How do we define and evaluate "good" automated feedback?

P. B. Johnson[a,1], **A. Neagu**[b], **M. Messer**[c], **K. Lundengard**[d], **P. Ramsden**[e]

[a] Imperial College London, UK, ORCID 0000-0001-7841-691X
[b] Imperial College London, UK, ORCID 0009-0004-0840-3776
[c] Imperial College London, UK, ORCID 0000-0001-5915-9153
[d] Imperial College London, UK, ORCID 0000-0003-3204-617X
[e] Imperial College London, UK, ORCID 0000-0002-7099-1415

Conference Key Areas*: Digital tools and AI in engineering education, Improving higher engineering education through researching engineering education*
Keywords*: Digital, automation, formative feedback, evaluation criteria, ecosystems*

[1] *Corresponding Author*
P. B. Johnson
peter.johnson@ic.ac.uk

ABSTRACT

Automated formative feedback has the potential to improve learning, especially for large cohorts. High-quality automated feedback requires rigorous and transparent testing and evaluation of the algorithms that generate feedback. Current solutions are rarely evaluated transparently. This workshop addressed this challenge by initiating community-sourced criteria by which we might evaluate automated feedback algorithms. By establishing these criteria, we aim to support transparent and rigorous evaluation of feedback algorithms, empowering educators to make informed decisions on the technology that they deploy to their students.

The workshop focussed on feedback at the task and process levels—i.e., feedback on student work such as homework or self-study. We reviewed criteria traditionally considered important when evaluating (manual) feedback. While existing criteria were agreed to remain applicable to automated feedback, additional aspects were identified, in particular the consistency of an algorithm. Discussions around evaluating against these criteria provided initial insights into the likely process by which a teacher will select an algorithm, and initial ideas of metrics to evaluate an algorithm. Further work is required to establish a more complete set of evaluation methods.

The criteria for evaluation provided by this workshop are immediately applicable and provide a valuable reference point for practitioners and researchers in automated feedback. The problem of evaluating against these criteria requires further definition. Motivation to address this challenge remains high, as it will facilitate the interchange of feedback algorithms, increasing impact and equity of access.

1 BACKGROUND AND RATIONALE

Formative feedback is one of the most impactful interventions in education (Hattie and Timperley, 2007; Hattie, 2009). Formative feedback is defined as informing (Black and Wiliam, 1998; Sadler, 1989):

- a goal ('Where I'm going?')
- progress towards the goal ('Where am I?')
- how to progress toward the goal ('Where shall I go next?').

Providing frequent formative feedback is challenging from a resource perspective, especially for large cohorts. Feedback can also be mis-targeted due to its association with summative assessment (Winstone and Boud, 2022). If formative feedback is provided it is often by student teaching assistants (Mirza et al., 2019; Wald and Harland, 2018; Riese et al., 2021), who lack experience both in teaching and within their domain (Wald and Harland, 2018; Kristiansen et al., 2023).

To address these challenges, *automation* has the potential to enhance the impact of formative feedback on tasks, while shifting teachers' efforts to higher level feedback such as on self-regulation of learning.

Automation can have a high impact by:

- improving the consistency, timeliness, and quality of task- and process-level feedback;
- enabling teachers to focus on higher levels of feedback

1.1 The need for transparent evaluation criteria

Despite its promise, automated feedback remains fragmented — developed within isolated platforms, lacking open or standard evaluation criteria, and rarely tested across diverse educational contexts. We have identified over 200 systems, in which automation algorithms are unique to the platform and are not transparently evaluated (Deeva et al. 2021). The fragmented and opaque nature of the algorithms is not conducive to responsible use of technology by teachers, or to economies of scale to achieve equity and efficiency.

Feedback algorithms may involve AI, or rules-based evaluations for example using computer algebra systems, or a hybrid of these technologies. Whatever the technology, we propose that any algorithm should be tested against pre-agreed criteria and that the tests and results should be transparently published.

There are currently no recognised criteria by which algorithms for automated feedback can be evaluated. As a community we need to agree on what the key criteria should be and how they can be evaluated.

Considering Hattie and Timperley's (2007) four levels of feedback – task, process, self-regulation, and self – this workshop focusses on formative feedback on tasks and processes. In other words, feedback on 'homework' or self-study.

1.2 Models of good formative feedback

Shute (2008) reviewed literature on feedback and identified distinct aspects of feedback that could be used to evaluate its effectiveness, and a list of 'Do's' and 'Don'ts'. The aspects and advice are given in Table 1.

Table 1. Shute's (2008) aspects of good formative feedback

Aspect	Description
Verification	Validity of student response
Elaboration	Explain validity, possibly with examples, hints or reasoning
Specificity	Enough detail to be actionable
Complexity & length	Matches the learner's needs
Goal-directedness	Relates clearly to a learning goal
Scaffolding	Guide next steps
Timing	'At the right moment'
Learner factors	Level, style, confidence, etc.
'Do's'	- Focus on task, not learner. - Specific, clear, simple, objective. - Link feedback to goals and gaps. - Give feedback after the learner has made an attempt. - Encourage reflection or improvement — not just correction.
'Don'ts'	- Give grades - Normative comparisons ("you're better than average"). - Discourage the learner. - Praise, not related to the work. - Interrupt the learner mid-task.

1.3 Applicability to automated feedback

While the literature includes well-reviewed models of good formative feedback that is delivered manually, there is a lack of literature on how to define 'good' feedback when it is delivered automatically. This workshop focusses on that question. We begin with the criteria listed in Table 1 and in each case ask:

- Is this criterion applicable for automated feedback? Should it be adapted in anyway? Are any criteria missing?
- How can we measure performance for each of these criteria? What are the key metrics and how can we evaluated them?

Our vision is for cross-platform algorithms for automated feedback to be tested on a large scale, against public data sets, and evaluated against pre-agreed metrics. Educators can then responsibly select feedback algorithms to deploy to their students. In this workshop we address the foundational question of which criteria the algorithms should be evaluated against, and how.

2 WORKSHOP OBJECTIVES

The aim of the workshop was to gather community input on *what* are the key criteria by which we evaluate *automated* feedback, and *how* we should evaluate (measure) against those criteria.

2.1 Target audience

The workshop targeted educators/lecturers/teachers who might configure the use of automated feedback. Other stakeholders, such as policy makers and support staff, were also welcome.

2.2 Expected learning outcomes

The purpose of the workshop was to collaboratively define the criteria by which automated feedback in engineering higher education should be evaluated. The expected outcome was a list of criteria, their relative priority, and a discussion of how the criteria can be evaluated. The outcome of the workshop was envisaged as the basis of a large scale survey to validate the criteria with the wider community, before starting to publish evaluations of feedback algorithms.

3 WORKSHOP DESIGN

3.1 Time plan

Table 1. Time plan

Run time	Activity	Notes
10 min	Introduction	Problem definition, theoretical framework
10 min	Group activity 1	Discuss criteria to include/modify/exclude
10 min	Discussion	Present arguments to the wider group
15 min	Group activity 2	Develop evaluation (testing) ideas, grouped by the agent under consideration: - Teacher experts - Automated testing - Learner feedback
10 min	Discussion and conclusions	Group contributions and synthesis

3.2 Interactivity

Apart from the introduction, all sessions were interactive. Group activities 1 & 2 were multiple small groups each around a table (e.g. 5 people). Each table was expected to facilitate their own discussions, but workshop hosts also worked the room ensuring all tables had the support they need and facilitated where needed. Discussions were with the whole group. Notes were made by the hosts, summarised at the end orally, and presented here in writing.

4 WORKSHOP RESULTS

Qualitative discussions highlighted the importance, in addition to evaluation criteria, of contextual information that accompanies any evaluation. For example, for a given algorithm, applicable task types, sources of ground truth, and deployment details such as whether used for formative and/or summative feedback, and how feedback is delivered (automated, semi-automated, enhanced manual marking).

Disambiguation of the term ‘algorithm’ was also discussed. The main focus of the workshop was on evaluation criteria, as discussed in the remainder of this section.

4.1 Task 1: selecting criteria based on Shute (2008)

The aspects of feedback identified by Shute (2008) were generally accepted. This was a key outcome of the workshop. Elaboration of these aspects was discussed.

The specific aspect of *verification/validation/correctness/accuracy* requires further clarification, including the data on which tests are conducted and whether or not grades are explicitly calculated/provided.

Adaptation to learner factors should consider aspects more pertinent to automation, such as using data on learner profiles or data on prior usage by a learner.

Consistency was proposed as an additional aspect to evaluate.

4.2 Task 2: evaluating the criteria

Evaluation is to empower teachers when choosing an algorithm. Discussion led to an expected behaviour by teachers who would initially filter algorithms by applicable task types, history of usage (how many courses, how many students), average response time, and other such quantitative measures. The decision from within that shortlist would then involve more qualitative judgement, likely strongly influenced by *teacher-user reviews*. For example, if Professor X at institution Y writes a positive review about using it in their class, this would give teachers confidence to select the algorithm.

Specific discussions around metrics for evaluation yielded some initial insights:

- **Accuracy:** defined relative to an ‘expert teacher’
- **Elaboration:** presence of hints/examples. Learner ratings.
- **Specificity:** task completion rate after feedback (evidence learners acted on it).
- **Complexity and length:** Average time learners spend reading/engaging with feedback. Dropout or ignore rates (too long/complex = higher skip rate).
- **Goal-directedness:** frequency of references to learning objectives (if provided).

Comprehensive conclusions were not reached; the discussions highlighted the challenge of defining methods to evaluate against the criteria. This problem needs more precise definition, and to identify specific aspects that require further research.

5 DISCUSSION AND CONCLUSION

The work of Schute (2008) was endorsed as criteria by which we evaluate automated feedback – with some adaptations and additions, most notably adding *consistency*. On evaluating against these criteria three key insights emerged: (1) a likely selection process for teachers when choosing an algorithm, involving quantitative filtering followed by consulting qualitative reviews; (2) the importance of teacher-user reviews; (3) some metrics for evaluation were identified, while others need further development. Further collaboration with the community will be required to achieve a more complete outcome – and motivation to continue such work was high as it will facilitate the interchange of feedback algorithms, increasing impact and equity of access.

REFERENCES

Black, P. and Wiliam, D. (1998). Assessment and classroom learning. Assessment in Education: principles, policy & practice, 5(1):7–74.

Deeva, G., Bogdanova, D., Serral, E., Snoeck, M., and De Weerdt, J. (2021). A review of automated feedback systems for learners: Classification framework, challenges and opportunities. Computers & Education, 162:104094.

Hattie, J. (2009). Visible Learning: A Synthesis of Over 800 Meta-analyses Relating to Achievement. Routledge.

Hattie, J. and Timperley, H. (2007). The power of feedback. Review of educational research, 77(1):81–112.

Kristiansen, N. G., Nicolajsen, S. M., and Brabrand, C. (2023). Feedback on student programming assignments: Teaching assistants vs automated assessment tool. Proceedings of the 23rd Koli Calling International Conference on Computing Education Research, page 1–10.

Mirza, D., Conrad, P. T., Lloyd, C., Matni, Z., and Gatin, A. (2019). Undergraduate teaching assistants in computer science. Proceedings of the 2019 ACM Conference on International Computing Education Research, page 31–40.

Riese, E., Lorås, M., Ukrop, M., and Effenberger, T. (2021). Challenges faced by teaching assistants in computer science education across europe. Proceedings of the 26th ACM Conference on Innovation and Technology in Computer Science Education V. 1, page 547–553.

Sadler, D. R. (1989). Formative assessment and the design of instructional systems. Instructional science, 18(2):119–144.

Shute, V. J. (2008). Focus on formative feedback. Review of Educational Research, 78(1):153–189.

Wald, N. and Harland, T. (2018). Rethinking the teaching roles and assessment responsibilities of student teaching assistants. Journal of Further and Higher Education, 44(1):43–53.

Winstone, N. E. and Boud, D. (2022). The need to disentangle assessment and feedback in higher education. Studies in higher education, 47(3):656–667.