels, and languages?	Moderator analysis; multi-site replication; context statements in all papers	Bias examination

Three cross-cutting commitments apply. Co-design with practitioners and multilingual learners, so that the problems studied are the ones classrooms have. Disaggregated reporting by language background in every AI-in-education study. Open evaluation of AI tools across languages and dialects, so that equity claims can be checked rather than asserted.

CONCLUSION

Critical readers of influential education books are, in effect, telling researchers what they need: evidence they can trust, guidance they can use, frameworks that avoid both hype and rejection, attention to the learners in front of them, and claims with stated limits. The corpus of 138 reviews across 20 books shows that credibility (30 reviews, 10 books) and implementation (29 reviews, 11 books) dominate, that polarization is confined to AI books (14 reviews, 4 books), and that multilingual learners are nearly invisible (12 reviews, 7 books, with only one review explicitly naming English language learners). For multilingual learners in STEM, the silence around language in these reviews is as important as anything said.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

The roadmap proposed here asks researchers to meet practitioners where they are, to report by language background, and to evaluate AI tools across languages and dialects. Doing so is how AI can reduce, not reproduce, inequity in STEM education. Future work should extend the method to post-2023 generative-AI books, to non-English reviews, and to direct studies with teachers of multilingual learners.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

References

Baker, R. S., and Hawn, A. (2022). Algorithmic bias in education. *International Journal of Artificial Intelligence in Education, 32*(4), 1052-1092.

Bender, E. M., Gebru, T., McMillan-Major, A., and Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? *Proceedings of FAccT '21*, 610-623.

Braun, V., and Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology, 3*(2), 77-101.

Cheung, K. C., and Ng, T. K. D. (2026). A systematic review of research on translanguaging in STEM education. *International Journal of Multilingualism, 23*(1), 28-45. https://doi.org/10.1080/14790718.2025.2512783

Coburn, C. E., and Penuel, W. R. (2016). Research-practice partnerships in education: Outcomes, dynamics, and open questions. *Educational Researcher, 45*(1), 48-54.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*(1), 37-46.

Fish, S. (1980). *Is there a text in this class? The authority of interpretive communities.* Harvard University Press.

Fock, A., and Siller, H.-S. (2025). Generative AI in secondary STEM education in the light of Human Flourishing: A scoping literature review. *International Journal of STEM Education, 12*, 1-18.

Franzke, a. s., Bechmann, A., Zimmer, M., Ess, C., and the Association of Internet Researchers. (2020). *Internet research: Ethical guidelines 3.0.* AoIR.

Garcia, O., and Wei. (2014). *Translanguaging: Language, bilingualism and education.* Palgrave Macmillan.

Gay, G. (2018). *Culturally responsive teaching: Theory, research, and practice* (3rd ed.). Teachers College Press.

Hofmann, V., Kalluri, P. R., Jurafsky, D., and King, S. (2024). AI generates covertly racist decisions about people based on their dialect. *Nature, 633*, 147-154.

Holmes, W., Bialik, M., and Fadel, C. (2019). *Artificial intelligence in education: Promises and implications for teaching and learning.* Center for Curriculum Redesign.

Holur, P., Shahsavari, S., Ebrahimzadeh, E., Tangherlini, T. R., and Roychowdhury, V. (2021). Modeling social readers: Novel tools for addressing reception from online book reviews. arXiv:2105.01150.

Iser, W. (1978). *The act of reading: A theory of aesthetic response.* Johns Hopkins University Press.

Jauss, H. R. (1982). *Toward an aesthetic of reception.* University of Minnesota Press.

Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., … Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences, 103*, 102274.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

Ladson-Billings, G. (1995). Toward a theory of culturally relevant pedagogy. *American Educational Research Journal, 32*(3), 465-491.

Liang, W., Yuksekgonul, M., Mao, Y., Wu, E., and Zou, J. (2023). GPT detectors are biased against non-native English writers. *Patterns, 4*(7), 100779.

Lincoln, Y. S., and Guba, E. G. (1985). *Naturalistic inquiry.* Sage.

Pérez, G., Henderson, T., Pierre, T., Marvez, G. R., Vasquez, A., Eshun, P., and Polanco Pino, Y. (2025). Subjective intelligence: A framework for generative AI in STEM education. *Education Sciences, 15*(12), 1571. https://doi.org/10.3390/educsci15121571

Reich, J. (2020). *Failure to disrupt: Why technology alone can't transform education.* Harvard University Press.

Selwyn, N. (2019). *Should robots replace teachers? AI and the future of education.* Polity.

Watters, A. (2021). *Teaching machines: The history of personalized learning.* MIT Press.

Williamson, B., and Eynon, R. (2020). Historical threads, missing links, and future directions in AI in education. *Learning, Media and Technology, 45*(3), 223-235.

Zawacki-Richter, O., Marín, V. I., Bond, M., and Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education: Where are the educators? *International Journal of Educational Technology in Higher Education, 16*, Article 39.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

Appendix A. Book Corpus

ID	Domain	Title and Author	Year	Reviews Analyzed	Notes
B01	STEM education	*Invent to Learn* - Martinez and Stager	2013	17	Three reviews carry uncertain attribution (B01-R12, B01-R13, B01-R14)
B02	STEM education	*STEM Lesson Essentials* - Vasquez, Sneider and Comer	2013	2	Thin evidence
B03	STEM education	*Successful K-12 STEM Education* - National Research Council	2011	2	Thin evidence; one item may be a continuation of the other
B04	STEM education	*Teaching STEM in the Early Years* - Moomaw	2013	1	Thin evidence
B05	STEM education	*Make It Stick* - Brown, Roediger and McDaniel	2014	26	One review contains a slur and is not quoted
B06	Math and science	*Mathematical Mindsets* - Boaler	2016	14	
B07	Math and science	*Building Thinking Classrooms in Mathematics* - Liljedahl	2020	12	One item is a summary of the method, not a review
B08	Math and science	*The Teaching Gap* - Stigler and Hiebert	1999	7	One Persian review; one item is a passage quoted from the book
B09	Math and science	*Ambitious Science Teaching* - Windschitl, Thompson and Braaten	2018	2	
B10	Equity	*The Brilliance of Black Children in Mathematics* - Davis and Jett (Eds.)	2014	0	Case study of scholarly critique; no platform reviews
B11	Equity	*Culturally Responsive Teaching and the Brain* - Hammond	2015	11	
B12	Equity	*Whistling Vivaldi* - Steele	2010	10	
B13	Equity	*Inferior* (2017) / *Superior* (2019) - Saini	2017/ 2019	6	Paired titles; one review contains profanity and is not quoted
B14	Equity	*Rough Draft Math* - Jansen	2020	1	Thin evidence; the review is from a teacher of 100% English language learners
B15	AI in education	*Brave New Words* - Khan	2024	14	One Portuguese review; one Indonesian review
B16	AI in education	*Teaching with AI* - Bowen and Watson	2024	2	
B17	AI in education	*Failure to Disrupt* - Reich	2020	4	One Spanish review
B18	AI in education	*The Alignment Problem* (2020) - Christian / *Unmasking AI* (2023) - Buolamwini	2020/2 023	3	Paired titles; attribution uncertain; one review contains offensive political content and is not quoted
B19	AI in education	*Distrust* - Smith	2023	2	
B20	AI in education	*Teaching Machines* - Watters	2021	2	
Total				**138 items**	**135 genuine reviews**

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License