

An Introduction to Pedagogical Evaluation of Automated Feedback (PEAF)

Peter B. Johnson¹, Marcus Messer^{1,*}, Camille Kandiko Howson¹, Alexandra Neagu¹, Jaromir Savelka², Prarthana Bhattacharyya³ and Simon Woodhead³

¹Imperial College London, London, UK

²Carnegie Mellon University, Pittsburgh, USA

³Eedi, London, UK

Abstract

Feedback is an essential aspect of students' learning. Automated feedback systems in the form of deterministic algorithms, machine learning models, and, more recently, generative AI have been successfully used to provide feedback to large cohorts across many domains, from essay writing to medicine to computer science. However, while there have been many automated feedback systems successfully deployed at scale, many fail to address whether the system supports learning. The Pedagogical Evaluation of Automated Feedback (PEAF) is research that investigates how automated feedback systems support learning, either by delivering feedback messages to students as information or by enabling the feedback process that allows students to close the gap between their actual and reference levels.

In this inaugural workshop, participants from the international community first reviewed examples of automated feedback and found that there is a diversity of views on what constitutes "good" feedback. The participants then discussed research challenges and directions derived from the submitted papers. They identified four key themes: privacy, whether automated feedback is equivalent to human feedback, the downstream effects of the feedback message, and learning over time. They highlighted future research directions, including sharing best practices for handling student privacy, developing pedagogically grounded benchmarks for evaluating automated feedback messages, examining how learning goals affect the downstream effects of feedback messages, and exploring issues when measuring learning over time.

Keywords

Automated Feedback, Automated Assessment, Pedagogy, Pedagogical, Evaluation, Feedback Quality

1. Introduction

Feedback is at the heart of education. For large cohorts, efforts to automate the role of the teacher in feedback have long been attempted. Early work began in the 1920s [1] and was popularised by Skinner in the 1950s [2, 3]. Since then, in the computer era, automated feedback has been researched across domains from computer science [4, 5] to essay writing [6] to medicine [7]. Computerised automation began with rules-based algorithms [8], and more recently, large language models (LLMs) have been increasingly employed [9, 10].

Automation typically evaluates a student's input against a set of rules or criteria, and provides a message back to the student [8]. While it is necessary to conduct technical evaluations when developing automation systems, there is also a need to evaluate the educational impact of the systems. Too often, we see research in educational contexts that fails to address the following question:

Does the system support learning?

PEAF 2026: The First International Workshop on Pedagogical Evaluation of Automated Feedback, June 28 2026, Seoul, South Korea

*Corresponding Author

✉ peter.johnson@imperial.ac.uk (P. B. Johnson); m.messer@imperial.ac.uk (M. Messer); c.howson@imperial.ac.uk (C. K. Howson); alexandra.neagu20@imperial.ac.uk (A. Neagu); jsavelka@andrew.cmu.edu (J. Savelka); prarthana.bhattacharyya@eedi.com (P. Bhattacharyya); simon.woodhead@eedi.com (S. Woodhead)

ORCID 0000-0001-7841-691X (P. B. Johnson); 0000-0001-5915-9153 (M. Messer); 0000-0003-1045-3655 (C. K. Howson); 0009-0004-0840-3776 (A. Neagu); 0000-0002-3674-5456 (J. Savelka); 0009-0009-0191-7148 (P. Bhattacharyya); 0000-0002-2192-9797 (S. Woodhead)



© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

We call research that addresses this question *pedagogical* evaluation of automated feedback (PEAF) [11]. In order to emphasise the importance of pedagogical evaluation, the first annual workshop on PEAF was created with this special focus.

The workshop aimed to explore in more depth how to evaluate the pedagogical quality of automated feedback. In this introductory paper we summarise the outcomes of the workshop. Firstly, the in-person only, day-long, interaction-driven approach emphasised community building. Secondly, framing of the workshop, which is reviewed in Section 2, encouraged clear theoretical constructs. Thirdly, the workshop brought together peer-reviewed research dedicated to PEAF, which we briefly review in Section 2.5 and follows in more detail in the remainder of the proceedings. Finally, we collaborated to identify research directions for the community, which we review in Section 3.

2. Theoretical Framing

A foundation for the evaluation of automated feedback is a definition of feedback. In this section, we explore a tension in the literature between feedback as information and feedback as a process. We resolve the tension by advocating for more specificity, identifying whether feedback *messages* or feedback *processes* are under discussion. By clarifying our definitions of feedback itself, we can be clearer when pursuing the pedagogical evaluation of automated feedback (PEAF).

Central to all definitions of feedback is the concept of

“...a gap between a current level of performance and some desired level...” [12, p. 154]

Information about the gap is essential but there is tension around whether the concept of feedback covers more than just information about the gap.

2.1. Feedback As Information

Hattie and Timperley’s [13] widely cited meta-analysis on the effects of feedback is explicit that

“feedback is conceptualized as information” [13, p. 81]

The emphasis on feedback as information excludes from the definition its effects, such as improved learning. While Hattie and Timperley [13] reviewed studies that report the impacts of feedback, and focus their research on those impacts, such impacts nevertheless fall outside the definition of ‘feedback’. Such a delineation is important for a community focusing on evaluating automated ‘feedback’. Is it sufficient for PEAF research only to evaluate the information provided to a student?

Other influential research on feedback broadly agrees with a definition focusing on information. Shute [12] reviewed studies on formative feedback to identify best practice, defining feedback as information [12, p. 154] and using it as such, for example, concluding that best practice is to provide timely, actionable, objective ‘feedback’. Lipnevich and Panadero’s [14] review of theoretical models found that 13 of the 14 models employed a ‘feedback as information’ view.

In addition to theoretical research, we can consider our day-to-day experiences in education. Teachers will refer to ‘giving’ students feedback and to students’ complaints that they do not ‘get’ enough feedback [15]. Students may not have ‘received’ their feedback yet or may find it unclear [16]. These lived experiences reflect a discourse employing the definition of feedback as information. If ‘feedback’ is used in practice to mean information, then on Wittgenstein’s meaning-as-use argument [17, Section 43] that is the definition of feedback.

2.2. Feedback as a Process

Information about a gap is not necessarily used by students, or if it is used, then it does not necessarily have a positive impact. Sadler [18] included a key additional criterion, defining feedback as

“...information about the gap between the actual level and the reference level...*used to alter the gap* in some way”

— Ramaprasad [19] quoted by Sadler [18, p. 120]. Emphasis added.

The key inclusion of ‘used to alter the gap’ broadens the definition of feedback. Such a distinction has important implications for evaluating feedback. If a feedback *message* follows Shute’s [12] criteria, is that sufficient to be evaluated as good feedback? There is a strong body of literature that says ‘no’; feedback must “focus ...on students’ actions in response” [20].

The concept of feedback as a wider process, of which the message is just a part, is not new, but Winstone et al. [21] have been influential in emphasising this difference in recent years. The focus has shifted from the feedback messages provided by teachers to how the student responds. For example, student feedback ‘literacy’ [22, 23] refers to their appreciation of feedback and their ability to engage.

The process-view is more explicitly invoked when referring to ‘feedback experiences’ [23, p. 1319], and feedback as a dialogue [24, 25], where students interact with the information provided and with other stakeholders in a social process of meaning-making.

2.3. Ambiguity and the Need for Specificity

The previously cited studies were selected for their explicit positioning of feedback either as information or as a wider process. Between those studies, however, is a wider literature that leaves ambiguity. For example, Black and Wiliam’s influential work on formative feedback¹ [26] notes the discrepancy between Kluger and DeNisi’s [27] large-scale meta review, which conceives of feedback as information, and Ramaprasad’s [19] process view, but nevertheless employs definitions with varying emphasis on information or process.

Nicol and MacFarlane’s [28] model of formative assessment is also ambiguous, starting with a definition of feedback as information before developing a model of constructing meaning, reconceptualising feedback as dialogue as opposed to transmission, and invoking Boud’s feedback loop.

The influential works of Black and Wiliam [26] and Nicol and MacFarlane [28] represent a wider literature that seems content with ambiguous definitions. For the emerging field of PEAf, it would help the community to clarify the definition. While there are obvious practical benefits to clearer definitions, they also shape research questions and the nature of the automation activities.

The apparent tension in the varying definitions can be resolved if researchers are careful to explicitly distinguish between feedback *messages* and feedback *processes*. Both are valid and valuable pursuits. We advocate for clarity to avoid confusion. Simply referring to ‘feedback’ is ambiguous. When it comes to pedagogical evaluation of automated feedback, we simply ask that researchers append clarification—messages or processes. We also suggest that, given the prevalence of an information view of feedback in practice, researchers in automation may want to focus on the less-obvious, more challenging, and potentially more valuable problem of evaluating what happened before or after a message was generated.

2.4. Automated Feedback

We now refine our focus to automation, which Skinner [2] defined as removing the teacher from the process. Automation itself referred to the generation of informative messages, but Skinner’s behaviourism was concerned with the effects of ‘reinforcement’, i.e. closing the gap, undoubtedly a process-oriented view of feedback.

Automation in the era of the computer has retreated from the process view, and this reframing has important implications for research. Jensen et al. [29] highlight that ‘nonconscious conceptualisations shape the research questions we ask and the online feedback practices we recommend’ [29]. For example, the very idea of ‘automating feedback’² is based on feedback as information. Jensen et al. [29] surveyed the corpus of e-learning research and found a predominantly feedback-as-information view.

¹Which was called ‘assessment for learning’ or ‘formative assessment’.

²Referred to by Jensen et al. [29] as ‘feedback input’ [to students].

While Jensen et al.'s work covered e-learning in general, literature specifically on automated feedback also shows an information-centric view. Evaluations report accuracy, specificity, or agreement with human raters [4]. Classifications of automated feedback sort systems by software architecture, message type and timing [5, 8, 30]. All of these approaches imply a view of feedback as information. While modern work on automated feedback conceptualises feedback as information, we advocate for PEAf to not just evaluate information, but to do so from an educational perspective, which means attending to the whole process.

2.5. Locating Research on PEAf

PEAf can mean pedagogical evaluation of messages, or of the wider process in which they are generated and acted upon. In the literature, most evaluations are focused on message accuracy within a specific domain, rather than pedagogical qualities; research considering pedagogical aspects is often limited to relatively shallow measurements of student opinions about feedback messages [4].

In contrast, this workshop brings together nine peer-reviewed papers that constitute the remainder of these proceedings and exemplify genuinely pedagogical evaluation. Five of the studies focused on feedback messages, but with the common strength that they evaluate the messages across multiple dimensions relevant to education. Of the other studies, one bridges message evaluation and fairness outcomes, while the others consider the wider process, including pre-message diagnosis, perceptions of messages, and effects on learning behaviour and outcomes.

In summary, PEAf answers the question of whether automated feedback supports learning. It does this by evaluating feedback messages from a pedagogical perspective, or by evaluating the wider feedback processes that narrow the gap between the present status and an agreed goal. Research focused on messages must acknowledge the limited scope of any research claims, and that the pedagogical value of the messages is only ultimately seen when the wider process is considered. Research into other parts of the process is less common, but is essential and therefore encouraged.

3. Research Directions

The PEAf workshop was a forum for discussion about research directions. We began by reviewing samples of automated feedback messages (replicated in Appendix A) and asked:

- Is this feedback?
- Is this good feedback?
- Why / why not?

The activity highlighted the diversity of views within the community, which spans all countries, all academic domains, and all levels of education. While participants were generally aligned on the role and importance of feedback, there were different perspectives on whether the feedback messages were “good”. For example, with Appendix A.1, some felt the message provided good scaffolding, helping sufficiently and avoiding doing ‘too much’ for the student, while others felt that the message was too vague. Clearly, evaluation is contextual, which has implications for research directions. We cannot expect context-free research findings in PEAf. We may, however, still have sufficiently common interests to share research directions. The aim of the workshop was to identify shared directions.

3.1. Participant Activities

To identify research directions for PEAf, participants were provided with a set of topics based on the peer-reviewed research presented at the workshop and formed groups based on their interests.

The topics selected by the participants were: AI feedback message generation, feedback message design, understanding the student, and trust and validity.

Participants were asked to identify research directions, with the following guiding questions:

- What challenges within this theme would I or others like to work on?
- Which challenges in this theme are the most urgent and the most important in the long run?
- Which challenges in this theme should be solved by academia and/or industry?
- What challenges within this theme have we been asking the wrong question about? What challenges within this theme would change practice if solved tomorrow?

Participants then ranked the sub-topics they had identified, and formed four groups through self-nomination. Participants added handwritten notes and post-it notes to flip-chart paper with the theme title, while discussing together. Participants could optionally join more than one theme during the time available.

3.2. Key Themes and Critical Discussion

In this section, we summarise the participants' discussions of the challenges within these key themes and potential avenues for future work.

Privacy Automating feedback messages inherently involves processing data related to students, leading to concerns about privacy, which participants advocated as a key underlying theme that needs prioritising when automating feedback messages.

Privacy was framed through legal, ethical, and technical lenses. Legal components are the incentives and penalties. Ethical aspects include self-governance, community standards, and collective voice or resistance. Technical aspects include data security, infrastructure, governance, and software design.

Privacy is not a direct research question, but an enabling layer for research. Participants expressed an interest in sharing best practice and co-developing standards, such as risk management protocols and public data sets.

Is Automated Feedback the Same as Human Feedback? The comparison of human-generated and automated feedback messages was discussed during the workshop. Participants highlighted that automated feedback differs from human feedback, prompting further discussion of how educational theory and educational practice, especially in automated feedback, differ.

The participants suggested that future work in the pedagogical evaluation of automated feedback messages could involve developing a pedagogically grounded benchmark, investigating how theory is applied in practice, and examining the design and technical limitations of current approaches to providing automated feedback messages.

Downstream of the Message Framed by Section 2, the step following message delivery was chosen for discussion — what is the impact of a feedback message? Participants proposed that in principle these studies should define goals, identify a 'gap' between the present situation and the goal, and study changes in the gap.

Goals can be complex, and more distal or long-term goals can be more contestable. On that basis the participants suggested that the community focus on proximal or short-term goals for early research gains. Well-defined proximal goals are often the least ambiguous, most immediately actionable, and most benefit from timely input — aligning with the strengths of automated feedback messages.

There was also a reflection on how realistic it is to measure changes in identified gaps. 'No measurable change' is a likely outcome in many cases, and we should embrace this as part of the uncertainty of research and development.

Learning Over Time How to assess and account for students' prior knowledge, and how to ensure comprehensible feedback messages. Across the contexts of the participants, no single approach to this issue was agreed. For example, knowledge graphs are used in educational data mining to represent students' knowledge in a structured format; however, some participants raised concerns about their limitations. Future research could explore the applicability of knowledge graphs and their limitations.

4. Discussion and Conclusions

At the first annual PEAf workshop, the focus on whether automated feedback supports learning yielded multiple benefits. Putting aside the technical challenges of automation enabled a critical and nuanced discussion about the wider purpose of automation. Our theoretical framing emphasises the need for specificity in research. Feedback messages can be automated as part of a wider process of student engagement and learning, which is where the ultimate impact is seen.

Workshop activities brought a diverse community together. Through sharing peer-reviewed research and discussing challenging topics within PEAf, we found common ground and explored differences stemming from our diverse educational contexts. Automated feedback is a complex area facing significant research challenges. Discussions identified some important research directions for the community, including enabling best practice in privacy, investigating how automated feedback differs from human feedback, identifying the hallmarks of good research on processes downstream of message delivery, and raising issues around the measurement of learning over time.

The research directions identified here do not claim to represent a sector-wide consensus or a rigorously derived set of priorities. They represent expert opinion, drawn from the experience and knowledge of researchers active in the field. Their value lies in guiding anyone seeking productive and, ultimately, useful research avenues, as agreed by a small but significant group of researchers committed to PEAf.

Research into messages is often more conclusive, but it must clearly limit the scope of its claims regarding educational impact. Evaluating the quality of a feedback message is necessary but not sufficient to fully evaluate the impact of automated feedback on students' learning.

We advocate for research that asks whether automated feedback messages support learning. We encourage studies beyond message accuracy, exploring the multi-dimensional evaluation of messages, and the broader process, of which message generation is just one part.

Acknowledgments

We would like to thank the workshop participants for their in-depth interaction, the authors who submitted and presented their work, and the program committee members for their reviews.

We would like to further thank the Imperial Global Connect Fund.

Declaration on Generative AI

During the preparation of this work, the author(s) used Claude Sonnet to review and comment on the paper's structure, writing and organisation. No generative AI content was incorporated into the manuscript. The author(s) reviewed all suggestions and take full responsibility for the publication's content.

References

- [1] S. Petrina, Sidney pressey and the automation of education, 1924-1934, *Technology and Culture* 45 (2004) 305–330. URL: <http://www.jstor.org/stable/40060743>.
- [2] B. F. Skinner, Teaching machines: From the experimental study of learning come devices which arrange optimal conditions for self-instruction., *Science* 128 (1958) 969–977.
- [3] A. Watters, Teaching machines: The history of personalized learning, mit Press, 2023.
- [4] M. Messer, N. C. Brown, M. Kölling, M. Shi, Automated grading and feedback tools for programming education: A systematic review, *ACM Transactions on Computing Education* 24 (2024) 1–43.
- [5] H. Keuning, J. Jeuring, B. Heeren, A systematic literature review of automated feedback generation for programming exercises, *ACM Transactions on Computing Education (TOCE)* 19 (2018) 1–43.

- [6] D. Ramesh, S. K. Sanampudi, An automated essay scoring systems: a systematic literature review, *Artificial Intelligence Review* 55 (2022) 2495–2527. doi:10.1007/s10462-021-10068-2.
- [7] F. Holderried, C. Stegemann-Philipps, A. Herrmann-Werner, T. Festl-Wietek, M. Holderried, C. Eickhoff, M. Mahling, A language model-powered simulated patient with automated feedback for history taking: Prospective study, *JMIR Medical Education* 10 (2024) e59213.
- [8] G. Deeva, D. Bogdanova, E. Serral, M. Snoeck, J. De Weerd, A review of automated feedback systems for learners: Classification framework, challenges and opportunities, *Computers & Education* 162 (2021) 104094.
- [9] H. Shi, V. Aryadoust, A systematic review of ai-based automated written feedback research, *ReCALL* 36 (2024) 187–209.
- [10] R. L. Moore, S. S. Lee, Harnessing generative ai (genai) for automated feedback in higher education: A systematic review, *Online Learning Journal* (2024).
- [11] M. Messer, P. B. Johnson, A. Neagu, C. Kandiko Howson, J. Savelka, S. Woodhead, The First International Workshop on Pedagogical Evaluation of Automated Feedback (PEAF 2026), in: *Artificial Intelligence in Education. Late Breaking Results, WideAIED, Practitioners, Industry and Policies, Blue Sky, Doctoral Consortium, FoL Workshops and Tutorials, FoL Invited Papers*, Springer Nature Switzerland, 2026, pp. 55–60.
- [12] V. J. Shute, Focus on Formative Feedback 78 (2008) 153–189. doi:10.3102/0034654307313795.
- [13] J. Hattie, H. Timperley, The power of feedback, *Review of Educational Research* 77 (2007) 81–112. doi:10.3102/003465430298487.
- [14] A. A. Lipnevich, E. Panadero, A review of feedback models and theories: Descriptions, definitions, and conclusions, in: *Frontiers in education*, volume 6, Frontiers Media SA, 2021, p. 720195.
- [15] E. Pitt, N. Winstone, Enabling and valuing feedback literacies, *Assessment and evaluation in higher education* 48 (2022) 149–157.
- [16] A. J. McLean, C. H. Bond, H. D. Nicholson, An anatomy of feedback: a phenomenographic investigation of undergraduate students' conceptions of feedback, *Studies in Higher Education* 40 (2015) 921–932.
- [17] L. Wittgenstein, *Philosophical Investigations*, 4 ed., Wiley-Blackwell, Chichester, 2009.
- [18] D. R. Sadler, Formative assessment and the design of instructional systems, *Instructional science* 18 (1989) 119–144.
- [19] A. Ramaprasad, On the definition of feedback, *Behavioral science* 28 (1983) 4–13.
- [20] N. Winstone, D. Carless, *Designing effective feedback processes in higher education: A learning-focused approach*, Routledge, 2019.
- [21] N. Winstone, D. Boud, P. Dawson, M. Heron, From feedback-as-information to feedback-as-process: a linguistic analysis of the feedback literature, *Assessment & Evaluation in Higher Education* 47 (2022) 213–230.
- [22] P. Sutton, Conceptualizing feedback literacy: Knowing, being, and acting, *Innovations in Education and Teaching International* 49 (2012) 31–40.
- [23] D. Carless, D. Boud, The development of student feedback literacy: Enabling uptake of feedback, *Assessment & evaluation in higher education* 43 (2018) 1315–1325.
- [24] D. Nicol, The power of internal feedback: Exploiting natural comparison processes, *Assessment & Evaluation in higher education* 46 (2021) 756–778.
- [25] R. Esterhazy, What matters for productive feedback? disciplinary practices and their relational dynamics, *Assessment & Evaluation in Higher Education* 43 (2018) 1302–1314.
- [26] P. Black, D. Wiliam, Assessment and classroom learning, *Assessment in Education: principles, policy & practice* 5 (1998) 7–74.
- [27] A. N. Kluger, A. DeNisi, The effects of feedback interventions on performance: a historical review, a meta-analysis, and a preliminary feedback intervention theory., *Psychological bulletin* 119 (1996) 254.
- [28] D. J. Nicol, D. Macfarlane-Dick, Formative assessment and self-regulated learning: A model and seven principles of good feedback practice, *Studies in higher education* 31 (2006) 199–218.
- [29] L. Jensen, M. Bearman, D. Boud, Understanding feedback in online learning—a critical review and

metaphor analysis (2021).

- [30] A. P. Cavalcanti, A. Barbosa, R. Carvalho, F. Freitas, Y.-S. Tsai, D. Gašević, R. F. Mello, Automatic feedback in online learning environments: A systematic literature review, *Computers and Education: Artificial Intelligence* 2 (2021) 100027.
- [31] P. B. Johnson, J. Fenton, P. Ramsden, R. Chatley, M. Ribera-Vicent, K. Lundengård, Formative feedback on engineering self-study: Towards 1 million times per year per cohort, in: 2025 IEEE Global Engineering Education Conference (EDUCON), IEEE, 2025, pp. 1–3.
- [32] H. McNichols, F. Ikram, A. Lan, The studychat dataset: Analyzing student dialogues with chatgpt in an artificial intelligence course, in: *Proceedings of the LAK26: 16th International Learning Analytics and Knowledge Conference, LAK '26*, Association for Computing Machinery, New York, NY, USA, 2026, pp. 53–63. URL: <https://doi.org/10.1145/3785022.3785029>. doi:10.1145/3785022.3785029.
- [33] J. Ye, S. Wang, J. Chen, H. Wang, D. Zou, Y. Zhu, J. Tang, H.-T. Zheng, R. Liu, H. Li, et al., Feanel: A benchmark for fine-grained error analysis in k-12 english writing, *arXiv preprint arXiv:2511.22883* (2025).

A. Feedback Examples

A.1. Example 1

Context: Homework from a higher education mechanical engineering course. Data from Lambda Feedback [31].

Question: Find the required number of belts

Student's response: 2.41

Teachers answer: 3

Feedback: “Close, remember only an integer number of belts exist in the real world.”

A.2. Example 2

Context: Chatbot dialogue from a college-level AI course. Data from StudyChat [32]

Student question: “how can i get all of the unique characters from a string”

Chatbot response:

To get all of the unique characters from a string in Python, you can use the following methods:

Method 1: Using a Set

A set automatically stores unique items. You can convert the string to a set and then, if needed, convert it back to a string or list.

Here's an example:

```
input_string = "hello world"
unique_characters = set(input_string)
print(unique_characters) # Output: {'o', 'r', ' ', 'l', 'h', 'e', 'w', 'd'}
```

Method 2: Using a Loop

You can also use a loop to manually collect unique characters.

```
input_string = "hello world"
unique_characters = []
```

A.3. Example 3

Context: K-12 English essay for English as a foreign language students, with a focus on grammar rules. Data from [33].

Student's sentence: “Mr. Cool Clothes Store’s clothes is very check now.”

Correct sentence: “Mr. Cool Clothes Store’s clothes are very cheap now.”

Feedback:

Subject-Verb – Severity 4

Error Explanation

1. Subject-Verb Agreement Error: the subject *<clothes>* is plural and requires the plural verb [are], so *<is>* is changed to [are].
2. Adjective Error: *<check>* means to examine; the adjective [cheap] indicating price should be used here, so *<check>* is changed to [cheap].