International Journal of Assessment Tools in Education
2026, Vol. 13, No. 1, 21–44
https://doi.org/10.21449/ijate.1692300
journal homepage: https://dergipark.org.tr/en/pub/ijate *Review Article*

Methodological gaps and domain imbalances in flipped learning reviews: A critical umbrella review

Abdullah Arslan [1*]

[1]Akdeniz University, School of Foreign Languages, Antalya, Türkiye

ARTICLE HISTORY
Received: May 5, 2025
Accepted: Sep. 29, 2025

Keywords:
Flipped learning,
Umbrella review,
Systematic review,
Educational outcomes,
Research synthesis.

Abstract: This critical umbrella review (UR) examined methodological rigor and outcome assessment practices in flipped learning (FL) research syntheses. A systematic analysis of 41 reviews (2017-2023) spanning STEM, medical/health sciences, and social sciences identified significant methodological concerns and domain imbalances in reported outcomes. The analysis revealed widespread inconsistencies in primary source selection, with most reviews excluding grey literature. Inadequate quality assessment practices were evident, with over half lacking formal risk of bias evaluation. The analysis also found variable approaches to publication bias assessment and potential language bias, with most reviews limiting inclusion to English-language publications. The analysis further reveals a pronounced imbalance in outcome domains. Cognitive outcomes dominated, particularly academic achievement. The affective and interpersonal domains received substantially less attention despite their theoretical importance to FL pedagogy. The paucity of specified cognitive metrics in numerous reviews was a matter of particular concern, whilst the near absence of metric reporting for affective and interpersonal outcomes was even more concerning. In addition, the psychometric properties of instruments were rarely critically evaluated. Although reviews consistently identified active learning, self-paced instruction, and increased interaction as key mechanisms for FL effectiveness, many failed to provide substantive exploration of how these mechanisms operate within specific disciplinary contexts. This critical synthesis highlights the need for methodological improvements in FL reviews, including comprehensive literature sourcing, rigorous quality assessment, expanded domain focus beyond cognitive outcomes, and deeper mechanism exploration to strengthen the evidence base guiding educational practice.

INTRODUCTION

Educational paradigms are undergoing profound transformation in response to 21st-century demands. The complex interplay of technological advancement, globalization, and evolving workforce requirements necessitates educational approaches that transcend traditional knowledge acquisition to develop critical competencies including problem-solving, communication, collaboration, and analytical thinking (Ramma *et al*., 2021). This shift reflects growing recognition that conventional educational models, which rely on passive information transfer, prove insufficient for preparing learners to navigate complex personal, professional, and social environments, as research increasingly demonstrates their inadequacy in fostering

*Corresponding author at: Abdullah ARSLAN abdullaharslan@akdeniz.edu.tr Akdeniz University, School of Foreign Languages, Antalya, Türkiye

e-ISSN: 2148-7456

the robust, adaptable skill sets required to meet contemporary challenges (Care, 2017; Fry *et al*., 2009; Huang *et al*., 2020; Shih & Huang, 2019; Taguma *et al*., 2018). As Zhu *et al*. (2020) argue, educational effectiveness now extends beyond content mastery to encompass adaptability, critical thinking, and collaborative capabilities—competencies poorly served by passive learning modalities. This recognition has catalysed widespread interest in innovative pedagogical frameworks that promote active engagement, experiential learning, and skill development across diverse educational contexts (Bialik & Fadel, 2015; Francis *et al*., 2020).

Among these pedagogical innovations, flipped learning (FL) has emerged as a particularly promising approach, offering a restructuring of the educational experience through strategic reallocation of learning activities. The past decade has witnessed unprecedented expansion in FL implementation across educational contexts from K-12 classrooms to tertiary institutions spanning disciplines as diverse as mathematics, medicine, engineering, and language acquisition. At its core, FL represents a pedagogical restructuring that strategically inverts the traditional instructional sequence. Foundational knowledge transmission shifts from synchronous classroom lectures to asynchronous pre-class engagement through multimedia resources, whilst valuable in-person class time is repurposed for interactive, application-focused activities including collaborative problem-solving, discussions, and personalised guidance (Bishop & Verleger, 2013). This pedagogical inversion fundamentally challenges conventional educational time-space configurations, reflecting broader shifts toward student-centred learning paradigms.

As FL implementations have proliferated globally, a substantial body of primary research has emerged examining its effectiveness, implementation challenges, and contextual adaptations. This expansion generated a complex, heterogeneous evidence base that educators, policymakers, and researchers increasingly rely upon for implementation guidance. The diversity of this primary research, spanning varied disciplines, educational levels, cultural contexts, and methodological approaches, has necessitated systematic synthesis efforts to consolidate findings and extract meaningful patterns. Consequently, systematic reviews (SRs) and meta-analyses (MAs) have become critical tools for navigating this complex research landscape, attempting to provide systematic assessments of FL effectiveness across diverse contexts.

However, a critical problem has emerged: whilst the number of FL reviews continues to grow exponentially, there has been no systematic examination of whether these syntheses themselves meet established methodological standards or provide reliable evidence for educational decision-making. This unchecked proliferation appears to create a paradox. The very tools meant to clarify the FL evidence base may themselves be contributing to confusion if they employ weak methodologies or incomplete outcome assessments.

The credibility and utility of these research syntheses fundamentally depend on their methodological rigor and transparency. Established frameworks such as PRISMA provide comprehensive guidelines for transparent reporting of review methodology and findings (Moher *et al*., 2015), whilst assessment tools like AMSTAR 2 offer structured criteria for evaluating methodological quality (Shea *et al*., 2017). As the corpus of FL reviews continues to expand, evaluating these syntheses against established methodological standards becomes increasingly important for discerning the reliability of their conclusions and recommendations.

High-quality SRs and MAs in education exhibit several essential methodological characteristics that serve as benchmarks for evaluation. First, they implement transparent and comprehensive literature search strategies, documenting databases consulted, search terms employed, date parameters, and language restrictions. They also consider unpublished or grey literature to minimise selection bias (Higgins *et al*., 2023). This comprehensiveness is particularly important given documented publication bias in educational research, where studies reporting statistically significant or positive outcomes are more likely to be published in peer-reviewed journals.

Second, rigorous reviews systematically assess the quality or risk of bias in included primary studies using validated tools appropriate to their research designs. This critical appraisal process recognises that synthesizing evidence without accounting for primary study quality can yield misleading conclusions that may inappropriately influence educational practice (Shea *et al*., 2017). In educational contexts, where research designs often include diverse methodologies beyond randomised trials, appropriate quality assessment becomes particularly challenging yet essential. Third, meta-analyses must employ appropriate statistical approaches for data synthesis, with justifiable choices regarding effect size metrics, heterogeneity assessment, and model selection (fixed-effect vs. random-effects) (Borenstein *et al*., 2009). Given the inherent variability in educational interventions and contexts, robust strategies for exploring and managing heterogeneity using measures like I^2, τ^2, subgroup analyses, or meta-regression are necessary for meaningful interpretation of aggregated effects. Finally, established techniques for detecting publication bias (e.g., funnel plots, Egger's test, fail-safe N calculations) are essential for assessing whether the available literature presents a balanced view of the intervention's effects (Sterne *et al*., 2011). This becomes particularly relevant for educational innovations like FL that have generated substantial enthusiasm, increasing pressure toward publication of positive findings.

Beyond methodological considerations, the value of FL reviews depends significantly on their conceptualisation and assessment of educational outcomes. Educational interventions like FL impact diverse outcome domains: cognitive (knowledge acquisition, conceptual understanding, critical thinking), affective (motivation, satisfaction, self-efficacy), and interpersonal (collaboration, communication skills, and engagement) (Bloom *et al*., 1956; Krathwohl *et al*., 1964). Comprehensive reviews should capture this multidimensional impact to fully represent FL's educational potential, particularly given its theoretical foundations in active learning and constructivist pedagogies that emphasise holistic development (Prevalla & Uzunboylu, 2019; Robertson, 2022; Rutkienė *et al*., 2022). Despite this, educational research, including FL studies, faces persistent challenges in outcome measurement. These include substantial variability in assessment approaches, limited standardisation of instruments, inconsistent reporting of psychometric properties, and frequent reliance on self-reported measures for non-cognitive outcomes (Leatherman & Cleveland, 2019; Ozenen, 2023; Roehling, 2018; Silverajah *et al*., 2022). High-quality reviews should acknowledge these limitations and, where possible, critically analyse their implications for interpreting findings regarding FL effectiveness.

Moreover, contemporary educational research increasingly recognises that moving beyond simple effectiveness comparisons to investigate causal mechanisms (explaining how FL produces outcomes) and moderating factors (determining when and for whom FL works best) is essential for developing nuanced understanding that can inform practice (Goedhart *et al*., 2019; Kapur *et al*. 2022; Lewis *et al*., 2020). Reviews that explore these dimensions provide more valuable guidance for educational practitioners than those simply aggregating effectiveness measures without addressing underlying processes.

Despite the proliferation of systematic reviews and meta-analyses on FL, a significant gap exists: no comprehensive synthesis has critically examined the methodological rigor and outcome assessment practices across these reviews. This gap is troubling for three critical reasons:

- Quality uncertainty: Without systematic evaluation, educators and policymakers cannot distinguish between methodologically robust reviews that provide reliable evidence and weak syntheses that may perpetuate flawed conclusions, misguiding educational practice (Tricco *et al*., 2018).
- Incomplete outcome mapping: Without understanding the patterns of outcome assessment across FL reviews, researchers cannot identify which learning dimensions remain

understudied or poorly measured, limiting comprehensive understanding of FL's educational impact.

- Methodological blind spots: The absence of meta-methodological analysis prevents the identification of systematic weaknesses in FL research synthesis, hindering improvements in future evidence generation and synthesis practices.

This umbrella review (UR) addresses these critical gaps by systematically evaluating the methodological quality and outcome reporting practices of FL research syntheses, thereby providing meta-methodological insights essential for improving future evidence synthesis in educational research. This study makes three distinctive contributions to the field:

- Comprehensive quality assessment: This work presents the first systematic evaluation of methodological rigor across FL reviews, establishing a baseline for quality and identifying key areas in need of improvement.
- Outcome assessment mapping: Educational outcomes in FL reviews are examined in terms of standardisation and measurement, uncovering consistent patterns, notable gaps, and opportunities for more holistic assessment approaches.
- Meta-methodological framework: A framework is proposed for evaluating educational research syntheses, offering guidance for future reviewers aiming to produce higher-quality syntheses and enabling readers to critically appraise existing reviews.

This critical UR examines the methodological underpinnings and outcome evaluations of the growing body of FL research syntheses, thereby assessing the quality of evidence guiding implementation decisions. Through systematic analysis of 41 comprehensive reviews spanning STEM, medical and health sciences, and social sciences, this review addresses the following research questions that guide the synthesis of FL reviews:

What are the general descriptive characteristics of FL reviews (e.g., publication year, disciplinary distribution, review type, learner type, and country/region)?

What methodological challenges are reported in FL reviews? (e.g., sources of primary studies, quality assessment of primary studies, publication bias, and language bias)?

What outcome-related issues are identified in FL reviews? (e.g., domain outcome focus, outcome metrics, and outcome mechanisms)?

By critically examining these dimensions, this review aims to illuminate both the strengths and limitations of current FL research syntheses, providing a meta-methodological perspective crucial for researchers, practitioners, and policymakers navigating the rapidly evolving FL evidence landscape. The findings will not only inform more rigorous future syntheses but also enable more critical consumption of existing FL evidence, ultimately supporting more informed educational decision-making.

METHOD

This UR was conducted in accordance with established methodological guidance for evidence synthesis. Whilst URs typically integrate findings from multiple systematic reviews (Aromataris *et al*., 2015), the PRISMA-ScR reporting framework was adopted (Tricco *et al*., 2018) to ensure comprehensive and transparent reporting, as its systematic approach to mapping evidence aligned with the study’s objective of identifying methodological patterns and gaps across FL reviews.

A critical UR methodology was employed to synthesise and appraise evidence drawn from existing systematic reviews and meta-analyses on FL. This design was deemed particularly appropriate given the proliferation of FL research syntheses, as it facilitates the integration of their findings into a higher-level, meta-methodological perspective (Becker & Oxman, 2011; Grant & Booth, 2009). The accelerating expansion of FL scholarship since 2012 (Bergmann &

Sams, 2012; Birgili, 2021) underscores the need for a structured and critical synthesis to inform evidence-based practice.

The Population–Concept–Context (PCC) framework guided the formulation of the research questions (Joseph *et al*., 2023; Peters *et al*., 2020):

- Population: Systematic reviews and meta-analyses on FL.
- Concept: Methodological quality indicators and outcome assessment practices.
- Context: Reviews published between 2017 and 2023 in STEM, medical/health, and social science fields.

The UR process followed the steps outlined by Becker and Oxman (2011) and Grant and Booth (2009):

- Identification and screening of relevant systematic reviews and meta-analyses.
- Extraction of descriptive, methodological, and outcome-related data.
- Critical appraisal of methodological quality using established frameworks (Hartling *et al*., 2014).
- Thematic synthesis of findings through iterative coding and consensus-building (Braun & Clarke, 2006; Kyngäs *et al*., 2020).

By combining descriptive mapping, methodological critique, and thematic outcome analysis, this review aims to highlight both the strengths and limitations of current FL evidence syntheses. In doing so, it provides direction for more rigorous future research and offers a stronger evidence base for pedagogical decision-making across diverse educational contexts.

Search Strategy

The search strategy was developed following PRISMA-ScR guidelines (Tricco *et al*., 2018) through a systematic, multiphase approach to ensure comprehensive identification of relevant literature. The selection of search terms followed a rigorous, iterative process to maximise search sensitivity whilst maintaining precision. Initially, a preliminary scoping review of seminal FL publications was conducted to identify commonly used terminology and conceptual variations in the field (Bergmann & Sams, 2012; Bishop & Verleger, 2013). This preliminary analysis revealed that whilst "flipped learning" and "flipped classroom" represent the dominant terminology, regional and disciplinary variations exist, with some European and technical education contexts preferring "inverted classroom" or "inverted learning."

Subsequently, the keyword selection was refined through three complementary approaches:

- Literature-derived terms: Analysis of index terms, keywords, and abstracts from highly cited FL reviews (Hew *et al*., 2021; Lo, 2020) identified core terminology patterns and conceptual synonyms
- Database thesaurus consultation: Medical Subject Headings (MeSH) and Education Resources Information Center (ERIC) descriptors were examined to identify controlled vocabulary relevant to FL concepts
- Expert validation: The preliminary search string was reviewed by educational technology researchers to ensure coverage of discipline-specific terminology variations

This systematic process resulted in the final search string: ("*flipped learning*" OR "*flipped classroom*" OR "*inverted classroom*" OR "*inverted learning*"), which balanced comprehensiveness with precision by capturing the primary conceptual variations whilst avoiding overly broad educational technology terms that would reduce search specificity.

A comprehensive search was conducted across four major academic databases selected for their coverage of educational research: Web of Science (multidisciplinary coverage with strong citation tracking capabilities), EBSCO (comprehensive education database aggregator including Education Research Complete), ERIC (specialised education database maintained by the US Department of Education), and SCOPUS (broad interdisciplinary coverage with extensive international journal representation).

The search strategy was adapted to each database's specific syntax requirements whilst maintaining conceptual consistency. All searches utilised the "All Fields" designation to capture relevant studies regardless of indexing variations. For example, the Web of Science search applying the keywords to all fields retrieved 2,824 articles (see Appendix).

The search was deliberately limited to English-language, open-access articles published between 2012 and 2024. The year 2012 was selected as the starting point based on bibliometric evidence indicating a notable surge in FL research following the publication of Bergmann and Sams' (2012) foundational text, which is widely recognised as catalysing mainstream adoption of this pedagogical approach (Birgili, 2021). The open-access limitation was implemented to ensure reproducibility and accessibility of the review findings, aligning with contemporary open science principles.

Recognising the limitations of database indexing for review-type articles (Holly, 2017), particularly the inconsistent application of publication type filters across databases, a multi-pronged supplementary search strategy was employed. First, given the absence of reliable automated filters for systematic reviews and meta-analyses across all selected databases, titles and abstracts of all retrieved records were manually screened for explicit review indicators (e.g., "*systematic review*," "*meta-analysis*," "*research synthesis*," "*umbrella review*"). Second, reference tracking was conducted on all included reviews to identify relevant syntheses not captured through database searching, following established snowballing procedures (Greenhalgh & Peacock, 2005). Third, targeted grey literature searches were performed through Google Scholar using truncated search strings to identify recent reviews not yet indexed in traditional databases.

The comprehensive search strategy yielded 7,419 initial records across all databases. Following automated and manual duplicate removal using reference management software, 2,946 unique records remained for screening. Two independent reviewers with expertise in educational research methodology conducted title and abstract screening using predetermined eligibility criteria, achieving substantial inter-rater agreement ($\kappa = 0.82$). This initial screening identified 118 potential reviews warranting full-text evaluation.

During full-text assessment, the two reviewers independently evaluated each article against detailed inclusion and exclusion criteria, with discrepancies resolved through discussion and, when necessary, consultation with a third reviewer. This rigorous evaluation process resulted in the inclusion of 41 systematic reviews and meta-analyses in the final UR synthesis. The complete selection process, including reasons for exclusion at the full-text stage, is detailed in Figure 1 (PRISMA flow diagram).

Inclusion / Exclusion Criteria

Rigorous eligibility criteria were applied during full-text screening. Only systematic reviews using qualitative, quantitative, or mixed methods were included, whilst primary studies (n = 2,828) were excluded. Non-systematic reviews (n = 25), including integrative, scoping, and narrative reviews were excluded, along with those addressing overly broad topics (n = 5), lacking clearly defined educational outcomes (n = 12), or focusing on non-educational contexts (n = 13). All reviews were evaluated using the JBI critical appraisal checklist (Aromataris *et al*., 2020), leading to the exclusion of four reviews. Nine reviews showing overlap in their primary studies were excluded to maintain synthesis integrity (Ortega *et al*., 2016). Futhermore, five reviews were excluded for not aligning with the operational definition of FL requiring both pre-class video/audio content and interactive in-class activities (Bishop & Verleger, 2013), and four more for lacking disciplinary outcome reporting.

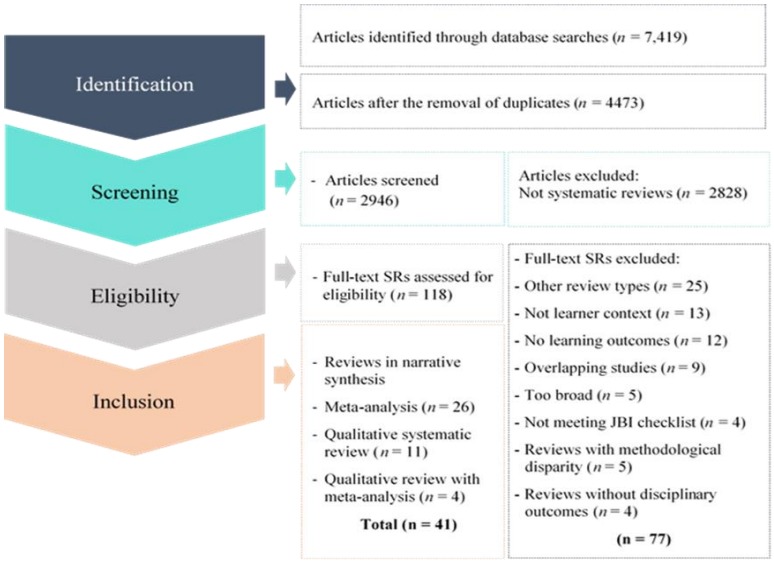

Figure 1. *PRISMA search methodology.*

Data Extraction & Synthesis

The PRISMA-ScR framework (Tricco *et al*., 2018) guided all aspects of this UR. The framework informed the structure of research questions using the Population-Concept-Context (PCC) approach and directed the search strategy, screening, and data extraction procedures. Inclusion criteria specified that reviews must: (a) employ systematic review or meta-analysis methodology, (b) examine FL in educational contexts, (c) report educational outcomes explicitly, and (d) be published between 2017-2023. Nine reviews with overlapping primary studies were excluded to maintain synthesis integrity.

Data extraction followed the JBI UR protocol (Aromataris *et al*., 2015) and incorporated the data-charting methodology from scoping review literature (Tricco *et al*., 2018). Two reviewers independently extracted data using a standardised form capturing three information categories: (a) descriptive characteristics, (b) methodological features based on PRISMA criteria (Moher *et al*., 2009), and (c) outcome-related issues. This dual extraction approach ensured reliability as recommended in PRISMA-ScR guidelines.

For RQ1 (descriptive characteristics), data including publication year, review type, disciplines, learner populations, and geographical regions were systematically extracted from all 41 reviews and summarised in Figure 2. For RQ2 (methodological issues), a critical appraisal framework based on established practices (Higgins *et al*., 2023; Kung *et al*., 2010) evaluated literature search scope, quality assessment approaches, and publication bias investigation methods, with findings presented in Figure 3.

For RQ3 (outcome-related issues), data extraction and synthesis were conducted using qualitative content analysis with MAXQDA 22.0, a validated tool for systematic reviews (Saldaña, 2021). The analysis used an inductive approach, allowing patterns and themes to emerge from the data rather than applying predetermined categories (Bengtsson, 2016). Both reviewers independently read all included studies multiple times before beginning formal coding. Sentences served as the unit of analysis to capture detailed insights, following best practices for educational research content analysis (Kyngäs *et al*., 2020). An initial coding framework was developed through open coding of 10% of the studies, then refined through pilot coding of 25% of the studies (n = 18), with adjustments made to improve clarity and coverage (Silver & Lewins, 2014). Three domains emerged from the data: cognitive, affective, and interpersonal. Whilst these align with established educational taxonomies (Bloom *et al*., 1956; Krathwohl *et al*., 1964), their identification was data-driven rather than predetermined. Two coders independently applied the emergent framework, resolving discrepancies through

consensus discussion following established qualitative reliability procedures (Kyngäs *et al.*, 2020; Neuendorf, 2018; Saldaña, 2016). Inter-rater agreement reached 95%, with Cohen's κ = 0.75, indicating substantial reliability. These domains, metrics, and mechanisms are summarised in Figure 4.

RESULTS

Descriptive Characteristics of the Reviews

Based on the descriptive analysis of 41 reviews on FL approach, several key characteristics emerge. The publication timeline, spanning from 2017 to 2023, reveals a significant increase in research synthesis in recent years, with the large majority (n = 31) published between 2020 and 2023, peaking with twelve reviews in 2020. The field representation shows a predominance in medical and health sciences (n = 13) and social sciences (n = 11), followed closely by interdisciplinary fields (n = 9) and STEM (n = 8).

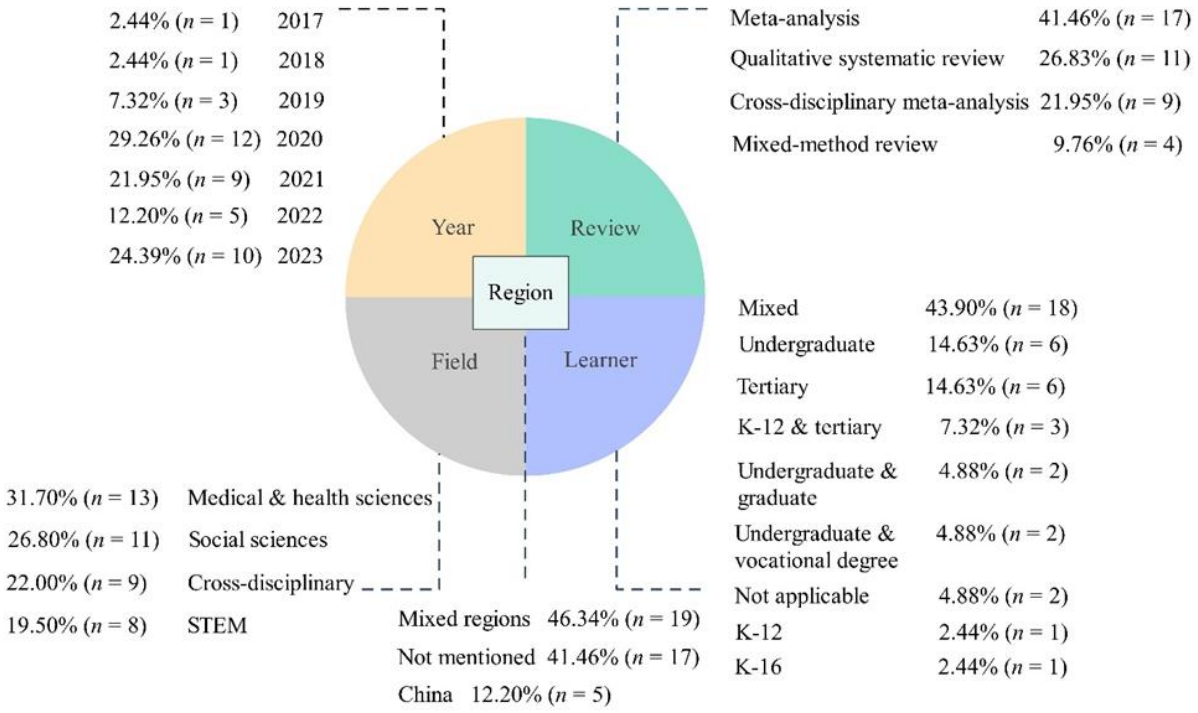

Figure 2. *The overview of the reviews by year, review type, learner type, field, and region.*

Methodologically, meta-analytical reviews were most prevalent (n = 26), significantly exceeding the number of qualitative (n = 11) and mixed-method reviews (n = 4). The included reviews predominantly investigated mixed learner populations (n = 18), with undergraduate (n = 6) and tertiary-level learners (n = 6). Less frequent reviews were focusing on K-12 (n = 1) and K-16 (n = 1) learner groups, whilst two reviews did not specify applicable learner types. Geographically, the scope was often international, with 19 reviews drawing studies from mixed countries/regions. A substantial number of reviews (n = 17) did not specify the country or region of the primary studies included, whilst five reviews focused exclusively on research conducted in China.

Methodological Issues in FL Reviews

The credibility and utility of systematic reviews and meta-analyses depend heavily on the methodological rigor employed both in the primary studies included in the synthesis process itself. Examining the methodological features reported across FL reviews reveals considerable variability and highlights several recurring challenges that impact the interpretation of their

findings. The key methodological aspects as reported in FL reviews cover sources of primary studies, quality assessment of primary studies, publication bias, and language restrictions.

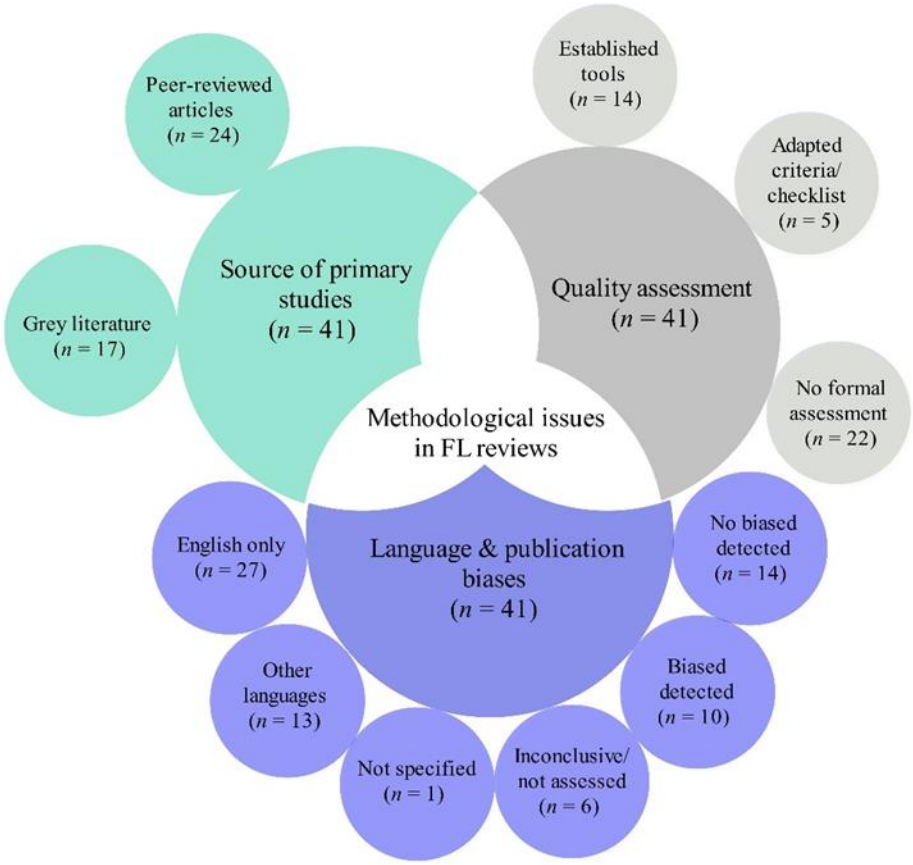

Figure 3. *Methodological issues in FL reviews.*

Concerning sources of primary studies, many reviews (*n* = 24) restricted inclusion primarily or exclusively to peer-reviewed journal articles (e.g., Akçayır & Akçayır, 2018; Aljaber *et al*., 2023). A significant number of reviews (*n* = 17) explicitly included grey literature such as theses, dissertations, and conference papers/proceedings, often to mitigate publication bias or capture a broader evidence base (e.g., Arslan, 2020; Aybirdi *et al*., 2023).

Reviews demonstrated diverse approaches to quality assessment of primary studies. Whilst some reviews (*n* = 14) employed established tools including MERSQI and JBI checklists (Aljaber *et al*., 2023), EPHPP (Cui *et al*., 2023), CASP (Li *et* al., 2020), NOS (Ge *et al*., 2020), and Cochrane RoB tool (Naing *et al*., 2023; Park & Suh, 2021; Xiao *et al*., 2022; Xu *et al*., 2019; Zhu, 2021), and others (*n* = 5) used adapted criteria (e.g., Tutal & Yazar, 2021; Vitta & Al-Hoorie, 2020), a substantial number of reviews (*n* = 22) reported no formal quality appraisal, relying instead on inclusion criteria as quality filters (e.g., Arslan, 2020; Güler *et al*., 2023) or mentioned assessment without providing details (e.g., Özbay & Çınar, 2021; Akçayır & Akçayır, 2018; Doğan *et al*., 2021). This lack of systematic quality assessment significantly undermines the reliability of many reviews' conclusions.

For publication bias in meta-analyses (*n* = 30), common techniques included funnel plots, statistical tests (Egger's regression, Begg's rank correlation), Trim-and-Fill method, Fail-Safe N calculations, and occasionally p-curve analysis. Findings were mixed, with 14 reviews concluding low risk of publication bias (e.g., Aybirdi *et al*., 2023; Bredow *et al*., 2021), 10 reviews finding evidence suggesting bias presence requiring statistical adjustment (e.g., Cui *et al*., 2023; Farmus *et al*., 2020), and six reviews presenting inconclusive assessments due to high heterogeneity (e.g., Naing *et al*., 2023).

Regarding language inclusion, most reviews (n = 27) limited inclusion to English-language publications (e.g., Turan, 2023). Yet, some demonstrated multilingual inclusivity, with six reviews including both English and Chinese studies (e.g., Xu *et al*., 2019), two incorporating English and Turkish articles (e.g., Doğan *et al*., 2021), one focusing on English and Korean based on search terms (Jang & Kim, 2020), and one stating no language restrictions (Naing *et al*., 2023). Conversely, some had narrower linguistic focus, with three reviews including only Chinese publications (e.g., Peng *et al*., 2022) and one limited to Korean studies (Park & Suh, 2021).

Outcome Related Issues in FL Reviews

The analysis of 41 reviews on FL practices emerged a total of 132 codes across cognitive, affective, and interpersonal learning domains within STEM, M&H, and SS fields. The cognitive domain accounted for half of the reported outcomes (n = 66). The affective domain, encompassing satisfaction, positive thoughts, self-confidence, and motivation, was less frequently addressed (n = 35). Similarly, the interpersonal domain, focusing on social engagement, interaction, active-in class time, and peer learning & collaboration, received the least attention (n = 31) across fields (see Figure 4).

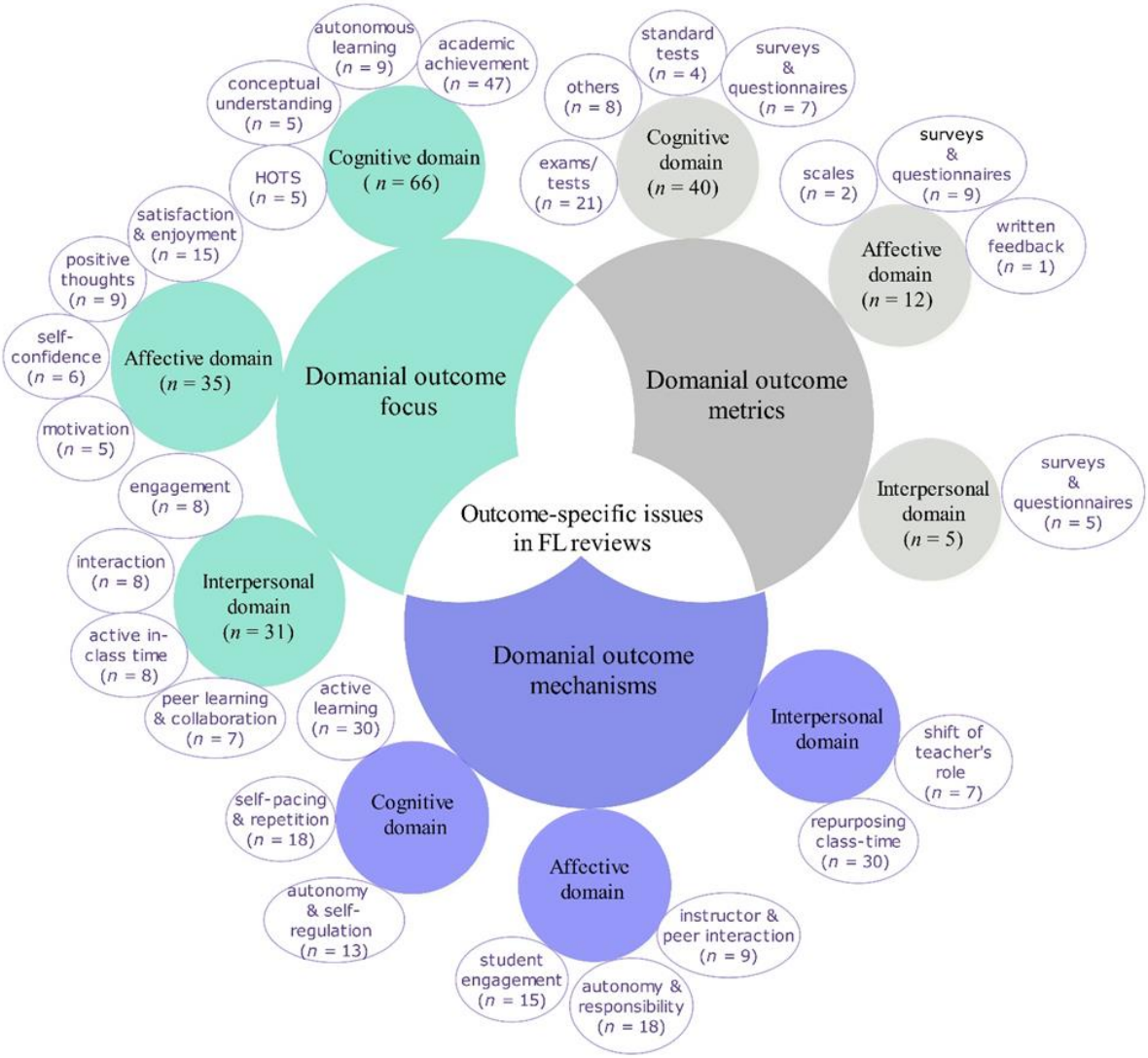

Figure 4. *Outcome related issues in FL reviews.*

Domain outcome focus

The cognitive domain dominated FL research outcomes across fields (n = 66), with academic achievement as the most studied outcome. In STEM disciplines, meta-analyses revealed positive effects in science (e.g., Doğan *et al*., 2023; g = 0.72, k = 30) and engineering (e.g., Strelan *et al*., 2020a; g = 0.72, k = 15). M&H sciences showed even stronger impacts in nursing (Xu et al., 2019; *SMD* = 1.79, k = 22) and pharmacy (e.g., Cui *et al*., 2023; *SMD* = 1.30, k = 22), whilst SS fields demonstrated moderate to strong effects in teaching EFL/ESL (e.g., Ni *et al*., 2023; *ES* = 0.68, k = 24) and general education (Jang & Kim, 2020; *ES* = 0.80, k = 26). Beyond achievement, FL fostered autonomous learning across disciplines, evidenced in mathematics, as found in 16 out of 97 studies reviewed by Cevikbas & Kaiser (2023), radiology (Ge *et al*., 2020; *OR* = 1.98, k = 5), Chinese medicine (Xiao *et al*., 2022; *RR* = 1.52, k = 5), and language education, as found in nine out of 22 studies reviewed by Linling & Abdullah (2023). Conceptual understanding also improved across fields (Akçayır & Akçayır, 2018; Cevikbas & Kaiser, 2023; Li *et al*., 2020), whilst higher-order thinking was particularly prominent in nursing education (Park & Suh, 2021; g = 0.87, k = 5).

The affective domain (n = 35) showed notable patterns across disciplines. Student satisfaction emerged as the most consistently documented outcome, with positive evidence in mathematics (Bredow *et al*., 2021; g = 0.53, k = 10), ICT (Strelan *et al*., 2020b; g = 0.81, k = 7), nursing (Xu et al., 2019; *ES* = 1.39, k = 6), health professions (Naing *et al*., 2023; *SMD* = 0.48, k = 8), radiology (Ge *et al*., 2020; *OR* = 1.70, k = 7), and language education, where seven out of 78 studies in Arslan's (2020) review showed satisfaction. Positive attitudes were documented across mathematics, noted in 27 out of 97 studies reviewed by Cevikbas & Kaiser (2023), nursing (Xu *et al*., 2019; *ES* = 1.58, k = 5), and language education reported in 23 studies (Arslan, 2020). Self-confidence received limited attention in STEM, highlighted in 14 out of 97 studies reviewed by Cevikbas & Kaiser (2023), and appeared only in pharmacy/anatomy education (Aljaber *et al*., 2023; Kazeminia *et al*., 2022) and SS reviews (n = 3). For example, in ELT, 10 studies out of 78 in a review noted that FL contributed to building self-confidence in language learners (Arslan, 2020), whilst six studies out of 22 in EFL contexts highlighted the approach's efficacy in enhancing self-confidence during the COVID-19 pandemic (Linling & Abdullah, 2023). Besides, five articles out of 71 in broader educational fields reported increased student confidence through FL practices (Akçayır & Akçayır, 2018). Learner motivation was moderately studied in STEM, largely absent from M&H, but extensively documented in language education, where 35 out of 78 studies in Arslan's (2020) review documented this.

The interpersonal domain (n = 31) revealed transformation in classroom dynamics across disciplines. Peer learning was prominent in STEM (n = 4), with enhanced collaboration in mathematics, identified in 23 out of 97 studies by Cevikbas & Kaiser (2023), and engineering (Lo & Hew, 2019). M&H sciences showed strong quantitative evidence (Xu *et al*., 2019; *ES* = 1.60, k = 10), whilst SS fields placed less emphasis, as noted in 14 out of 78 studies reviewed by Arslan (2020). Active in-class time was featured prominently in STEM (n = 4), less in M&H (n = 1), and moderately in SS (n = 3). Classroom interaction displayed an inverse pattern compared to STEM. It received greater emphasis in M&H, as evidenced by Xu *et al*. (2019; *ES* = 1.41, k = 8), and in SS, where Linling & Abdullah (2023) reported interaction in eight out of 22 studies. Student engagement varied, being moderately represented in STEM, observed in 28 out of 97 studies by Cevikbas & Kaiser (2023), and M&H, where Li *et al*. (2020) reported engagement in six studies out of 56 studies, and Aljaber *et al*. (2023) observed engagement in five out of 35 studies, but dominating in SS contexts qualitative reviews (n = 3) on ELT (Arslan, 2020; Linling & Abdullah, 2023; Turan & Akdag-Cimen, 2020), particularly noted by Arslan (2020) in 19 out of 78 studies. Reviews showcase the effectiveness of FL practices in boosting student engagement, as evidenced by 19 studies out of 78 studies (Arslan, 2020), 11 out of 43 (Turan & Akdag-Cimen, 2020), 11 out of 22 studies (Linling & Abdullah, 2023), in ELT contexts and 10 out of 71 studies (Akçayır & Akçayır, 2018) in educational fields.

Despite their theoretical importance, non-cognitive outcomes showed substantial limitations in both breadth and depth across all fields. M&H sciences notably lacked motivation studies, creating a critical gap in understanding how FL shapes motivational dynamics in healthcare education. These evidentiary gaps suggest that whilst FL shows promising potential for enhancing non-cognitive outcomes, current research provides an incomplete foundation for understanding its comprehensive impact.

Domain outcome metrics

The synthesis of outcome metrics across 41 reviews revealed significant patterns and limitations. Cognitive outcomes were assessed using diverse quantitative metrics in 18 reviews, including GPAs, various tests/exams, grades, quizzes, and homework (e.g., Akçayır & Akçayır, 2018; Aljaber *et al*., 2023; Cui *et al*., 2023; Lo & Hew, 2019). Health-related disciplines employed skill-based assessments (n = 7), including practical skills tests and OSCEs (Cui *et al*., 2023; Li *et al*., 2022). Less comprehensive formats included quizzes (n = 4), homework (n = 2), MCQs (n = 2), and subjective evaluations (n = 4) (e.g., Ge *et al*., 2020; Peng *et al*., 2022). Notably, 19 reviews failed to explicitly report cognitive outcome metrics (e.g., Arslan, 2020; Aybirdi *et al*., 2023; Turan, 2023), and most offered limited critical analysis regarding reliability and validity of assessment instruments.

In the affective domain, survey instrument predominated, including general questionnaires (n = 4) (e.g., Aljaber *et al*., 2023; Kazeminia *et al*., 2022), Likert scales (Kazeminia *et al*., 2022), attitude scales (Tutal & Yazar, 2021), satisfaction scales (Naing *et al*., 2023), and written feedback (Aljaber *et al*., 2023). Most significantly, 33 reviews failed to report affective outcome metrics (e.g., Doğan *et al*., 2021; Park & Suh, 2021). Reviews typically did not discuss psychometric quality, potential self-report biases, or whether standardised instruments were employed, raising questions about the trustworthiness of positive affective findings.

Interpersonal outcomes represented the most underreported domain with minimal methodological diversity. Limited reports included qualitative surveys on interaction (Kazeminia *et al*., 2022), subjective evaluations of cooperation/teamwork (n = 2) (Ge *et al*., 2020; Xiao *et al*., 2022), and questionnaires addressing communication skills (n = 2) (Li et al., 2022; Peng *et al*., 2022). Although five reviews highlighted increased interaction as a benefit, they provided almost no critical analysis of measurement quality. Most strikingly, 36 reviews failed to report interpersonal outcome metrics (e.g., Farmus *et al*., 2020; Güler *et al.*, 2023; Turan, 2023), revealing a striking incongruity: whilst FL reviews consistently emphasise collaborative processes, it simultaneously neglects rigorous assessment of these dimensions.

Domain outcome mechanisms

Across reviews (n = 30), active learning during repurposed class time emerged as the primary mechanism driving cognitive benefits (e.g., Akçayır & Akçayır, 2018; Arslan, 2020; Park & Suh, 2021). This mechanism operates by shifting focus to activities like problem-solving and discussions, promoting deeper cognitive processing and higher-order thinking skills in many reviews (n = 15) (e.g., Aljaber *et al.*, 2023; Doğgan *et al*., 2021). Self-paced learning via pre-class materials was identified as an enabling mechanism in many reviews (n = 18) (e.g., Arslan, 2020; Cui *et al*., 2023; Peng *et al*., 2022), allowing students to master foundational knowledge before class. Some reviews (n = 13) highlighted autonomy development as both prerequisite and outcome (e.g., Cevikbas & Kaiser, 2022; Turan, 2023). Nevertheless, many reviews (n = 11) were critiqued for lacking depth when discussing these mechanisms, with connections remaining superficial or speculative (e.g., Aljaber et al., 2023; Cui *et al*., 2023).

For affective outcomes, many reviews (n = 15) identified active engagement as driving increased motivation and satisfaction (e.g., Cevikbas & Kaiser, 2022; Kazeminia *et al*., 2022). The autonomy offered by self-paced learning was frequently cited (n = 12) as enhancing satisfaction (e.g., Aljaber *et al*., 2023; Zheng *et al*., 2020), whilst responsibility and autonomy were mechanisms for increased self-confidence in some reviews (n = 6) (e.g., Akçayır &

Akçayır, 2018; Jang & Kim, 2020). Increased interaction was proposed (*n* = 9) as contributing to a more positive environment (e.g., Bredow *et al*., 2021; Lo *et al*., 2017). However, some reviews (*n* = 5) did not explore how these mechanisms operated, leaving connections general and complex factors underexplored (e.g., Farmus *et al*., 2020; Turan & Akdag-Cimen, 2020).

For interpersonal outcomes, the most frequent mechanism across reviews (*n* = 30) was repurposing class time for increased interaction and collaboration (e.g., Ge *et al*., 2020; Linling & Abdullah, 2023). By moving content delivery outside class, instructors implemented activities like group work and discussions, enhancing student-student and student-teacher interaction (e.g., Doğan *et al*., 2021; Sopamena *et al*., 2023). The teacher's shift to facilitator was also identified as an interpersonal mechanism (*n* = 7) (e.g., Akçayır & Akçayır, 2018; Doğan *et al*., 2021). In contrast, many reviews (*n* = 11) failed to provide substantive insights into the specific nature of these interactions or how collaboration effectively contributes to learning outcomes (e.g., Chen *et al*., 2023; Ni et al., 2023).

DISCUSSION and CONCLUSION

The synthesis of 41 reviews across fields on FL spanning 2017-2023 uncovered crucial findings that transcend methodological issues. These findings address core questions regarding FL's conceptualisation, implementation, and assessment as an educational innovation. Whilst methodological rigor remains fundamental to any evidence synthesis, the patterns that emerged from this UR highlight distinctive challenges that carry particular significance for FL's theoretical underpinnings and practical application.

The synthesis revealed systematic methodological opacity across FL reviews that undermines evidence credibility. Methodological transparency exists on a continuum describing the degree of detail and disclosure about research steps and decisions. Methodological opacity in reviews reflects broader systemic challenges across research domains, fundamentally representing a fundamental threat to educational knowledge advancement (Aguinis *et al*., 2020). The current UR synthesis revealed that educational intervention in FL reviews suffers from methodological transparency. This pattern mirrors what Campbell *et al*. (2019) identified as alarming gaps in methodological reporting, with 95% of reviews in their study failing to describe their narrative synthesis methods adequately. Similarly, Castro-Gil and Correa (2021) found that roughly 73% of reviews on blended learning fell into the two lowest quartiles of methodological transparency. That said, the present findings extended beyond simple reporting omissions to reveal the failure to capture FL's multifaceted nature as a pedagogical innovation. Whilst Templier and Paré (2018) highlighted that transparency varied considerably across six key review steps from problem formulation to data analysis, the present synthesis illuminated an additional critical dimension: the transparency of intervention implementation details. This is in line with what Paré *et al.* (2016) described as the need to make the review process a non-"black box" experience, yet reviews face the unique challenge of needing transparency at multiple levels—both in their methodological procedures and in their documentation of the pedagogical interventions under investigation.

The inadequate reporting of intervention details represents a critical transparency failure paralleling issues in related educational research. Recent review (Kager *et al*., 2024) on methodological transparency in lesson study research revealed significant underreporting challenges particularly in the observation and reflection stages. The study also identified that high variability and multiple adaptations of lesson study create substantial barriers to meaningful evidence synthesis, as inadequate descriptions of conceptualisation and enactment render research difficult to understand and replicate. This methodological opacity became particularly questionable because FL educational interventions inherently depend on multiple interrelated contextual factors, including pre-class material quality, in-class activity design, student preparation levels, and instructor facilitation skills. The confluence of insufficient review methodology transparency and inadequate intervention documentation creates an urgent need for fundamentally reconceptualised reporting standards in educational reviews. Without

enhanced frameworks that extend beyond existing guidelines like ENTREQ (Tong *et al*., 2012) to systematically capture pedagogical design decisions, technological affordances, and contextual moderators, educational research will perpetuate methodological shortcomings. These shortcomings have been identified by Mesquida *et al*. (2025) as threats to reproducibility, transparency, and statistical validity in evidence synthesis.

Building on these transparency concerns, another critical concern involves the exclusion of grey literature in FL reviews, which represents a critical methodological limitation that undermines evidence validity and comprehensiveness. The analysis converged with a paradoxical finding that whilst most systematic reviews conduct comprehensive grey literature searches, few incorporate these materials into final analysis. This exclusion created systematic bias in effect size estimation (Hartling *et al*., 2017). To illustrate, published studies showed 9% greater treatment effects compared to grey literature trials, inflating FL effectiveness by 15-30% (Hopewell *et al*., 2007; Schmucker *et al*., 2017). Beyond statistical bias, grey literature exclusion compromises ecological validity and practical implementation understanding. Converging evidence from research highlights the importance of grey literature in reviews. Adams *et al*. (2017) emphasised that grey literature sources enable reviews to incorporate relevant contemporary material in dynamic and applied topic areas where academic research lags, whilst Zhang *et al*. (2021) argued that they provide essential practitioner insights for emerging topics, and some findings existing exclusively in grey literature (Kamei *et al*., 2021). The inclusion is of great significance as reviews may overlook implementation challenges, resource requirements, and contextual adaptations documented in conference proceedings and institutional reports rather than peer-reviewed journals. At this point, grey literature exclusion might be especially critical under conditions where interventions like FL exhibit complexity, outcome measurement lacks standardised consensus, contextual factors critically influence implementation, and research-based evidence remains scarce (Benzies *et al*., 2006).

Furthermore, the synthesis revealed that omitting formal quality assessment in reviews represents methodological failure that undermines evidence synthesis reliability and utility, contradicting well-established consensus that quality assessment constitutes an essential systematic review component. Congruent with the research synthesis, research underlines the fact that this omission is characterised as uncritical acceptance of systematic review results, carrying significant risks for educational policy and practice, and is particularly troubling given that educational intervention reviews appear disconnected from methodological advances (Luchini *et al*., 2021; Shea *et al*., 2017). Page *et al*. (2021) describe this process as sophisticated shifts from assessing "quality" to assessing "certainty" in evidence bodies through advanced tools like ROBIS and AMSTAR 2. Inadequate quality assessment undermines evidence validity (Whiting *et al*., 2017) and becomes an obstacle in comprehensive methodological and evidence quality appraisal (Harrison *et al.*, 2021), particularly challenging for complex pedagogical frameworks like FL that involve multiple components, diverse contexts, and varied outcomes demanding sophisticated quality appraisal approaches. Zeng *et al*. (2015) noted that including selection, performance, detection, attrition, and reporting biases may offer false confidence to educators and policymakers. Addressing this issue requires fundamental changes. These include mandatory quality assessment protocols, development of multi-dimensional assessment tools capable of evaluating both methodological quality and intervention-specific factors, and investment in validating quality assessment tools specifically designed for complex pedagogical interventions like FL.

Moving to another significant issue, the findings from FL reviews that restricted their searches to English-language publications align with broader patterns where approximately 38% (Stern & Kleijnen, 2020) to 46.4% (Helbach *et al*., 2022) of systematic reviews impose such restrictions. Such practice may distort understanding of pedagogical effectiveness of an approach across diverse cultural contexts (Grégoire *et al*., 1995). For instance, Egger *et al*. (1997) found that only 35% of German-language articles reported statistically significant results compared to 62% of English-language articles, suggesting disproportionate publication of

positive results in English literature. Furthermore, Jüni *et al.* (2002) demonstrated that non-English language trials showed treatment effect estimates averaging 16% more beneficial compared to English-language trials, which indicates that English-dominated evidence bases may systematically overstate effect sizes whilst underrepresenting implementation challenges. The cultural dimensions of this exclusion are particularly concerning for educational interventions inherently culturally embedded practices where local context fundamentally shapes implementation and effectiveness. Language restrictions create systematic bias in reviews, with Pham *et al.* (2005) demonstrating effect estimates approximately 63% smaller in English-only versus language-inclusive reviews of complementary medicine interventions, whilst Wang *et al.* (2015) found such restrictions reduce methodological quality scores and overall credibility. These impacts become particularly concerning for educational research given Walpole's (2019) evidence that approximately 10-15% of relevant studies exist exclusively in non-English databases, suggesting reviews systematically exclude critical insights about cultural moderators, implementation barriers, and contextual factors determining intervention success or failure. Addressing this bias requires fundamental changes in evidence synthesis approaches. These include comprehensive multilingual search strategies, transparent reporting of language decisions, investment in translation resources, and explicit journal guidelines recognising the culturally embedded nature of pedagogical practices.

With respect to domain outcomes, the striking dominance of cognitive ones in FL reviews represented a profound theoretical-empirical disconnect that fundamentally undermines understanding of pedagogical innovation considering that FL is explicitly grounded in constructivist and social learning theories (Abeysekera & Dawson, 2015). This pattern, which extends beyond what was revealed in the current UR, might be the manifestation of deep structural features embedded within educational research traditions, institutional priorities, and measurement paradigms. Affirming the patterns identified in the synthesis, Franco *et al.* (2014) pointed out that systematic publication bias toward statistically significant results in educational research creates academic reward systems that inadvertently shape outcome selection toward easily quantifiable cognitive measures. Easterbrook *et al.* (1991) quantified this bias, underlining that studies with statistically significant results were 2.32 times more likely to be published than those with non-significant findings, creating systematic exclusion of nuanced investigations into affective and interpersonal outcomes. An alternative explanation for this dominance might be the theory-measurement disconnect that analogically exemplifies a process termed "learnification," which refers to the reduction of complex educational processes to quantifiable learning outcomes (Biesta, 2010).

The synthesis of assessment instruments for reporting domain outcomes in reviews revealed limited instrument analysis and heavy reliance on self-report tools, presenting fundamental threats to the validity of FL evidence synthesis. This concern found theoretical grounding in Clark and Watson's (2019) assertion that validity belongs to measurement interpretations, not measures, requiring empirical testing of hypothesised relationships between constructs and observable manifestations—validation processes partially absent in FL reviews. In addition, Murphy and Hall (2024) identified this failure as a dangerous assumption that "neither rigorous research design, nor advanced statistics, nor large samples can correct" false inferences from faulty measurements. The CoDiL framework developed by Reinhold *et al.* (2024) revealed that cognitive assessment validity depends on carefully theorised connections between behavioural indicators and underlying processes, yet FL reviews seldom investigate whether their measurement approaches correspond with constructivist principles. Recent research (Sancar-Tokmak & Dagli, 2025) provides clear evidence of the quantitative impact of these validity issues, demonstrating that theoretically grounded studies produced positive outcomes at a rate of 56%—significantly exceeding the 29% rate observed in atheoretical approaches. This finding reveals that measurement validity fundamentally influences both effect size estimation and the conclusions reached through evidence synthesis.

In line with the findings of the synthesis, research highlights that the reliance on self-report measures for affective and interpersonal outcomes represents significant validity threats. Kreitchmann *et al*. (2019) demonstrated that such measures are susceptible to social desirability and acquiescent responding biases, consequently introducing systematic errors that compromise assessment fairness and validity. Furthermore, Coman and Richardson (2006) stated that 60% of studies inappropriately compared self-report disability measures with performance-based functional limitation measures, representing fundamental construct mismatches, whilst only 44% adequately reported psychometric properties. Their review found correlations between self-report and performance measures ranging from -0.72 to 0.71, thus emphasising that self-reporting reflects complex perceptions influenced by affective function, personality, and culture, with stronger correlations observed only when instruments measured identical constructs. Ekinci *et al*. (2023) similarly identified that many educational scales lack philosophical and theoretical grounding, thereby compromising the validity and reliability of measurement instruments. Moreover, recent research underscores critical gaps in validated instruments for technology-enhanced learning environments. Specifically, Alqahtani *et al.* (2023) emphasised that FL require specialised measurement approaches capturing both in-class and out-of-class learning experiences whilst maintaining rigorous validation standards. In addition, Urcola-Pardo *et al.* (2024) demonstrated that robust assessment tools require systematic multi-phase approaches including item selection, expert panel validation, and psychometric testing. Larsen *et al*. (2025) explicitly documented that this reliance on self-report instruments ultimately constrains researchers' ability to understand motivational regulation and learning design relationships, thereby limiting the development of effective interventions in real-time learning contexts.

The inadequate exploration of causal mechanisms linking FL to outcomes represents a fundamental issue emerging from the synthesis. The superficial treatment of theoretical frameworks including constructivism, self-determination theory, and active learning principles without substantive mechanistic investigation reveals deeper methodological problems. Aeschbach *et al*. (2021) documented these transparency challenges, finding that only 22.62% of measured constructs were clearly defined, merely 13.1% were situated within theoretical frameworks, just 19.05% of measurements provided selection rationales, and only 4.54% of self-developed measures demonstrated construct validity evidence. The absence of mechanistic clarity has critical implications for understanding intervention effectiveness. Reinhold *et al*. (2024) underscored that learning activities must serve as theoretical and empirical mediators, functioning as "links" between students' external behavioural interactions and internal cognitive processes. Without establishing these connections, reviews cannot explain which specific mechanisms generate positive outcomes or identify the conditions under which these mechanisms operate effectively. This mechanistic gap manifests in inadequate reporting of implementation details.

In addition, in a recent study Chen *et al*. (2025) found that numerous studies failed to provide detailed descriptions of pedagogical practices, thereby limiting analysis of specific implementation mechanisms and exposing critical construct validity gaps within FL literature. They further emphasised that future research must prioritise comprehensive reporting of instructional design features to facilitate nuanced meta-analytic investigations, underscoring the necessity of construct clarity for developing valid measures of educational interventions and their learning outcomes. Moreover, Luo and Gan (2022) accentuated three crucial implications for educational measurement: first, cross-sectional validation studies inherently limit causal inference, necessitating longitudinal or experimental designs; second, the cultural and contextual specificity of validated instruments requires careful consideration when applying measures across diverse populations; and third, multi-method approaches that combine quantitative validation with qualitative exploration are essential for enhancing both validity and practical utility of educational measurement instruments.

In short, this UR of 41 systematic reviews on FL (2017-2023) revealed that the evidence base for this widely adopted pedagogical innovation has methodological challenges. These include methodological opacity, grey literature exclusion, language restrictions, cognitive outcome dominance, measurement validity issues, and inadequate mechanistic understanding. These challenges collectively undermine the credibility and utility of current FL research synthesis.

RECOMMENDATIONS

The critical evidence in this UR covering methodological and domain issues necessitates comprehensive reforms in how FL research is conducted, synthesised, and reported. These recommendations address not isolated methodological issues but rather interconnected challenges that collectively compromise the evidence base for this globally implemented pedagogical innovation. The following recommendations, grounded in the specific patterns and issues identified in this synthesis, outline essential changes needed across multiple levels of the research enterprise to establish a credible and actionable evidence base for FL practices.

Researchers must adopt comprehensive reporting standards extending beyond procedural details to include implementation characteristics, following models like the ENTREQ statement and explicitly specifying review processes, sources, and rationales. Educational researchers must move beyond token searches to meaningful incorporation, with review protocols explicitly specifying search strategies, quality assessment criteria, and integration methods, including routine sensitivity analyses.

Mandatory protocols should be established for all educational intervention reviews. Journals should require explicit reporting as publication conditions and employ multi-dimensional assessment tools evaluating both methodological quality and intervention-specific factors. Comprehensive search strategies must include multilingual databases and international collaboration. These strategies should include explicit documentation of languages included and excluded, requiring investment in translation resources despite practical constraints.

Researchers must develop instruments aligned with FL's theoretical foundations, creating assessments for critical thinking and collaborative skills beyond traditional knowledge-recall measures. This development should address self-report bias through methodological triangulation. Process-oriented methodologies are needed to track learning over time, including ecological momentary assessment, learning analytics, and mixed-methods approaches, with scale development targeting mechanistic constructs rather than just outcomes.

These recommendations represent not merely methodological refinements, but fundamental reconceptualisation of how educational interventions like FL are evaluated and understood. Such comprehensive transformation may enable FL research to achieve the rigor necessary.

LIMITATIONS

This UR has several methodological limitations that should be considered when interpreting its findings. First, the restriction to English-language, open-access publications may have excluded relevant reviews published in other languages or behind paywalls. Given the global implementation of FL across diverse educational contexts, particularly in non-English speaking regions of Asia, Europe, and Latin America, this linguistic restriction limits the comprehensiveness of the synthesis and may introduce cultural bias toward Western educational perspectives.

Second, the exclusion of grey literature reviews (scoping reviews, narrative reviews, integrative reviews) focused solely on systematic reviews and meta-analyses. Whilst this decision enhanced methodological consistency, it may have excluded valuable insights from other synthesis approaches that could provide complementary perspectives on FL implementation and effectiveness. Additionally, the requirement for reviews to align with Bishop and Verleger's (2013) operational definition of FL requiring both pre-class video/audio content and interactive

in-class activities may have excluded reviews examining variations or adaptations of FL that deviate from this specific conceptualisation.

Third, the data extraction and coding processes, whilst following established protocols and achieving substantial inter-rater reliability (Cohen's $\kappa = 0.75$), remain subject to interpretive decisions inherent in qualitative thematic analysis. The emergent categorization of outcomes into cognitive, affective, and interpersonal domains, though data-driven and aligning with established educational taxonomies, represents one possible organisational framework among several that could have been applied.

Fourth, the heterogeneity included reviews spanning diverse disciplines, educational levels, and geographical contexts whilst providing comprehensive scope, also complicates direct comparisons and may obscure discipline-specific or context-dependent patterns. The aggregation of findings across such diverse reviews assumes a level of comparability that may not fully account for the unique characteristics of different educational settings.

Finally, this UR inherits the limitations of its constituent reviews. Quality assessment revealed that over half of the included reviews lacked formal risk of bias evaluation, many restricted searches to English-language publications, and most excluded grey literature. These methodological weaknesses in the primary reviews necessarily constrain the reliability and generalisability of this synthesis. The review can only be as robust as the evidence it synthesises, and the identified methodological gaps in the reviewed studies represent fundamental limitations that cannot be overcome through UR methodology alone.

Despite these limitations, this UR provides the first extensive critical analysis of FL research synthesis, offering valuable insights into methodological patterns, outcome assessment practices, and areas requiring improvement in future evidence synthesis efforts.

Declaration of Conflicting Interests and Ethics

This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the author.

Orcid

Abdullah Arslan https://orcid.org/0000-0002-3979-6371

REFERENCES

Abeysekera, L., & Dawson, P. (2015). Motivation and cognitive load in the flipped classroom: Definition, rationale and a call for research. *Higher Education Research & Development*, *34*(1), 1–14. https://doi.org/10.1080/07294360.2014.934336

Adams, R.J., Smart, P., & Huff, A.S. (2017). Shades of grey: Guidelines for working with the grey literature in systematic reviews for management and organizational studies. *International Journal of Management Reviews*, *19*(4), 432-454. https://doi.org/10.1111/ijmr.12102

Aeschbach, L.F., Perrig, S.A.C., Weder, L., Opwis, K., & Brühlmann, F. (2021). Transparency in measurement reporting: A systematic literature review of CHI PLAY. *Proceedings of the ACM on Human-Computer Interaction, 5(CHI PLAY), 233*. https://doi.org/10.1145/3474660

Aguinis, H., Ramani, R.S., & Alabduljader, N. (2020). Best-practice recommendations for producers, evaluators, and users of methodological literature reviews. *Organizational Research Methods*, *26*(1), 46–76. https://doi.org/10.1177/1094428120943281

Alqahtani, T.M., Yusop, F.D., & Halili, S.H. (2023). Content validity of the constructivist learning in higher education settings (CLHES) scale in the context of the flipped classroom in higher education. *Humanities and Social Sciences Communications*, *10*(1). https://doi.org/10.1057/s41599-023-01754-3

Aromataris, E., Fernandez, R., Godfrey, C.M., Holly, C., Khalil, H., & Tungpunkom, P. (2015). Summarizing systematic reviews: Methodological development, conduct and reporting of an umbrella review approach. *International Journal of Evidence-based Healthcare*, *13*(3), 132–140. https://doi.org/10.1097/xeb.0000000000000055

Aromataris, E., Fernandez, R., Godfrey, C., Holly, C., Khalil, H., & Tungpunkom, P. (2020). Chapter 10: Umbrella reviews. *JBI Reviewer's Manual*. https://doi.org/10.46658/jbirm-17-08

Becker, L.A., & Oxman, A.D. (2008). Overviews of reviews. *Cochrane Handbook for Systematic Reviews of Interventions*, 607-631. https://doi.org/10.1002/9780470712184.ch22

Bengtsson, M. (2016). How to plan and perform a qualitative study using content analysis. *NursingPlus Open, 2*, 8-14. https://doi.org/10.1016/j.npls.2016.01.001

Bergmann, J., & Sams, A. (2012). *Flip your classroom: Reach every student in every class every day*. International society for technology in education.

Benzies, K.M., Premji, S., Hayden, K.A., & Serrett, K. (2006). State-of-the-evidence reviews: Advantages and challenges of including grey literature. *Worldviews on Evidence-Based Nursing, 3*(2), 55–61. https://doi.org/10.1111/j.1741-6787.2006.00051.x

Bialik, M., Fadel, C., Trilling, B., Nilsson, P., & Groff, J. (2015). *Skills for the 21st century: What should students learn*. Centre for Curriculum Redesign.

Biesta, G. (2010). *Good education in an age of measurement: Ethics, politics, democracy.* Paradigm Publishers.

Birgili, B., Seggie, F.N., & Oğuz, E. (2021). The trends and outcomes of flipped learning research between 2012 and 2018: A descriptive content analysis. *Journal of Computers in Education, 8*(3), 365-394. https://doi.org/10.1007/s40692-021-00183-y

Bishop, J., & Verleger, M. (2013). Testing the flipped classroom with model-eliciting activities and video lectures in a mid-level undergraduate engineering course. *In 2013 IEEE Frontiers in Education Conference (FIE)* (pp. 161–163). https://doi.org/10.1109/fie.2013.6684807

Bloom, B.S. (Ed.), Engelhart, M.D., Furst, E.J., Hill, W.H., & Krathwohl, D.R. (1956). *Taxonomy of educational objectives: The classification of educational goals. Handbook 1: Cognitive domain*. David McKay.

Borenstein, M., Hedges, L.V., Higgins, J.P., & Rothstein, H.R. (2009). *Introduction to meta-analysis*. John Wiley & Sons. https://doi.org/10.1002/9780470743386

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology, 3*(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

Campbell, M., Katikireddi, S.V., Sowden, A., & Thomson, H. (2019). Lack of transparency in reporting narrative synthesis of quantitative data: A methodological assessment of systematic reviews. *Journal of Clinical Epidemiology, 105*,1-9. https://doi.org/10.1016/j.jclinepi.2018.08.019

Care, E. (2017). 21st century skills: From theory to action. In E. Care, P. Griffin, & M. Wilson (Eds.), *Assessment and teaching of 21st century skills: Research and applications* (pp. 3–17). Springer.

Castro-Gil, R., & Correa, D. (2021). Transparency in previous literature reviews about blended learning in higher education. *Education and Information Technologies, 26*(3), 3399-3426. https://doi.org/10.1007/s10639-020-10406-x

Chen, H.J., Chen, C.H., & Wu, W.C.V. (2025). Effects of flipped learning on language learning outcomes: A meta-analysis investigating moderators. *SAGE Open, 15*(2). https://doi.org/10.1177/21582440251331298

Clark, L.A., & Watson, D. (2019). Constructing validity: New developments in creating objective measuring instruments. *Psychological Assessment, 31*(12), 1412-1427. https://doi.org/10.1037/pas0000626

Coman, L., & Richardson, J. (2006). Relationship between self-report and performance measures of function: A systematic review. *Canadian Journal on Aging / La Revue Canadienne Du Vieillissement, 25*(3), 253–270. https://doi.org/10.1353/cja.2007.0001

Creswell, J.W., & Plano Clark, V.L. (2017). *Designing and conducting mixed methods research (3rd ed.)*. SAGE Publications.

Easterbrook, P.J., Gopalan, R., Berlin, J.A., & Matthews, D.R. (1991). Publication bias in clinical research. *The Lancet, 337*(8746), 867-872. https://doi.org/10.1016/0140-6736(91)90201-y

Egger, M., Zellweger-Zähner, T., Schneider, M., Junker, C., Lengeler, C., & Antes, G. (1997). Language bias in randomised controlled trials published in English and German. *The Lancet, 350*(9074), 326-329. https://doi.org/10.1016/S0140-6736(97)02419-7

Ekinci, F., Bektaş, O., Karaca, M., & Yiğit, K.N. (2023). The flipped learning perception scale: A validity and reliability study. *Education and Information Technologies, 28*(11), 14139–14166. https://doi.org/10.1007/s10639-023-11731-7

Francis, N., Morgan, A., Holm, S., Davey, R., Bodger, O., & Dudley, E. (2020). Adopting a flipped classroom approach for teaching molar calculations to biochemistry and genetics students. *Biochemistry and Molecular Biology Education, 48*(3), 220-226. https://doi.org/10.1002/bmb.21328

Franco, A., Malhotra, N., & Simonovits, G. (2014). Publication bias in the social sciences: Unlocking the file drawer. *Science, 345*(6203), 1502-1505. https://doi.org/10.1126/science.1255484

Fry, H., Ketteridge, S., & Marshall, S. (2009). *A handbook for teaching and learning in higher education: Enhancing academic practice*. Routledge. https://doi.org/10.4324/9781315763088

Grégoire, G., Derderian, F., & Le Lorier, J. (1995). Selecting the language of the publications included in a meta-analysis: Is there a tower of babel bias? *Journal of Clinical Epidemiology, 48*(1), 159–163. https://doi.org/10.1016/0895-4356(94)00098-b

Goedhart, N.S., Westrhenen, N.B., Möser, C., & Zweekhorst, M. (2019). The flipped classroom: Supporting a diverse group of students in their learning. *Learning Environments Research, 22*(2), 297-310.

Grant, M.J., & Booth, A. (2009). A typology of reviews: An analysis of 14 review types and associated methodologies. *Health Information and Libraries Journal, 26*(2), 91-108. https://doi.org/10.1111/j.1471-1842.2009.00848.x

Greenhalgh, T., & Peacock, R. (2005). Effectiveness and efficiency of search methods in systematic reviews of complex evidence: Audit of primary sources. *British Medical Journal, 331*(7524), 1064–1065. https://doi.org/10.1136/bmj.38636.593461.68

Harrison, R., Jones, B., Gardner, P., & Lawton, R. (2021). Quality assessment with diverse studies (QuADS): An appraisal tool for methodological and reporting quality in systematic reviews of mixed- or multi-method studies. *BMC Health Services Research, 21*(1). https://doi.org/10.1186/s12913-021-06122-y

Hartling, L., Vandermeer, B., & Fernandes, R.M. (2014). Systematic reviews, overviews of reviews and comparative effectiveness reviews: A discussion of approaches to knowledge synthesis. *Evidence-Based Child Health: A Cochrane Review Journal, 9*(2), 486–494. https://doi.org/10.1002/ebch.1968

Hartling, L., Featherstone, R., Nuspl, M., Shave, K., Dryden, D.M., & Vandermeer, B. (2017). Grey literature in systematic reviews: A cross-sectional study of the contribution of non-English reports, unpublished studies and dissertations to the results of meta-analyses in child-relevant reviews. *BMC Medical Research Methodology, 17*(1). https://doi.org/10.1186/s12874-017-0347-z

Helbach, J., Pieper, D., Mathes, T., Rombey, T., Zeeb, H., Allers, K., & Hoffmann, F. (2022). Restrictions and their reporting in systematic reviews of effectiveness: An observational study. *BMC Medical Research Methodology, 22*(1). https://doi.org/10.1186/s12874-022-01710-w

Hew, K.F., Bai, S., Huang, W., Dawson, P., Du, J., Huang, G., & Thankrit, K. (2021). On the use of flipped classroom across various disciplines: Insights from a second-order meta-analysis. *Australasian Journal of Educational Technology, 37*(2), 132-151.

Higgins, J.P.T., Thomas, J., Chandler, J., Cumpston, M., Li, T., Page, M.J., & Welch, V.A. (Eds.). (2023). *Cochrane handbook for systematic reviews of interventions (Version 6.4)*. Cochrane. www.training.cochrane.org/handbook

Holly, C., Salmond, S., & Saimbert, M. (Eds.). (2017). *Comprehensive systematic review for advanced practice nursing*. https://doi.org/10.1891/9780826131867

Hopewell, S., McDonald, S., Clarke, M.J., & Egger, M. (2007). Grey literature in meta-analyses of randomized trials of health care interventions. *Cochrane Database of Systematic Reviews, (2)*, MR000010. https://doi.org/10.1002/14651858.MR000010.pub3

Huang, H.L., Chou, C.P., Leu, S., You, H.L., Tiao, M.M., & Chen, C.H. (2020). Effects of a quasi-experimental study of using flipped classroom approach to teach evidence-based medicine to medical technology students. *BMC Medical Education, 20*(1), https://doi.org/10.1186/s12909-020-1946-7

Joseph, J., Firmin, S., Oseni, T., & Stranieri, A. (2023). Decoding employee ambidexterity: Understanding drivers, constraints, and performance implications for thriving in the evolving work landscapes-A scoping review. *Heliyon, 9*(12), e22493. https://doi.org/10.1016/j.heliyon.2023.e22493

Jüni, P., Holenstein, F., Sterne, J., Bartlett, C., & Egger, M. (2002). Direction and impact of language bias in meta-analyses of controlled trials: Empirical study. *International Journal of Epidemiology, 31*(1), 115-123. https://doi.org/10.1093/ije/31.1.115

Kager, K., Kalinowski, E., Jurczok, A., & Vock, M. (2024). A systematic review of transparency in lesson study research: How do we report on the observation and reflection stages? *Frontiers in Education*, 9. https://doi.org/10.3389/feduc.2024.1322624

Kamei, F., Pinto, G., Wiese, I., Ribeiro, M., & Soares, S. (2021). What evidence we would miss if we do not use grey literature? *Proceedings of the 15th ACM / IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM), 1-11*. https://doi.org/10.1145/3475716.3475777

Kapur, M., Hattie, J., Grossman, I., & Sinha, T. (2022). Fail, flip, fix, and feed – Rethinking flipped learning: A review of meta-analyses and a subsequent meta-analysis. *Frontiers in Education*, 7. https://doi.org/10.3389/feduc.2022.956416

Krathwohl, D.R., Bloom, B.S., & Masia, B.B. (1964). *Taxonomy of educational objectives: The classification of educational goals. Handbook II: Affective domain*. David McKay.

Kreitchmann, R.S., Abad, F.J., Ponsoda, V., Nieto, M.D., & Morillo, D. (2019). Controlling for response biases in self-report scales: Forced-choice vs. psychometric modeling of Likert items. *Frontiers in Psychology*, 10. https://doi.org/10.3389/fpsyg.2019.02309

Kung, J., Chiappelli, F., Cajulis, O.O., Avezova, R., Kossan, G., Chew, L., & Maida, C.A. (2010). From systematic reviews to clinical recommendations for evidence-based health care: validation of revised assessment of multiple systematic reviews (R-AMSTAR) for grading of clinical relevance. *The Open Dentistry Journal, 4*(1), 84-91. https://doi.org/10.2174/1874210601004010084

Kyngäs, H., Kääriäinen, M., & Elo, S. (2020). The trustworthiness of content analysis. In Kyngäs, H., Mikkonen, K., Kääriäinen, M. (Eds.), *The application of content analysis in nursing science research* (pp. 41– 48). Springer. https://doi.org/10.1007/978-3-030-30199-6_5

Larsen, J.N., Maxwell, K.M.D., & Khalil, M. (2025). Exploring the link between motivational regulation and learning design with learning analytics. *Journal of Learning Analytics, 12*(1), 215–237. https://doi.org/10.18608/jla.2025.8469

Leatherman, J.L., & Cleveland, L.M. (2019). Student exam performance in flipped classroom sections is similar to that in active learning sections, and satisfaction with the flipped classroom hinges on attitudes toward learning from videos. *Journal of Biological Education, 54*(3), 328–344. https://doi.org/10.1080/00219266.2019.1575266

Lewis, C.C., Boyd, M.R., Walsh-Bailey, C., Lyon, A.R., Beidas, R., Mittman, B., Aarons, G.A., Weiner, B.J., & Chambers, D.A. (2020). A systematic review of empirical studies examining mechanisms of implementation in health. *Implementation Science, 15*(1). https://doi.org/10.1186/s13012-020-00983-3

Lo, C.K. (2020). Systematic reviews on flipped learning in various education contexts. In Zawacki-Richter, O., Kerres, M., Bedenlier, S., Bond, M., Buntins, K. (eds) *Systematic reviews in educational research*. Springer. https://doi.org/10.1007/978-3-658-27602-7_8

Luchini, C., Veronese, N., Nottegar, A., Shin, J.I., Gentile, G., Granziol, U., Soysal, P., Alexinschi, O., Smith, L., & Solmi, M. (2021). Assessing the quality of studies in meta-research: Review/guidelines on the most important quality assessment tools. *Pharmaceutical Statistics, 20(*1), 185–195. https://doi.org/10.1002/pst.2068

Luo, S., & Gan, Z. (2022). Validation of a measure of flipped English learning readiness and examination of its relationships with instructional practices, learning self-efficacy, and learning beliefs. *Frontiers in Psychology*, 13. https://doi.org/10.3389/fpsyg.2022.846781

Mesquida, C., Warne, J., & Lakens, D. (2025). A scoping review of the transparency, reporting practices and methodological rigor of meta-analyses published in Sports Medicine. https://doi.org/10.51224/srxiv.576

Moher, D., Liberati, A., Tetzlaff, J., Altman, D.G., & PRISMA Group (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *PLoS Medicine, 6*(7), e1000097. https://doi.org/10.1371/journal.pmed.1000097

Moher, D., Shamseer, L., Clarke, M., Ghersi, D., Liberati, A., Petticrew, M., Shekelle, P., Stewart, L.A., & PRISMA-P Group. (2015). Preferred reporting items for systematic review and meta-analysis

protocols (PRISMA-P) 2015 statement. *Systematic Reviews, 4*(1), 1. https://doi.org/10.1186/2046-4053-4-1

Murphy, B.A., & Hall, J.A. (2024). How a strong measurement validity review can go astray: A look at Higgins *et al.* (2024) and recommendations for future measurement-focused reviews. *Clinical Psychology Review*, 114, 102506. https://doi.org/10.1016/j.cpr.2024.102506

Neuendorf, K.A. (2018). Content analysis and thematic analysis. *Advanced Research Methods for Applied Psychology*, 211–223. https://doi.org/10.4324/9781315517971-21

Ortega, A., Lopez-Briz, E., & Fraga-Fuentes, M.D. (2016). From qualitative reviews to umbrella reviews. In Biondi-Zoccai, G. (Ed.), *Umbrella reviews* (pp. 21-41). Springer. https://doi.org/10.1007/978-3-319-25655-9_3

Ozenen, G. (2023). Self-assessment and learning outcome evaluation of interior architecture students using flipped versus traditional classroom education models. *SAGE Open, 13*(4). https://doi.org/10.1177/21582440231209891

Page, M.J., McKenzie, J.E., Bossuyt, P.M., Boutron, I., Hoffmann, T.C., Mulrow, C.D., Shamseer, L., Tetzlaff, J.M., Akl, E.A., Brennan, S.E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M.M., Li, T., Loder, E.W., Mayo-Wilson, E., McDonald, S., ... Moher, D. (2021). The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ, 372*, n71. https://doi.org/10.1136/bmj.n71

Paré, G., Tate, M., Johnstone, D., & Kitsiou, S. (2016). Contextualizing the twin concepts of systematicity and transparency in information systems literature reviews. *European Journal of Information Systems, 25*(6), 493–508. https://doi.org/10.1057/s41303-016-0020-3

Peters, M.D., Marnie, C., Tricco, A.C., Pollock, D., Munn, Z., Alexander, L., ... & Khalil, H. (2020). Updated methodological guidance for the conduct of scoping reviews. *JBI evidence synthesis, 18*(10), 2119-2126. https://doi.org/10.11124/jbies-20-00167

Pham, B., Klassen, T.P., Lawson, M.L., & Moher, D. (2005). Language of publication restrictions in systematic reviews gave different results depending on whether the intervention was conventional or complementary. *Journal of Clinical Epidemiology, 58*(8), 769-776.e2. https://doi.org/10.1016/j.jclinepi.2004.08.021

Prevalla, B., & Uzunboylu, H. (2019). Flipped learning in engineering education. *TEM Journal, 8*(2), 446-450. https://doi.org/10.18421/tem82-46

Ramma, Y., Bholoa, A., Jawaheer, S., Gunness, S., Wah, H.T.Y.L.K., Gopee, A.K., & Authelsingh, D. (2021). Teaching and learning science in the 21st century: A study of critical thinking of learners and associated challenges. In J. Naidoo (Ed.), *Teaching and learning in the 21st century* (pp.139–156). https://doi.org/10.1163/9789004460386_009

Reinhold, F., Leuders, T., Loibl, K., Nückles, M., Beege, M., & Boelmann, J.M. (2024). Learning mechanisms explaining learning with digital tools in educational settings: A cognitive process framework. *Educational Psychology Review, 36*(1). https://doi.org/10.1007/s10648-024-09845-6

Robertson, W.H. (2022). The constructivist flipped classroom. *The Journal of College Science Teaching, 52*(2). https://doi.org/10.1080/0047231x.2022.12290644

Roehling, P.V. (2018). *Flipping the college classroom: An evidence-based guide*. Palgrave. https://doi.org/10.1007/978-3-319-69392-7

Rutkienė, A., Kaçar, I.G., Karakuş, E., Baltaci, H.S., Altun, M., Şahintaş, Z.A., Barendsen, R. M., Wierda, R., & García, B.C. (2022). The impact of flipped learning on students' engagement and satisfaction development: A cross-country action research study. *Pedagogika, 147*(3), 201-220. https://doi.org/10.15823/p.2022.147.12

Saldaña, J. (2016). *The coding manual for qualitative researchers (3rd ed.)*. Sage Publications.

Sancar-Tokmak, H., & Dagli, Z. (2025). A systematic review of theoretical foundations and learning effects in gamified flipped classroom research. *Evaluation Review, 49*(5), 880–913. https://doi.org/10.1177/0193841x251320438

Schmucker, C.M., Blümle, A., Schell, LK., Schwarzer, G., Oeller, P., Cabrera, L., von Elm, E., Briel, M., & Meerpohl, J.J. (2017). Systematic review finds that study data not published in full text articles have unclear impact on meta-analyses results in medical research. *PLOS ONE, 12*(4), e0176210. https://doi.org/10.1371/journal.pone.0176210

Shea, B.J., Reeves, B.C., Wells, G., Thuku, M., Hamel, C., Moran, J., Moher, D., Tugwell, P., Welch, V., Kristjansson, E., & Henry, D.A. (2017). AMSTAR 2: A critical appraisal tool for systematic

reviews that include randomised or non-randomised studies of healthcare interventions, or both. *BMJ, 358*, j4008. https://doi.org/10.1136/bmj.j4008

Shih, H.J., & Huang, S.C. (2019). College students' metacognitive strategy use in an EFL flipped classroom. *Computer Assisted Language Learning, 33*(7), 755-784. https://doi.org/10.1080/09588221.2019.1590420

Silver, C., & Lewins, A. (2014). *Using software in qualitative research: A step-by-step guide*. https://doi.org/10.4135/9781473906907

Silverajah, V.S.G., Wong, S.L., Govindaraj, A., Khambari, M.N.M., Rahmat, R.W.B.O.K., & Deni, A.R.M. (2022). A systematic review of self-regulated learning in flipped classrooms: Key findings, measurement methods, and potential directions. *IEEE Access, 10(*08), 20270. https://doi.org/10.1109/ACCESS.2022.3143857

Sterne, J.A.C., Sutton, A.J., Ioannidis, J.P.A., Terrin, N., Jones, D.R., Lau, J., Carpenter, J., Rücker, G., Harbord, R.M., Schmid, C.H., Tetzlaff, J., Deeks, J.J., Peters, J., Macaskill, P., Schwarzer, G., Duval, S., Altman, D.G., Moher, D., & Higgins, J.P.T. (2011). Recommendations for examining and interpreting funnel plot asymmetry in meta-analyses of randomised controlled trials. *BMJ, 343*, d4002. https://doi.org/10.1136/bmj.d4002

Stern, C., & Kleijnen, J. (2020). Language bias in systematic reviews: You only get out what you put in. *JBI Evidence Synthesis, 18*(9), 1818–1819. https://doi.org/10.11124/jbies-20-00361

Taguma, M., Feron, E., & Lim, M.H. (2018). Future of education and skills 2030: Conceptual learning framework. OECD Publishing. Retrieved February 10, 2025, from https://www.oecd.org/education/2030-project/about/documents/

Templier, M., & Paré, G. (2018). Transparency in literature reviews: An assessment of reporting practices across review types and genres in top IS journals. *European Journal of Information Systems, 27*(5), 503–550. https://doi.org/10.1080/0960085x.2017.1398880

Tong, A., Flemming, K., McInnes, E., Oliver, S., & Craig, J. (2012). Enhancing transparency in reporting the synthesis of qualitative research: ENTREQ. *BMC Medical Research Methodology, 12*(1). https://doi.org/10.1186/1471-2288-12-181

Tricco, A.C., Lillie, E., Zarin, W., O'Brien, K.K., Colquhoun, H., Levac, D., ... & Straus, S.E. (2018). PRISMA extension for scoping reviews (PRISMA-ScR): Checklist and explanation. *Annals of Internal Medicine, 169*(7), 467-473. https://doi.org/10.7326/m18-0850

Urcola-Pardo, F., Subiron-Valera, A.B., Anton-Solanas, I., Orkaizagirre-Gomara, A., Torres-Enamorado, D., & Gonzalez-Sanz, J.D. (2024). Design and validation of the flipped-learning assessment scale for undergraduate nursing education. *Nursing Open, 11*(12). https://doi.org/10.1002/nop2.70112

Walpole, S.C. (2019). Including papers in languages other than English in systematic reviews: Important, feasible, yet often omitted. *Journal of Clinical Epidemiology, 111*, 127–134. https://doi.org/10.1016/j.jclinepi.2019.03.004

Wang, Z., Brito, J., Tsapas, A., Griebeler, M., Alahdab, F., & Murad, M.H. (2015). Systematic reviews with language restrictions and no author contact have lower overall credibility: a methodology study. *Clinical Epidemiology, 243*. https://doi.org/10.2147/clep.s78879

Whiting, P., Wolff, R., Mallett, S., Simera, I., & Savović, J. (2017). A proposed framework for developing quality assessment tools. *Systematic Reviews, 6*, 1-9. https://doi.org/10.1186/s13643-017-0604-6

Zeng, X., Zhang, Y., Kwong, J.S.W., Zhang, C., Li, S., Sun, F., Niu, Y., & Du, L. (2015). The methodological quality assessment tools for preclinical and clinical studies, systematic review and meta-analysis, and clinical practice guideline: A systematic review. *Journal of Evidence-Based Medicine, 8*(1), 2–10. https://doi.org/10.1111/jebm.12141

Zhang, H., Mao, R., Huang, H., Dai, Q., Zhou, X., Shen, H., & Rong, G. (2021). Processes, challenges and recommendations of gray literature review: An experience report. *Information and Software Technology, 137*, 106607. https://doi.org/10.1016/j.infsof.2021.106607

Zhu, L., Lian, Z., & Engström, M. (2020). Use of a flipped classroom in ophthalmology courses for nursing, dental and medical students: A quasi-experimental study using a mixed-methods approach. *Nurse Education Today, 85*, 104262. https://doi.org/10.1016/j.nedt.2019.104262

APPENDIX

The Search of the Web of Science Database