IJTE International Journal of Technology in Education
www.ijte.net

A Systematic Literature Review of Automated Feedback Generation in Education

Yajie Song [1], Yimei Zhang [2], Maria Cutumisu [3] *

[1] McGill University. Canada, 0009-0007-5910-6319
[2] McGill University. Canada, 0000-0002-4955-6726
[3] McGill University. Canada, & Mila - Quebec Artificial Intelligence Institute, Montreal, QC, Canada, 0000-0003-2475-9647
* Corresponding author: Maria Cutumisu (maria.cutumisu@mcgill.ca)

Article Info

Article History

Received:
23 July 2025

Revised:
8 December 2025

Accepted:
17 January 2026

Published:
1 April 2026

Keywords

Adapted learning
Automated feedback generation
Systematic literature review
Data-driven education
Feedback
Technology-enhanced learning

Abstract

Feedback that is individualized and immediate is essential to improving learning outcomes but providing it to every learner is difficult. Automatic feedback generation (AFG) aims to alleviate this problem, especially with technology-enhanced learning environments. This systematic literature review of AFG in education, following the PRISMA framework, examines 34 peer-reviewed publications. The findings revealed that the reviewed studies (1) gained momentum after 2019; (2) often used secondary cognitive data to evaluate AFG approaches; (3) mainly targeted the computer science domain; (4) frequently combined multiple methods to generate feedback; (5) employed multiple performance evaluations; and (6) mostly provided written feedback aimed at correcting student errors. This review also highlighted several gaps, including the lack of (1) in-depth cognitive and affective data from user studies to evaluate feedback and understand how students interpret it; (2) research on feedback use and strategies to close the feedback loop; (3) AFG systems for ill-defined domains with strong transferability; (4) elaborated feedback that scaffolds problem-solving rather than providing answers; (5) feedback using multiple modalities and valences; and (6) integration of learning theories in AFG design. This review advances our understanding of current AFG practices, evaluates and extends conceptual frameworks of AFG, and provides insights for future AFG design and evaluation.

Citation: Song, Y., Zhang, Y., & Cutumisu, M. (2026). A systematic literature review of automated feedback generation in education. *International Journal of Technology in Education (IJTE), 9*(2), 512-556. https://doi.org/10.46328/ijte.5347

ISSN: 2689-2758 /

Introduction

Learning is a complex process that requires active construction of knowledge through social interactions with knowledgeable others (Vygotsky, 1978). Effective learning is essential in education because it enables learners to transfer the learned knowledge from the classroom to real-world practice. Among all the interventions that promote learning, feedback is one of the most effective, helping learners address their misconceptions and achieve their intended learning goals (Hattie & Timperley, 2007). Feedback is defined as information that compares a learner's actual performance with a pre-set goal to reinforce learning (Molloy & Boud, 2014). A meta-analysis of 435 feedback studies reveals a medium effect of feedback on student learning achievement (Wisniewski et al., 2019). However, their moderation analysis suggests that the effectiveness of feedback on learning was impacted by the feedback forms and the conveyed information (Wisniewski et al., 2019). For example, Wiener (1954) argued that feedback contributes to learning only when it is used to modify student performance rather than merely to criticize it. Wiener's argument reflects the dichotomy in teachers' and students' perceptions of feedback effectiveness. Specifically, although teachers claim to use feedback strategies effectively to enhance student learning and motivation, empirical data from observations, interviews, and documentary sources do not support these claims (Murtagh, 2014). For example, students reported that instructors often only provide phatic or evaluative feedback that acknowledges receipt of the work and judges the student's work, respectively, without giving constructive guidance for students to follow (Murtagh, 2014). Therefore, feedback should be carefully designed and evaluated to promote effective learning.

Challenges in the Provision of Effective Feedback

One major obstacle to providing effective feedback is the difficulty of scaling instructors' manual efforts to large classes (Kim et al., 2016). As class size increases, instructors struggle to maintain consistency in providing feedback, whereas students report lower satisfaction with the feedback they receive (Sondergaard & Thomas, 2004). Moreover, teacher feedback is often delayed due to limited teaching resources. Concomitantly, students find delayed feedback irrelevant, as the passage of time leads to forgetting the content (Poulos & Mahony, 2008). Even when feedback is timely, students may not use it effectively due to its varied quality and their limited feedback literacy, which hinders them from using feedback to regulate their learning processes (Jonsson, 2013). In summary, providing high-quality, individualized, and immediate feedback has been an ongoing challenge in educational practice (Pardo et al., 2019).

Automated Feedback Generation as a Solution

Providing immediate and individualized feedback is essential to scaffold students' agency to reflect, engage, and regulate their learning processes. To address the challenges of providing and using manual feedback, numerous automated feedback generation (AFG) systems have been developed. Automated feedback (AF) is defined as "automatically generated and delivered by a technology-enhanced learning environment" (Serral Asensio et al., 2019, p. 1). AFG is promising, given the wide use of information and communication technologies in current educational practice. For example, according to the Programme for International Student Assessment 2018 data, only 27.5% of valid responses worldwide indicate that computers are unavailable at home, while 19% indicate

that computers are unavailable at school. Furthermore, new learning technology-based products enable students and teachers to "interact synchronously and asynchronously" (Larrondo et al., 2021, p. 1). Therefore, AFG has the potential to deliver timely, scalable, and personalized feedback that supports student learning across various contexts.

Limitations of Existing AFG Reviews and Contributions of the Current Study

The rapid development of AFG research makes it difficult for researchers and practitioners to identify the most appropriate systems for specific tasks or contexts. Thus, an up-to-date and comprehensive literature review is necessary to support informed decisions in this regard (Lasserson et al., 2019). For this reason, many AFG reviews have been published. However, the existing reviews are limited in scope, often focusing on specific domains and constrained methodological approaches. Regarding the domain, Shadiev and Feng (2023) reviewed studies on AF in language learning. Keuning et al. (2018) focused on AFG tools for programming. Finally, Larrondo et al. (2021) restricted their AFG review to open-ended engineering problems. Regarding the focus areas, Cavalcanti et al. (2021) limited their review to AF in online AFG systems. Also, Buckingham Shum et al. (2023) focused on teacher feedback literacy in the implementation of AFG tools. Huang et al. (2025) provided a broad review of how technology enhances feedback generation, delivery, and usage across diverse domains.

Regarding the methodology, several reviews have imposed restrictions on data sources, search processes, and screening procedures. Most reviews used only one or two databases, except Cavalcanti et al. (2021), who used six. All reported searches applied time-span restrictions. Some reviews did not report inclusion criteria (Buckingham Shum et al., 2023; Larrondo et al., 2021). In addition, some reviews included only studies from high-impact journals (e.g., Shadiev & Feng, 2023) or applied domain-specific inclusion criteria, such as certain classes of programming exercises (Keuning et al., 2018) or fully automated feedback (Deeva et al., 2021). Moreover, existing coding frameworks for evaluating AFG tools also have limitations in scope, design, and application. For example, Keuning et al. (2018) introduced a coding framework specific to programming with some highly specialized coding categories, and investigated whether AFG systems supported program transformations between programming languages. Deeva et al. (2021) proposed the systematic TAF-ClaF (Technologies for Automated Feedback - Classification Framework) based on data from one database with a time restriction. The completeness of the TAF-ClaF deserves further investigation. Thus, the current review will evaluate several subcategories of this framework. Finally, Serral Asensio et al. (2019) proposed a framework to capture many characteristics of AFG systems, but they provided only two coded examples based on their framework. The effectiveness and completeness of their framework also need further evaluation.

Despite these limitations, previous reviews offer useful contributions. For example, Shadiev and Feng (2023) reported the accuracy rates for each of the reviewed AFG systems. Keuning et al. (2018) investigated whether each of the reviewed AFG systems considered the student model when adapting the feedback. Deeva et al. (2021) focused on the domain model and expert knowledge when examining how those systems generate AF. Serral Asensio et al. (2019) noted the importance of examining whether the AFG system supported learner control. Buckingham Shum et al. (2023) emphasized the importance of teacher control in the design of AFG systems. However, one critical issue that has largely been overlooked in previous AFG reviews is ethics. Most studies did

not discuss how educational technologies may reinforce educational inequalities (Martínez-Alemán et al., 2015). Identifying, understanding, and addressing ethical issues is critical and pressing in research, especially when introducing a new technology-based tool into educational practice. Regan and Jesse (2019) identified several ethical issues arising from the explosion of educational technologies, including information privacy violations, hampered student autonomy through personalized learning, contentious ownership of information, and algorithm discrimination that learned to classify students based on their race, gender, and socioeconomic status. Discussing and addressing these issues is critical to guaranteeing the effective application of the newly introduced educational technologies. Therefore, the current review will also aim to understand how each included study discusses the relevant ethical issues.

The Current Study

The current review differs from previous work in several important ways. First, it employs four databases, whereas most previous reviews used fewer. Second, it does not impose a time restriction on the search, unlike all previous reviews, which applied a specific time span. Third, it does not set specific restrictions on the characteristics of AFG systems during screening, whereas previous reviews applied several exclusion criteria. Fourth, it codes the included studies using the well-established Message, Implementation, Student, Context, and Agents (MISCA) framework, which is introduced below. Finally, it identifies the essential requirements for effective AFG systems, including context characteristics, feedback design and evaluation, feedback characteristics, and related ethical issues, which are areas that were not fully addressed in previous reviews.

Research Questions

The current review aims to summarize the state of the art in AFG research in education and identify research gaps. Thus, it is guided by the following research questions:

1. What are the context characteristics of existing AFG systems in education?
2. What are the foundational designs and evaluations of AFG systems in education?
3. What type of feedback is provided by AFG systems in education?
4. What ethical issues are identified or need to be addressed in educational AFG systems?

The rest of the current review is organized as follows. First, the theoretical framework guiding the analysis, interpretation, and discussion of the included studies is introduced. Second, the applied systematic literature review methodology is outlined, followed by the results that address the four research questions. Then, the findings are discussed and interpreted. Finally, implications and suggestions for future AFG studies are provided.

Theoretical Framework

As an educational intervention and research topic, feedback has significantly evolved since Thorndike (1927) proposed the Law of Effect. Grounded in behaviorist theory, Thorndike's Law of Effect states that external positive or negative reinforcement shapes learning (Thorndike, 1927). Building on this behaviorist foundation, several influential feedback frameworks and models emerged, including Sadler's (1989) seminar work in

formative assessment and Kluger and DeNisi's (1996) feedback intervention theory. Sadler (1989) outlined three conditions to make feedback effective: learners need to 1) set a learning aim/standard; 2) compare their actual performance with the learning aim/standard; and 3) take actions to reduce the gap between their actual performance and the learning aim/standard. Kluger and DeNisi (1996) also discussed ways to support feedback effectiveness through feedback content. Specifically, they stated that feedback should guide learners to focus their attention on how to solve the problem rather than on their emotions.

In the late 1950s, cognitive theory became the main focus of learning theories. It emphasized the complex cognitive processes involved in learning (Ertmer & Newby, 2013), which also informed several feedback theories or models, such as Kulhavy and Stock's (1989) three-cycle feedback model, Bangert-Drowns et al.'s (1991) five-stage model, and Butler and Winne's (1995) self-regulated learning (SRL) theory. In the three-cycle feedback model, Kulhavy and Stock (1989) stated that feedback improved student learning in three cycles: 1) learners respond to a task using their prior knowledge; 2) learners receive and process feedback to revise their responses; and 3) learners compare their revised answers to the task criteria until they reach the task criteria. Bangert-Drowns et al. (1991) developed a five-stage feedback model to explain the feedback process, showing some similarities to Kulhavy and Stock's (1989) three-cycle feedback model. Specifically, the five stages include the learners': 1) initial knowledge state; 2) retrieval or search of their initial knowledge state when solving the task; 3) generation of their responses to the task; 4) evaluation of their responses by comparing them with the received feedback; and 5) revising of their initial knowledge state based on the evaluation (Bangert-Drowns et al., 1991). Butler and Winne (1995) explained the feedback process from the perspective of SRL, which contains: 1) internal feedback that stems from learners monitoring their own task-solving strategies and progress; and 2) external feedback that stems from outside (e.g., teachers, peers, or AFG systems) to guide learners in improving their task-solving knowledge, strategies, and performance.

In contrast to behaviorism and cognitivism, which assume that knowledge exists objectively outside the learner, constructivism assumes that knowledge is the meaning learners construct from their own experiences (Ertmer & Newby, 2013). Constructivism informs several feedback theories or models, including Evans's (2013) feedback landscape model and Carless and Boud's (2018) feedback literacy model. Evans (2013) stated that feedback effectiveness is moderated by 12 variables shared by learners and instructors (i.e., ability, personality, gender, culture, social and cultural capital, prior learning experiences, attributions, perceived task support, capacity to navigate learning communities, beliefs about learning, cognitive styles, and perceived roles within learning communities). The effectiveness of feedback is also mediated by three instructor-specific variables: awareness of students' other academic contexts, alignment with other modules, and knowledge of students and the extent of adaptation. Carless and Boud (2018) proposed that students develop feedback literacy through four interrelated processes: 1) appreciating the value of feedback; 2) making judgments based on received feedback and self-assessment; 3) managing the emotions that feedback may trigger; and 4) taking action based on the feedback.

Message, Implementation, Student, Context, and Agents (MISCA) Feedback Framework

Although the above overview includes only a few feedback theories and models, it nonetheless illustrates the wide

range of theoretical frameworks available to researchers and instructors. Lipnevich and Panadero (2021) systematically selected and reviewed 14 feedback theories and models to facilitate the selection of the most appropriate feedback theory or model for researchers and practitioners seeking to understand feedback in educational settings. They examined each model with respect to its definition of feedback, theoretical framework, model description, pictorial representation, empirical evidence, and feedback elements, including the information contained, the gap between actual performance and target, the underlying influential process, the agents involved, and the learners' internal process of feedback. Based on these review results, Panadero and Lipnevich (2022) proposed an integrative model of feedback, the MISCA framework, which organizes feedback around five key components: message, implementation, student, context, and agent.

Message

First, the message component of the MISCA framework refers to the information or content communicated to the learner (Panadero & Lipnevich, 2022). For example, Kulhavy and Stock (1989) classified feedback into two types: verification and elaboration. Verification provides a simple judgment of whether the student's answer is correct or incorrect, whereas elaboration offers additional information beyond a dichotomous judgment. Moreover, Kulhavy and Stock (1989) further classified elaboration feedback into three dimensions: type (e.g., process-related or evaluation feedback), form (e.g., written or audio feedback), and load (e.g., a simple score, a brief explanation of an error, or detailed narrative feedback of explanations and suggestions). Concomitantly, Bangert-Drowns et al. (1991) classified feedback into three categories: intentionality (e.g., formal feedback used in instruction or informal feedback that still affects learning), target (e.g., feedback aimed at influencing affect, supporting SRL, or correcting student knowledge), and content (e.g., load, form, and type). The sub-categories of feedback content share the same definition as those defined by Kulhavy and Stock (1989) for elaboration feedback. As various feedback models have proposed different ways to classify feedback messages, Panadero and Lipnevich (2022) proposed a comprehensive classification that included four components: verification, valence, load, and type of information. Three components (verification, load, and type of information) were introduced above. The fourth component, valence, refers to whether the content, the expected outcome, and the emotion triggered by feedback are positive or negative (Panadero & Lipnevich, 2022). Moreover, the type of information can contain many sub-categories, such as knowledge of results feedback versus cognitive feedback (Butler & Winne, 1995), rewarding versus punishing feedback (Tunstall & Gsipps, 1996), response-continent versus bug-related feedback (Mason & Bruning, 2001), knowledge of task constraints versus knowledge about concepts (Narciss, 2008).

Implementation

Second, the implementation component of the MISCA framework indicates 1) the function of the feedback and 2) how learners process the feedback they receive (Panadero & Lipnevich, 2022). Based on their analysis of 14 feedback models or theories, Panadero and Lipnevich (2022) classified feedback functions into three types: learning/performance (i.e., feedback aims to enhance student learning or performance), motivation/affect (i.e., feedback aims to increase positive emotions and reduce negative ones), and SRL (i.e., feedback aims to support learners in regulating their behavior, emotion, metacognition, and motivation). Regarding the internal processing

of feedback, Panadero and Lipnevich (2022) positioned six models that addressed how learners internally process feedback on a continuum. On the left, models such as Kluger and DeNisi (1996) and Kulhavy and Stock (1989) focus on feedback as the trigger for learner action. On the right, models such as those from Bangert-Drowns et al. (1991) and Butler and Winne (1995) emphasize the learner, viewing feedback as one of many elements processed during task completion. In the middle, models such as Lipnevich et al. (2016) and Narciss (2008) assign equal importance to both feedback and learner characteristics.

Student

Third, the student component, which is also central to the MISCA framework, indicates the impact of student characteristics on feedback effectiveness (Panadero & Lipnevich, 2022). It is suggested that feedback should be tailored to individual differences, such as cognitive ability, prior knowledge, SRL skills, and personality traits, to maximize its effectiveness (Panadero & Lipnevich, 2022).

Context

Fourth, the instructional context component of the MISCA framework indicates the conditions or settings in which feedback is provided, which should be conducive to achieving its intended effectiveness (Panadero & Lipnevich, 2022). The instructional context component contains two parts: 1) the pedagogical approach of feedback and 2) the presentation or delivery format of the feedback (Panadero & Lipnevich, 2022). Regarding the pedagogical approach of feedback, Panadero and Lipnevich (2022) positioned six models along a continuum based on how they addressed the implementation and delivery of feedback. On the left of the continuum, Carless and Boud (2018), as well as Sadler (1989) discussed general pedagogical principles for implementing and delivering feedback, such as the conditions required for feedback to be effective, as previously discussed. On the right of the continuum, Mason and Bruning (2001) provided a list of specific pedagogical suggestions for implementing and delivering feedback, focusing on concrete guidelines tailored to learner characteristics and instructional settings. For example, they recommended using response-contingent feedback to explain why a response is incorrect (Mason & Bruning, 2001). In the middle of the continuum, Evans (2013), Hattie and Timperley (2007), and Nicol and Macfarlane-Dick (2006) proposed models that are neither overly general nor highly specific to guide feedback pedagogy practice. For example, Nicol and Macfarlane-Dick (2006) outlined seven principles for delivering feedback to enhance SRL. Hattie and Timperley (2007) stated that instructors should address the three questions to make feedback effective: Where am I going? How am I going? And where to next? Regarding the presentation or delivery format of the feedback, Panadero and Lipnevich (2022) adopted the categories proposed by Narciss and Huth (2004), which include: (1) immediate versus delayed feedback, (2) single versus multiple feedback, (3) adaptive versus nonadaptive feedback, and (4) unimodal versus multimodal formats (e.g., written or audio).

Agent

Finally, the agent component of the MISCA framework specifies the source of feedback and how different agents interact when delivering it (Panadero & Lipnevich, 2022). Regarding the agents, Panadero and Lipnevich (2022)

identified four main sources of feedback generation: teachers, peers, computers, and the learners themselves. Moreover, this review focuses on AF generated by computers. Therefore, the interaction among multiple feedback agents is beyond the scope of this review.

Application of the MISCA Framework to This Review

As discussed above, the MISCA framework is a comprehensive model for guiding feedback design and application, serving as the theoretical foundation for the current review's analysis and understanding of the included AFG studies. Specifically, the five components of the MISCA framework (i.e., message, implementation, student, context, and agent) will guide the current review to understand what AF was generated and how it was delivered.

Feedback Message

Following the category of feedback message proposed by Panadero and Lipnevich (2022), the current review applies the following labels to analyze the feedback message in the included studies.

- Verification: feedback that judges the answers as correct or incorrect.
- Valence: feedback that is positive or negative according to its tone or intended outcome.
- Load: the amount of information provided in the feedback, ranging from simple scores to detailed explanations.
- Type of information: as Panadero and Lipnevich (2022) did not provide an exhaustive list for this category, the current review uses thematic coding to identify and categorize the types of information provided in AF.

Feedback Implementation

Feedback Function. Following the MISCA framework, the current review uses the following labels to identify the intended purpose of AF in the included studies.

- Learning/performance: Feedback is designed to enhance learning or performance.
- Motivation/affect: Feedback is designed to increase positive emotions and reduce negative ones.
- SRL: Feedback is designed to support SRL processes.

*Feedback Processing***.** Based on the continuum proposed by Panadero and Lipnevich (2022), the current review classifies the feedback design in each study as one of the following:

- Feedback-centered: Feedback is the primary trigger of learner action.
- Student-centered: Learner characteristics primarily shape how feedback is processed.
- Balanced: Equal emphasis is placed on the feedback message and learner characteristics.

Student Characteristic

The current review uses thematic coding to examine whether the generation of AF considers any student characteristics, such as cognitive ability, prior knowledge, SRL skills, or personality traits, to enhance its effectiveness (Panadero & Lipnevich, 2022).

Feedback Context

Feedback Pedagogy. The current review uses thematic coding to examine whether the generation of AF includes or applies pedagogy guidelines for the implementation or delivery of feedback.

Feedback Presentation. Following the feedback delivery category developed by Narciss and Huth (2004), the review will examine whether the feedback is: 1) immediate or delayed; 2) single or multiple; 3) adaptive or nonadaptive; and 4) unimodal (e.g., written) or multimodal (e.g., written and audio).

Feedback Agent

Although the current review focuses on AF, some studies generate AF through collaboration between computers and human experts (Deeva et al., 2021). Therefore, the agent component of the MISCA framework is retained in this review to capture such collaborative feedback generation.

Methods

The current review adopts a systematic literature review approach to address the research questions. The current review used the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) framework to guide the systematic literature review process (Page et al., 2021). PRISMA outlines three main phases of systematic reviews: identification, screening, and inclusion. In the identification phase, records are collected from each selected database, and then duplicates or ineligible records are removed. In the screening phase, each record is examined based on its abstract and title against the pre-defined exclusion criteria. In the inclusion phase, the full text of each record is reviewed to assess its eligibility for inclusion in the review. The PRISMA process employed in this review is summarized in Figure 1. The web-based systematic review software Covidence (2025) was used to implement the PRISMA process.

Identification of Relevant Records

The current review aimed to provide a comprehensive understanding of AFG systems in education. To guarantee that the search results are relevant, the search string was selected to reflect the main research target: automated feedback generation. Therefore, the final employed search string was: “automated feedback generation” OR “automatic feedback generation.” The search was conducted within the full text of four educational databases without domain or time-span restrictions. In total, we obtained 109 records from the search conducted on May 1,

2025. Figure 1 summarizes the search results per database. After deleting the replication records (N = 40), 69 studies remained for abstract and title screening.

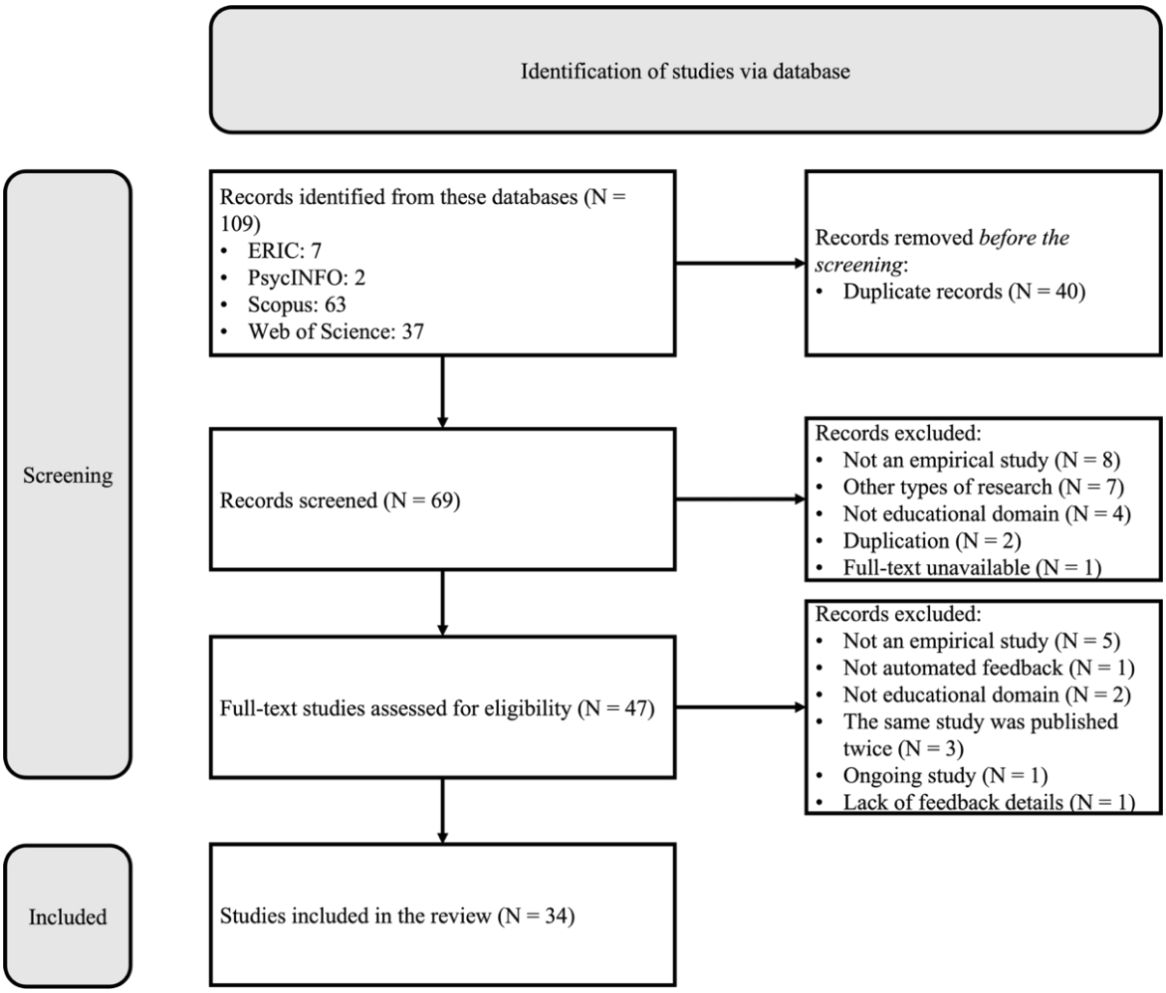

Figure 1. Selection Process Based on the PRISMA Framework

Four databases were selected: Education Resources Information Center (ERIC), PsycINFO, Scopus, and Web of Science. ERIC primarily provides education literature, whereas PsycINFO specializes in psychological literature. ERIC and PsycINFO provide indexes of journals, books, and gray literature, and the latter also stores conference proceedings and multimedia. Scopus and Web of Science provide records in social and computer science, including journals, books, and conference proceedings.

Selection of Records

In this step, the reviewers conducted a few rounds of exclusion of ineligible records in Covidence (2025) based on the pre-defined inclusion and exclusion criteria, as shown in Table 1. In the first round, two authors independently assessed each record's eligibility based on its title and abstract. In the second round, two authors read the full text to determine the final eligibility decision for each record. Figure 1 illustrates the entire process. In the end, 34 records remained for the final coding.

Table 1. Selection Criteria

Selection Criterion	Inclusion	Exclusion
Type of record	Peer-reviewed publications	Other types of research (e.g., gray literature, such as magazines, theses, dissertations, patents, and technical reports)
Domain	Any subject domain in education	Non-educational domains
Recency	If an AFG system was featured in several publications by the same authors, the most recent one was selected	
Whether automated feedback?	Automated process of feedback generation	Non-automated feedback or no details of automated feedback generation
Feedback details	Detailed descriptions of feedback that address multiple dimensions of the MISCA framework	Lack of detailed descriptions of feedback that address multiple dimensions of the MISCA framework
Methodology	Empirical studies (qualitative, quantitative, or mixed methods designs)	Non-empirical studies
Written language	English	Non-English

Data Extraction

This review developed a codebook to systematically encode information from the included records. The abbreviations for each code used in results, tables, and figures are capitalized and provided in parentheses. The codebook was developed based on previous literature and refined by coding ten randomly selected records.

This section summarizes the coding columns for the research questions. The coding strategy was designed to accommodate the varying levels of detail and information across different coding columns; the code for each column is summarized in Table 2. Two reviewers independently coded a random sample of 15% of the studies (N = 6). The interrater reliability between the two reviews was substantial (Cohen's kappa = 0.79; Landis & Koch, 1977). Given the substantial level of agreement, one reviewer independently coded the remaining included studies.

Table 2. Coding Categories for Each Research Question

Research Question	Coding Category	Coding Label
1. What are the context characteristics of	Author country/region	Open coding
	Publication year	Open coding

Research Question	Coding Category	Coding Label
existing AFG systems in education?	Publication venue	Journal Article (JA), Conference Proceedings (CP), Book Chapter (BC)
	Data type	AFF; COG
	Data source	FIR; SEC
	Educational domain	11, 14, 16, 23, 27, 40, 42, 51, 52
2. What are the foundational designs and evaluations of AFG	AFG System	Open coding
	Method for generating AF	COM; DAT; MLE; NLP; RUL; TEM
	Evaluation of AFG system	BEH; CON; EXP; MEA; SUR; TES
3. What type of feedback is provided by AFG systems in education?	Feedback message	VER; VAL LOA-L; LOA-H INF-PER; INF-COR; INF-TAS; INF-CON; INF-MIS; INF-PRO
	Feedback function	LEA; MOT; SRL
	Feedback processing	FEE; STU; BAL
	Student characteristic of feedback	Open coding
	Feedback pedagogy	Open coding
	Feedback presentation	IMM; DEL; SIN; MUL; ADA; NON; UNI; MULTI
	Feedback agent	COMP; HUM
4. What ethical issues are identified or need to be addressed in educational AFG systems?	Whether it identified ethical issues	Y; N
	Whether it needs to consider ethical issues	Y; N

What are the Context Characteristics of Existing AFG Systems in Education?

Data Type

Many included records in this review used first- or second-hand data reflecting participants' cognitive (COG)

knowledge to evaluate the built AFG system or method. Cognitive data are often obtained from specific tasks, such as electricity questions (Dzikovska et al., 2014), linked list items (Fossati et al., 2015), programming tasks (Kim et al., 2016), and essays (Lu & Cutumisu, 2021). Several included studies also used surveys or interviews to collect affective (AFF) data that reflect participants' perceptions, opinions, and attitudes toward received AF. For example, Fossati et al. (2015) used a survey to understand learners' experiences with the provided feedback, including their perceptions of its helpfulness and whether using the AFG system was interesting.

Data Source

The classification of first-hand (FIR) and second-hand (SEC) was used to understand how the included studies obtained their data. First-hand (FIR) data means the researchers collected data for their study themselves, whereas second-hand (SEC) data means the researchers used existing public data in their studies.

Educational Domain

In terms of the educational domain, this review adopted the Classification of Instructional Programs Canada 2021 (CIP), created and maintained by the Department of Education in Canada and the United States, which provides a "standard classification system for all post-secondary degree programs" (Rose, 2023, p. 1). The code used in the current review includes 11 (Computer and information sciences and support services), 14 (Engineering), 16 (Indigenous and foreign languages, literature, and linguistics), 23 (English language and literature/letters), 27 (Mathematics and statistics), 40 (Physical sciences), 42 (Psychology), 51 (Health professions and related programs), and 52 (Business, management, marketing, and related support services).

What are the Foundational Designs and Evaluations of AFG Systems in Education?

Method for Generating AF

Previous reviews of AF used different classification frameworks when summarizing the feedback generation model, including data-driven, expert-driven, and mixed (Deeva et al., 2021); model tracing, constraint-based modeling, and historical student data (Keuning et al., 2018); as well as comparisons with desired solution, dashboard, Natural Language Processing (NLP), ontology, graphs, and neural network (Cavalcanti et al., 2021).

The current review comprehensively organized and summarized the methods employed in AFG research. First, four approaches (i.e., data-driven, model tracing, historical student data, and graphs) rely on students' problem-solving process data to generate feedback. Thus, this review groups these approaches under the data-driven category. Second, expert-driven approaches could be classified into rule-based (i.e., constraint-based modeling), template-based (i.e., dashboard and ontology), and comparison with desired answers. These subcategories are retained to preserve the fine-grained distinctions among expert-based methods. Third, the neural network approach was renamed to 'other machine learning methods' to include machine learning algorithms beyond neural networks.

Data-driven methods imply that the AFG model relies on student data to generate feedback. For example, Marwan et al. (2019) used historical student data to identify the correct solution and generate abstract syntax trees (ASTs) for the solutions, which were then used to provide feedback by identifying common nodes between the ASTs and new submissions. Expert-driven methods rely on expert knowledge to identify rules, templates, and correct solutions. For example, Dzikovska et al. (2014) created several tutoring tactics (i.e., rules) and employed different ones based on the diagnosis of the conversation between the tutor and the learner. NLP and other machine-learning methods refer to the methods that use NLP or other algorithms to generate feedback. For example, Dzikovska et al. (2014) used an NLP algorithm to analyze natural language conversations in their system to provide adaptive feedback. In contrast, Bhatia et al. (2018) used a deep neural network to correct students' programming syntactic errors.

The current review aims to extract more detailed information about feedback generation methods and therefore uses the following classifications: Compare with true answers (COM), Data-driven (DAT), Large language model (LLM), Expert-driven rule-based (RUL), Expert-driven template-based (TEM), NLP-based (NLP), and other machine learning-based methods (MLE). It is worth noting that a single study may combine multiple methods to generate AF. For example, Lu and Cutumisu (2021) used an NLP algorithm to augment a feedback template to generate AF. COM indicates that the study compared student answers to pre-defined true answers and used the comparison results to generate feedback. For example, Heo et al. (2023) corrected the student programming submission by comparing it with previously collected true answers.

DAT refers to generating feedback by mining historical data (Deeva et al., 2021). For example, Jia et al. (2022) collected 484 group project reports with instructor-provided feedback for each project. They fed the projects into the language model using supervised learning to provide feedback for unseen data. LLM refers to the use of large language models, such as ChatGPT, to generate AF. For example, Behzad et al. (2024) introduced a corpus of English essays paired with feedback generated by both humans and ChatGPT-4. Their findings suggested that LLM feedback can be irrelevant or inaccurate, prompting them to also release human-generated feedback for comparison. RUL refers to providing feedback by matching learner performance against predefined rules derived from expert knowledge (Deeva et al., 2021). For example, Bhatia et al. (2018) proposed five rewrite rules in the error model that were the basis for correcting student programming submissions.

TEM indicates that expert knowledge was used to derive a template to facilitate the AFG process. For example, Obaido et al. (2020, p. 7) developed templates for providing feedback, such as "Do you mean the {attribute} in the Table 1 | Table 2 | ...| Table N table?" After receiving a learner's input, the value of the {attribute} is filled in and delivered to the learner. NLP and MLE refer to generating feedback based on NLP and other ML algorithms, respectively. As discussed in the literature review section, NLP indicates that AFG systems need to analyze or diagnose learners' natural language elicited to provide tailored feedback. For instance, NLP was used to correct student responses by iteratively replacing the most important words and generating new responses through paraphrasing (Filighera et al., 2022). MLE indicates that all other machine learning algorithms are combined. For example, a supervised neural classifier was used to predict the correctness of the open-ended programming answers by analyzing static source code to detect potential errors, such as incorrect commands, missing function

parameters, or wrong outputs (Vittorini & Galassi, 2023).

Evaluation of Generated AF

Deeva et al. (2021) classified the methods used to evaluate AFG systems into Surveys (SUR), Control group experiments (CON), Comparison of pre-test and post-test (TES), analysis of student Behavior (BEH), and Expert evaluation or comparison with expert solutions (EXP). Moreover, the original definition of SUR (Student surveys) was modified in the current review to include the surveys aimed at instructors when evaluating the AFG system. Besides the existing evaluation methods for generated AF, MEA (i.e., measure the capability of the AFG system) was added because many included studies only evaluate whether the developed system or method could generate feedback for the collected cognitive data. For example, Ahmed et al. (2022) found that their system successfully generated feedback for 58.40% of student submissions.

What Type of Feedback is Provided by AFG Systems in Education?

Feedback Message

Following the categories of feedback messages proposed by Panadero and Lipnevich (2022), the current review applies the following labels to examine the nature of the AF in terms of its content: Verification (VER), Valence (VAL), Load (LOA), and Type of information (INF). As the definitions of these labels have been provided in the theoretical framework section, they will not be repeated here to avoid redundancy. Panadero and Lipnevich (2022) did not provide a specific sub-category for LOA (load) and INF (type of information). To enable a more fine-grained analysis of the included studies, the current review applied low (LOA-L) and high (LOA-H) sub-categories for the LOA label. Specifically, LOA-L refers to feedback that verifies or corrects student responses without providing explanations. For example, Bhatia et al. (2018) offered direct repairs of student programs as feedback without explaining the errors. In contrast, LOA-H refers to feedback that includes both verification and explanation. For example, Dzikovska et al. (2014) designed AF to both verify learners' problem-solving behavior (e.g., "Right, Bulb A is contained in a closed path. Keep trying" [p. 294]) and provide further guidance (e.g., "Here's a hint. Your answer should mention a battery" [p. 294]). Regarding INF (type of information), the current review adopted the classification defined by Narciss (2008), which includes the following categories:

- INF-PER (knowledge of performance): A summary of overall task performance (e.g., 15 correct out of 20 items).
- INF-RES (knowledge of result/response): Indicates whether an individual's answer is correct or incorrect. This category is excluded from the current review as it overlaps with the VER (verification) label in the feedback message.
- INF-COR (knowledge of the correct results): Provides the correct answer.
- INF-TAS (knowledge about task constraints): Explains the rules, conditions, or requirements of the task.
- INF-CON (knowledge about concepts): Offers information about the key concepts involved in the task.
- INF-MIS (knowledge about mistakes): Offers information about student errors, such as their type, location, or cause.

- INF-PRO (knowledge about how to proceed): Provides the procedural guidance or hints for next steps.
- INF-MET (knowledge about metacognition): Encourages reflection on problem-solving. This category is excluded from the current review due to overlapping with the MOT (motivation/affect) and SRL (self-regulated learning) labels under the feedback function.

Feedback Function

Guided by the MISCA framework, the current review uses the following labels to identify the intended purpose of AF in the included studies: Learning/performance (LEA), Motivation/affect (MOT), and self-regulated learning (SRL).

Feedback Processing

Based on the MISCA framework, the current review classifies whether feedback design is 1) feedback-centered (FEE), 2) student-centered (STU), or 3) balanced between feedback and learner characteristics (BAL).

Student Characteristics

Panadero and Lipnevich (2022) did not provide an exhaustive list of student characteristics involved in the design and application of feedback. Thus, the current review uses thematic coding to examine whether the generation of AF takes into account student characteristics (e.g., prior knowledge or SRL skills).

Feedback Pedagogy

The current review uses thematic coding to examine whether the generation of AF follows or incorporates pedagogical guidelines for its implementation or delivery.

Feedback Presentation

Based on the MISCA framework, the review examines whether the feedback is: 1) immediate (IMM) or delayed (DEL); 2) single (SIN) or multiple (MUL); 3) adaptive (ADA) or nonadaptive (NON); and 4) unimodal (UNI) or multimodal (MULTI).

Feedback Agent

If the included records require input from experts or instructors in providing AF, the feedback agent is classified as human (HUM); otherwise, it is classified as computer (COMP). For example, Ahmed et al. (2022) employed the COM (compare with true answer) method to provide AF for programming tasks, which requires experts or instructors to provide reference solutions for the programming tasks used.

What Ethical Issues are Identified or Need to be Addressed in Educational AFG Systems?

Ethical Issues Identified

If the included study discussed the ethical issues in providing AF, code Yes (Y); otherwise, code No (N). For example, Jia et al. (2022) employed data-driven and NLP-based methods in providing AF for the project report and evaluated whether the system would generate improper or offensive language and whether the generated AF contained students' private information.

Ethical Issues to be Addressed

If the included record does not discuss the ethical issue, while the authors of this review believe it is necessary to discuss it, the code is Yes (Y); otherwise, No (N). For example, Obaido et al. (2020) provided AF based on users' voice inputs and collected voice data, but did not discuss data storage or protection.

Results

All the coding results are presented in Table A1 and Table A2 in Appendix A. This section presents the answers to our four research questions.

What are the Context Characteristics of Existing AFG Systems in Education?

Most of the reviewed empirical AFG studies (N = 15; 44%) were authored by researchers from the United States, as shown in Figure 2. Other frequently represented author countries include China, South Korea, Germany, the United Kingdom, and Italy.

Figure 3 illustrates the publication-year trend of the included studies from 2006 to 2024. Over half of the included studies were published after 2019, peaking in 2024 with nine publications, indicating that research on AFG has recently gained significant momentum.

Figure 4 presents the distribution of publication venues among the included studies. Most included AFG studies were published in conference proceedings (N = 23), with fewer published as journal articles (N = 11), indicating that AFG research is more frequently disseminated through conferences.

Figure 5 illustrates the types of data used to evaluate AFG systems in the included studies. Most studies (N = 22) used only cognitive data (e.g., task performance) to evaluate the capability of AFG systems or methods in generating AF. A few studies (N = 2) incorporated affective data (e.g., user satisfaction) to understand users' experiences and perceptions with the generated AF. Nine studies used both student cognitive and affective data to evaluate the performance of AFG systems and to examine user experiences. For cognitive data, the reviewed studies commonly used new or existing programming tasks (Bhatia et al., 2018), projects (Jia et al., 2022), essay-writing tasks (Toma et al., 2021), and tests (Fossati et al., 2015) to evaluate the capability of the proposed AFG

systems or methods in generating AF.

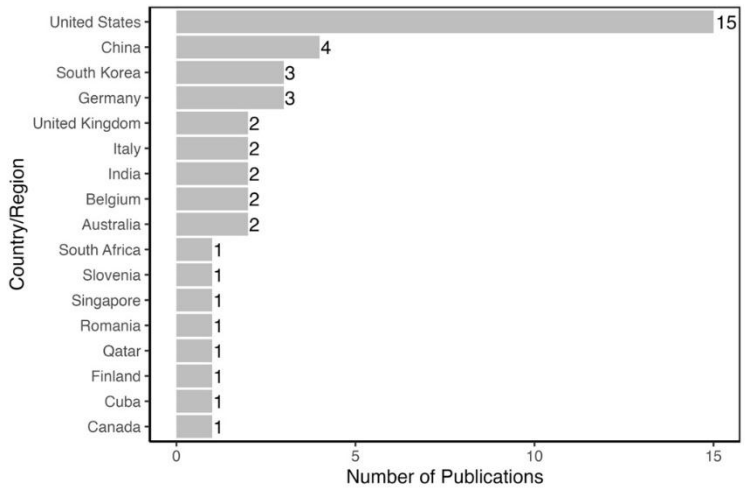

Figure 2. Distribution of Included Studies by Author Country or Region

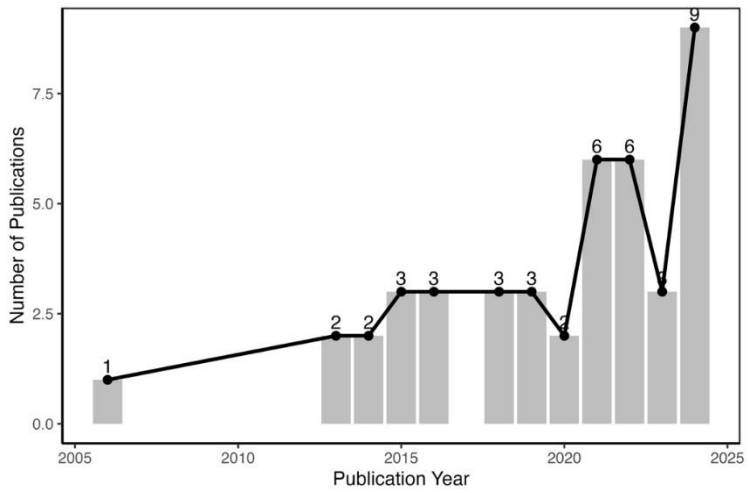

Figure 3. Distribution of Included Studies by Publication Year

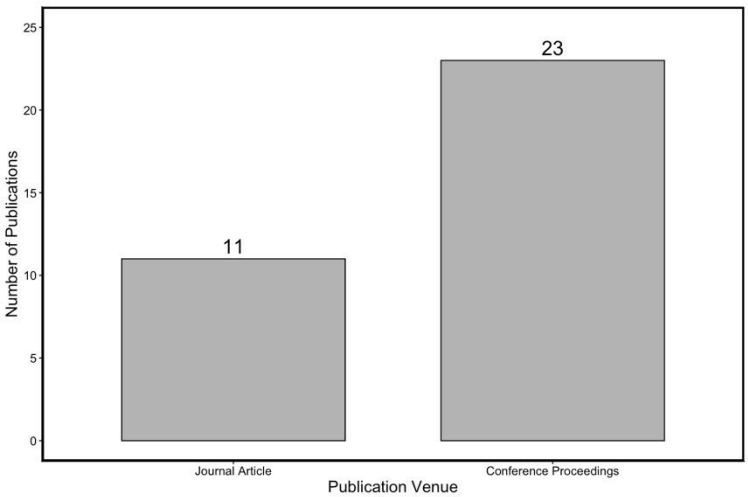

Figure 4. Distribution of Included Studies by Publication Venue

For affective data, surveys and questionnaires were frequently used to collect users' opinions or attitudes regarding the usability and satisfaction of AF. For example, Ahmed et al. (2022) used affective data to evaluate AF generated by the AFG system Verifix. Tutors were asked to report their experiences and perceptions of the quality of the AF, their willingness to use it when supporting students, its effectiveness in grading student submissions, and their confidence in using the AF generated by Verifix (Ahmed et al., 2022).

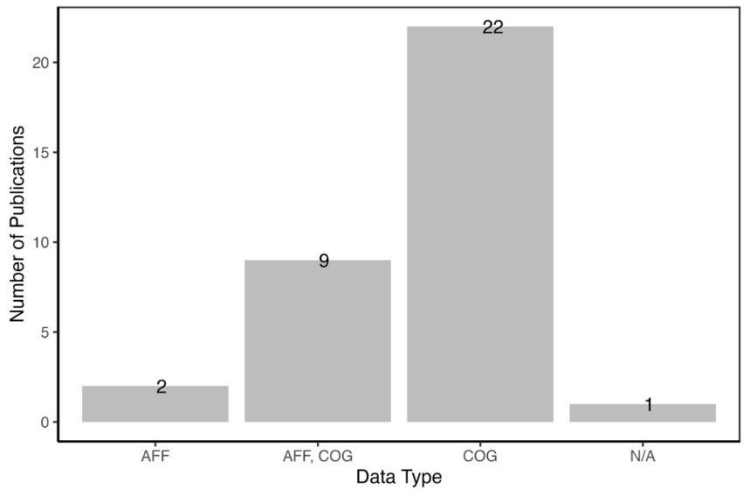

Figure 5. Distribution of Included Studies by Data Type. AFF = affective data; COG = cognitive data.

Figure 6 presents the data sources used to develop or evaluate the proposed AFG methods and systems in the included studies. Most studies collected first-hand data (N = 16) or used second-hand data (N = 14), with three studies using both. Among the 14 studies that employed second-hand data, nine focused on programming tasks, which benefited from abundant publicly available student programming submissions, allowing researchers to evaluate the proposed AFG methods and systems without collecting new data.

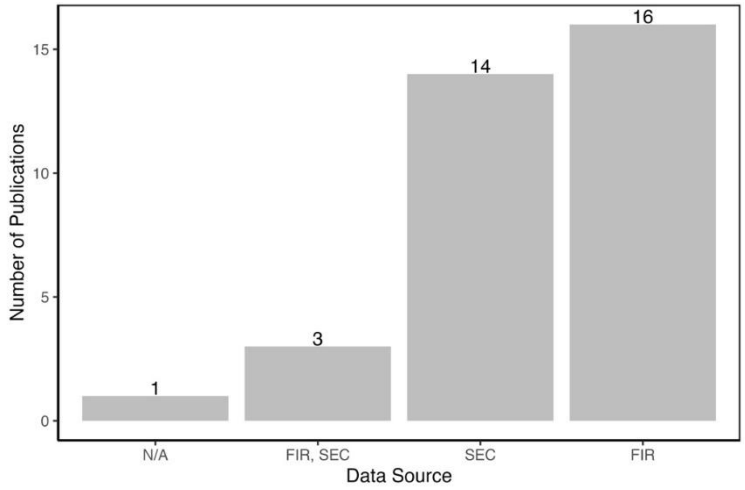

Figure 6. Distribution of Included Studies by Data Source. FIR = first-hand data; SEC = second-hand data.

Figure 7 summarizes the educational domain targeted for the included AFG studies. Most studies (N = 19) focused on providing AF for computer science (i.e., Computer and informative sciences and support services). The

remaining studies were distributed across various domains, including counselor education (i.e., Psychology), essay writing (i.e., English language and literature/letters), financial statement analysis (i.e., Business, management, marketing, and related support services), Physics science, data analysis (i.e., Mathematics and statistics), sign language (i.e., Indigenous and foreign languages, literature, and linguistics), and engineering.

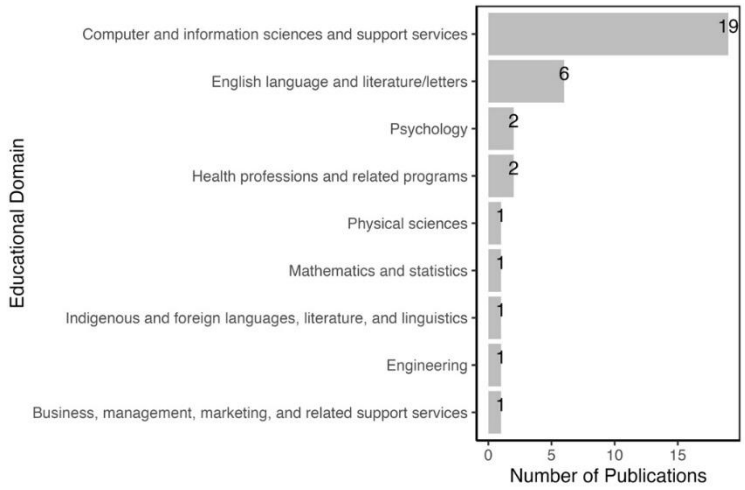

Figure 7. Distribution of Included Studies by Educational Domain

What are the Foundational Designs and Evaluations of AFG Systems in Education?

Figure 8 summarizes the AFG methods used in the included studies. The horizontal bar plot at the bottom left of Figure 8 shows the overall frequency of each method. Rule-based methods (RUL; N = 19), comparison with true answers (COM; N = 17), and machine learning-based approaches (MLE; N = 12) were the most frequently applied methods in generating AF.

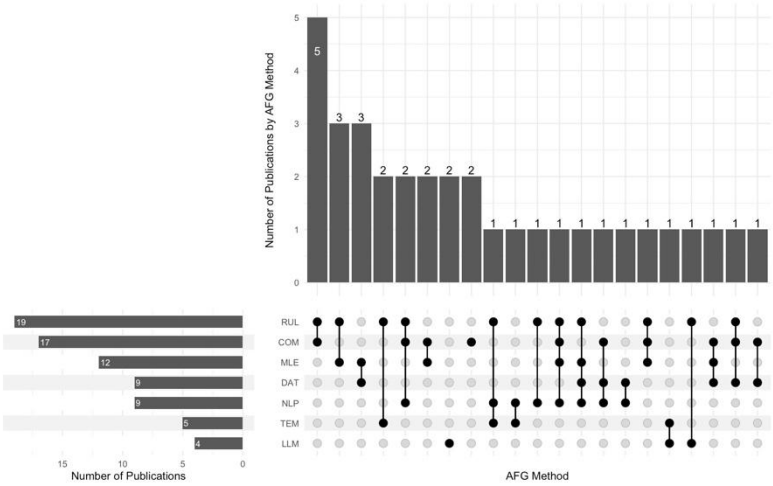

Figure 8. Distribution of Included Studies by AFG Method

The vertical bar plot at the top of Figure 8 shows the number of studies for each unique method combination, whereas the dot matrix below indicates the specific methods involved in each combination. The vertical bar plot

suggests that some studies combined two or more methods, such as RUL with COM (N = 5) or RUL with MLE (N = 3). For example, Dzikovska et al. (2014) integrated NLP and rule-based methods in the BEETLE II (Basic Electricity and Electronics Tutorial Learning Environment) system to generate feedback in physics. After diagnosing student responses, the system used predefined tutoring rules, and the tutorial planner selected the appropriate rule based on NLP analysis.

Figure 9 presents the evaluation methods used to assess the generated AF in the included studies. The horizontal bar plot at the bottom left shows the overall frequency of each evaluation method. The most frequently applied methods for evaluating the generated AF were measuring the AFG system's capability (MEA; N = 30), using surveys (SUR; N = 11), and analyzing student behavior (BEH; N = 7). The vertical bar plot at the top of Figure 9 presents the number of studies for each unique combination of evaluation methods, whereas the dot matrix below indicates the specific evaluation methods involved in each combination.

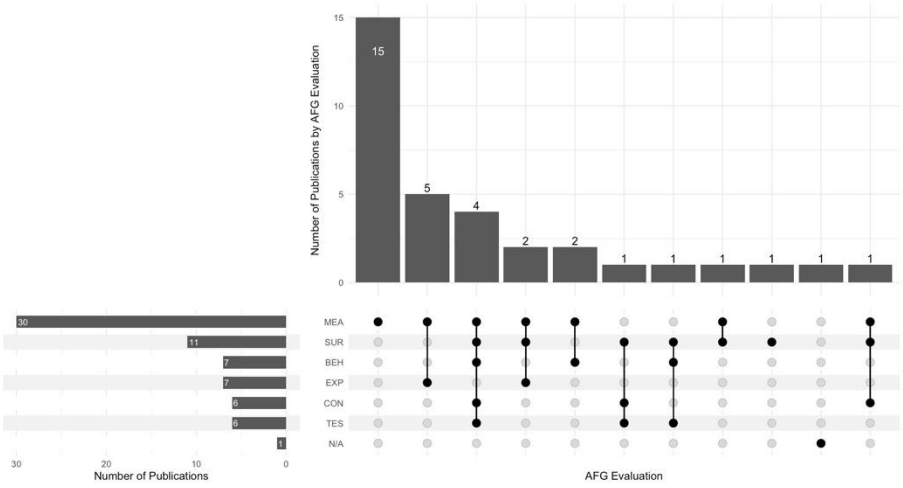

Figure 9. Distribution of Included Studies by Evaluation of AFG

Similarly, the vertical bar plot suggests that several studies combined multiple evaluation methods, such as MEA with EXP (expert evaluation or comparison with expert solutions; N = 5) or MEA with BEH (N = 2). For example, Guid et al. (2019) used a combination of surveys and pre- and post-tests to collect data. They also analyzed student behavior to evaluate the effectiveness of the proposed AFG method in supporting learning about financial statement argumentation. Their findings showed improvements in both student performance and confidence.

What Type of Feedback is Provided by AFG Systems in Education?

Figure 10 presents the distribution of feedback messages in the included studies based on the MISCA framework. The horizontal bar plot at the bottom left shows that 20 studies provided verification feedback. Moreover, seven studies incorporated valence, indicating a positive or negative tone or intended outcome in the feedback. The vertical bar plot in Figure 10 shows that the included AFG studies often combined multiple feedback message characteristics when generating AF. For example, 10 studies incorporated LOA-L (low information load), VER (verification), INF-MIS (knowledge about mistakes), and INF-COR (knowledge of the correct results). One such

example is the feedback generated in Rivers and Koedinger (2013), which included all four elements. For instance, the feedback "In the return statement, the string '23456789YJQKA' should be '23456789TJQKA'" (p. 8) indicates the correctness of the response (VER), identifies the error (INF-MIS), provides the corrected answer (INF-COR), and does so without explanation (LOA-L). The next most frequently appearing combination of feedback message characteristics includes INF-PRO (knowledge about how to proceed), LOA-H (high information load), and VAL (valence), appearing in five studies.

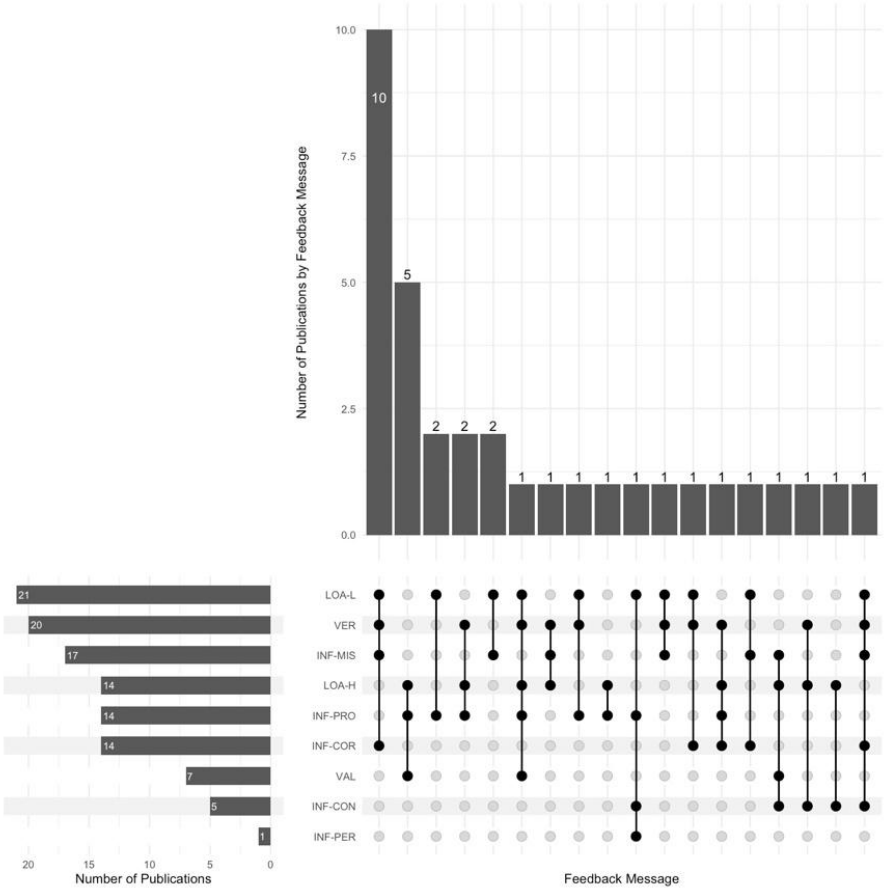

Figure 10. Distribution of Included Studies by Feedback Message

Figure 11 shows the distribution of the feedback functions in the included studies. Most studies (N = 27) generated AF primarily to enhance learning or performance, neglecting MOT (motivation/affect) or SRL functions. Only four studies were identified that designed AF to support all three functions: improving learning or performance (LEA), increasing positive emotions (MOT), and promoting SRL processes (SRL). One such example is Haut et al. (2023), who proposed an AFG system, SOPHIE, to help doctors practice communication skills with patients. Regarding learning support, results indicated that participants who used SOPHIE significantly outperformed the control group in communication with patients (Haut et al., 2023). Regarding emotional support, SOPHIE tracked the participants' emotional changes throughout the conversation and presented an ideal emotional trend for comparison. For SRL support, Haut et al. (2023) stated that the users could reflect on their communication behavior by reviewing actionable feedback. For example, when a user said, "It's very normal to feel down given this news," the system suggested: "You could have said 'I'm sorry that this happened to you. I wish that there was

no such thing as cancer at all" (p. 2).

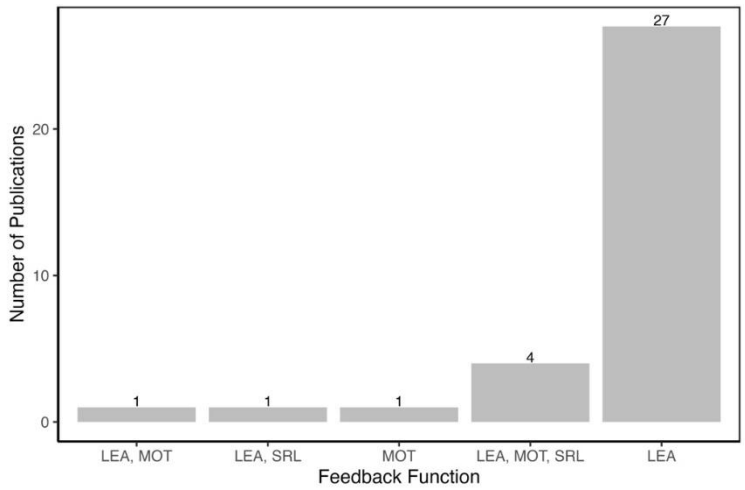

Figure 11. Distribution of Included Studies by Feedback Function. LEA = learning/performance; MOT = motivation/affect; SRL = self-regulated learning.

Figure 12 illustrates the distribution of feedback processing approaches across the included studies. Most studies (N = 27) adopted a feedback-centered (FEE) design, where feedback serves as the primary trigger for learner action. For example, Ahmed et al. (2022) generated verifiably correct program repairs as feedback without considering the individual learner characteristics, indicating that the feedback itself initiates student behavior and guides the learning process. Four studies adopted a student-centered (STU) design, where learner characteristics shape the processing of feedback. For example, Beltrami et al. (2006) designed an AFG system to support self-paced learning and adapt to different learning needs, allowing learners to take agency in interpreting and acting on feedback.

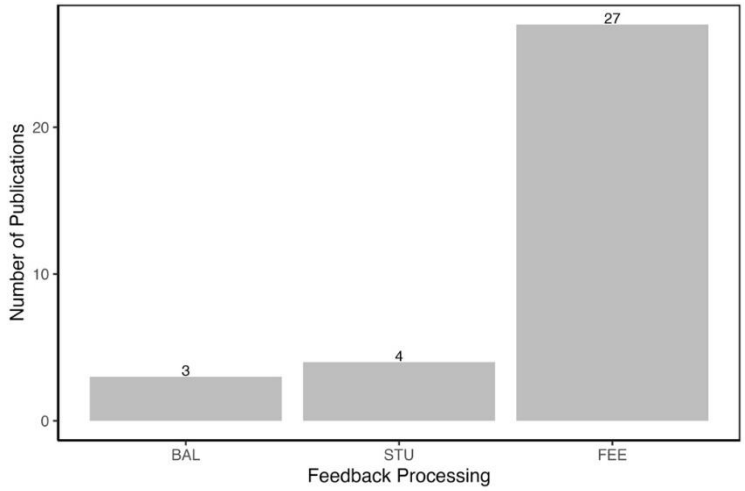

Figure 12. Distribution of Included Studies by Internal Processing of Feedback

Moreover, three studies emphasized both the feedback message and learner characteristics, reflecting a balanced feedback processing design. For example, Fossati et al. (2015) generated procedural feedback to support students'

problem-solving. In addition, the system considered students' uncertainty and behavior (e.g., hesitation time and undo behavior) when delivering feedback. Specifically, the delay before presenting feedback was longer for students with good behavior and shorter for those with less effective behavior.

Figure 12 also shows that seven of the included studies considered student characteristics in the design and generation of AF. These characteristics include learning needs and pace (Beltrami et al., 2006), historical problem-solving behavior (Fossati et al., 2015), problem-solving responses (Dzikovska et al., 2014), domain knowledge (Guid et al., 2019), learner reflection (Haut et al., 2023; Rudolph et al., 2024), and well-being context and responses (Kylvaja et al., 2019). Regarding feedback pedagogy practices, ten studies explicitly addressed how to enhance the impact of feedback by drawing on established theoretical models or applying structured instructional strategies for its implementation and delivery, as shown in Table 3.

Specifically, Rudolph et al. (2024) outlined several key principles of effective feedback when designing their AFG system. For example, one principle is "Avoiding vague statements" (p. 504), emphasizing the importance of clarity and specificity in feedback delivery. Dai et al. (2024) applied Hattie and Timperley's (2007) model to prompt ChatGPT to generate feedback targeting its key components: feeding up, feeding back, feeding forward, as well as its levels: the task, process, self-regulation, and self. The results indicated that ChatGPT-4 outperformed human instructors in generating feedback aligned with these effective components. Dzikovska et al. (2014) employed a conceptual change approach in their system, BEETLE II, to address student misconceptions through AF. In BEETLE II, students first make predictions that often reveal misunderstandings. The system then engages them in a dialogue that helps identify discrepancies and guides them toward revising their conceptual understanding. Haut et al. (2023) designed SOPHIE to improve doctors' communication skills with patients. The feedback content and delivery in SOPHIE were structured around the MVP (Medical Situation, Values, Plan) and 3E's (Empower, Explicit, Empathize) communication paradigm, ensuring feedback was aligned with effective communication practices.

Vittorini and Galassi (2023) designed their AFG system, rDSA, to support *learner-controlled and iterative feedback*. Students could interact with the system multiple times before submitting their final solution. rDSA provided only a suggestion when a mistake was first detected. Then, it released the full solution only if the same error was repeated (Vittorini & Galassi, 2023). Behzad et al. (2024) used a *rubric-based prompt* to guide LLMs in generating automated feedback for English essay writing. The use of rubric-based prompts helps to produce more targeted, structured, and criterion-aligned feedback. Fossati et al. (2015) developed iList by analyzing *effective human tutoring practices* and incorporating these into the AFG design. The system includes proactive feedback strategies, such as anticipating student mistakes and guiding them accordingly. Nayak et al. (2024) applied Bloom's Taxonomy to structure automated feedback for SQL tasks. For instance, at the "apply" level, students were expected to use SQL clauses correctly, and feedback was aligned with that cognitive expectation. Finally, Ruiz and Snoeck (2021) developed FENIkS using a structured *taxonomy of feedback types*, focusing on purpose (corrective, explanatory, formative), timing (on-demand), and learner control (students choose when and where to view feedback).

Table 3. Identified Feedback Pedagogy Practices

Study	Identified Feedback Pedagogy Practice
Rudolph et al., 2024	Key principles of effective feedback
Dai et al., 2024	Hattie and Timperley's (2007) model
Dzikovska et al., 2014	Conceptual change approach
Haut et al., 2023	MVP and 3E's communication paradigm
Vittorini & Galassi, 2023	Iterative and learner controlled Feedback
Behzad et al., 2024; Hossain et al., 2021	Rubrics-based feedback
Fossati et al., 2015	Human tutoring analysis
Nayak et al., 2024	Bloom's taxonomy
Ruiz & Snoeck, 2021	Taxonomy of feedback types

Figure 13 presents the distribution of feedback presentation across the included studies. The horizontal bar plot at the bottom left shows that 31 studies provided adaptive feedback for various student responses or behaviors, whereas only two provided nonadaptive feedback. Regarding the quantity of feedback, 27 studies generated a single feedback message in response to learners' actions, whereas six provided multiple feedback messages. Regarding modality, 25 studies used a single modality (e.g., text-only), whereas seven employed multiple modalities (e.g., text and visuals). Lastly, 24 studies delivered delayed feedback, whereas 10 provided immediate feedback.

The vertical bar plot in Figure 13 shows that the included AFG studies often combined multiple feedback presentation characteristics when generating AF. The most common combination (N = 19) was adaptive feedback presented as a single message, using a single modality, and delivered with a delay. For example, Bhatia et al. (2018) used second-hand data, which prevented the provision of immediate feedback. Instead, they offered program repair as feedback, such as modifying "return e-= 1" to "return b * recPower(b, e-1)." The feedback adaptively corrected student errors and was delivered as a single written message with a delay.

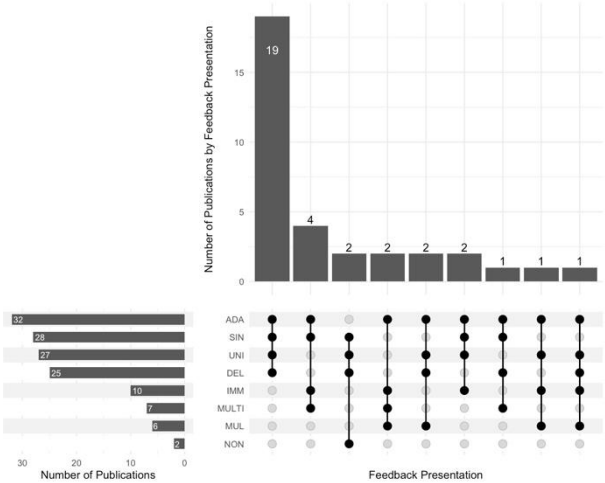

Figure 13. Distribution of Included Studies by Feedback Presentation

Among the 34 included studies, 10 required assistance from experts or instructors when generating AF, such as providing correct solutions for the task (Kim et al., 2016) and identifying the target for the training data (Jia et al., 2022). For example, Singh et al. (2013) required instructors to apply their domain knowledge to define an error model outlining potential student errors. This error model, along with the reference implementation of the programming task, was then used to generate automated feedback.

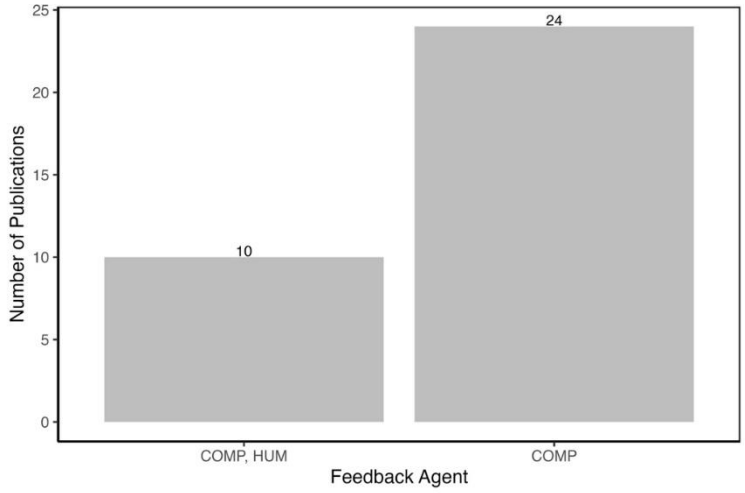

Figure 14. Distribution of Included Studies by Feedback Agent. COMP = computer; HUM = human.

What Ethical Issues are Identified or Need to be Addressed in Educational AFG Systems?

Among the reviewed studies, only five discussed the identified ethical issues. For example, Jia et al. (2022) used a pre-trained language model to generate feedback that might contain inappropriate words or privacy information. They manually inspected the generated feedback and found no ethical violations. In contrast, studies such as Hossain et al. (2021) and Obaido et al. (2020) collected video and voice data, respectively, but did not discuss how the data were managed, stored, or protected, leaving ethical concerns unaddressed.

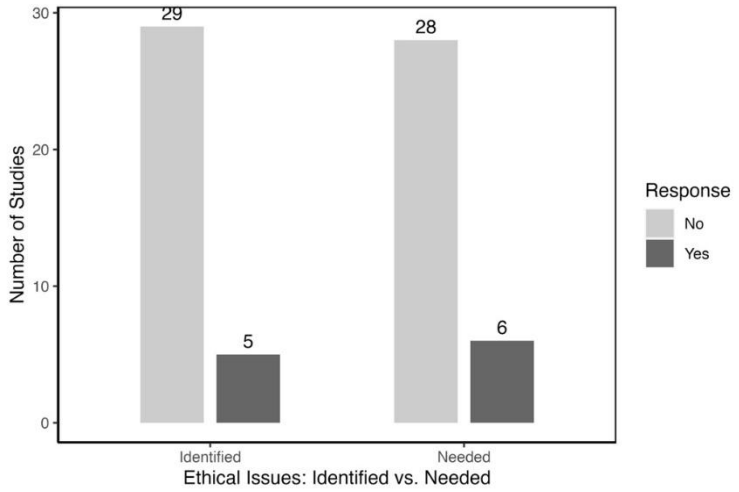

Figure 15. Distribution of Included Studies by Ethical Issues

Discussion

What are the Context Characteristics of Existing AFG Systems in Education?

The findings on the characteristics of existing AFG systems in education show that most of the included studies were authored by researchers based in the United States, pointing to the US's leadership in AFG research. This finding aligns with the results reported by Cavalcanti et al. (2021). Moreover, compared with the decrease in publications from 2017 to 2019 reported by Deeva et al. (2021), we observed a rise in publications after 2019, suggesting that advances in artificial intelligence in education have prompted AF research and increased demand for scalable feedback. Most included studies analyzed cognitive data to evaluate the performance of the proposed AFG method or system, reflecting a primary focus on learning outcomes. Such emphasis indicates that current AFG research prioritizes the feedback on students' cognitive understanding over their attitudes toward the received AF. The underlying assumption of this practice is that learners are expected to use the received AF to improve their cognitive understanding. However, this assumption cannot be guaranteed, as Sychev (2022) found that nearly half of the students (207/464) did not use the provided AF. In this case, none of the AFG systems will be effective if students decide not to use the feedback. Therefore, future studies should investigate factors influencing students' attitudes toward AF, develop strategies to address the non-use of AF, and integrate these factors and strategies into the design of the proposed AFG system and method.

Regarding data collection methods, the findings show that simple surveys or informal interviews were often used to collect affective information from students and teachers in the included studies. However, these methods may not fully capture how students interact with the receiving AF. When learners choose not to engage with the feedback, the provided feedback becomes merely "dangling data." This was also highlighted by Murtagh (2014), who found that teachers and students held different perspectives on the same feedback, with students often finding it unhelpful. To avoid this problem, future studies could also explore more in-depth affective data to gain a comprehensive understanding of students' experiences, perceptions, and performance when receiving AF during learning. For example, using think-aloud could help to understand how students ascribe meaning to the received AF (Máñez et al., 2019).

Moreover, the included studies often use the existing secondary data to evaluate the capability of the proposed AFG method or system (e.g., Bhatia et al., 2018; Kang et al., 2019). On the one hand, such cost-effective practices enable researchers to develop and evaluate their AFG system or methods without the need for significant time and financial investment in new data collection. Additionally, secondary data analysis provides a benchmark for comparison with other studies that analyze the same dataset. On the other hand, relying solely on secondary cognitive data to evaluate the AFG system's effectiveness in enhancing learning outcomes is insufficient to understand how students perceive, respond to, and process AF in practice. The effectiveness of interventions would be minimized when excluding users from the evaluation process (Wolcott & McLaughlin, 2024). Future studies could collect more experiences and reflections from involved stakeholders when evaluating the proposed AFG system and method.

Finally, most reviewed studies targeted the computer science domain, consistent with prior findings (Cavalcanti

et al., 2021; Deeva et al., 2021). Such concentration may be attributed to the easier availability of secondary data in computer science compared to other fields. Specifically, among the 17 studies that used secondary data, 11 focused on computer science domains. Moreover, many studies used programming problems that were not open-ended and generated AF by comparing student answers to correct solutions (e.g., Ahmed et al., 2022). In contrast, generating AF for open-ended questions is more challenging and costly because of the variability of possible answers (Anna Filighera et al., 2022). Therefore, future research may consider developing new AFG systems and methods tailored to ill-defined questions or domains.

What are the Foundational Designs and Evaluations of AFG Systems in Education?

Seven main methods were identified in generating AF: rule-based, template-based, comparison with true answers, data-driven, NLP, LLMs, and other machine learning-based techniques. The frequent application of rule-based methods and comparison with true answers aligns with previous findings (Cavalcanti et al., 2021; Deeva et al., 2021). However, their widespread application also highlights several limitations. For example, rule-based, template-based, and comparison with true answers heavily rely on experts to create rules, templates, and reference true answers. Such reliance might limit the quality of the generated feedback, because expert domain knowledge might not capture all possible problem-solving paths. Consequently, some generated feedback might be far from the learners' current reasoning process (Filighera et al., 2022). For example, although Piech et al. (2015) found that one expert annotation could serve as a reference for up to 214 other student programs, relying on fixed expert annotations might limit feedback diversity and adaptability. Moreover, the required templates, rules, and reference solutions expand rapidly as the problems become more complex (Sychev, 2022). The performance of data-driven methods depends on large and high-quality data sets. Therefore, their effectiveness is questionable in contexts where such data are limited or not available (Vittorini & Galassi, 2023). Overall, the traditional methods for generating AF remain prevalent. However, the associated limitations highlight the need for more flexible, scalable approaches to feedback generation.

Moreover, the included studies often combined multiple methods to generate AF (e.g., Dzikovska et al., 2014; Lu & Cutumisu, 2021), indicating the potential synergies among different approaches to enhance feedback quality. Therefore, future AFG research could investigate the most frequently used combinations of AFG methods to better understand their effectiveness and applicability across contexts. The reviewed studies have often focused on evaluating the capability of the proposed AFG systems or methods by reporting feedback generation accuracy. However, one limitation of this method is that it does not reveal the actual effect of the generated AF in facilitating the learning process, as user studies were typically lacking. Among the reviewed AFG studies that conducted user studies, surveys are the most frequently used method to collect users' affective responses and attitudes toward AF, aligning with previous findings (Deeva et al., 2021). However, evaluations on long-term learning outcomes were largely neglected, which might obscure the true effectiveness of these AFG methods. Therefore, future AFG research should consider using multiple evaluation methods to better understand how the generated AF impacts the learning process and outcomes.

Figure 16 presents the distribution of AFG methods and evaluation approaches across domains in the included

studies. For example, in essay writing (English language and literature/letters), LLMs are often used to generate AF, whereas measuring the capability of the AFG method is a common evaluation approach. Future AFG studies could consider adopting the most frequently used AFG methods and evaluation approaches tailored to specific domains.

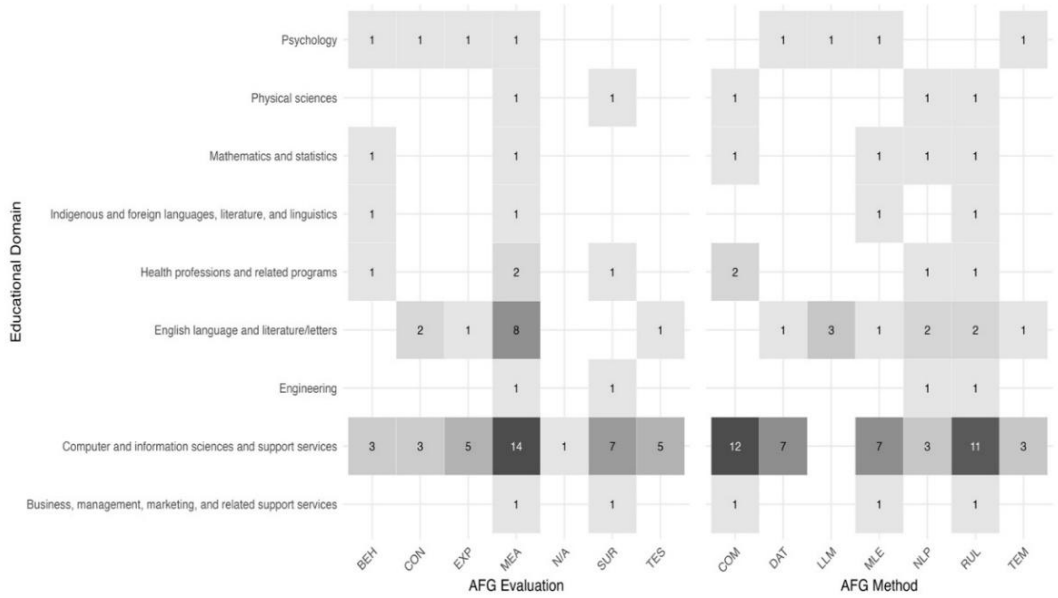

Figure 16. Distribution of AFG Methods and Evaluation Across Educational Domains

What Type of Feedback is Provided by AFG Systems in Education?

Feedback Message and Function

Figure 17 presents the distribution of feedback characteristics across educational domains in the included studies. The prevalence of low-information-load, knowledge about mistakes, and verification feedback in the computer science domain suggests that this domain emphasizes simple corrective support. Most reviewed studies provide corrective feedback, indicating student errors and the corresponding solutions, consistent with previous findings (Deeva et al., 2021), which found that over 90% of reviewed studies provided corrective feedback. In contrast, feedback focused on how to proceed, and a high information load was more evident in the essay-writing domain, aligning with previous findings (Yu et al., 2020). Such a difference might reflect the distinct nature of the two domains. For example, students learning how to code often need feedback to locate and fix specific bugs (Bhatia et al., 2018). In contrast, students learning how to write essays typically require more abstract guidance, such as developing a strong argument (Wingate, 2012). Future AFG research could consider tailoring feedback messages to the specific needs and characteristics of the target domain.

Additionally, providing knowledge about mistakes or corrected solutions may not always enhance learning. For example, Sychev (2022) found that simple corrective feedback did not promote meaningful learning and that improved student performance does not necessarily indicate actual learning gains. Such feedback may enable

students to game the system by focusing on correct answers rather than deeper understanding. Therefore, future AFG studies should investigate feedback strategies that support meaningful learning.

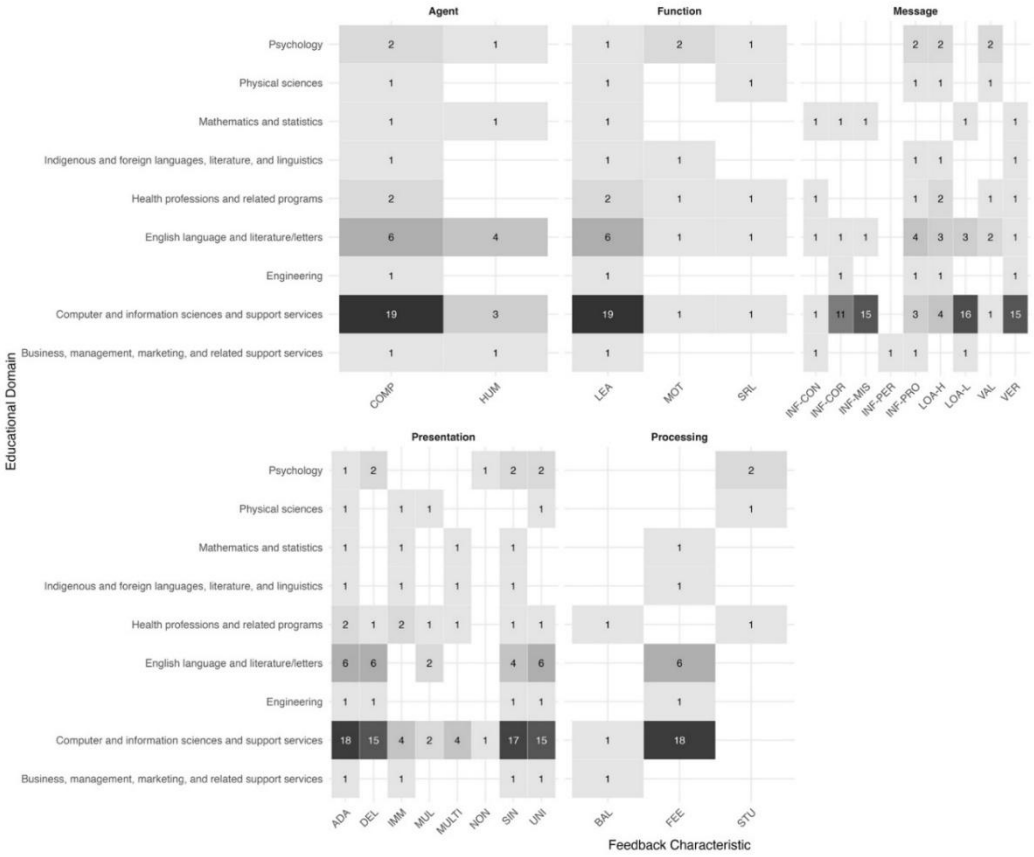

Figure 17. Distribution of Feedback Characteristics Across Educational Domains

The most prevalent feedback valence in the included studies was neutral rather than positive or negative. However, Máñez et al. (2019) used the think-aloud method to understand how students process complex formative feedback. They found that students paid more attention to feedback highlighting their errors than to feedback indicating success. Future AFG studies could investigate how different feedback valences influence student attention, interpretation, and learning outcomes.

The potential of feedback to enhance teaching and learning is well established, including its significant impact on student learning (Wisniewski et al., 2020), benefits for student SRL skills (Bellhäuser et al., 2022), and influence on student motivation and engagement (Yu et al., 2020). However, the current review found that most included AFG studies emphasized the learning function of feedback, whereas they relatively neglected its motivational and SRL functions. Therefore, future AFG research should explore the full range of feedback functions, rather than focusing solely on learning outcomes.

Feedback Processing and Student Characteristics of Feedback Design

Although constructivism, which emphasizes student agency in constructing knowledge, has been widely adopted

as a learning theory (Ertmer & Newby, 2013), the current review found that most included studies adopted a behaviorist perspective by positioning feedback as the primary driver of learning behavior and outcomes. Prior research has shown that student agency significantly influences academic performance and perceived learning experience (Luo et al., 2019). As shown in the results, only seven of the included studies considered student characteristics in feedback design. Future AFG research could further explore how supporting learner agency can enhance the effectiveness of automated feedback.

Feedback Pedagogy, Presentation, and Agent

As shown in the results, 10 of the included studies drew on various theories or guidelines to inform feedback implementation. The specific applications of these pedagogical practices are not repeated here to avoid redundancy. However, most of the included studies lacked a pedagogical foundation and generated AF directly. Feedback developed without a sound theoretical basis may be less effective in supporting learning. For example, Kluger and DeNisi (1996) found that over one-third of the feedback decreased student performance, which they attributed to whether the feedback directed students' attention to task-solving or to their emotions. Fleckenstein et al. (2023) reported significant heterogeneity in the effect sizes of feedback on student writing performance (Q = 285.89, df = 83, p < .001), which they suggested might be related to feedback duration. Wisniewski et al. (2020) also found significant heterogeneity (Q = 7,339, df = 993, p < .001), which they concluded may result from the use of different forms of feedback. Therefore, future AFG research should ensure that the generated feedback is grounded in established feedback theories or guidelines to enhance its educational effectiveness.

The dominance of adaptive, delayed, single-modality, and single-message feedback reveals gaps in the use of multimodal and immediate feedback practices. First, although written feedback is the most common modality, previous research has found that students prefer a combination of feedback modalities to support learning (Ice et al., 2010). Additionally, a systematic literature review on engaging educational games suggests that clear feedback is more effective when delivered via multiple modalities (Laine & Lindberg, 2020). Therefore, future studies could consider incorporating multiple feedback modalities to motivate learners.

Second, although immediate feedback has been found to be more effective than delayed feedback for enhancing learning (Jia et al., 2022), 24 of the reviewed studies provided feedback only after task completion. The delay in feedback delivery may hinder students from using it to learn, as there is a time gap between the task-solving behavior and the time of receiving feedback. Therefore, future AFG research should consider the design of immediate feedback to better support real-time learning. The identified purpose of employing expert knowledge (e.g., providing correct solutions for the task, building templates or rules for feedback, and defining error models) aligns with previous findings (Deeva et al., 2021). This consistent finding of the need for expert knowledge in developing AFG systems and methods suggests that purely data-driven or artificial intelligence-based AFG systems are still under development in current research. As noted by Deeva et al. (2021), only 19.3% of the studies they reviewed were fully data driven. Therefore, further investigation is needed to automate complex domain knowledge for effective feedback generation.

What Ethical Issues are Identified or Need to be Addressed in Educational AFG Systems?

Most of the reviewed studies provided simple corrective feedback, which did not raise ethical concerns. However, as shown in the results, two studies provided feedback based on voice and video data but did not discuss the required data management and protection procedures. Moreover, Behzad et al. (2024) crawled data from the Essay Forum platform, but did not provide an ethical statement regarding user consent. Dai et al. (2024) used ChatGPT to generate feedback, but did not discuss how the ChatGPT-generated content was controlled to avoid bias or misinformation. Haut et al. (2023) applied sentiment scoring to detect users' emotions, which may carry cultural or linguistic biases depending on the characteristics of the training data used. Kylvaja et al. (2019) grouped students into different profiles based on their responses to a well-being survey. However, labeling based on well-being data might lead to unintended negative consequences if the labels are inaccurate or not used carefully. Therefore, future AFG research should clearly report how users' private data are protected and implement measures to decrease potential bias in feedback generation.

Identified Limitations in Reviewed Studies

Transferability of AFG Systems or Methods

The included studies often focused on programming and essay-writing domains, while students needed feedback across all the educational domains they encountered. For this reason, the transferability of the AFG system or method is critical to decreasing time and money investments. However, each reviewed study needs to be adapted to transfer to new domains. For example, Fossati et al. (2015) extracted a procedural knowledge model from student historical data to understand the correct problem-solving path and used this information to provide AF. When using this data-greedy method in new domains, large historical student data from those new domains are needed to identify a representative procedural knowledge model. Future AFG systems should consider the transferability function of the systems during the design stage.

Lacking Participants' Details, Theoretical Foundations, and Student Models

Most studies that used second-hand cognitive data often did not report details about the participant country, sample size, age, or grades. Lacking such information undermines the utility of AFG systems and limits the effectiveness of AF in scaffolding learners, potentially deterring instructors from using these systems. To maximize the impact of AF on learning processes and outcomes, educational and learning theories and models are essential for guiding the design, implementation, and evaluation of AFG systems. A thorough understanding of how these theories interact with student traits and learning environments is essential to fully unlock the potential of AF on student learning. However, only ten studies considered the pedagogical practices when designing or proposing AFG systems or methods. The lack of theoretical foundations in existing AFG studies suggests that future studies should pay attention to the alignment of feedback provision with learning, cognitive, and self-regulation theories.

One of the most significant advantages of the AFG system is that it provides tailored feedback to learners, which requires tracking information on student performance. However, only a few studies have developed student

models within their AFG systems or methods that track and store learners' progress as they solve a task. Therefore, future AFG systems could incorporate the student model into their design to compensate for the disadvantages of feedback based on predefined templates and rules.

Closing the Feedback Loop

Sadler (1989) argued that feedback is merely "dangling data" when users do not understand or monitor how it impacts learners' performance. Therefore, monitoring how students perceive, process, and respond to the feedback they receive is essential to closing the feedback loop (Boud & Molloy, 2013). Thus, completing the feedback loop is critical to achieving feedback's full efficacy. However, only Toma et al. (2021) explicitly discussed the process of closing the feedback loop. Specifically, their AFG system allows students to integrate the received AF into essay writing and revise their work, with feedback generated and used iteratively. Future AFG systems should evaluate the use of generated AF more deeply to assess its quality.

Learner Control in AFG Systems

Students should be regarded as active learners during the learning process, whereas AFG systems should serve as facilitators rather than knowledge providers. However, the most common form of learning control in existing AFG systems is to allow students to request feedback at any time. Although beneficial, such a characteristic may also be exploited to game the system and obtain the true answer without learning (Marwan et al., 2019). Ruiz and Snoeck (2021) extended learner control by allowing students to select the knowledge content for which they wanted to receive feedback. Future AFG systems should explore how to support learner agency and be aware of potential system-gaming behaviors.

Contributions

Theoretically, the current review expanded the conceptual frameworks of AF proposed in previous research (Deeva et al., 2021) by evaluating the generated AF and verifying the categories for feedback characteristics from the MISCA framework. The current review also provides a comprehensive overview of empirical AFG studies in education, highlighting both consistencies and discrepancies with previous studies. Methodologically, the current review employed a systematic literature review following the PRISMA framework, ensuring the rigor, transparency, and utility of the results. It also expanded several coding columns based on the MISCA framework, providing a reference to future review studies on this topic that aim to consolidate findings given the rapid increase of AFG research. Practically, the current review identified several evidence-based practices in AFG studies that require further attention or should be avoided in future studies. Such research gaps in the AFG methods and evaluation provide valuable insights for future researchers and users in developing and applying AFG systems.

Implications

Theoretically, the current review identifies several issues in the existing AFG studies and could inform future

AFG studies on the system design and feedback provision. Future studies should focus on other methods that are not limited by domain expertise or the quality of historical student data. The educational and learning theories should inform the design of the feedback purpose, nature, and provision timing to ensure it aligns with the individual's knowledge progression and state. The design and evaluation of generated AF could consider closing the feedback loop to increase the possibility of feedback utilization in the learning process.

Practically, the current review identified multiple research gaps, including the need to collect non-cognitive data, tailor feedback to ill-defined questions or domains, develop more flexible and scalable feedback generation approaches, automate complex domain knowledge, design immediate feedback, incorporate multiple feedback modalities, draw on established feedback theories or guidelines, address ethical considerations, and support learner agency in AFG studies. Future studies may consider these aspects when setting out to develop more effective, engaging, and ethically sound AFG systems, thereby optimizing their impact on education.

Conclusion

This systematic literature review of empirical AFG studies in education provides a comprehensive overview of the characteristics of existing AFG systems. The present review extends existing AFG reviews by focusing on the design and evaluation features of AFG systems and the characteristics of the feedback provided, based on the MISCA framework. Theoretically, this review informs future AFG studies by summarizing the limitations of the AFG methods and evaluation procedures currently used. Practically, this review summarizes the current development of AFG and provides directions for future research on AFG. The findings contribute to optimizing AFG's impact on future education and learning. Finally, this review also identifies several research gaps, including the need for broader target educational domains, grounding in educational and learning theory, student modeling, ethical considerations, and enhanced learner control in AFG system development. Future AFG research may consider focusing on these areas to develop more effective, engaging, and ethically sound systems.

References

Ahmed, U., Fan, Z., Yi, J., Al-Bataineh, O., & Roychoudhury, A. (2022). Verifix: Verified repair of programming assignments. *ACM Transactions on Software Engineering and Methodology*, *31*(4), 1–31. https://doi.org/10.1145/3510418

Bangert-Drowns, R. L., Kulik, C.-L. C., Kulik, J. A., & Morgan, M. (1991). The instructional effect of feedback in test-like events. *Review of Educational Research*, *61*(2), 213–238.

Behzad, S., Kashefi, O., & Somasundaran, S. (2024). LEAF: Language learners' English essays and feedback corpus. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, *2*, 433–442. https://doi.org/10.18653/v1/2024.naacl-short.36

Bellhäuser, H., Liborius, P., & Schmitz, B. (2022). Fostering self-regulated learning in online environments: Positive effects of a web-based training with peer feedback on learning behavior. *Frontiers in Psychology*, *13*, 813381. https://doi.org/10.3389/fpsyg.2022.813381

Beltrami, P., Crescenzi, P., Gensini, G., Innocenti, A., Lippi, P., & Saccone, N. (2006). Automatic feedback generation in scenario-based e-learning with an application to the healthcare sector. *Journal of E-Learning and Knowledge Society*, *2*(2), 229–240. https://doi.org/10.20368/1971-8829/716

Bhatia, S., Kohli, P., & Singh, R. (2018). Neuro-symbolic program corrector for introductory programming assignments. In *Proceedings of the 40th International Conference on Software Engineering*, 60–70. https://doi.org/10.1145/3180155.3180219

Boud, D., & Molloy, E. (2013). Rethinking models of feedback for learning: The challenge of design. *Assessment & Evaluation in Higher Education*, *38*(6), 698–712. https://doi.org/10.1080/02602938.2012.691462

Buckingham Shum, S., Lim, L.-A., Boud, D., Bearman, M., & Dawson, P. (2023). A comparative analysis of the skilled use of automated feedback tools through the lens of teacher feedback literacy. *International Journal of Educational Technology in Higher Education*, *20*(1), 40. https://doi.org/10.1186/s41239-023-00410-9

Butler, D. L., & Winne, P. H. (1995). Feedback and self-regulated learning: A theoretical synthesis. *Review of Educational Research*, *65*(3), 245–281. https://doi.org/10.3102/00346543065003245

Carless, D., & Boud, D. (2018). The development of student feedback literacy: Enabling uptake of feedback. *Assessment & Evaluation in Higher Education*, *43*(8), 1315–1325. https://doi.org/10.1080/02602938.2018.1463354

Cavalcanti, A. P., Barbosa, A., Carvalho, R., Freitas, F., Tsai, Y.-S., Gašević, D., & Mello, R. F. (2021). Automatic feedback in online learning environments: A systematic literature review. *Computers and Education: Artificial Intelligence*, *2*, 100027. https://doi.org/10.1016/j.caeai.2021.100027

Covidence (2025). *Covidence systematic review software*. [Computer software]. Veritas Health Innovation. www.covidence.org

Dai, W., Tsai, Y.-S., Lin, J., Aldino, A., Jin, H., Li, T., Gašević, D., & Chen, G. (2024). Assessing the proficiency of large language models in automatic feedback generation: An evaluation study. *Computers and Education: Artificial Intelligence*, *7*, 100299.

Deeva, G., Bogdanova, D., Serral, E., Snoeck, M., & De Weerdt, J. (2021). A review of automated feedback systems for learners: Classification framework, challenges and opportunities. *Computers & Education*, *162*, 104094. https://doi.org/10.1016/j.compedu.2020.104094

Dzikovska, M., Steinhauser, N., Farrow, E., Moore, J., & Campbell, G. (2014). BEETLE II: Deep natural language understanding and automatic feedback generation for intelligent tutoring in basic electricity and electronics. *International Journal of Artificial Intelligence in Education*, *24*(3), 284–332. https://doi.org/10.1007/s40593-014-0017-9

Ertmer, P. A., & Newby, T. J. (2013). Behaviorism, cognitivism, constructivism: Comparing critical features from an instructional design perspective. *Performance Improvement Quarterly*, *26*(2), 43–71. https://doi.org/10.1002/piq.21143

Evans, C. (2013). Making sense of assessment feedback in higher education. *Review of Educational Research*, *83*(1), 70–120. https://doi.org/10.3102/0034654312474350

Filighera, A., Tschesche, J., Steuer, T., Tregel, T., & Wernet, L. (2022). Towards generating counterfactual examples as automatic short answer feedback. In M. M. Rodrigo, N. Matsuda, A. I. Cristea, & V. Dimitrova (Eds.), *Artificial Intelligence in Education* (Vol. 13355, pp. 206–217). Springer International

Publishing. https://doi.org/10.1007/978-3-031-11644-5_17

Fossati, D., Di Eugenio, B., Ohlsson, S., Brown, C., & Chen, L. (2015). Data driven automatic feedback generation in the iList intelligent tutoring system. *Technology, Instruction, Cognition and Learning*, *10*(1), 5–26.

Guid, M., Možina, M., Pavlič, M., & Turšič, K. (2019). Learning by arguing in argument-based machine learning framework. In A. Coy, Y. Hayashi, & M. Chang (Eds.), *Intelligent Tutoring Systems* (Vol. 11528, pp. 112–122). Springer International Publishing. https://doi.org/10.1007/978-3-030-22244-4_15

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, *77*(1), 81–112. https://doi.org/10.3102/003465430298487

Haut, K., Wohn, C., Kane, B., Carroll, T., Guigno, C., Kumar, V., Epstein, R., Schuber, L., & Hoque, E. (2023). Validating a virtual human and automated feedback system for training doctor-patient communication skills. In *2023 11th International Conference on Affective Computing and Intelligent Interaction (ACII)*, 1–8. https://doi.org/10.1109/ACII59096.2023.10388213

Heo, J., Jeong, H., Choi, D., & Lee, E. (2023). REFERENT: Transformer-based feedback generation using assignment information for programming course. In *2023 IEEE/ACM 45th International Conference on Software Engineering: Software Engineering Education and Training*, 101–106. https://doi.org/10.1109/ICSE-SEET58685.2023.00035

Hossain, S., Kamzin, A., Amperayani, V. N. S. A., Paudyal, P., Banerjee, A., & Gupta, S. K. S. (2021). Engendering trust in automated feedback: A two step comparison of feedbacks in gesture based learning. In I. Roll, D. McNamara, S. Sosnovsky, R. Luckin, & V. Dimitrova (Eds.), *Artificial Intelligence in Education* (pp. 190–202). Springer International Publishing. https://doi.org/10.1007/978-3-030-78292-4_16

Huang, W., Stephens, J. M., & Brown, G. T. L. (2025). Feedback assisted by technology: A systematic review of empirical research. *International Journal of Technology in Education*, *8*(2), 421–444. https://doi.org/10.46328/ijte.1061

Ice, P., Swan, K., Diaz, S., Kupczynski, L., & Swan-Dagen, A. (2010). An analysis of students' perceptions of the value and efficacy of instructors' auditory and text-based feedback modalities across multiple conceptual levels. *Journal of Educational Computing Research*, *43*(1), 113–134. https://doi.org/10.2190/EC.43.1.g

Jia, Q., Young, M., Xiao, Y., Cui, J., Liu, C., Rashid, P., & Gehringer, E. (2022). Automated feedback generation for student project reports: A data-driven approach. *Journal of Educational Data Mining*, *14*(3), 132–161. https://doi.org/10.5281/zenodo.7304954

Jonsson, A. (2013). Facilitating productive use of feedback in higher education. *Active Learning in Higher Education*, *14*(1), 63–76. https://doi.org/10.1177/1469787412467125

Keuning, H., Jeuring, J., & Heeren, B. (2018). A systematic literature review of automated feedback generation for programming exercises. *ACM Transactions on Computing Education*, *19*(1), 1–43. https://doi.org/10.1145/3231711

Kim, D., Kwon, Y., Liu, P., Kim, I. L., Perry, D. M., Zhang, X., & Rodriguez-Rivera, G. (2016). Apex: Automatic programming assignment error explanation. *ACM SIGPLAN Notices*, *51*, 311–327. https://doi.org/10.1145/2983990.2984031

Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a

meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin*, *119*(2), 254–284. https://doi.org/10.1037/0033-2909.119.2.254

Kulhavy, R. W., & Stock, W. A. (1989). Feedback in written instruction: The place of response certitude. *Educational Psychology Review*, *1*(4), 279–308. https://doi.org/10.1007/BF01320096

Kylvaja, M., Kumpulainen, P., & Konu, A. (2019). Application of data clustering for automated feedback generation about student well-being. In *Proceedings of the 1st ACM SIGSOFT International Workshop on Education Through Advanced Software Engineering and Artificial Intelligence*, 21–26. https://doi.org/10.1145/3340435.3342720

Laine, T. H., & Lindberg, R. S. N. (2020). Designing engaging games for education: A systematic literature review on game motivators and design principles. *IEEE Transactions on Learning Technologies*, *13*(4), 804–821. https://doi.org/10.1109/TLT.2020.3018503

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159–174. JSTOR. https://doi.org/10.2307/2529310

Larrondo, P., Frank, B., & Ortiz, J. (2021). The state of the art in providing automated feedback to open-ended student work. In *Proceedings of the Canadian Engineering Education Association (CEEA)*, 1–8. https://doi.org/10.24908/pceea.vi0.14854

Lasserson, T. J., Thomas, J., & Higgins, J. P. (2019). Starting a review. In J. P. Higgins, J. Thomas, J. Chandler, M. Cumpston, T. Li, M. J. Page, & V. A. Welch (Eds.), *Cochrane handbook for systematic reviews of interventions* (pp. 1–12). Wiley Online Library. https://doi.org/10.1002/9781119536604.ch1

Lipnevich, A. A., Berg, D. A., & Smith, J. K. (2016). Toward a model of student response to feedback. In *Handbook of Human and Social Conditions in Assessment* (pp. 169–185). Routledge.

Lipnevich, A. A., & Panadero, E. (2021). A review of feedback models and theories: Descriptions, definitions, and conclusions. *Frontiers in Education*, *6*, 720195. https://doi.org/10.3389/feduc.2021.720195

Lu, C., & Cutumisu, M. (2021). Integrating deep learning into an automated feedback generation system for automated essay scoring. In *Proceedings of The 14th International Conference on Educational Data Mining (EDM21)*, 573–579. https://educationaldatamining.org/edm2021/

Luo, H., Yang, T., Xue, J., & Zuo, M. (2019). Impact of student agency on learning performance and learning experience in a flipped classroom. *British Journal of Educational Technology*, *50*(2), 819–831. https://doi.org/10.1111/bjet.12604

Máñez, I., Vidal-Abarca, E., Kendeou, P., & Martínez, T. (2019). How do students process complex formative feedback in question-answering tasks? A think-aloud study. *Metacognition and Learning*, *14*(1), 65–87. https://doi.org/10.1007/s11409-019-09192-w

Marwan, S., Lytle, N., Williams, J. J., & Price, T. (2019). The impact of adding textual explanations to next-step hints in a novice programming environment. In *Proceedings of the 2019 ACM Conference on Innovation and Technology in Computer Science Education*, 520–526. https://doi.org/10.1145/3304221.3319759

Mason, B. J., & Bruning, R. (2001). *Providing feedback in computer-based instruction: What the research tells us?* (No. 9). Center for Instructional Innovation.

Molloy, E. K., & Boud, D. (2014). Feedback models for learning, teaching and performance. In J. M. Spector, M. D. Merrill, J. Elen, & M. J. Bishop (Eds.), *Handbook of Research on Educational Communications and Technology* (pp. 413–424). Springer New York. https://doi.org/10.1007/978-1-4614-3185-5_33

Murtagh, L. (2014). The motivational paradox of feedback: Teacher and student perceptions. *The Curriculum Journal*, *25*(4), 516–541. https://doi.org/10.1080/09585176.2014.944197

Narciss, S. (2008). Feedback strategies for interactive learning tasks. In *Handbook of research on educational communications and technology* (3rd Edition, pp. 125–143). Routledge.

Narciss, S., & Huth, K. (2004). How to design informative tutoring feedback for multi-media learning. In H. M. Niegemann, D. Leutner, & R. Brünken (Eds.), *Instructional Design for Multimedia Learning* (pp. 181–195). Waxmann.

Nayak, S., Agarwal, R., Khatri, S. K., & Mohammadian, M. (2024). Student outcome assessment on structured query language using rubrics and automated feedback generation. *International Journal of Advanced Computer Science and Applications*, *15*(3), 728. https://doi.org/10.14569/IJACSA.2024.0150374

Nicol, D., & Macfarlane-Dick, D. (2006). Formative assessment and self-regulated learning: A model and seven principles of good feedback practice. *Studies in Higher Education*, *31*(2), 199–218. https://doi.org/10.1080/03075070600572090

Obaido, G., Ade-Ibijola, A., & Vadapalli, H. (2020). TalkSQL: A tool for the synthesis of SQL queries from verbal specifications. In *2020 2nd International Multidisciplinary Information Technology and Engineering Conference (IMITEC)*, 1–10. https://doi.org/10.1109/IMITEC50163.2020.9334088

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. *International Journal of Surgery*, *88*, 105906. https://doi.org/10.1016/j.ijsu.2021.105906

Panadero, E., & Lipnevich, A. A. (2022). A review of feedback models and typologies: Towards an integrative model of feedback elements. *Educational Research Review*, *35*, 100416. https://doi.org/10.1016/j.edurev.2021.100416

Pardo, A., Jovanovic, J., Dawson, S., Gašević, D., & Mirriahi, N. (2019). Using learning analytics to scale the provision of personalised feedback. *British Journal of Educational Technology*, *50*(1), 128–138. https://doi.org/10.1111/bjet.12592

Piech, C., Huang, J., Nguyen, A., Phulsuksombati, M., Sahami, M., & Guibas, L. (2015). Learning program embeddings to propagate feedback on student code. In *Proceedings of the 32nd International Conference on Machine Learning*, 1093–1102. http://proceedings.mlr.press/v37/piech15.html

Poulos, A., & Mahony, M. J. (2008). Effectiveness of feedback: The students' perspective. *Assessment & Evaluation in Higher Education*, *33*(2), 143–154. https://doi.org/10.1080/02602930601127869

Rivers, K., & Koedinger, K. R. (2013). Automatic generation of programming feedback: A data-driven approach. In *The First Workshop on AI-Supported Education for Computer Science (AIEDCS 2013)*, *50*, 50–59.

Rose, K. J. (2023). Using classification of instructional program codes in human resource development. *Human Resource Development Review*, *22*, 428–444. https://doi.org/10.1177/15344843231184101

Rudolph, E., Seer, H., Mothes, C., & Albrecht, J. (2024). Automated feedback generation in an intelligent tutoring system for counselor education. In *Proceedings of the 19th Conference on Computer Science and Intelligence Systems (FedCSIS)*, 501–512. https://doi.org/10.15439/2024F1649

Ruiz, J., & Snoeck, M. (2021). Automatic feedback generation for supporting user interface design. In H. Fill, M.

VanSinderen, & L. Maciaszek (Eds.), In *Proceedings of the 16th International Conference on Software Technologies (ICSOFT 2021)* (pp. 23–33). https://doi.org/10.5220/0010513400230033

Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional Science*, *18*(2), 119–144. https://doi.org/10.1007/BF00117714

Serral Asensio, E., Ruiz, J., Elen, J., & Snoeck, M. (2019). Conceptualizing the domain of automated feedback for learners. In *Proceedings of the XXII IberoAmerican Conference on Software Engineering, CIbSE 2019*, 223–236.

Shadiev, R., & Feng, Y. (2023). Using automated corrective feedback tools in language learning: A review study. *Interactive Learning Environments*, 1–29. https://doi.org/10.1080/10494820.2022.2153145

Singh, R., Gulwani, S., & Solar-Lezama, A. (2013). Automated feedback generation for introductory programming assignments. In *Proceedings of the 34th ACM SIGPLAN Conference on Programming Language Design and Implementation*, 15–26. https://doi.org/10.1145/2462156.2462195

Sondergaard, H., & Thomas, D. (2004). Effective feedback to small and large classes. In *34th Annual Frontiers in Education, 2004. FIE 2004.*, F1E-9. https://doi.org/10.1109/FIE.2004.1408573

Sychev, O. (2022). Write a line: Tests with answer templates and string completion hints for self-learning in a CS1 course. In *Proceedings of the ACM/IEEE 44th International Conference on Software Engineering: Software Engineering Education and Training*, 265–276. https://doi.org/10.1109/ICSE-SEET55299.2022.9794157

Thorndike, E. L. (1927). The law of effect. *American Journal of Psychology*, *39*(1/4), 212–222. JSTOR. https://doi.org/10.2307/1415413

Toma, I., Marica, A., Dascalu, M., & Trausan-Matu, S. (2021). Readerbench—Automated feedback generation for essays in Romanian. *University Politehnica of Bucharest Scientific Bulletin Series C-Electrical Engineering and Computer Science*, *83*(2), 21–34.

Tunstall, P., & Gsipps, C. (1996). Teacher feedback to young children in formative assessment: A typology. *British Educational Research Journal*, *22*(4), 389–404. https://doi.org/10.1080/0141192960220402

Vittorini, P., & Galassi, A. (2023). rDSA: An intelligent tool for data science assignments. *Multimedia Tools and Applications*, *82*(9), 12879–12905. https://doi.org/10.1007/s11042-022-14053-x

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes* (M. Cole, Ed.). Harvard University Press; WorldCat.

Wiener, N. (1954). *The human use of human beings: Cybernetics and society*. New Haven: Houghton Mifflin.

Wingate, U. (2012). 'Argument!' helping students understand what essay writing is about. *Journal of English for Academic Purposes*, *11*(2), 145–154. https://doi.org/10.1016/j.jeap.2011.11.001

Wisniewski, B., Zierer, K., & Hattie, J. (2019). The power of feedback revisited: A meta-analysis of educational feedback research. *Frontiers in Psychology*, *10*, 1664–1708. https://doi.org/10.3389/fpsyg.2019.03087

Wisniewski, B., Zierer, K., & Hattie, J. (2020). The power of feedback revisited: A meta-analysis of educational feedback research. *Frontiers in Psychology*, *10*, 1664–1708. https://doi.org/10.3389/fpsyg.2019.03087

Wolcott, M. D., & McLaughlin, J. E. (2024). Exploring user experience (UX) research methods in health professions education. *Currents in Pharmacy Teaching and Learning*, *16*(2), 144–149. https://doi.org/10.1016/j.cptl.2023.12.010

Yu, S., Jiang, L., & Zhou, N. (2020). Investigating what feedback practices contribute to students' writing

motivation and engagement in Chinese EFL context: A large scale study. *Assessing Writing*, *44*, 100451. https://doi.org/10.1016/j.asw.2020.100451

Appendix. Coding Results

A.1 Coding Results 1

Table A1. Description of Included Studies – Part 1

Study	Author Country/ Region	Publication Year	Publication Venue	Data Type	Data Source	Educational Domain	AFG System	Method for Generating AF	Evaluation of AFG
Ahmed et al., 2022	Singapore, South Korea	2022	JA	AFF, COG	FIR, SEC	11	Verifix	COM	SUR, MEA
Behzad et al., 2024	United States	2024	CP	COG	FIR	23	LEAF	LLM	MEA, EXP
Beltrami et al., 2006	Italy	2006	JA	N/A	N/A	51	SBLC	COM	N/A
Bhatia et al., 2018	India, United Kingdom, United States	2018	CP	COG	SEC	11	N/A	MLE, RUL	MEA
Choi et al., 2021	South Korea	2021	CP	COG	FIR	11	MUNCK	COM, RUL	MEA
Dai et al., 2024	Australia, United States	2024	JA	COG	FIR	23	N/A	LLM	EXP, MEA, SUR
Dzikovska et al., 2014	United Kingdom, United States	2014	JA	AFF, COG	FIR	40	BEETLE II	NLP, RUL, COM	SUR, TES, BEH, MEA, CON
Fossati et al., 2015	Qatar, United States	2015	JA	AFF, COG	FIR	11	iList	DAT, MLE	SUR, TES, BEH, CON, MEA
Gao et al., 2016	United States	2016	CP	COG	FIR	11	N/A	COM, RUL	BEH, MEA
Guid et al., 2019	Slovenia	2019	CP	AFF, COG	FIR	52	N/A	MLE, RUL, COM	TES, CON, SUR, BEH, MEA
Haut et al., 2023	United States	2023	CP	AFF, COG	FIR	51	SOPHIE	COM, NLP, RUL	CON, BEH, SUR, TES, MEA
Heo et al., 2023	South Korea	2023	CP	COG	SEC	11	REFERENT	COM, DAT, NLP,	MEA
Hossain et al., 2021	United States	2021	CP	COG	FIR, SEC	16	ASLHelp	RUL, MLE	EXP, MEA
Ifflander et al., 2015	Germany	2015	CP	COG	FIR	11	PABS	RUL, COM	MEA, BEH
Jia et al., 2022	United States	2022	JA	COG	FIR	23	Insta-Reviewer	DAT, NLP	EXP, MEA
Kang et al., 2019	United States	2019	JA	COG	SEC	14	N/A	RUL, NLP	MEA
Kim et al., 2016	United States	2016	CP	AFF, COG	FIR, SEC	11	APEX	COM, MLE	SUR, CON, MEA
Kylvaja et al., 2019	Finland	2019	CP	AFF	SEC	42	N/A	DAT, MLE	MEA

Study	Author Country/ Region	Publication Year	Publication Venue	Data Type	Data Source	Educational Domain	AFG System	Method for Generating AF	Evaluation of AFG
Lu & Cutumisu, 2021	Canada	2021	CP	COG	SEC	23	N/A	TEM, NLP	MEA, EXP
Nayak et al., 2024	Australia, India	2024	JA	COG	FIR	11	ASQGS	COM, MLE	MEA
Obaido et al., 2020	South Africa	2020	CP	AFF	FIR	11	TalkSQL	RUL, TEM, NLP	SUR
Paaßen et al., 2016	Germany	2016	CP	COG	SEC	11	N/A	COM, DAT, MLE	MEA
Rivers & Koedinger, 2013	United States	2013	CP	COG	SEC	11	N/A	DAT, MLE	MEA
Rudolph et al., 2024	Germany	2024	CP	AFF, COG	FIR	42	N/A	LLM, TEM	SUR, EXP, MEA
Ruiz & Snoeck, 2021	Belgium, Cuba	2021	CP	AFF, COG	FIR	11	FENIkS	TEM, RUL	SUR, CON, TES
Singh et al., 2013	United States	2013	CP	COG	SEC	11	N/A	RUL, COM	MEA
Toma et al., 2021	Romania	2021	JA	COG	SEC	23	ReaderBench	RUL, MLE	MEA
Van Praet et al., 2024	Belgium	2024	CP	COG	FIR	11	ASSIST	RUL, TEM	MEA
Vittorini & Galassi, 2023	Italy	2023	JA	AFF, COG	FIR	27	rDSA	COM, MLE, RUL, NLP	SUR, TES, BEH
Wang et al., 2020	China	2020	JA	COG	SEC	11	N/A	DAT, COM, RUL	MEA
Xie et al., 2024	China	2024	CP	COG	SEC	11	BRAFAR	COM, RUL	MEA
Zhan & Hsiao, 2022	United States	2022	CP	COG	SEC	11	LEGO NLPr	NLP, RUL, DAT, MLE	MEA
Zhang et al., 2022	United States, China	2022	CP	COG	SEC	11	CLEF	DAT, COM	MEA
Zhong et al., 2024	China	2024	CP	COG	SEC	23	N/A	LLM, RUL	EXP, MEA

Notes. From left to right, the column headings in Table A1 indicate (1) the APA in-text citation; (2) the author countries; (3) publication year; (4) publication venue; (5) type of the analyzed data; (6) source of the analyzed data; (7) adopted educational domain; (8) name of the proposed AFG system or system; (9) method used for automated feedback generation; and (10) the evaluation of the proposed AFG system or method.

A.2 Coding Results 2

Table A2. Description of Included Studies – Part 2

Study	Feedback Message	Feedback Function	Feedback Processing	Student Characteristic	Feedback Pedagogy	Feedback Presentation	Feedback Agent	Ethical Issues Discussion	Ethical Issues Discussion Necessity
Ahmed et al., 2022	LOA-L, INF-MIS	LEA	FEE	N/A	N/A	DEL, SIN, NON, UNI	COMP, HUM	No	No
Behzad et al., 2024	VAL, LOA-H, INF-CON, INF-MIS	LEA	FEE	N/A	Using rubrics	DEL, MUL, ADA, UNI	COMP, HUM	Yes	Yes
Beltrami et al., 2006	VER, LOA-H, INF-CON	LEA	STU	Learning needs and paces	N/A	IMM, DEL, MUL, ADA, UNI	COMP	No	No
Bhatia et al., 2018	VER, LOA-L, INF-MIS, INF-COR	LEA	FEE	N/A	N/A	DEL, SIN, ADA, UNI	COMP	No	No
Choi et al., 2021	LOA-L, INF-MIS, INF-COR	LEA	FEE	N/A	N/A	DEL, SIN, ADA, UNI	COMP	No	No
Dai et al., 2024	VAL, LOA-H, INF-PRO	LEA, MOT, SRL	FEE	N/A	Hattie and Timperley's (2007) model	DEL, SIN, ADA, UNI	COMP, HUM	No	Yes
Dzikovska et al., 2014	VAL, LOA-H, INF-PRO	LEA, SRL	STU	Student problem-solving responses	Conceptual change approach	IMM, MUL, ADA, UNI	COMP	No	No
Fossati et al., 2015	VER, VAL, LOA-L, LOA-H, INF-PRO	LEA, MOT, SRL	BAL	Student historical problem-solving behavior	Human tutoring analysis	IMM, MUL, ADA, MULTI	COMP	No	No
Gao et al., 2016	VER, LOA-L, INF-MIS	LEA	FEE	N/A	N/A	DEL, SIN, ADA, UNI	COMP	No	No
Guid et al., 2019	LOA-L, INF-PER, INF-PRO, INF-CON	LEA	BAL	Student domain knowledge	N/A	IMM, SIN, ADA, UNI	COMP, HUM	No	No
Haut et al., 2023	VAL, LOA-H, INF-PRO	LEA, MOT, SRL	BAL	Learner reflection	MVP and 3E's communication paradigm	IMM, SIN, ADA, MULTI	COMP	Yes	Yes
Heo et al., 2023	VER, LOA-L, INF-MIS, INF-COR	LEA	FEE	N/A	N/A	DEL, SIN, ADA, UNI	COMP	No	No
Hossain et al., 2021	VER, LOA-H, INF-PRO	LEA, MOT	FEE	N/A	Using rubrics	IMM, SIN, ADA, MULTI	COMP	No	Yes

Study	Feedback Message	Feedback Function	Feedback Processing	Student Characteristic	Feedback Pedagogy	Feedback Presentation	Feedback Agent	Ethical Issues Discussion	Ethical Issues Discussion Necessity
Ifflander et al., 2015	VER, LOA-H, INF-MIS	LEA	FEE	N/A	N/A	IMM, SIN, ADA, UNI	COMP	No	No
Jia et al., 2022	LOA-H, INF-PRO	LEA	FEE	N/A	N/A	DEL, SIN, ADA, UNI	COMP, HUM	Yes	No
Kang et al., 2019	VER, LOA-H, INF-COR, INF-PRO	LEA	FEE	N/A	N/A	DEL, SIN, ADA, UNI	COMP	No	No
Kim et al., 2016	VER, LOA-L, INF-MIS, INF-COR	LEA	FEE	N/A	N/A	DEL, SIN, ADA, UNI	COMP, HUM	No	No
Kylvaja et al., 2019	VAL, LOA-H, INF-PRO	MOT	STU	Learner well-being context and responses	N/A	DEL, SIN, NON, UNI	COMP, HUM	Yes	Yes
Lu & Cutumisu, 2021	LOA-L, INF-PRO	LEA	FEE	N/A	N/A	DEL, SIN, ADA, UNI	COMP, HUM	No	No
Nayak et al., 2024	VER, LOA-H, INF-PRO	LEA	FEE	N/A	Bloom's taxonomy	DEL, SIN, ADA, UNI	COMP	No	No
Obaido et al., 2020	LOA-H, INF-CON	LEA	FEE	N/A	N/A	IMM, MUL, ADA, MULTI	COMP	No	Yes
Paaßen et al., 2016	LOA-L, INF-MIS	LEA	FEE	N/A	N/A	DEL, SIN, ADA, UNI	COMP	No	No
Rivers & Koedinger, 2013	VER, LOA-L, INF-MIS, INF-COR	LEA	FEE	N/A	N/A	DEL, SIN, ADA, UNI	COMP	No	No
Rudolph et al., 2024	VAL, LOA-H, INF-PRO	LEA, MOT, SRL	STU	Learner reflection	Key principles of effective feedback	DEL, SIN, ADA, UNI	COMP	Yes	No
Ruiz & Snoeck, 2021	VER, LOA-L, INF-PRO	LEA	FEE	N/A	Taxonomy of feedback types	IMM, SIN, ADA, MULTI	COMP	No	No
Singh et al., 2013	VER, LOA-L, INF-MIS, INF-COR	LEA	FEE	N/A	N/A	DEL, SIN, ADA, UNI	COMP, HUM	No	No
Toma et al., 2021	LOA-L, INF-PRO	LEA	FEE	N/A	N/A	DEL, MUL, ADA, UNI	COMP	No	No
Van Praet et al., 2024	VER, LOA-L, INF-MIS, INF-COR	LEA	FEE	N/A	N/A	DEL, SIN, ADA, UNI	COMP	No	No

Study	Feedback Message	Feedback Function	Feedback Processing	Student Characteristic	Feedback Pedagogy	Feedback Presentation	Feedback Agent	Ethical Issues Discussion	Ethical Issues Discussion Necessity
Vittorini & Galassi, 2023	VER, LOA-L, INF-MIS, INF-COR, INF-CON	LEA	FEE	N/A	Iterative and learner controlled feedback	IMM, SIN, ADA, MULTI	COMP, HUM	No	No
Wang et al., 2020	VER, LOA-L, INF-MIS, INF-COR	LEA	FEE	N/A	N/A	DEL, SIN, ADA, UNI	COMP	No	No
Xie et al., 2024	VER, LOA-L, INF-MIS, INF-COR	LEA	FEE	N/A	N/A	DEL, SIN, ADA, UNI	COMP	No	No
Zhan & Hsiao, 2022	VER, LOA-L, INF-MIS, INF-COR	LEA	FEE	N/A	N/A	DEL, SIN, ADA, MULTI	COMP	No	No
Zhang et al., 2022	VER, LOA-L, INF-MIS, INF-COR	LEA	FEE	N/A	N/A	DEL, SIN, ADA, UNI	COMP	No	No
Zhong et al., 2024	VER, LOA-L, INF-COR	LEA	FEE	N/A	N/A	DEL, SIN, ADA, UNI	COMP	No	No

Note. From left to right, the column headings in Table A2 indicate (1) the APA in-text citation; (2) feedback message; (3) feedback function; (4) feedback processing (5) student characteristics of feedback design; (6) feedback pedagogy consideration; (7) feedback presentation; (8) feedback agent; (9) ethical issues discussed; and (10) ethical issues that need to be discussed but neglected.