Backward evaluation in peer assessment: A scoping review

Kamila Misiejuk [a,b,*], Barbara Wasson [a,b]

[a] Department of Information Science & Media Studies, University of Bergen, PO Box 7800, N-5020, Bergen, Norway
[b] Centre for the Science of Learning & Technology (SLATE), University of Bergen, PO Box 7800, N-5020, Bergen, Norway

ABSTRACT

Implementing backward evaluation as part of the peer assessment process enables students to react to the feedback they receive on their work within one peer assessment activity cycle. The emergence of online peer assessment platforms has brought new opportunities to study the peer assessment process, including backward evaluation, through the digital data that the use of these systems generates. This scoping review provides an overview of peer assessment studies that use backward evaluation data in their analyses, identifies different types of backward evaluation and describes how backward evaluation data have been used to increase understanding of peer assessment processes. The review contributes to a mapping of backward evaluation terminology and shows the potential of backward evaluation data to give new insights on students' perceptions of what is useful feedback, their reactions to the feedback received and its consequences for feedback implementation.

1. Introduction

Backward evaluation (BE) (also called back-review or back-evaluation) is defined as 'the feedback that an author provides to a reviewer about the quality of the review' (Luxton-Reilly, 2009, p. 226). BE can be a part of *peer assessment* (PA), which is commonly defined as 'an arrangement for learners to consider and specify the level, value, or quality of a product or performance of other equal-status learners' (Topping, 2009, p. 20–21). Fig. 1 shows a PA process that includes BE. The common PA practice includes a student (author) developing an artefact that is later reviewed by a peer (reviewer) who gives feedback to the artefact developer (author). This feedback should be reflected on and can be used to improve the original artefact. BE is an additional step in the PA process that entails a student (author) giving feedback to and/or rating the feedback that they received on their work from the peer (reviewer), who should then reflect on the quality of the feedback provided.

From the BE receiver perspective, BE is a way to ensure that students actively process the feedback that they receive and should lead to increased student engagement and reflection, as well as changes in behaviour (Cook, 2019; Winstone, Nash, Parker, & Rowntree, 2017; Yuan & Kim, 2015). BE providers have the opportunity to improve their evaluative judgement skills of what constitutes useful feedback by evaluating the feedback they receive on their work (Tai, Ajjawi, Boud, Dawson, & Panadero, 2018), and they are also exposed to their peers' reactions to the feedback that they provide. Hence, BE is an accountability measurement to encourage students to give more useful feedback and have higher commitment to the PA task (Luxton-Reilly, 2009; Patchan, Schunn, & Clark, 2018). Potter et al. (2017) indicate BE as one of the approaches that can help students give more meaningful feedback. Giving feedback is a difficult task, especially for novices, since it is a complex process and requires students 'to recognize limitations of given answers and to

formulate clear explanations about how to improve them' (Potter et al., 2017, p. 90). Disadvantages of BE are an increased workload for students and retaliation/bias in evaluating the feedback that they receive; for example, positive feedback encourages positive BE and vice versa, and receivers often lack the skills to determine the usefulness of feedback. Additionally, it is difficult to ensure that students read their BEs (Patchan et al., 2018). Despite its challenges, BE has the potential to enhance and improve the PA process and might be a valuable step to add while designing PA activities. From a research perspective, BE provides an opportunity to gain more insights into student feedback receiving skills, as well as the interplay between the roles of the feedback receiver and feedback provider (Adewoyin, Araya, & Vassileva, 2016; Mulliner & Tucker, 2017; Patchan et al., 2018).

The advent of online PA platforms has enabled the enhancement of PA activity with new features that would be time-consuming and logistically difficult to perform in offline learning environments. BE features have been a part of online PA platforms since their emergence, often implemented as giving 'likes', ratings or comments. An early systematic review of PA tools by Luxton-Reilly (2009) lists three platforms that facilitate some form of BE: PEARS, developed in 2005; SWoRD, developed in 2007; and Aropä, developed in 2007. Nowadays, PA is a common element in online spaces such as Masssive Open Online Courses (MOOCs), Learning Management Systems (LMSs) and educational platforms developed specifically for PA (Gamage, Staubitz, & Whiting, 2021; O'Brien, Forte, Mackey, & Jacobson, 2017). These tools open new possibilities to enhance the PA process with not only new features but with the emergence of new research fields to gain insights into student learning processes from educational big data (Misiejuk, Wasson, & Egelandsdal, 2021; Romero & Ventura, 2020). In these environments, BE is an additional but integrated step in the PA process that can be utilised by an instructor during PA activity design.

In this scoping review, we focus on a particular type of BE that is 1) a step in the PA process conducted on an online platform, i.e., not a survey after the activity, and 2) is given by a peer and directed to another peer to help them develop their feedback skills. The present study seeks to answer the following research questions:

- RQ1: What are the characteristics of the studies employing backward evaluation in peer assessment?
- RQ2: How is backward evaluation conducted (platform, backward evaluation features, etc.)?
- RQ3: What did the analyses of the backward evaluation data reveal?

2. Background

2.1. Peer assessment

Peer assessment (PA) is an activity in which peers evaluate each other's work (Topping, 1998). PA can be summative (students' evaluations contribute to the final grades of other students) or formative (students' evaluations help improve other students' performance) (Patchan et al., 2018; Topping, 1998). As PA facilitates student dialogue about their learning, stimulates student self-monitoring and self-evaluating skills and helps students improve their performance in different phases of PA assignments, it can be categorised as sustainable assessment (Boud & Molloy, 2013).

A number of literature reviews on PA has been published over the years, including two meta-analyses that found a positive effect of PA on student performance (Double, McGrane, & Hopfenbeck, 2020; Li, Xiong, Hunter, Guo, & Tywoniw, 2020). Van Zundert, Sluijsmans, and Van Merriënboer (2010) focussed on the different variables that support effective PA. Training and experience in PA can on the one hand help improve PA's psychometric qualities, such as reliability and validity, and on the other hand increase students' positive attitudes towards PA. Moreover, domain-specific skills have the potential to improve through revisions following a PA activity. In addition, the development of PA skills helps with academic achievement. Two meta-analyses considered a comparison between student and teacher grading. Li, Xiong, Zang, and KornhaberLyuChungK.Suen (2016) found a moderately strong correlation between peer and teacher grades, whereas Falchikov and Goldfinch (2000) related a higher validity of PA with the design of the PA activity. Aspects such as clear criteria and more guidance led to higher agreement between teacher and student grading. A systematic
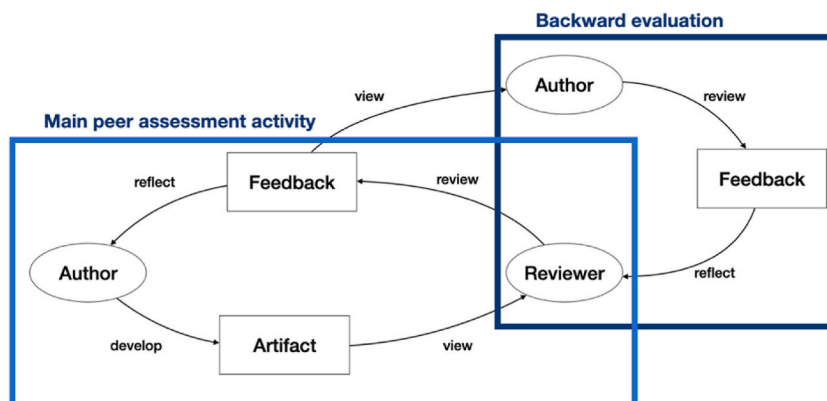

Fig. 1. A model of the peer assessment process including backward evaluation (adapted from Indriasari, Luxton-Reilly, & Denny, 2020).

literature review by Indriasari et al. (2020) focussed on the use of gamification in PA and reported positive effects of gamification on student engagement. The effects of PA depend on the design and framing of PA activities, as well as the organisational limitations and pedagogical goals of a course (Topping, 1998; Van den Berg, Admiraal, & Pilot, 2006). In a series of three experiments, Hicks, Pandey, Fraser, and Klemmer (2016) showed how different kinds of questions in rubrics, the structure of a task or the way artefacts are presented led to different results in terms of feedback quality and the focus of the reviewer.
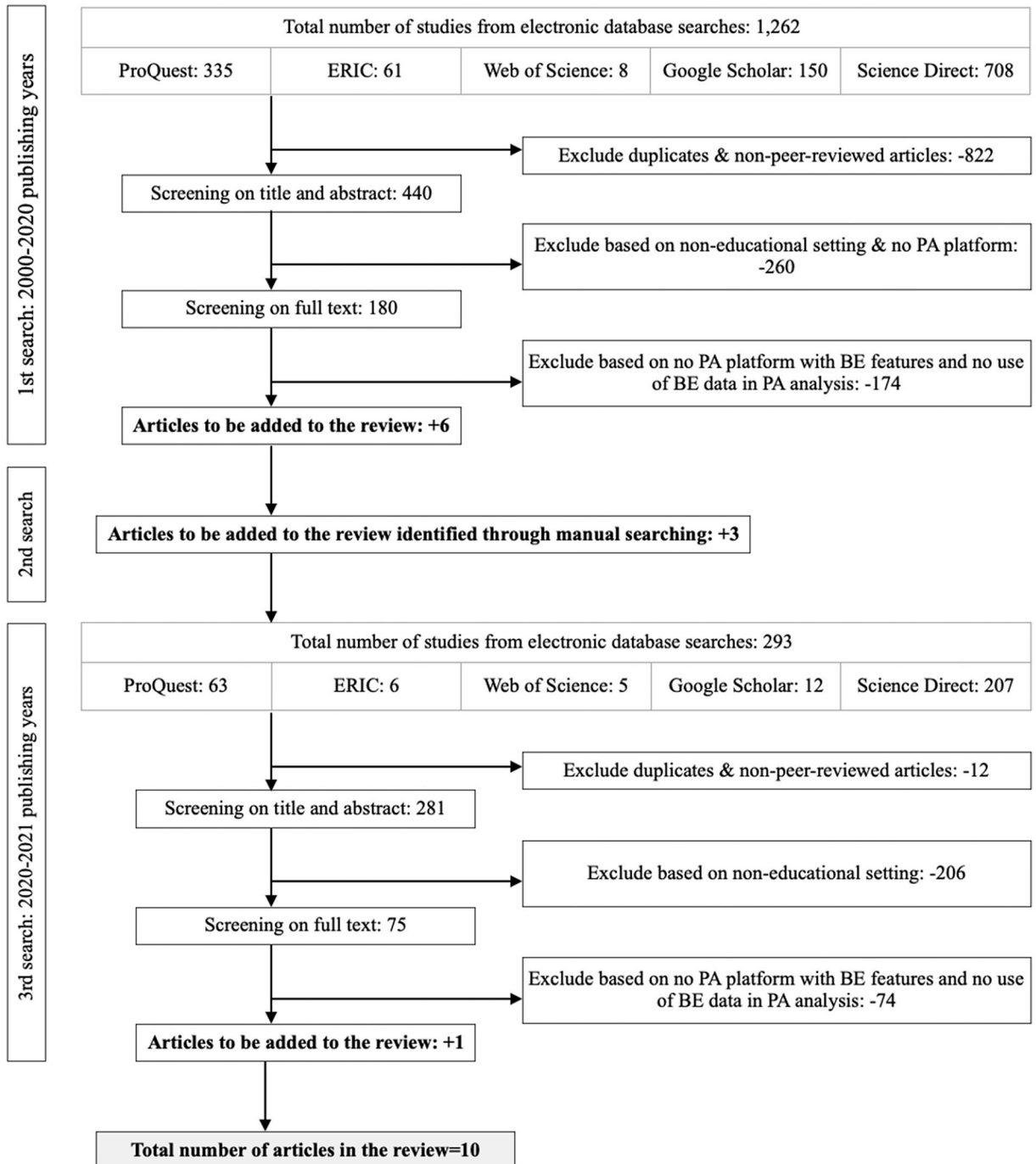

Fig. 2. Inclusion/exclusion process.

2.2. Backward evaluation to increase peer feedback quality

The inclusion of *backward evaluation* (BE) as an accountability element in the *peer assessment* (PA) process helps not only with feedback quality (Luxton-Reilly, 2009; Patchan et al., 2018) but also with the development of evaluative judgment (Tai et al., 2018). In this way, students reflect on the quality of the feedback they receive, improve their own feedback skills and finally turn these skills towards their own work, improving the quality of their current and future texts.

Previous research on student perceptions of feedback quality has mostly focussed on data collected through surveys rather than data collected from authentic PA activities (Wu & Schunn, 2021; e.g., Mostert & Snowball, 2013; Loretto, DeMartino, & Godley, 2016). Survey data collection is often limited to analysing general attitudes towards PA and the overall helpfulness of the feedback. Moreover, a survey might take place outside of the PA activity and not be directed at the peer who gave feedback, instead only being meant for research purposes. If students do not receive any feedback on their feedback, they remain unaware of their own feedback skills and do not have the opportunity to advance their evaluative judgement skills using the BE they received on the feedback they gave to others.

Common measures of the quality of PA are validity, which compares peer feedback to instructor feedback (Fernández-Toro & Furnborough, 2014; Kubincová, Dropcová, & Homola, 2016), and reliability, defined as examining the distribution of peer grades among peers. A newer approach is automated metareviews that automatically analyse the quality of feedback and present this evaluation to the student (Nguyen, Xiong, & Litman, 2017; Ramachandran, Gehringer, & Yadav, 2017; Yadav & Gehringer, 2016). This is a promising method in large online classes or MOOCs; however, it does not enable students to engage in developing their feedback skills through rating other peers' feedback. Another method to increase student engagement with feedback is including rebuttals or appeal letters in the PA activity. These tasks resemble an academic peer review in which students write letters accepting or rejecting the feedback they receive (Gordijn, Broekhans, Dunn, & Ubacht, 2018; Harland, Wald, & Randhawa, 2017; Zhou, Zheng, & Tai, 2020). The rebuttals, however, are not addressed directly to the peers that gave the feedback but to the instructors, which means that students are not presented with feedback on their feedback skills from their peers. Additionally, students might discuss the feedback they receive either in live discussions on an online PA platform (Yang, Badger, & Yu, 2006) or synchronous online discussions (Cevik, Haşlaman, & Çelik, 2015; Zheng, Cui, Li, & Huang, 2018). However, these implementations of BE are outside of the scope of this review, as data from a dynamic dialogue about feedback present a different level of complexity than a one-way written comment.

3. Methodology

The goal of a scoping review is to show 'the breadth and depth of a field' (Levac, Colquhoun, & O'Brien, 2010, p. 1) and is especially useful to investigate emerging topics or research fields. Though analytical steps in the search process are followed, a wide variety of studies might be included, and the selection of articles does not include a quality assessment (Wilson, Anafi, Roh, & Errasti-Ibarrondo, 2020, pp. 1–7). This study follows the steps and recommendations for scoping reviews as described by Levac et al. (2010).

The search string was built with terms describing 'peer assessment' in combination with terms and synonyms used for 'backward evaluation' (see Appendix A for the full search string). The search was conducted in March 2020 across five databases (ProQuest, Google Scholar, ERIC, Web of Science and Science Direct) and restricted by publishing year (2000–2020), and it resulted in 1262 articles. As depicted in Fig. 2, the articles were screened in three rounds using the inclusion criteria listed in Table 1. Step 1 involved the removal of duplicates, non-English articles and non-peer-reviewed articles, leaving 440 articles. In step 2, we read the titles and abstracts and excluded another 260 articles due to non-educational settings and lack of a *peer assessment* (PA) platform. During step 3, the full text of 180 articles was read, and studies were included if they used a PA platform offering *backward evaluation* (BE) features and used BE data in their analyses. This resulted in the exclusion of another 174 articles, leaving us with six articles.

As only six articles were included after the third screening, a non-systematic hand search was completed in order to identify additional articles that might have been missed in the original search. Fourteen searches in Google Scholar were carried out using the individual names of tools offering BE functionalities identified in the systematic reviews by Luxton-Reilly (2009) and Patchan et al. (2018). The search string included 'backward evaluation/feedback/grading' AND the name of the tool (SWoRD/Peerceptiv, Peergrade, CrowdGrader, Blackboard, Virtual Learning Community, MobiusSLIP, PeerGrader, PECASSE, Eli Review, Aropä, peerScholar, Peer-Wise and PEARS) for the publishing years 2000–2020. The hand search resulted in the addition of three articles that reported on BE studies and fulfilled the inclusion criteria.

In April 2021, an additional search using the original search string was administered in the same databases as the original search, with the only difference being that the publishing years were set to 2020–2021. Two hundred and ninety-three articles were found and screened using the same criteria as in the original search. Screening of titles and abstracts led to the exclusion of 218 articles. The full text analysis of the remaining 75 articles resulted in an inclusion of one additional paper to the scoping review. Thus, the final dataset in this scoping review includes 10 articles and 11 studies as Van der Pol, Van den Berg, Admiraal, and Simons (2008) report on two

Table 1
Inclusion criteria.

The article was published in English between 2000 and 2021.
The article was peer-reviewed.
The article reports on an empirical study of the implementation of peer assessment with backward evaluation.
The peer assessment activity was conducted on an online platform that offers backward evaluation features.
The peer assessment activity took place in an educational setting.
Backward evaluation data were used in the peer assessment analysis.

studies in their article (see Table 3).

Although we attempted to find all articles addressing empirical studies using BE in PA, there is always the possibility that we missed some due to divergent terminology. Moreover, the inclusion criteria significantly narrowed the scope of the analysis; for example, some studies would mention BE as a part of their PA design but would not include BE results in their analyses; these articles were not included in this review (e.g., Cho & Schunn, 2007; Park & Cho, 2017; Wu & Schunn, 2021). Zou, Schunn, Wang, and Zhang (2018) carried out a survey with students that included a question about BE and used it to examine student attitudes towards PA and BE after a PA activity on the SwoRD/Peerceptiv platform; however, since they did only use the survey results and not the platform data, it was not included in the review.

To answer the research questions, the coding scheme, as depicted in Table 2, was developed, mapping 1) the main focus of a paper, 2) the context of its PA study (e.g., discipline, educational level), 3) the characteristics of the study (e.g., sample size, PA platform used), 4) BE implementation in the PA activity and 5) the results of the BE implementation. The results of the coding are presented in Tables 3 and 4.

4. Results

Empirical studies that include *backward evaluation* (BE) are sparse. Five of the articles used in this study were published in conference proceedings, and five articles were published in journals. Though 10 articles are included in this scoping review, 11 studies on BE were identified; Van der Pol et al. (2008) report on two separate studies: study 1 with a sample of 27 college students and study 2 with 38 college students.

4.1. RQ1: what are the characteristics of the studies employing backward evaluation in peer assessment?

The study characteristics mapped include BE terminology, educational level, discipline, and sample size. The breadth of terminology used to describe BE in the articles included in this review was analysed (see Table 3). 'Back-review' is used in three articles (Cho & Kim, 2007; Nelson & Schunn, 2009). Patchan et al. (2018) introduce back-review with two synonyms: 'double-loop feedback' and 'metareviewing'. Wu and Schunn (2020a) describe BE as both back-review and 'back-evaluation'. The latter term can be also found in Adewoyin et al. (2016). Misiejuk et al. (2021) use 'backward evaluation'. de Alfaro and Shavlovsky (2016) write about 'review feedback', while Cho & Schunn, 2007 refer to BE as the more specific 'helpfulness rating'. Tsivitanidou and Ioannou (2019) do not use a specific term but describe different activities as being in a 'react phase'. Similarly, Van der Pol et al. (2008) define BE through its measurement, such as agreement and importance, under the umbrella term of 'reception of feedback'.

Eight studies were conducted in the context of higher education. de Alfaro and Shavlovsky (2016) and Misiejuk et al. (2021) conducted studies with datasets from multiple universities and high schools. Adewoyin et al. (2016) focussed on professional learning, specifically on teacher professional development, whereas Wu and Schunn (2020a) analysed data from a secondary school. BE was implemented in multiple discipline settings in four studies, and in physics in three other studies. Nelson and Schunn (2009) collected data in a history course, while Wu and Schunn (2020a) collected data in a writing course. Study 1 by Van der Pol et al. (2008) was in a health care course, and study 2 was in educational science. The biggest sample size was analysed by de Alfaro and Shavlovsky (2016), who included data from 23,762 students. The second biggest dataset was used by Misiejuk et al. (2021), representing 7,660 records, followed by Cho and Kim (2007) with 617 participants. Three studies had datasets with 100–300 students, and five studies included data from fewer than 100 students.

In summary, there are few peer-reviewed articles published on using BE on *peer assessment* (PA) platforms. Most of these articles focus on higher education (9/11 studies) and only one discipline (7/11 studies). Moreover, the majority of these studies (8/11 studies) have relatively small sample sizes, varying from 21 to 300 students, while one study had over 23,000 students. Nine different terms were used in the articles to describe BE.

4.2. RQ2: how is backward evaluation conducted (platform, backward evaluation features, etc.)?

To determine the different types of BE implementation, we examined the tools/platforms used to conduct BE, the characteristics of BE analysed in a given study, whether the BE activity was an obligatory part of the PA activity and how BE was framed and defined (see Table 3).

Five studies used the SWoRD tool, later renamed Peerceptiv (peerceptiv.com), which was developed at the University of

Table 2
Coding scheme.

Code	Description	Criteria
Focus	What is the focus of the paper?	
Study characteristics	How is peer assessment implemented? What is the sample size? How is backward evaluation described?	Backward evaluation terminology; educational level; discipline; sample size
Backward evaluation implementation	What platform is used? How is backward evaluation integrated into the peer assessment activity?	Platform name; backward evaluation types; obligatory or voluntary participation in the activity?
Findings about backward evaluation	What were the results of the backward evaluation implementation? How are the backward evaluation data used in the analysis?	

Table 3
Studies and BE implementation characteristics.

Article	Peer assessment platform	Educational level	Discipline	n	Backward evaluation type	Backward evaluation obligatory?	Backward evaluation scale	Backward evaluation comment coding	Other
Cho and Kim (2007)	SWoRD/Peerceptiv	Higher education	Multiple disciplines	617	Scale, comment	Yes	7-point star scale	Not coded	–
Cho & Schunn, 2007	SWoRD/Peerceptiv	Higher education	Physics	87	Scale, comment	Yes	7-point helpfulness scale	Not coded	–
Study 1 by Van der Pol et al. (2008)	Virtual Learning Community	Higher education	Health care	27	Scale, comment	No	4-point importance scale	3-point level of agreement scale (do not agree, partly agree, completely agree) Validation: Cohen's kappa, two raters	–
Study 2 by Van der Pol et al. (2008)	Blackboard; Annotation system	Higher education	Educational science	38	Scale, comment	No	5-point usefulness scale	3-point level of agreement scale (do not agree, partly agree, completely agree) Validation: Not specified	–
Nelson and Schunn (2009)	SWoRD/Peerceptiv	Higher education	History	24	Scale, comment	Yes	7-point helpfulness scale	2-point level of agreement scale (not agreed, agreed) 2-point level of feedback understanding (not understood, understood) Validation: Cohen's kappa, two raters	–
de Alfaro and Shavlovsky (2016)	CrowdGrader	Higher and secondary education	Multiple disciplines	23,762	Scale	Yes	5-star helpfulness rating	–	–
Adewoyin et al. (2016)	Non-commercial (in-house developed)*	Teacher professional development	Mathematics, music and language	284	Scale	Yes	7-point Likert scale*	–	–
Patchan et al. (2018)	SWoRD/Peerceptiv	Higher education	Physics	287	Scale	Yes	5-point helpfulness scale	–	–
Tsivitanidou and Ioannou (2019)	Peergrade	Higher education	Physics	21	Likes, flags, comment, scale	Only scale obligatory	4-point Likert scale	Not coded	Likes, flags
Wu and Schunn (2020a)	SWoRD/Peerceptiv	Secondary education	Writing	185	Scale, comment	Yes	5-point helpfulness scale	2-point level of agreement scale (not agreed or partially agreed, agreed) 2-point level of feedback understanding scale (not understood or partially understood, understood) Validation: Cohen's kappa, multiple raters	–
Misiejuk et al. (2021)	Peergrade	Higher and secondary education	Multiple disciplines	7,660	Scale, comment, improvement suggestions	Not specified as no context information provided in dataset	5-point usefulness scale	Three codes: accepting, defending, gratitude	Improvement suggestions (kindness, justification, constructivity, relevance, specificity)

*This information was not specified in the publications but was gathered from communication with the author(s).

Table 4
Study focus, backward evaluation terminology, and main findings.

Study focus	Article	Term used for backward evaluation	Main findings
feedback uptake	Study 1 by Van der Pol et al. (2008)	reception of feedback	High feedback uptake if 1) feedback included recommendations for revision, 2) feedback focussed on the content and style of the draft or 3) there was a high backward evaluation importance rating.
	Study 2 by Van der Pol et al. (2008)	reception of feedback	High backward evaluation usefulness rating if high agreement with feedback. High feedback uptake if 1) feedback included an analysis of an issue, an evaluation or a revision recommendation, 2) feedback focussed on the content and style of the draft or 3) there was high agreement with the feedback giver.
	Nelson and Schunn (2009)	back-review	High feedback uptake if high understanding of the problem described in the feedback. High problem understanding if 1) feedback included a solution, 2) feedback included a location of a problem or a solution or 3) feedback included a summary. Low feedback understanding if feedback included problem explanation.
	Wu and Schunn (2020a)	back-review, back-evaluation	High feedback uptake if 1) high agreement with problems or constructive comments in feedback comments, 2) high understanding of problems or constructive comments in feedback comments, 3) high agreement with explanations in feedback comments, 4) feedback comments included explanations of problems or 5) feedback comments included hedges for problems and suggestions or solutions. Low feedback uptake if 1) feedback comments included high praises or 2) feedback comments included hedges. High feedback understanding if 1) feedback comments included a solution or 2) longer feedback comments included a problem or a solution. Low feedback agreement if 1) feedback comments included high praise or 2) the first draft was of high quality. High problem agreement if feedback comments included mitigating praise.
improvement of writing skills	Cho & Schunn, 2007	helpfulness rating	Longer feedback comments predict higher helpfulness ratings. High writing performance in the final draft if students gave more helpful feedback.
learning analytics insights into BE	Tsivitanidou and Ioannou (2019)	react phase	More backward evaluation comments if low agreement with feedback.
	Misiejuk et al. (2021)	backward evaluation	When feedback was perceived as having high usefulness, 1) the backward evaluation comments contained less gratitude, 2) the backward evaluation comments contained mostly confusion, criticism or disagreement or 3) most suggestions are for feedback to have been more constructive and/or just. The higher the perceived usefulness of feedback, the more the backward evaluation comments contained 1) gratitude and/or praise, 2) error acknowledgment or 3) intention of revision.
quality of peer feedback	Patchan et al. (2018)	back-review	If students think their reviewing grade is influenced by the helpfulness of their feedback, 1) feedback is more helpful, 2) more criticisms, solutions and localised comments are included in the feedback and 3) feedback is more reliable.
tit-for-tat strategy	Cho and Kim (2007)	back-review	Low chance of tit-for-tat strategy if cognitive interface design is implemented.
	de Alfaro and Shavlovsky (2016)	review feedback	Evidence of tit-for-tat strategy based on helpfulness ratings independent of subject area.
	Adewoyin et al. (2016)	back-evaluation	Longer feedback comments do not predict higher backward evaluation ratings. No evidence of tit-for-tat strategy based on helpfulness ratings.

Pittsburgh's Learning Research and Development Centre as an online peer and self-assessment platform that provides the ability to integrate it with an LMS. Peergrade (peergrade.io), an online PA platform developed at the Technical University of Denmark, was used by Tsivitanidou and Ioannou (2019) and by Misiejuk et al. (2021). de Alfaro and Shavlovsky (2016) used CrowdGrader (crowdgrader. org), a peer grading platform that runs on GoogleCloud. The first study by Van der Pol et al. (2008) used Virtual Learning Community (vlc.uchicago.edu) from the University of Chicago, an online platform facilitating PA that lacks BE features. This resulted in students responding to the feedback that they received by including their BE comments in the final versions of their written assignments. Study 2 by Van der Pol et al. (2008) was conducted on two platforms: a popular LMS called Blackboard (blackboard.com) and an annotation system developed by Van der Pol, Admiraal, and Simons (2006) to support 'anchored discussions' that display 'both artefact and discussion in a linked, yet independent manner' (p. 343). Discussion forums were adopted to facilitate PA and BE in both the annotation system and Blackboard. Adewoyin et al. (2016) used a non-commercial platform. It is important to note that depending on the tool, different BE types and measurements are available. At the same time, instructors have the opportunity to customise the settings. Finally, though BE was used as a part of a PA activity, not all available BE data collected might have been used in a research study.

The most popular BE method was a scale-comment combination that was found in six studies (Cho & Kim, 2007; Cho & Schunn, 2007; study 1 and study 2 by; Van der Pol et al., 2008; Nelson & Schunn, 2009; Wu & Schunn, 2020a). de Alfaro and Shavlovsky (2016), Patchan et al. (2018) and Adewoyin et al. (2016) used only a scale. Students in the Tsivitanidou and Ioannou (2019) study could not only use a comment and a scale in their BE activity but had the opportunity to 'like' comments that they appreciated or 'flag' comments that they disagreed with and wanted an instructor to intervene for. Misiejuk et al. (2021) included not only comments and a scale in their analysis but also a multiple-choice question on improvement suggestions in five categories: kindness ('The feedback is too harsh and uses harsh language'.), justification ('The feedback should be more justified and give more arguments for the decisions'.), constructivity ('The feedback should be more constructive and propose things to improve'.), relevance ('The feedback does not feel relevant to my hand-in or addresses the wrong things'.) and specificity ('The feedback should be more specific and point to concrete things that can be improved'.).

Studies that implemented scales to measure BE differ both in the scale range and measurement type. The scale ranges varied from seven points (Adewoyin et al., 2016; Cho & Schunn, 2007; Cho & Kim, 2007; Nelson & Schunn, 2009) to four points (Tsivitanidou & Ioannou, 2019; study 1 by; Van der Pol et al., 2008). The most popular range was five points, as implemented in five studies (study 2 by Van der Pol et al., 2008; de Alfaro & Shavlovsky, 2016; Patchan et al., 2018; Wu & Schunn, 2021; Misiejuk et al., 2021). BE was measured using a variety of concepts: five studies focussed on helpfulness (Cho & Schunn, 2007; Nelson & Schunn, 2009; de Alfaro & Shavlovsky, 2016; Patchan et al., 2018; Wu & Schunn, 2021), while two studies focussed on usefulness (study 1 by Van der Pol et al., 2008; Misiejuk et al., 2021). One study focussed on the importance of the feedback (study 2 by Van der Pol et al., 2008). Adewoyin et al. (2016) used 10 BE questions (e.g., 'Was the feedback constructive?') that students answered using a Likert scale from 'worst' to 'best'. Cho and Kim (2007) used a star scale.

Not all studies that included comments as part of the BE activity used these data in their analysis. To code the BE comments, studies used two main codes: agreement with the feedback and understanding of the feedback. The level of agreement was coded either on a three-point scale ('do not agree', 'partly agree', 'completely agree') (study 1 and 2 by Van der Pol et al., 2008) or on a two-point scale ('not agreed'/'not agreed or partially agreed', 'agreed') (Nelson & Schunn, 2009; Wu & Schunn, 2020a). The level of understanding was applied only in two studies and measured on a two-point scale: 'not understood'/'not understood or partially understood', 'understood' (Nelson & Schunn, 2009; Wu & Schunn, 2020a). Misiejuk et al. (2021) used three codes to analyse the data: 1) accepting (defined as praise, error acknowledgment or intention of revision), 2) defending (defined as confusion, criticism or disagreement) and 3) gratitude. The BE comment coding was validated using Cohen's Kappa with the help of two or more raters (study 1 by Van der Pol et al., 2008; Nelson & Schunn, 2009; Wu & Schunn, 2020a; Misiejuk et al., 2021). Study 2 by Van der Pol et al. (2008) did not report the code validation method employed.

In seven studies, BE was an obligatory part of the PA activity (Cho & Kim, 2007; Cho & Schunn, 2007; Nelson & Schunn, 2009; Adewoyin et al., 2016; de Alfaro & Shavlovsky, 2016; Patchan et al., 2018; Wu & Schunn, 2021). In Tsivitanidou and Ioannou's (2019) study, students were only required to use the scale to rate the helpfulness of the feedback; comments, likes and flagging were voluntary. BE participation was voluntary in two studies (study 1 and 2 in Van der Pol et al., 2008). The participation requirements were not specified in one study due to the lack of context information in the dataset (Misiejuk et al., 2021).

In summary, most studies (9/11 studies) used platforms focussed on facilitating PA, though many LMSs used in higher education nowadays can also be used to conduct PA and might include BE features. Using a scale to measure BE was the most popular method (10/11 studies), followed by BE comments (8/11 studies). In most studies, students were asked to grade the helpfulness of the feedback (5/8 studies), and their BE comments were examined to determine if students agreed with the feedback provided (5/8 studies). In most of the studies (8/11 studies), BE, or part of BE, was an obligatory part of the PA activity.

4.3. RQ3: what did the analyses of backward evaluation data reveal?

We examined the focus and main findings of the studies included in this review to show how and why BE data are used in research (see Table 4). Three main research aims were discovered in the analysis: a tit-for-tat strategy, feedback uptake and insights from learning analytics into BE.

A tit-for-tat strategy is broadly defined as 'an individual [reacting] to an opponent by repeating the opponent's action' (Cho & Kim, 2007, p. 210). In the context of PA, this refers to a situation in which students react positively to positive feedback and negatively to negative feedback. This fosters a competitive rather than a collaborative learning environment, and it might compromise the validity of PA (Cho & Kim, 2007). There are three studies in this review that used BE data to detect if students used tit-for-tat strategies

(Adewoyin et al., 2016; Cho & Kim, 2007; de Alfaro & Shavlovsky, 2016). Cho and Kim (2007) compared two interface designs to determine which one was better at mitigating tit-for-tat. Adewoyin et al. (2016) examined if including BE in a PA activity would encourage students to engage in tit-for-tat. de Alfaro and Shavlovsky's (2016) research focussed on errors in peer grading in a big dataset collected in the CrowdGrader tool, and part of their analysis considered if grades diverting from a consensus are caused by tit-for-tat; CrowdGrader uses the BE rating as part of the overall grades that students receive on their assignments.

Interestingly, de Alfaro and Shavlovsky (2016) found evidence for tit-for-tat, whereas Adewoyin et al. (2016) did not. This opens up an opportunity for further research in factors mediating tit-for-tat, especially considering findings from Cho and Kim (2007) that tit-for-tat can be mitigated through interface design.

Feedback uptake usually refers to feedback that triggers revisions in the final draft, or more generally, feedback implementation. Four studies in this review focussed on feedback uptake (study 1 and study 2 in Van der Pol et al., 2008; Nelson & Schunn, 2009; Wu & Schunn, 2020a). Study 1 and study 2 by Van der Pol et al. (2008) focussed on the relationship between assignment revision, BE and the nature of feedback—including the feedback's function (analysis, explanation, evaluation, revision) and aspect (content, structure, writing style). BE metrics were defined as feedback importance in study 1 and feedback agreement in study 2 by Van der Pol et al. (2008). Nelson and Schunn (2009) coded their feedback using the following categories: 1) type of feedback (praise, problem/solution, summary), 2) scope of the problem/solution (global, local), 3) type of affective language (mitigation-compliment, mitigation-other), 4) localisation of the problem/solution (localised, not localised), 5) type of problem/solution (problem, solution, both) and explanation of the problem (absent, content) and finally 6) explanation of the solution (absent, content). Student reactions to comments addressing a problem or solution were the basis of their BE analyses that examined if students understood and/or agreed with the feedback provided. The analysis aimed to identify mediators of feedback uptake. A similar study was conducted by Wu and Schunn (2020a), who coded their feedback based on 1) type of feedback (praise, summary, implementable comments), 2) feedback features (identification, explanation, suggestion, solution, mitigating praise, hedges) and 3) scope of implementable feedback (high-level, low-level). As in Nelson and Schunn (2009), reaction to the feedback was coded based on students' agreement with and understanding of a problem or a solution in the feedback. Finally, the implementation of feedback was mapped in the students' final drafts. Two types of findings are reported in studies on feedback uptake: 1) what influences feedback uptake and 2) which elements of feedback influence feedback uptake mediators, such as feedback agreement or understanding.

Learning analytics is a field focussed on 'the measurement, collection, analysis and reporting of data about learners and their contexts, for purposes of understanding and optimising learning and the environments in which it occurs' (Buckingham Shum & Ferguson, 2012, p. 4). Two studies examined the potential insights into BE that might be discovered using learning analytics. Tsivi-tanidou and Ioannou (2019) focussed on how the data from an online PA platform that included BE could help better elucidate the PA process. Misiejuk et al. (2021) explored two research questions: 1) the relationship of different measures of BE (scale, comment and improvement suggestions) and 2) the relationship between BE and rubric characteristics.

Two studies (Cho & Schunn, 2007; Patchan et al., 2018) focussed neither on feedback uptake nor on tit-for-tat strategy. To examine the improvement of writing skills during the PA activity, Cho & Schunn, 2007 divided students into a high helpful group (students whose feedback was rated as more helpful), and a low helpful group (students whose feedback was rated as less helpful); both groups had similar writing skills at the beginning of the study. The quality of their writing in the final assignment was compared. Patchan et al. (2018) examined the effects of BE on PA reliability and the quality of feedback comments. A percentage of a student's final grade was determined by the quality of the feedback comments rated using a reviewing grade. Students were assigned to three groups: 1) reviewing grade depends on the PA reliability, i.e., consistency of feedback ratings among students, 2) reviewing grade depends on the helpfulness of their feedback and 3) reviewing grade depends both on reliability and helpfulness of the feedback that students give. Two hypotheses were tested: 1) the direct accountability hypothesis, stating that 'the reviewing grades will directly affect the quality of peer assessment' (Patchan et al., 2018, p. 2268) and 2) the depth-of-processing hypothesis, assuming that accountability through reviewing grades will not only trigger deep processing in feedback comments but also improve peer rating reliability. To measure feedback comment quality, the amount of feedback (volume of feedback, number of comments, number of long comments) and feedback features (number of criticism comments, number of solutions, number of localised comments) were used.

Findings regarding three main aspects of BE are described in the studies included in this scoping review: BE rating, feedback understanding and feedback agreement.

Cho & Schunn, 2007 found that those with high helpfulness ratings provided longer feedback comments, while Adewoyin et al. (2016) reported that longer feedback comments did not predict higher BE ratings. Students give feedback that is rated as more helpful if they think their reviewing grade is influenced by the helpfulness of their feedback (Patchan et al., 2018). Moreover, Cho & Schunn, 2007 found that students whose feedback was rated as more helpful exhibited higher writing performance in their final drafts, and study 1 by Van der Pol et al. (2008) found that higher BE importance ratings influenced high feedback uptake. Misiejuk et al. (2021) found that students rating the feedback that they received on their work as not useful at all rarely showed gratitude and mostly expressed confusion, criticism or disagreement with the feedback when writing their BE comments. Moreover, they mainly suggested that the feedback should have been more constructive and/or just. The BE comments of students who perceived the feedback as extremely useful included gratitude, gratitude mixed with praise, error acknowledgment or intention of revision or only praise, error acknowledgment or intention of revision. No significant results were found for the relationship between BE and rubric characteristics.

Wu and Schunn (2020a) found that students understand feedback better if a feedback comment includes a solution, or if it is a longer comment describing either a problem or a solution. A solution, location of a problem or a solution or a summary in feedback comments increases a student's understanding of a problem; however, including an explanation in the feedback comment has a negative effect on feedback understanding (Nelson & Schunn, 2009). Students that better understood a problem (Nelson & Schunn, 2009; Wu & Schunn, 2020a) or received constructive comments (Wu & Schunn, 2020a) were more likely to implement the feedback.

Students are not likely to agree with feedback if it includes a lot of praise or if they wrote high-quality first drafts (Wu & Schunn, 2020a). If students do not agree with the feedback, they write more BE comments (Tsivitanidou & Ioannou, 2019). However, if feedback comments include mitigating praise, students are more likely to agree with a problem described in the feedback comment (Wu & Schunn, 2020a). Moreover, if they rate feedback as very useful, they are more likely to agree with it (study 2 by Van der Pol et al., 2008). High agreement with the feedback giver (study 2 by Van der Pol et al., 2008), with problems or constructive comments or with explanations in feedback comments (Wu & Schunn, 2020a) results in higher feedback uptake.

In summary, BE data are mostly used to determine tit-for-tat strategies (3/11 studies) or to inform how student perception influences feedback uptake (4/11 studies). The findings from the studies included in this scoping review show the potential of BE measures to examine a variety of research questions, ranging from student perception of particular feedback features to mediators of feedback implementation.

5. Discussion and conclusions

This paper offers a scoping review of *backward evaluation* (BE) in *peer assessment* (PA) research, focussing on study characteristics, BE characteristics and the use of BE data. Although we found relatively few empirical studies on PA that also use BE, they offer new insights into different aspects of the PA process. Including BE in a PA activity opens new opportunities for understanding not only students' perceptions of what useful feedback is and how they react to the feedback received but also what consequences their reactions have on actually implementing the feedback.

This analysis shows that research on BE in PA is focussed on higher education and is conducted on relatively small sample sizes of students. Helpfulness of and agreement with feedback are the most popular BE measurements that give new insights into how students perceive and process feedback. Moreover, BE data can help determine if students engage in tit-for-tat strategies and the extent of feedback uptake. Finally, there are some examples of BE data being analysed using innovative techniques from the learning analytics field to discover new insights into the PA process.

Two studies in this review showed potential for more experimental research using BE data. Patchan et al. (2018) created learning environments with different accountability systems to test their influence on student behaviour, and Cho and Kim (2007) designed different types of interfaces and tested them using BE data.

Interestingly, only one study used an LMS to facilitate its PA process (study 2 by Van der Pol et al., 2008), while others used online platforms that focus only on PA. The lack of studies using data from LMSs and MOOCs could be due to many reasons. Nowadays, many MOOCs and LMSs provide PA and BE functionalities, and PA platforms can be integrated in an LMS or a MOOC, though it depends on the instructor's decision. Future work could consider these platforms to collect the data on PA and BE to follow individual student feedback giving and receiving patterns or progress over long periods of time, rather than in a single course setting. Moreover, it could open new possibilities to investigate additional context data that can be collected about the students and the general inclusion of PA in learning design.

Furthermore, this scoping review showcased the variety and diversity of terminology that describes BE: back-review, double-loop feedback, metareviewing, back-evaluation, review feedback, helpfulness rating, react phase, backward evaluation or reception of feedback. This might hinder knowledge production on BE or might make it harder to find relevant articles. It also indicates there is a need to establish a common vocabulary to describe BE. Thus, this scoping review contributes to a mapping of BE terms. This points to a limitation of this scoping view as there is the possibility that we missed some relevant research.

Finally, this scoping review makes a significant contribution to PA research as it is the first literature survey to address BE in PA, and in particular the use of BE data in empirical study analysis. The potential for further innovation and development of a PA activity using online platforms and techniques such as BE has been highlighted. This scoping review shows that BE data can be used to answer new research questions and to gain new insights into student feedback perception and processing. Furthermore, the results encourage practitioners to include BE in their PA learning designs in order to give students the opportunity to improve their own feedback skills and develop their evaluative skills through the recognition of feedback quality.

Author contribution

Kamila Misiejuk: Conceptualization, Visualization, Writing – original draft, Writing – review & editing, Data curation, Formal analysis, Methodology, Software, Barbara Wasson: Conceptualization, Supervision, Writing – review & editing, Writing – original draft.

Acknowledgements

Appendix A. Search string

('Peer feedback' OR 'Peer review' OR 'Peer grading' OR 'Peer evaluation' OR 'Peer assessment' OR 'Peer rating') AND ('Feedback-to-feedback' OR 'review of a review' OR 'Reciprocal assessment' OR 'Reciprocal evaluation' OR 'Reciprocal feedback' OR 'Reciprocal grading' OR 'Reciprocal review*' OR 'Back-review assessment' OR 'Back-review evaluation' OR 'Back-review feedback' OR 'Back-

review grading' OR 'Back-review review*' OR 'double-loop assessment' OR 'double-loop evaluation' OR 'double-loop feedback' OR 'double-loop grading' OR 'double-loop review*' OR 'Backwards assessment' OR 'Backward assessment' OR 'Backwards evaluation' OR 'Backward evaluation' OR 'Backwards feedback' OR 'Backward feedback' OR 'Backwards grading' OR 'Backward grading' OR 'Backwards review*' OR 'Backward review*' OR Metareview* OR 'Meta-review*' OR 'Meta-feedback' OR Metafeedback OR Meta-grading OR 'Meta-grading' OR 'Meta-assessment')

References

Adewoyin, O., Araya, R., & Vassileva, J. (2016). Peer review in mentorship: Perception of the helpfulness of review and reciprocal ratings. In *Proceedings of the 13th international conference on intelligent tutoring systems* (pp. 286–293). Cham: Springer.

de Alfaro, L., & Shavlovsky, M. (2016). Dynamics of peer grading: An empirical study. In T. Barnea, & M. C.&M. Feng (Eds.), *Proceedings of the 9th international conference on educational data mining* (pp. 62–69). Raleigh, NC: International Educational Data Mining Society.

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

Buckingham Shum, S., & Ferguson, R. (2012). Social learning analytics. *Educational Technology & Society, 15*(3), 3–26.

Cevik, Y. D., Haşlaman, T., & Çelik, S. (2015). The effect of peer assessment on problem solving skills of prospective teachers supported by online learning activities. *Studies In Educational Evaluation, 44*, 23–35.

Cho, K., & Kim, B. (2007). Suppressing competition in a computer-supported collaborative learning system. In *Proceedings of the 12th international conference on human-computer interaction* (pp. 208–214). Berlin, Heidelberg: Springer.

Cho, K., & Schunn, C. D. (2007). Scaffolded writing and rewriting in the discipline: A web-based reciprocal peer review system. *Computers & Education, 48*(3), 409–426.

Cook, A. (2019). *Using interactive learning activities to address challenges of peer feedback systems (Doctoral dissertation)*. Pittsburgh, PA, USA: Carnegie Mellon University.

Double, K. S., McGrane, J. A., & Hopfenbeck, T. N. (2020). The impact of peer assessment on academic performance: A meta-analysis of control group studies. *Educational Psychology Review, 32*, 481–509.

Falchikov, N., & Goldfinch, J. (2000). Student peer assessment in higher education: A meta-analysis comparing peer and teacher marks. *Review of Educational Research, 70*(3), 287–322.

Fernández-Toro, M., & Furnborough, C. (2014). Feedback on feedback: Eliciting learners' responses to written feedback through student-generated screencasts. *Educational Media International, 51*(1), 35–48.

Gamage, D., Staubitz, T., & Whiting, M. (2021). Peer assessment in MOOCs: Systematic literature review. *Distance Education, 42*(2), 268–289. https://doi.org/10.1080/01587919.2021.1911626

Gordijn, J., Broekhans, B., Dunn, K., & Ubacht, J. (2018). Increasing the effect of peer review. In *Proceedings of the 11th annual international conference of education, research and innovation* (pp. 3640–3650). Seville, Spain: International Association of Technology.

Harland, T., Wald, N., & Randhawa, H. (2017). Student peer review: Enhancing formative feedback with a rebuttal. *Assessment & Evaluation in Higher Education, 42*(5), 801–811.

Hicks, C. M., Pandey, V., Fraser, C. A., & Klemmer, S. (2016). Framing feedback: Choosing review environment features that support high quality peer assessment. In *Proceedings of the 2016 CHI conference on human factors in computing systems (CHI '16)* (pp. 458–469). New York, NY: ACM. https://doi.org/10.1145/2858036.2858195.

Indriasari, T. D., Luxton-Reilly, A., & Denny, P. (2020). Gamification of student peer review in education: A systematic literature review. *Education and Information Technologies, 25*, 5205–5234.

Kubincová, Z., Dropčová, V., & Homola, M. (2016). Students' acceptance of peer review in computer science course. *EAI Endorsed Transactions on e-Learning, 3*(10), e6. https://doi.org/10.4108/eai.11-4-2016.151153

Levac, D., Colquhoun, H., & O'Brien, K. K. (2010). Scoping studies: Advancing the methodology. *Implementation Science, 5*(69), 1–9.

Li, H., Xiong, Y., Hunter, C. V., Guo, X., & Tywoniw, R. (2020). Does peer assessment promote student learning? A meta-analysis. *Assessment & Evaluation in Higher Education, 45*(2), 193–211.

Li, H., Xiong, Y., Zang, X., Kornhaber, L., Lyu, M., Chung, Y., et al. (2016). Peer assessment in the digital age: A meta-analysis comparing peer and teacher ratings. *Assessment & Evaluation in Higher Education, 41*(2), 245–264.

Loretto, A., DeMartino, S., & Godley, A. (2016). Secondary students' perceptions of peer review of writing. *Research in the Teaching of English, 51*(2), 134–161. https://www.jstor.org/stable/24889912.

Luxton-Reilly, A. (2009). A systematic review of tools that support peer assessment. *Computer Science Education, 19*(4), 209–232.

Misiejuk, K., Wasson, B., & Egelandsdal, K. (2021). Using learning analytics to understand student perceptions of peer feedback. *Computers in Human Behavior, 11*, 10665. https://doi.org/10.1016/j.chb.2020.106658. ISSN 0747-5632.

Mostert, M., & Snowball, J. D. (2013). Where angels fear to tread: Online peer-assessment in a large first-year class. *Assessment & Evaluation in Higher Education, 38*(6), 674–686.

Mulliner, E., & Tucker, M. (2017). Feedback on feedback practice: Perceptions of students and academics. *Assessment & Evaluation in Higher Education, 42*(2), 266–288.

Nelson, M. M., & Schunn, C. D. (2009). The nature of feedback: How different types of peer feedback affect writing performance. *Instructional Science, 37*(4), 375–401.

Nguyen, H., Xiong, W., & Litman, D. (2017). Iterative design and classroom evaluation of automated formative feedback for improving peer feedback localization. *International Journal of Artificial Intelligence in Education, 27*(3), 582–622.

O'Brien, K., Forte, M., Mackey, T., & Jacobson, T. (2017). Metaliteracy as pedagogical framework for learner-centered design in three MOOC platforms: Connectivist, Coursera and Canvas. *Open Praxis, 9*(3), 267–286.

Park, J., & Cho, K. (2017). Toward the integration of peer reviewing and computational linguistics approaches. *Journal of Educational Computing Research, 55*(1), 123–144.

Patchan, M. M., Schunn, C. D., & Clark, R. J. (2018). Accountability in peer assessment: Examining the effects of reviewing grades on peer ratings and peer feedback. *Studies in Higher Education, 43*(12), 2263–2278.

Potter, T., Englund, L., Charbonneau, J., MacLean, M. T., Newell, J., & Roll, I. (2017). ComPAIR: A new online tool using adaptive comparative judgement to support learning with peer feedback. *Teaching & Learning Inquiry, 5*(2), 89–113.

Ramachandran, L., Gehringer, E. F., & Yadav, R. K. (2017). Automated assessment of the quality of peer reviews using natural language processing techniques. *International Journal of Artificial Intelligence in Education, 27*(3), 534–581.

Romero, C., & Ventura, S. (2020). Educational data mining and learning analytics: An updated survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 10*(3). https://doi.org/10.1002/widm.1355

Tai, J., Ajjawi, R., Boud, D., Dawson, P., & Panadero, E. (2018). Developing evaluative judgement: Enabling students to make decisions about the quality of work. *Higher Education, 76*(3), 467–481.

Topping, K. (1998). Peer assessment between students in colleges and universities. *Review of Educational Research, 68*(3), 249–276.

Topping, K. (2009). Peer assessment. *Theory into Practice, 48*(1), 20–27.

Tsivitanidou, O., & Ioannou, A. (2019). What do educational data, generated by an online platform, tell us about reciprocal web-based peer assessment?. In *Proceedings of the 14th European conference on technology enhanced learning* (pp. 600–603). Cham: Springer.

Van Zundert, M., Sluijsmans, D., & Van Merriënboer, J. (2010). Effective peer assessment processes: Research findings and future directions. *Learning and Instruction, 20*(4), 270–279.

Van den Berg, I., Admiraal, W., & Pilot, A. (2006). Design principles and outcomes of peer assessment in higher education. *Studies in Higher Education, 31*(3), 341–356.

Van der Pol, J., Admiraal, W., & Simons, P. R. J. (2006). The affordance of anchored discussion for the collaborative processing of academic texts. *International Journal of Computer-Supported Collaborative Learning, 1*(3), 339–357.

Van der Pol, J., Van den Berg, B. A. M., Admiraal, W. F., & Simons, P. R. J. (2008). The nature, reception, and use of online peer feedback in higher education. *Computers & Education, 51*(4), 1804–1817.

Wilson, D. M., Anafi, F., Roh, S. J., & Errasti-Ibarrondo, B. (2020). *A scoping research literature review to identify contemporary evidence on the incidence, causes, and impacts of end-of-life intra-family conflict.* Health Communication. https://doi.org/10.1080/10410236.2020.1775448

Winstone, N. E., Nash, R. A., Parker, M., & Rowntree, J. (2017). Supporting learners' agentic engagement with feedback: A systematic review and a taxonomy of recipience processes. *Educational Psychologist, 52*(1), 17–37.

Wu, Y., & Schunn, C. D. (2020a). From feedback to revisions: Effects of feedback features and perceptions. *Contemporary Educational Psychology, 60*, 1–17.

Wu, Y., & Schunn, C. D. (2021). The effects of providing and receiving peer feedback on writing performance and learning of secondary school students. *American Educational Research Journal, 58*(3), 1–35. https://doi.org/10.3102/0002831220945266

Yadav, R. K., & Gehringer, E. F. (2016). Metrics for automated review classification: What review data show. In *State-of-the-Art and future directions of smart learning* (pp. 333–340). Singapore: Springer.

Yang, M., Badger, R., & Yu, Z. (2006). A comparative study of peer and teacher feedback in a Chinese EFL writing class. *Journal of Second Language Writing, 15*(3), 179–200.

Yuan, J., & Kim, C. (2015). Effective feedback design using free technologies. *Journal of Educational Computing Research, 52*(3), 408–434.

Zheng, L., Cui, P., Li, X., & Huang, R. (2018). Synchronous discussion between assessors and assessees in web-based peer assessment: Impact on writing performance, feedback quality, meta-cognitive awareness and self-efficacy. *Assessment & Evaluation in Higher Education, 43*(3), 500–514.

Zhou, J., Zheng, Y., & Tai, J. H. M. (2020). Grudges and gratitude: The social-affective impacts of peer assessment. *Assessment & Evaluation in Higher Education, 45*(3), 345–358.

Zou, Y., Schunn, C. D., Wang, Y., & Zhang, F. (2018). Student attitudes that predict participation in peer assessment. *Assessment & Evaluation in Higher Education, 43*(5), 800–811.