Computers in Human Behavior 117 (2021) 106658

Contents lists available at ScienceDirect

Computers in Human Behavior

journal homepage: http://www.elsevier.com/locate/comphumbeh

Full length article

Using learning analytics to understand student perceptions of peer feedback

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

ARTICLE INFO

Keywords:
Peer assessment
Feedback
Backward evaluation
Learning analytics

ABSTRACT

Peer assessment (PA) is the process of students grading and giving feedback to each other's work. Learning analytics is a field focused on analysing educational data to understand and improve learning processes. Using learning analytics on PA data has the potential to gain new insights into the feedback giving/receiving process. This exploratory study focuses on backward evaluation, an under researched aspect of peer assessment, where students react to the feedback that they received on their work. Two aspects are analysed: 1) backward evaluation characteristics depending on student perception of feedback that they receive on their work, and 2) the relationship between rubric characteristics and backward evaluation. A big dataset (N = 7,660 records) from an online platform called Peergrade was analysed using both statistical methods and Epistemic Network Analysis. Students who found feedback useful tended to be more accepting by *acknowledging their errors*, *intending to revise their text*, and *praising its usefulness*, while students who found the feedback less useful tended to be more defensive by expressing that they were *confused about its meaning*, *critical towards its form and focus*, and in disagreement with the claims. Moreover, students mostly suggested feedback improvement in terms of feedback *specificity*, *justification* and *constructivity*, rather than *kindness.* The paper concludes by discussing the potential and limitations of using LA methods to analyse big PA datasets.

1. Introduction

Over the last three decades, *Formative Assessment* (FA) has received increasing attention and several studies have shown that FA practices can enhance student performance considerably (Black & Wiliam, 1998; Double et al., 2018; Evans, 2013; Hattie & Timperley, 2007; Jonsson, 2013; Shute, 2008). Unlike summative assessment, FA is not about grading or certification, but activities undertaken by teachers or students that provide information used to adapt teaching/studying to meet students' needs (Wiliam, 2011). FA also promotes a dynamic view of students as agents who should be actively involved in assessment practices through goal setting, peer assessment, and self-assessment (Black & Wiliam, 1998; Black & Wiliam, 2009; Nicol & Macfarlane-Dick, 2006; Sadler, 1989).

Some actors have argued that *Peer Assessment* (PA) is a particularly useful FA practice because students need to develop their own assessment competence to better recognise quality, understand assessment criteria, and self-assess their own work (Sadler, 2009; Sadler 2010). This encompasses that students can benefit from both receiving feedback from their peers and constructing feedback on the work of others, and some studies have found that giving feedback is just as effective, or more so, for improving writing performance as receiving feedback (Graner, 1987; Lundstrom & Baker Smemoe, 2009). Studies have also found that PA can have just as big an impact on student performance as assessments made by the teacher (see Double et al., 2018 for a meta-analysis on PA). Thus, PA stands out as a good alternative to teacher assessment, particularly in large classes where the teacher is not able to provide assessment for each individual student.

1.1. Students' experience of feedback

Some issues have been found in relation to how students experience and use feedback. Studies have found that students prefer teacher feedback compared with peer feedback (Jacobs, Curtis, Braine, & Huang, 1998; Nelson & Carson, 1998; Tsui & Ng, 2000; Zhang, 1995), and peers are sometimes perceived as less competent feedback providers than the teacher (Kaufman & Schunn, 2011). This indicates that there might be a trust issue when it comes to students' perception of feedback from peers. Several studies have also found that there is often a discrepancy between students' reception and use of feedback, referred

[*] Corresponding author. Centre for the Science of Learning & Technology (SLATE), University of Bergen, PO Box 7800, N-5020, Bergen, Norway.
E-mail address: kamila.misiejuk@uib.no (K. Misiejuk).

https://doi.org/10.1016/j.chb.2020.106658
Received 16 January 2020; Received in revised form 3 November 2020; Accepted 6 December 2020
Available online 14 December 2020

to as “the feedback gap” (Evans, 2013; Jonsson, 2013). A review by Jonsson (2013) concluded that this gap relates to student’s understanding of the feedback, as well as strategies and opportunities to use the feedback purposefully. For these reasons, investigating how students experience feedback is important from both an educational and research perspective.

1.2. Learning analytics

Learning Analytics (LA) is a field that tries to make sense of educational data in order to understand and improve learning processes (Long & Siemens, 2011) and is most often used on large datasets (Misiejuk & Wasson, 2017). LA opens new opportunities to shift the focus from the transmission of feedback information towards “actively supporting learners to gain impact through effective feedback processes” through, for example, more timely feedback or monitoring the uptake of feedback across subjects and time (Ryan, Gašević, & Henderson, 2019, p. 218). In particular, LA has the potential to improve a PA activity through methods, such as automatically classifying feedback given by students based on chosen criteria (e.g., a reviewer’s reputation), using predictive analytics to indicate feedback accuracy according to, for example, student’s domain knowledge, or clustering and visualizing feedback for the instructor to indicate which feedback needs their involvement (Wahid, Chatti, & Schroeder, 2016). At the same time, the analysis of large datasets poses new challenges, such as the automated coding of written peer feedback, or a limited interpretation of the analysis results in an educational context due to lack of contextual data (Mangaroska & Giannakos, 2018; Xiong, Litmaan, & Schunn, 2012). Moreover, research on feedback in data-rich environments requires a new conceptualisation of feedback. To address this, new feedback models are proposed, such as the model for data-supported feedback modeling the feedback process and data trails available to use for predictive algorithms by Pardo (2018). However, this promising work is still in early stages, and we were not able to use it in this study.

Some LA and PA research was conducted on facilitating dialogic peer feedback with LA (Er, Dimitriadis, & Gašević, 2019), and the effects of gamification on peer feedback (Huang, Hwang, Hew, & Warning, 2019). Divjak and Maretić (2017) developed a mathematical model to calculate grades in PA that can be used in assessment analytics. Other studies focused on writing analytics and examined how to augment peer feedback with automated feedback (Shibani, 2017), or used text analytics to examine the influence of different types of feedback messages on students’ writing performance (Cheng, Liang, & Tsai, 2015). Thus, using LA to understand how students experience peer feedback is a promising avenue. To the best of our knowledge, however, there are no studies that use LA to understand PA where the focus is on how students experience feedback.

1.3. Quality of LA data

Researchers agree that the quality of the results of LA on big data is dependent on the quality of the questions asked (e.g., Kitchin, 2013; Prinsloo & Slade, 2017). Big data is often collected by the private sector “as an auxiliary function of their core business” in order to “improve business processes and to document organization activities” (Buchanan, Gesher, & Hammer, 2015, p. 93). Roschelle and Krumm (2016) warn about mistaking “the ability of a system to collect abundant data with its ability to provide meaningful and useful measures” (p. 7) and notice that “many commercial online learning environments that students interact with do not track, or log, useful data” (p. 5).

LA researchers are usually not involved in the development of educational tools that are in widespread use in schools and universities. Thus, the data they analyse is that which is generated by the tools, as decided by the tool developer and not by the researcher who will use the data. This means that data is not always collected to gain insights into a specific educational question, but for other reasons, such as to optimize user experience. Another important aspect of educational big data coming from the private sector is that it often has to be combined with other data sources “to enrich the set of attributes to be studied” (Buchanan et al., 2015, p. 94). Buchanan et al. (2015) call this kind of dataset “massive but lean” (p. 94). Further, Krumm, Means, and Bienkowski (2018) argue:

> “The data a researcher eventually analyzes depends upon the business rules of the database as well as the informal rules around how individuals input and make use of data within these systems” (p. 27).

In particular, working with exhaust big data, which was collected as a by-product of the primary task, can be challenging. The data generated may be messy and dirty (Kitchin, 2014), however, for many researchers it is a reality to work this kind of data. In the ideal situation, a researcher would have a full control over the learning environment and be able to determine the kind and format of data that is going to be collected. Generally, this it is not the case.

In this study we analyse a dataset provided by a commercial PA platform, where we did not influence the data collection. The implications of this on the data analysis and findings are addressed in the discussion.

The paper is organised as follows. First, a short literature review of relevant research on feedback, peer assessment, and backward evaluation is presented. Next, the research questions, the research method, and details of the dataset are presented. An analysis and discussion of findings follows before we conclude.

2. Previous research

Research shows that feedback can have a considerable impact on student learning (Black & Wiliam, 1998; Evans, 2013; Hattie & Timperley, 2007; Kluger & DeNisi, 1996; Shute, 2008). Feedback interventions have been found to be particularly effective when they raise the students’ awareness of how to improve (feed forward) in relation to their current level of performance (feed back) and the learning intentions (feed up) (Black & Wiliam, 1998, 2009, p. 2009; Hattie & Timperley, 2007; Nicol & Macfarlane-Dick, 2006; Sadler, 1989). Nevertheless, feedback does not always result in student improvement, and may in some cases inhibit learning rather than promote it. Variations in the effect of feedback have been related to content, form and timing of the feedback, and studies have indeed found variations based on these factors (Hattie & Timperley, 2007; Kluger & DeNisi, 1996; Shute, 2008).

Another variation in the effectiveness of feedback is related to how individual students perceive and use feedback (Bloxham & Campbell, 2010; Carless, Salter, Yang, & Lam, 2010; Hattie & Gan, 2011; Higgins, Hartley, & Skelton, 2001; Nicol & Macfarlane-Dick, 2006; Sadler, 2010). In the literature, there are numerous examples of students failing to make use of the feedback they are given (see Evans, 2013; Jonsson, 2013 for reviews on the topic.). This discrepancy is commonly referred to as the “feedback-gap”. In a review, Jonsson (2013) found that students’ use of, or lack of use, is related to their understanding of the information. To strengthen this ability (to interpret feedback) it has been suggested that students need make their own assessment experiences through the assessment of peers (Sadler, 2009, 2010).

2.1. Peer assessment

Peer Assessment (PA) is an “arrangement in which individuals consider the amount, level, value, worth, quality, or success of the products or outcomes of learning of peers of similar status” (Topping, 1998, p. 250). PA can be qualitative (e.g., writing feedback comments), quantitative (e.g., assigning a grade) or a mixture of both (Patchan, Schunn, & Clark, 2018). When feedback is given for formative purposes it is generally agreed that feedback should not only be passively received, but also lead to improvement (Dawson et al., 2019; Evans, 2013; Jonsson, 2013).

Although PA is performed by the students themselves, studies have found that PA appears to be just as effective as teacher assessment when it comes to enhancing students' academic achievement (Double et al., 2018). This is perhaps surprising since teachers usually have more experience with both assessment and the content of a course. As several authors have noted (i.e., Double et al., 2018; Sadler, 2009; Topping, 2009; Tai, Ajjawi, Boud, Dawson, & Panadero, 2018), however, PA has some potential benefits over teacher assessment when it comes to both providing and receiving feedback.

As feedback providers students can develop their own assessment competence to better understand assessment criteria, recognise what is understood as quality in a particular field, and thus become better to interpret feedback and self-assess their own work in the future (Sadler, 2009(Sadler, 2010). As feedback receivers, students can get feedback from peers that is given in a language that is close to their own and with a level of complexity that is well adapted to their subject understanding (Topping, 2009). This might be particularly useful for undergraduate students where the difference in the competence of the teachers and the students can be a barrier for providing feedback adapted to the students' zone of proximal development (Hrepic, Zollman, & Rebello, 2007; Nicol, 2009).

Building on the work of Sadler and others, Tai et al. (2018) relate PA to the development of students Evaluative Judgement abilities. *Evaluative Judgement* is defined as the ability to evaluate the quality of own or other's work and is an important aspect of PA (Tai et al., 2018). Its goal is to develop an instinct for good and bad quality output. As a higher-level cognitive ability, evaluative judgement positions students as active participants in the PA process, where they use their critical thinking abilities to assess the quality of the work and are expected to justify their assessment. To develop evaluative judgement skills, students need to not only be exposed to work repeatedly, but also become familiar with the quality criteria as stated in the PA rubric (Tai et al., 2018).

Engaging in PA also seems to have an affective advantage in terms of self-efficacy. Feedback promoting self-efficacy leads to better self-regulation and more effort devoted to the task (Hattie & Timperley, 2007), and several studies have shown that PA correlates positively with self-efficacy (Baleghizadeh & Mortazavi, 2014; Ertmer et al., 2010; Liu, Lu, Wu, & Tsai, 2016). The positive findings on PA and self-efficacy have been explained by the increased opportunity for observational learning and peer-modeling (Double et al., 2018). This is likely to be related to the processes of both receiving and providing feedback since students get exposed to various ways in which their peers have solved a task when assessing others as well as receiving advice on their own work when receiving feedback. Engaging in such activities might boost the students' confidence in their own ability to meet the requirements of a course (Baleghizadeh & Mortazavi, 2014). This might be particularly useful for overcoming the feedback gap, since there is evidence that assessment enhances performance when self-efficacy is high and impedes performance when self-efficacy is low (Beckmann, Beckmann, & Elliott, 2009; Birney, Beckmann, Beckmann, & Double, 2017; Kluger & DeNisi, 1996).

2.2. Backward evaluation

For feedback to be successful, it needs to be actionable, lead a student to reflection and change in behaviour, however, it is difficult to ensure that a student will not only be a passive feedback recipient (Cook, 2019; Winstone, Nash, Parker, & Rowntree, 2017; Yuan & Kim, 2015). *Backward Evaluation* (BE) refers to students' evaluation of the peer feedback that they received on their work and is one of the methods that should increase student engagement (Luxton-Reilly, 2009). Thus, students are enabled to tell their peers (as well as the teacher) how they experienced the feedback. From a research perspective, it is an opportunity to gain more insight into student feedback receiving skills, and the interplay between roles as a feedback receiver and a feedback provider (Mulliner & Tucker, 2017; Adewoyin, Araya, & Vassileva, 2016; Patchan et al., 2018). Past research on student perception of feedback was limited to self-reports (Ryan et al., 2019). Due to technological developments it is possible to collect detailed data on student's digital behaviour and embed BE in the PA process on a digital platform in the form of scales (quantitative) or student comments (qualitative).

Only a few PA studies include BE in their analysis, typically as a helpfulness scale or a free-text comment. BE data is used to determine tit-for-tat behaviour by students in PA (Adewoyin et al., 2016; Cho & Kim, 2007; de Alfaro & Shavlovsky, 2016), or to examine the mediators of feedback implementation (Nelson & Schunn, 2009; Van der Pol, Van den Berg, Admiraal, & Simons, 2008; Wu & Schunn, 2020). Other examples are using BE to 1) examine if a student's belief that their feedback will be judged based on its helpfulness rather than its consistency with respect to other student's feedback influences feedback quality (Patchan et al., 2018), or 2) determine improvement in student's writing skills (Cho, Schunn, & Kwon, 2007).

BE comments are commonly analysed in the context of students either *agreeing* and/or *understanding* the feedback that they received. Van der Pol et al. (2008) conducted two studies in which students graded the feedback that they received using an importance score (study 1 with 27 students) and a helpfulness score (study 2 with 38 students), while BE comments were coded based on student's level of agreement with the feedback. Their first study found that a higher perceived importance of feedback on their work by students corresponded with more revisions in their written work, while the second study showed that students agreed more with the feedback that they perceived as useful. Student's agreement with the feedback, and not perceived feedback usefulness, correlated with higher rate of revision. Wu and Schunn (2020) conducted a study with 185 students. In addition to a score measuring feedback helpfulness, an extended BE comment coding that included both agreement with the feedback and how well students understood the feedback, was used. Student understanding and agreement with feedback were found to be significant predictors of revision. Feedback with concrete solutions contributed to a higher understanding of feedback, and feedback including mitigating praise predicted agreement with the problem. However, a higher number of praise comments predicted lower agreement with the feedback and a lower revision rate.

2.3. Rubrics

A *rubric* is defined as "a simple assessment tool that describes levels of performance on a particular task" (Hafner & Hafner, 2003, p. 1509). In the PA context, where students are not the experts, a rubric has two main purposes: improve student's feedback skills; and, teach them how to evaluate work within a certain discipline. As Nilson (2003) noticed the quality of feedback does not only depend on student's skills, but also the feedback questions that students are asked. Previous research on rubrics in PA focused on the amount of guidance necessary in a rubric. For example, Ashton and Davies (2015) compared two groups in a MOOC writing course; one group was guided only by the rubric, and the other one with an additional instructional section and a series of sub-questions aiming to enhance student's understanding of the rubric. Similarly, in a face-to-face setting Gielen and De Wever (2015) examined three levels of PA structuring through added instructions and guiding questions to the rubric. Other studies explore the validity or reliability of singular rubric. For example, De Wever, Van Keer, Schellens, and Valcke (2011) investigated the intra-group reliability of the same rubric used in two groups, the first group without previous instruction on the rubric and only one PA activity in a wiki environment, and the second group informed about the rubric before the activity and performing the PA twice during a semester. We found no research that looks at how student BE might provide insight into a rubric's quality.

2.4. Filling the research gaps

In this exploratory study we work with a dataset provided by an online PA platform and explore the variables and methods that can be used to expand knowledge of PA and identify the limitations of our approach. The main goal of our research is to explore how we can use LA to gain insight into PA, in particular in BE, which is an important indicator of how students perceive the feedback they have received. We extend previous research on BE in PA by gaining a better understanding of the relationship between the usefulness of feedback, improvement suggestions, and comments on the feedback, and by exploring the relationship between rubric characteristics and feedback perception.

Based on this background we have two research questions. The first research question is:

RQ1: *What is the relationship between student's perception of the usefulness of feedback, improvement suggestions, and comments on the feedback?*

To investigate if there is a relationship between the number and type of questions in a rubric and the student's perception of feedback, we ask:

RQ2: *What is the relationship between rubric characteristics and student's perception of the usefulness of feedback?*

3. Methodology

3.1. Dataset

Peergrade (peergrade.io) is an online PA platform that affords the opportunity for students to evaluate the usefulness of the feedback they receive by 1) assigning a numerical feedback grade (score), 2) selecting from a list of improvement suggestions, and 3) giving free-text comments. Data from these three functionalities provides an opportunity to gain more insight into how students experience feedback from their peers, and which characteristics of the feedback that students find useful.

As depicted in Fig. 1, a typical PA activity on the Peergrade platform starts with a teacher creating an assignment and a corresponding rubric according to which a student should evaluate another student's work (hand-in). The rubric can include boolean, numerical, and free-text questions. After finishing the assignment, students upload their work (hand-in) to the Peergrade platform. In the next step, students typically receive 3–5 hand-ins on which they should give feedback according to the rubric that the teacher has created. Finally, students receive feedback from 3 to 5 peers on their own hand-in and conduct BE by scoring the feedback on their hand-in, selecting improvement suggestions, and writing a comment. Table 1 shows the feedback grade–the numerical score scale of 1-5–that indicates student perceived feedback usefulness, and the multiple-choice improvement suggestions scale–with five suggestions–that indicates how the feedback that students receive on their work could have been improved.

In this study we use an anonymised Peergrade dataset collected across many institutions that used the tool between 2015 and 2017. The dataset has 10,197 unique student IDs and 6,329 unique course titles, but does not contain the student hand-ins, due to consent issues. We do not have any context information about the integration of the PA activity in course structure, nor its pedagogical context. While several courses have over 300 students participating in a PA activity, most courses have less than 30 students. The median number of students in a course is 15, while the average is 24 students. It is important to note that the number of students refers only to the number of participants that are visible in a particular PA activity and may not reflect the overall number of students in a course. From our own experience with university instructors using Peergrade we know that some student feedback is given from a group of students and not an individual student, thus what appears to be a single feedback might actually come from a group.

Table 1
Description of feedback grades and improvement suggestions in Peergrade.

Feedback grade	1 (FG1)	Not useful at all
	2 (FG2)	Not very useful
	3 (FG3)	Somewhat useful, although it could have been more elaborate
	4 (FG4)	Very useful, although minor things could have been better
	5 (FG5)	Extremely useful, constructive and justified
Improvement suggestions	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.
	specificity	The feedback should be more specific and point to concrete things that can be improved.

3.2. Methods

The research method involved data pre-processing and data analysis. Data pre-processing included both cleaning and coding of the data and was conducted using Python. Since we did not have control over the data collection, a major task was to understand the data structure and content, and what it represents. This was particularly challenging since the data had very limited context information. Thus, the variables used in our analysis had to be chosen based on their availability in the dataset and their potential for use in LA methods. These include some variables that have been used in earlier studies related to form and perceived use of feedback (recall section 2).

In order to gain insight into the dataset, we applied descriptive statistics and examined the distribution of dependent and independent variables. It was decided to conduct Spearman rank correlation, since it is more appropriate for correlation of ordinal variables than standard methods, such as Pearson correlation (Mukaka, 2012).

To select variables for the regression analysis, we conducted backwards stepwise regression that starts the analysis with all available independent variables, and with each iteration removes the least significant variable (Healy, 1995).

The dependent variable in the current study, the feedback grade (FG), is an ordinal categorical variable with five levels (recall Table 1). The recommended method to model an ordinal dependent variable is ordinal logistic regression, since metric methods might distort the analysis results (Liddell & Kruschke, 2018).

The statistical analysis was conducted in *R 3.5.0* using various packages, such as the *ggplot2* package for data visualisation (*v3.1.1*;

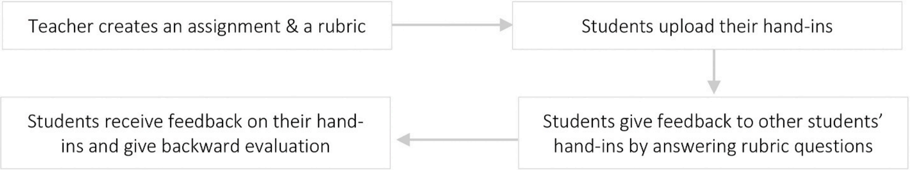
Teacher creates an assignment & a rubric
Students upload their hand-ins
Students give feedback to other students' hand-ins by answering rubric questions
Students receive feedback on their hand-ins and give backward evaluation

Fig. 1. Peer Assessment activity in Peergrade.

Wickham, 2016), the *sjPlot* package for Spearman Rank Correlation (*v2.7.0*; Lüdecke, 2019), and the *MASS* package for ordinal logistic regression and stepwise regression (*v7.3-51.4*; Venables & Ripley, 2002).

Moreover, *Epistemic Network Analysis* (ENA) was used to analyse and visualize the data. Epistemic Networks are "mathematical representations of the patterns of connections among Codes in the epistemic frame of a Discourse" ((Shaffer, 2017), p. 333). ENA models the connections between different concepts and projects them onto a two-dimensional space as a nondirectional network. This enables comparison between the groups by subtracting the edge weights of networks. Moreover, a statistical comparison of the variance explained by two axes and the goodness of fit of a particular model is possible (Shaffer & Ruis, 2017). ENA was conducted using the ENA web tool (epistemicnetwork.org). Though a usual application of ENA would model coded concepts, in this study we decided to model the variables available in our dataset with a goal to gain insights into students' choices regarding the number and kinds of improvement suggestions depending on their feedback perception. The motivation to apply ENA in this context is to visualize the students' use of improvement suggestions, and, thus, explore which insights can be gained from using this novel method.

3.3. Data pre-processing

Peergrade provided an anonymised dataset in multiple JSON files. The relevant variables were extracted into a CSV file.

Due to challenges in working with multiple languages, the Peergrade dataset was first parsed for BE comments in English. Twenty-five languages were detected but only English entries were retained, which resulted in a dataset with 10,197 unique student IDs and 6,329 unique course titles.

Dependent and independent variables in the dataset were preprocessed as follows. The dependent variable for both RQ1 and RQ2 is *feedback grade* (*f_grade*) and the numerical feedback grade given by the students (recall Table 1) was coded as an ordinal categorical variable. *F_grade* is the simplest variable to indicate students' perception of the feedback that they received. In Peergrade the scale measures *feedback usefulness*, however, in previous research *feedback helpfulness* (Cho et al., 2007; de Alfaro & Shavlovsky, 2016; Patchan et al., 2018; Wu & Schunn, 2020), or *feedback importance* (Van der Pol et al., 2008) can be found to measure BE.

The independent variables used to answer RQ1 include: *BE comments*, *BE comment length*, *part-of-speech tagging*, *sentiment analysis,* and *improvement suggestions*.

BE comments are free text comments where students can express their reaction to feedback that they received on their work. Previous studies coded their BE comments using either the level of agreement with the feedback comment and/or the level of understanding of the feedback comment using either 2- or 3-points scale (Nelson & Schunn, 2009; Van der Pol et al., 2008; Wu & Schunn, 2020). In this study, we decided to code the data using a bottom-up approach, where the coding categories emerged from looking at the data. The coding scheme was validated by two researchers that coded a random sample of 10% of the whole dataset and achieved an inter-rater reliability of Cohen's kappa of at least $\kappa = 0.8$ for every code. After this simple automatic coding was used. BE comments were coded using string matching into three suggestions: *accepting, defending,* and *gratitude* (see Table 2 for coding examples). The unit of analysis was one BE comment, which means that every comment could be coded with one or more category. As a result, *BE comments* were dummy coded with one of five variables: *only accepting* (*acc*), *only defending* (*def*), *only gratitude* (*grat*), *accepting-defending* (*acc_def*), *accepting-gratitude*, (*acc_grat*), *defending-gratitude* (*def_grat*), and *accepting-defending-gratitude* (*acc_def_grat*). BE comments that could not be coded due to their incomprehensibility (e.g., "tjlkdlfjsldkfj", "this is blank" or "giff all the points"), were removed from the dataset.

BE comment length (*BE_c_length*) was measured as the number of characters and was normalized to a 0–1 range. *BE_c_length* was used in previous research to predict the BE helpfulness rating (Cho et al., 2007; Adewoyin et al., 2016).

Table 2
BE comments coding examples.

Code	Description	Examples
Accepting (acc)	BE comments expressing praise, error acknowledgment, or intention of revision	"Great feedback! The comments in response to yes/no questions were particularly helpful." "You're right, there is a lot of depth I could have added. I'm in the process of growing as a writer and your advice will definitely help." "I will fix my mistakes, use more evidence and check over my essay better for the next time."
Defending (def)	BE comments expressing confusion, criticism, or disagreement	"I don't really understand the second one because what do they mean by "better paragraphs"?" "It lacked any form of elaboration. Very brief." "But we did have different lighting in the pictures."
Gratitude (grat)	BE comments expressing gratitude	"Thanks the grader's time and efforts for the grading."

Part-of-speech tagging (*p_of_speech*), that is the grammatical properties of BE comments, were extracted using the spaCy Python package. In the current study, we focused on three main tags: verbs (*verbs*), nouns (*nouns*), and adjectives (*adjs*). These tags were counted per BE comment, and are represented as a proportion of all words in a BE comment. *P_of_speech* is among the NLP features most commonly used to automatically detect a particular type of peer feedback comment, for example, helpful comments or suggestions within the feedback comments, using predictive models (Nguyen & Litman, 2014; Zingle et al., 2019). In this study, we decided to include *p_of_speech* to explore not only what students wrote in their BE comments, but also how they expressed themselves.

Sentiment analysis (*sentiment*) was conducted on BE comments using the Vader sentiment analyser (Hutto & Gilbert, 2014). Sentiment scores ranged from −1 (negative) to 1 (positive). Every BE comment has one sentiment score. Piech et al. (2013) used *sentiment* and *BE_c_length* to determine students' commenting style as a part of developing algorithms to reduce student biases and reliabilities in MOOCs PA.

Improvement suggestions (*impr_suggs*) refers to what students selected from a list of improvement suggestions (recall Table 1). Students could choose none, one, or many from five suggestions: *constructivity, specificity, kindness, justification,* and *relevance.* The *number of improvement suggestions* (*#_of_impr_suggs*) is a numerical variable that ranges from 0 to 5 and corresponds to the number of improvement suggestions selected by a student. *#_of_impr_suggs* was normalized to 0–1 for the statistical analysis. *#_of_impr_suggs* was transformed to a binary variable with two levels: *fewCat* (1–3 suggestions), and *manyCat* (4–5 suggestions) for the ENA. *Impr_suggs* is a unique PA platform feature found in Peergrade–we are not aware of previous research including this variable.

Two independent variables were included in the analysis of data related to the rubric design RQ2:

Question type (*q_type*) describes the type of question in a rubric: *numeric, boolean,* or *text.* The percentage of each type of questions per rubric was calculated.

# of questions (*#_of_qs*) refers to the number of questions per rubric. For the ordinal logistic regression and correlation analysis, it was normalized to 0–1. We have not found previous research that has investigated the rubric design and its relationship to PA, so these variables have been chosen as we feel that they clearly describe a rubric.

4. Analysis

After data pre-processing and removal of observations with missing values, the final dataset was n = 7,660 records. This section describes the analysis using descriptive statistics, Spearman rank correlation, ordinal logistic regression, and Epistemic Network Analysis.

4.1. Descriptive statistics

Table 3 shows the means, standard deviations, and median for the numerical variables included in the study, and the frequencies and percentages for each level of the categorical variables. Fig. 2 visualizes the distribution of each variable.

The majority of students (almost 60%) graded feedback *extremely useful* (FG5 = 0.32), or *very useful* (FG4 = 0.27) (see Fig. 2a). Only 18% of all feedback grades were *not useful at all* (FG1 = 0.09), or *not very useful* (FG2 = 0.1). As depicted in Fig. 2b, most BE comments were coded with only one category. *Defending* comments are the most frequent type of comment *(def* = 0.29) followed by *accepting* comments (*acc* = 0.28). The least frequent combination of codes was *defending and gratitude* (*def_*grat = 0.016) and *accepting, defending and gratitude* (*acc_def_*grat = 0.023). In contrast, the most popular combination of codes was *accepting and gratitude* (*acc_*grat = 0.13).

The density plot, see Fig. 2c, shows that the distribution of sentiment scores for BE comments is skewed towards positive (over 0) and neutral scores (around 0). As depicted in Fig. 2d, the majority of BE comments are short. The median text length is 69 characters, and the average is 104 characters. The shortest comment is 7 characters, and the longest is 2,735 characters. Moreover, most used part of speech is *verb* (mean = 0.205, median = 0.205) followed by *noun* (mean = 0.168, median = 0.158) (see Fig. 2e).

75% of students did not choose any improvement suggestion and only 3% chose four, whereas 1% selected all five improvement suggestions, as shown in Fig. 2f. The most popular improvement suggestion was *specificity* (25.08%), followed by *constructivity* (22.23%), and *justification* (17.26%).

Numerical and *text* questions were proportionally the most used questions per rubric (see Fig. 2g). The mean number of questions per rubric is 7.97. The shortest rubric has only 1 question, whereas the longest rubric has 64 questions (see Fig. 2h).

The proportion of *gratitude* and *accepting* comments are highest for FG5 (*grat* = 0.098; *acc_grat* = 0.078) and FG1 (*grat* = 0.001; *acc_grat* = 0.0003) as depicted in Fig. 3. Moreover, the proportion of *accepting* comments is the highest for *FG5* (*acc* = 0.125), whereas the proportion of *defending* comments is the highest for FG1 (*def* = 0.072). The highest proportion of comments coded with more than one code is for FG3 (*acc_def* = 0.26; *def_grat* = 0.06; *acc_def_grat* = 0.07), and FG4 (*acc_def* = 0.25; *def_grat* = 0.05; *acc_def_grat* = 0.009).

4.2. Spearman rank correlation

Spearman rank correlation results are listed in Table 4. Although most independent variables show a statistically significant relationship with feedback grade, no variables show *very strong* (*rho* = .8–1.0) or *strong* relationships (*rho* = 0.60-0.79). *#_of_impr_suggs* has a moderate negative relationship with FG5 (*rho* = −0.453, *p*=<.001), and a weak positive relationship with FG1 (*rho* = 0.214, *p*=<.001), FG2 (*rho* = 0.228, *p*=<.001), and FG3 (*rho* = 0.236, *p*=<.001). Only *defending* coded BE comments (*def*) show a weak positive relationship with FG1 (rho = 0.362, *p*=<.001) and FG2 (*rho* = −0.264, *p*=<.001), and a weak negative relationship with FG5 (*rho* = −0.390, *p*=<.001). BE comments coded as both *accepting* and *gratitude* (*acc_grat)* have a weak positive relationship with FG5 (*rho* = 0.235, *p*=<.001). *Constructivity*, *justification* and *specificity* have weak negative relationships with FG5 (*constructivity*, *rho* = −0.284, *p*=<.001; *justification, rho* = −0.231, *p*=<.001; *specificity, rho* = −0.284, *p*=<.001), while *relevance* has a weak positive relationship with FG1 (*rho* = 0.292, *p*=<.001). The *sentiment* has a weak positive relationship with FG5 (*rho* = 0.275, *p*=<.001), and a weak negative relationship with FG1 (*rho* = −0.268, *p*=<.001).

Adjs has a very weak negative relationship with FG1 (*rho* = −0.066, *p*=<.001), and a very weak positive relationship with FG5 (*rho* = 0.059, *p*=<.001). *Nouns* has a very weak negative relationship with FG3 (*rho* = −0.026, *p*=<.05), while *verbs* has a very weak negative relationship with FG5 (*rho* = −0.127, *p*=<.001), and a very weak positive relationship with FG1 (*rho* = 0.046, *p*=<.001), FG2 (*rho* = 0.069, *p*=<.001), and FG3 (*rho* = 0.066, *p*=<.001).

Boolean has a very weak negative relationship with FG3 (*rho* = 0.024, *p*=<.05), and a very weak positive relationship with FG5 (*rho* = −0.027, *p*=<.01), while *text* has very weak negative relationship with FG2 (*rho* = −0.035, *p*=<.01) and FG3 (*rho* = −0.037, *p*=<.01), and a very weak positive with FG5 (*rho* = 0.049, *p*=<.001). *#_of_qs* has a very weak negative relationships with FG1 (*rho* = −0.035, *p*=<.01) and a very weak positive relationship with FG2 (*rho* = 0.025, *p*=<.05).

Finally, *BE_c_length* has a very weak negative relationship with FG5 (*rho* = −0.136, *p*=<.001), and a very weak positive relationship with FG2 (*rho* = 0.083, *p*=<.001), FG3 (*rho* = 0.077, *p*=<.001), and FG4 (*rho* = 0.027, *p*=<.01).

4.3. Ordinal logistic regression

In order to select variables for the ordinal logistic regression, a stepwise regression using backward elimination was carried out. The

Table 3
Descriptive statistics of dependent (f_grade) and independent (BE_comment, impr_suggs, p_of_speech, q_type, BE_c_length, sentiment, #_of_impr_suggs, #_of_qs) variables.

		Freq/%			Mean/SD/Median
f_grade	1 (FG1)	674/8.80	p_of_speech	adjs	0.111/0.102/0.099
	2 (FG2)	771/10.06		nouns	0.168/0.109/0.158
	3 (FG3)	1,622/21.17		verbs	0.205/0.106/0.205
	4 (FG4)	2,091/27.30	q_type	boolean	0.2535/0.327/0.00
	5 (FG5)	2,502/32.66		numerical	0.3461/0.355/0.25
BE_comment	acc	2,131/27.82		text	0.4005/0.391/0.25
	def	2,196/28.67	BE_c_length		104.3/127.64/69.0
	grat	1,495/19.52	sentiment		0.358/0.426/0.44
	acc_def	580/7.57	#_of_impr_suggs		0.792/1.01/1.00
	acc_grat	961/12.55	#_of_qs		7.973/6.03/7.00
	def_grat	123/1.61			
	acc_def_grat	174/2.27			
impr_suggs	constructivity	1703/22.23			
	justification	1323/17.27			
	kindness	372/4.86			
	relevance	748/9.77			
	specificity	1921/25.08			

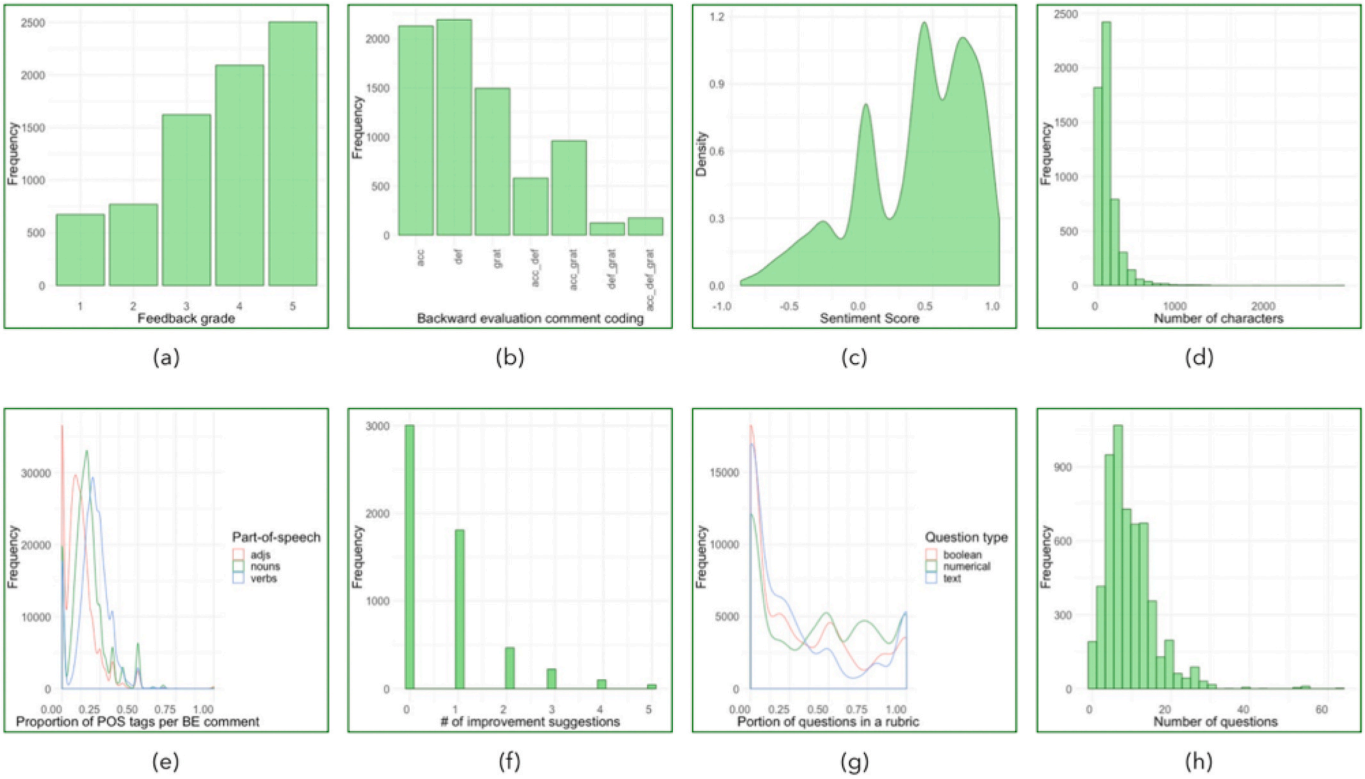

Fig. 2. Distributions of dependent and independent variables: (a) feedback grade, (b) BE comment codes, (c) sentiment score, (d) BE comment length, (e) part-of-speech tags, (f) number of selected improvement suggestions, (g) question type, (h) number of questions.

first model included all variables as listed in Table 3 and resulted in the Akaike information criterion (AIC) of 18420.73. After five iterations the final model had an AIC of 18413.19. With 15 selected variables from the final model of stepwise regression, an ordinal logistic regression (see Table 5) was run.

To check for the absence of multicollinearity, a generalised variance inflation factor (GVIF) was applied on the ordinal regression model (Fox & Weisberg, 2011). Three variables in the final model have GVIF values higher 5 (text, GVIF = 38.389; BE_comment, GVIF = 8.496; specificity, GVIF = 15.201), which indicates some multicollinearity and possible bias in the final model (see Table 6). To ensure that the Parallel Regression Assumption holds, a Brant test (Brant, 1990) was conducted. The test was successful, and the results are shown in Table 7.

The results of the ordinal logistic regression show that BE comments coded as *def* ($\beta = -2.30$, $p \leq .001$) *def_grat* ($\beta = -1.28$, $p \leq .001$) or *acc_def* ($\beta = -1.18, p \leq .001$; $\beta = -0.74, p \leq .001$) indicate that students are more likely to find feedback less useful in comparison with the baseline, i.e., BE comments coded with *acc*. Moreover, if BE comments were coded with *acc_grat* ($\beta = 0.67, p \leq .001$), or *grat* ($\beta = 0.19, p \leq .01$), there is a higher likelihood of perceiving feedback as more useful rather than not useful in comparison with the baseline, i.e., BE comments coded with *acc*.

The selection of an *impr_suggs* by a student predicts a higher likelihood that a student will find feedback less useful than more useful (*relevance*, $\beta = -1.23$, $p \leq .001$; *constructivity*, $\beta = -0.84$, $p \leq .001$; *specificity*, $\beta = -0.82$, $p \leq .001$; *kindness*, $\beta = -0.71$, $p \leq .001$; *justification*, $\beta = -0.6$, $p \leq .001$).

Unsurprisingly, a higher *sentiment* of a BE comment predicts that the students will find feedback more useful than less useful ($\beta = 0.94$, $p \leq .001$). Furthermore, a longer *BE_comment_length* indicates that students are more likely to perceive feedback as less useful ($\beta = -0.88$), however, this result is not statistically significant (p=.077).

The higher proportion of *text* questions per rubric predicts positive feedback perception more than negative feedback perception (β = 0.001, $p \leq .05$), and more *#_of_qs* per rubric makes students more likely to perceive feedback as more useful rather than less useful (*#_of_qs*, β = 0.39), however, this result is not statistically significant (p = .093).

4.4. Epistemic Network Analysis

In order to provide more insights into RQ1, ENA was used to model the relationships between the different improvement suggestions (*kindness*, *constructivity*, *specificity*, *relevance*, *justification,* recall Table 1) and the number of selected improvement suggestions (*#_of_impr_suggs*) grouped by the feedback grade. For this model *#_of_impr_suggs* was coded as *fewCat* for those where 1–3 suggestions were selected, and as *manyCat* for those where 4–5 suggestions were selected. The connections between *manyCat* or *fewCat* and individual improvement suggestions show which individual improvement suggestions were chosen based on the total number of suggestions selected, whereas the connections between the individual improvement suggestions indicate how often particular suggestions were chosen together.

As depicted in Fig. 4, five graphs for each feedback grade were constructed. A single BE activity, in which a student would write a BE comment, grade the feedback and choose improvement suggestions, comprises a unit of analysis. The stanza window was set to 1, since BE comments do not build a dialogue between each other. The edge line width represents the strength of the connection between the two codes, which is calculated through co-occurrence of codes. For better readability the edge weights were scaled by 2, and the model was rotated by FG1 and FG5. The means of the networks are the representation of the network's centroid for each feedback grade and are depicted by squares in the network space. Means rotation refers to a reduction of dimensions in order to position both means along a common axis to maximize the variance between the means of the two groups (Marquart et al., 2019). As the confidence intervals of the feedback grade centroids do not overlap, it indicates that there are statistically significant differences among the groups. 9.5% of the variance on the x-axis and 24.5% of the

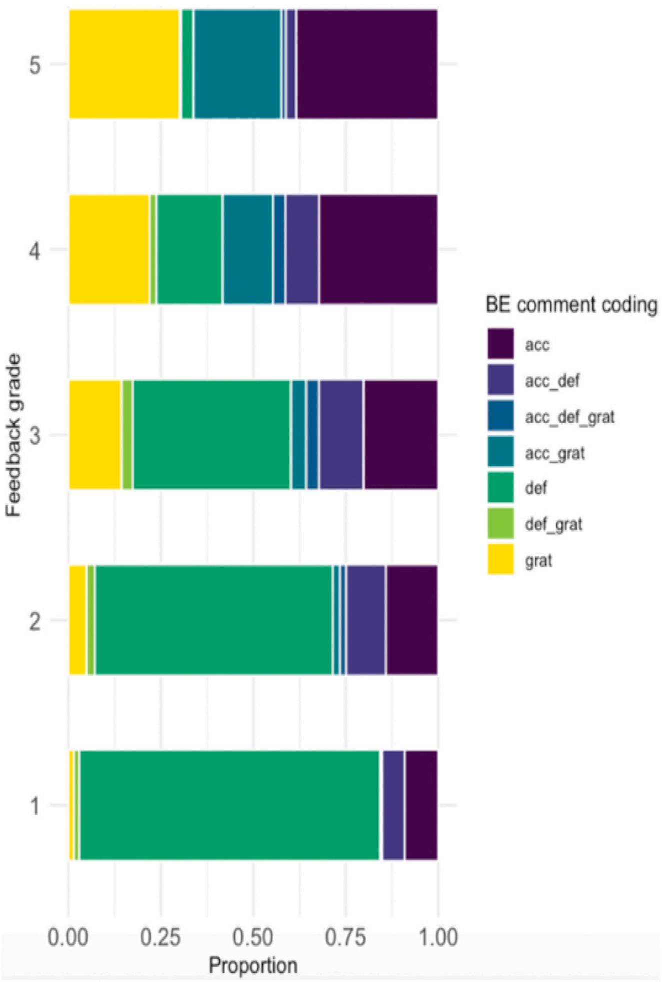

Fig. 3. The proportion of BE comment codes per feedback grade.

variance on the y-axis are explained by this model.

FG5 has the strongest connections between *fewCat-constructivity* (0.06), and *fewCat-specificity* (0.05) (See Fig. 4e). Similarly, the strongest relationships in FG4 are between *fewCat-specificity* (0.20), and *fewCat-constructivity* (0.14). Moreover, there is also a strong connection between *fewCat-justification* (0.09) (see Fig. 4d).

Fig. 4c shows that FG3 has not only strong connections between *fewCat* and almost all improvement suggestions (*specificity*, 0.19; *constructivity*, 0.14; *justification*, 0.10; *relevance*, 0.05), but also some strong relationships among the improvement suggestions themselves are visible: *specificity-constructivity* (0.09), *specificity-justification* (0.07), and *constructivity-justification* (0.05).

Similar to FG3, FG2 has strong connections between *fewCat* and almost all individual improvement suggestions, though the strength ranking is different (*relevance*, 0.11; *constructivity,* 0.10; *justification*, 0.10, *specificity*, 0.09) as depicted in Fig. 4b. The strongest connections among individual improvement suggestions are same as in FG3, however, the connections are stronger: *specificity-constructivity*, (0.15), *specificity-justification* (0.08), and, finally, *constructivity-justification* (0.08). Furthermore, FG2 builds strong connections between *manyCat* some individual improvement suggestions (*constructivity*, 0.07; *specificity*, 0.06; *justification*, 0.06).

As shown in Fig. 4a visualizing the plot for FG1, *fewCat* builds strong connections with all individual improvement suggestions (*relevance*, 0.10; *constructivity*, 0.09; *specificity*, 0.05; *justification*, 0.05; *kindness*, 0.05). Moreover, the following strong connections between individual improvement suggestions are prominent in this network: *specificity-constructivity* (0.14), *specificity-justification* (0.10), *constructivity-justification* (0.09), *specificity-relevance* (0.08), *relevance-constructivity* (0.08), and *relevance-justification* (0.07). The strong connections with *manyCat* were formed with every individual improvement suggestion, with the exception of *kindness:* (*specificity*, 0.10; *constructivity*, 0.10*; justification*, 0.09; *relevance*, 0.08).

Interestingly, *kindness* and *relevance* do not build many strong connections with other variables in plots for all feedback grades. *Relevance* can be found in FG3 and FG2 plots only in a strong connections with *fewCat*, and more prominently, in FG1 plot with *fewCat, manyCat,* and *specificity,* while *kindness* has only one strong connection with *fewCat* in FG1 plot.

Table 4
Spearman rank correlation between the levels of the dependent variable and independent variables (statistically significant moderate (rho = 0.60-0.79) and weak (rho = 0.20-0.39) relationships in bold).

		f_grade				
variable		1 (FG1)	2 (FG2)	3 (FG3)	4 (FG4)	5 (FG5)
BE_comment	acc	−0.130***	−0.102***	−0.089***	0.060***	0.164***
	def	**0.362*****	**0.264*****	0.163***	−0.146***	−**0.390*****
	grat	−0.141***	−0.123***	−0.067***	0.038***	0.187***
	acc_def	−0.017	0.040***	0.088***	0.035**	−0.126***
	acc_grat	−0.115***	−0.107***	−0.133***	0.019	**0.235*****
	def_grat	−0.003	0.016	0.056***	0.013	−0.069***
	acc_def_grat	−0.041***	−0.013	0.041***	0.046***	−0.046***
p_of_speech	adjs	−0.066***	−0.017	−0.012	0.002	0.059***
	nouns	−0.008	−0.007	−0.026*	0.011	0.022
	verbs	0.046***	0.069***	0.066***	−0.002	−0.127***
q_type	boolean	−0.002	0.006	0.024*	0.008	−0.027**
	numerical	−0.012	0.017	0.014	−0.010	0.013
	text	−0.003	−0.035**	−0.037**	0.020	0.049***
impr_suggs	constructivity	0.176***	0.145***	0.101***	−0.053***	−**0.284*****
	justification	0.154***	0.149***	0.117***	−0.064***	−**0.231*****
	kindness	0.194***	0.054***	0.002	−0.062***	−0.094***
	relevance	**0.292*****	0.149***	0.011	−0.105***	−0.182***
	specificity	0.119***	0.106***	0.163***	0.002	−**0.284*****
BE_c_length		−0.016	0.083***	0.077***	0.027**	−0.136***
sentiment		−**0.268*****	−0.168***	−0.093***	0.080***	**0.275*****
#_of_impr_suggs		**0.214*****	**0.228*****	**0.236*****	−0.031**	−**0.453*****
#_of_qs		−0.035**	0.025*	−0.005	−0.010	0.018

Statistically significant results in bold; ***$p \leq .001$; **$p \leq .01$; *$p \leq .05$.

Table 5
Results of the final model.

variable	Coeff.	SE	t-value	p value	OR	2.5%	97.5%
def	−2.305739	0.0658437	−35.018	0.000	0.0997	0.0876	0.1134
def_grat	−1.278116	0.1652677	−7.734	0.000	0.2786	0.2014	0.3850
relevance	−1.225436	0.0786071	−15.589	0.000	0.2936	0.2516	0.3425
acc_def	−1.181518	0.0884596	−13.357	0.000	0.3068	0.2579	0.3648
BE_comment_length	−0.877574	0.4968655	−1.766	0.077	0.4158	0.1562	1.0976
constructivity	−0.841965	0.0537310	−15.670	0.000	0.4309	0.3878	0.4787
specificity	−0.819465	0.0513720	−15.952	0.000	0.4407	0.3985	0.4874
acc_def_grat	−0.744511	0.1436610	−5.182	0.000	0.4750	0.3585	0.6297
kindness	−0.712994	0.1110787	−6.419	0.000	0.4902	0.3941	0.6091
justification	−0.595775	0.0595742	−10.001	0.000	0.5511	0.4904	0.6194
text	0.001134	0.0005291	2.144	0.032	1.0011	1.0001	1.0022
grat	0.185280	0.0649461	2.853	0.004	1.2036	1.0599	1.3672
#_of_qs	0.390543	0.2326673	1.679	0.093	1.4778	0.9376	2.334
acc_grat	0.671298	0.0782942	8.574	0.000	1.9568	1.6794	2.2828
sentiment	0.935929	0.2711685	3.451	0.001	2.5496	1.5041	4.3566
1|2	−4.255347	0.1531434	−27.787	0.000			
2|3	−3.034043	0.1480717	−20.490	0.000			
3|4	−1.371810	0.1438068	−9.539	0.000			
4|5	0.351695	0.1423876	2.470	0.014			

Abbrevations: Coeff. - Regression coefficient; SE - standard error; OR - odds ratio***P ≤ .001.

Table 6
GVIF results.

	GVIF	Df	GVIF (Adewoyin et al., 2016)
BE_comment	8.496	6	1.195
#_of_qs	1.905	1	1.380
BE_comment_length	1.018	1	1.009
text	38.389	1	6.196
sentiment	1.076	1	1.038
kindness	1.359	1	1.166
justification	1.434	1	1.197
constructivity	1.236	1	1.112
relevance	1.522	1	1.234
specificity	15.201	1	3.899

1 GVIF*^(1/(2*Df)).

Table 7
Brant test results.

variable	X2	df	probability
Omnibus	390.96	45	0
acc_def	9.77	3	0.02
acc_def_grat	12.75	3	0.01
acc_grat	21.89	3	0
def	25.57	3	0
def_grat	6.59	3	0.09
grat	28.46	3	0
#_of_qs	3.98	3	0.27
BE_comment_length	10.02	3	0.02
text	11.83	3	0.01
sentiment	5.36	3	0.15
kindness	9.99	3	0.02
justification	59.8	3	0
constructivity	16.31	3	0
relevance	1.88	3	0.6
specificity	148.68	3	0

5. Results

The main goal of our research is to explore how we can use LA to gain insight into PA, in particular BE in PA. In the current study we asked two research questions and analysed the Peergrade big dataset using descriptive statistics, Spearman rank correlation, and ENA. Stepwise regression was used to build the ordinal logistic regression to analyse the relationship between the ordinal dependent variable, feedback grade, and independent variables characterising BE and rubrics.

RQ1: *What is the relationship between student's perception of the usefulness of feedback, improvement suggestions, and comments on the feedback?*

When students perceived the feedback, they received on their work as *not useful at all* (FG1), they rarely expressed gratitude in their BE comments, but rather would voice confusion, criticism, or disagreement (*def*). This was also confirmed by the more likely negative sentiment score of the BE. Furthermore, the correlation analysis showed that they used less adjectives, and more verbs in their responses to feedback. Students selected more improvement suggestions and, in particular, *relevance.* This finding was expanded by the ENA, where *relevance* and *constructivity* had stronger connections with students selecting 1–3 improvement suggestions, while students that selected 4–5 improvement suggestions preferred mostly *specificity, constructivity* or *justification.* Furthermore, *specificity* was chosen mostly in combination with either *constructivity* or *justification.*

Similar to FG1, students that graded feedback as *not very useful* (FG2), had also expressed only *defending,* negative sentiment and used more verbs in their BE comments. Moreover, they were more likely to select improvement suggestions. Specifically, they selected 1–3 improvement suggestions, such as *relevance, constructivity, justification* or *specificity,* or a combination of *specificity* and *constructivity.*

Somewhat useful graded feedback (FG3), was accompanied by the BE comments coded with more than one code. However, as in the case of FG1 and FG2, the sentiment score of the BE comments was more likely to be negative, and a similar trend using more verbs was found. Moreover, students were less likely to use nouns in their comments. As for FG2, there is a positive correlation between FG3 and the selection of improvement suggestions and, in particular, the *specificity-constructivity* combination was the most popular choice among students. If they selected 1–3 improvement suggestions, the suggestions chosen were mostly *specificity* and *constructivity* followed by *justification.*

Though BE comments for feedback rated as *very useful* (FG4) are also among the ones with the highest proportion of comments with more than one code, they showed a positive sentiment score, rather than a negative sentiment score as was the case in FG3. In addition, students mostly selected 1–3 improvement suggestions, such as *specificity, constructivity,* and *justification*, which is the same pattern found in FG3.

Students grading the feedback as *extremely useful* (FG5), expressed most *gratitude*, gratitude mixed with praise, error acknowledgment, or intention of revision or only *accepting* in their BE comments compared to other feedback grades. Moreover, they were less likely to voice confusion, criticism, or disagreement in their BE comments, and their BE comments were more likely to have a positive sentiment score. In contrast to FG1, FG2 and FG3, these students were less likely to use

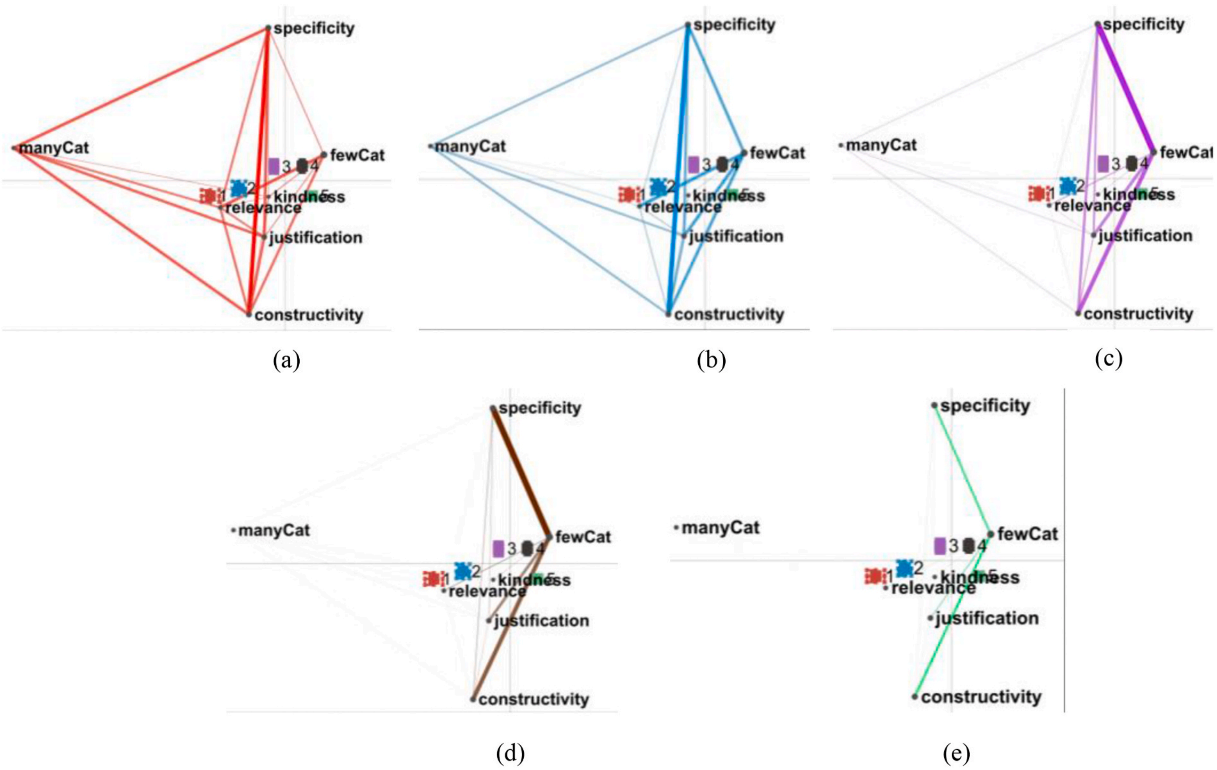

Fig. 4. ENA model plotting improvement suggestions with the number of improvement suggestions: (a) Plot for feedback grade 1 (FG1), (b) Plot for feedback grade 1 (FG2), (c) Plot for feedback grade 3 (FG3), (d) Plot for feedback grade 4 (FG4), (e) Plot for feedback grade 5 (FG5) NOTE: FG=Feedback Grade; FG1 = Feedback Grade 1, etc.

verbs, while in comparison to FG1, they were also more likely to use more adjectives in their BE comments. Furthermore, students used less improvement suggestions, when the found feedback *extremely useful.* ENA for FG5 showed only two moderately strong connections between the selection of 1–3 suggestions and *constructivity* or *specificity*. In addition, the correlation analysis showed that popular improvement suggestions for all other grades, *constructivity, justification* or *specificity* were less likely to be selected for FG5.

Generally, if students expressed any confusion, criticism, or disagreement in their BE comment–even if they also expressed gratitude, or praise, error acknowledgment, or intention of revision in the same comment–they were more likely to find feedback less useful. On the other hand, if students expressed praise, error acknowledgment or intention of revision alone or together with gratitude, or gratitude only, there was a higher likelihood of perceiving feedback as more useful. Similarly, Van der Pol et al. (2008) found that the more students agreed with the feedback, the more useful they would grade it. The selection of an improvement suggestion by a student predicted a higher likelihood that a student will find feedback less useful than more useful. Unsurprisingly, a higher sentiment score of BE comment predicted that the students will find feedback more useful. Furthermore, students writing a longer BE comment were more likely to have found feedback less useful, however, this result was not statistically significant. This finding corresponds to Adewoyin et al. (2016) who found that longer comments do not predict higher BE ratings.

RQ2: *What is the relationship between rubric characteristics and student's perception of the usefulness of feedback?*

The analysis for RQ2 did not show very interesting results. Only small differences were found between rubric characteristics according to student perception. The regression analysis showed that with more questions per rubric, the more students perceive the feedback as less useful rather than useful, although this finding was not statistically significant. For feedback graded *not useful at all* (FG1), there was a negative relationship with the number of questions, however, *not very useful* feedback (FG2) was positively correlated with the number of questions in a rubric. No statistically significant correlation results were found for other grades. The *text* questions had a negative relationship with both *not very useful* feedback (FG2) and *somewhat useful* graded feedback (FG3), and a positive relationship with the *extremely useful* feedback (FG5). The *boolean* questions were negatively correlated with *somewhat useful* feedback (FG3), and positively correlated *extremely useful* feedback (FG5). It is worth noticing that all correlations mentioned above are very weak. How to improve the analysis is addressed in the section on future work.

6. Discussion and conclusion

Our results contribute both to PA, especially its BE aspect, and the use of LA to analyse large PA datasets.

How students interpret and respond to feedback is determined both by the interaction between external conditions (e.g., social and material context, visualisation, and content of the feedback), and internal conditions of the students (e.g., motivation, beliefs, pre-understanding). Hence, different students in different feedback situations will interpret and use feedback in various ways. With this in mind, our findings do indicate some commonalities when it comes to student experience of feedback that is helpful and feedback that is perceived as unwarranted or incomprehensible.

Students who rated the feedback from their peers as useful tended to be more accepting of the feedback by *acknowledging their errors*, signalling that they intend to *revise their text*, and/or *praising the usefulness* of the feedback. On the other hand, students who rated the feedback as not useful tended to be more defensive in their response by expressing that they were *confused about its meaning, critical towards its form and focus,* and/or in *disagreement the claims.*

This shows that students who found the feedback more useful generally experience that the feedback made sense to them, appropriately addressed problems in their text (feedback) and were useful for improvement of their text or/and their competence as a writer (feedforward). Students, who on the other hand, rated the feedback as not useful, generally experienced the feedback as incomprehensible, unjust, or simply not useful.

Moreover, this finding poses an interesting question: Is the process of disagreeing with the feedback and trying to defend one's own work useful from a pedagogical perspective, even if the student does not perceive it as such? And if so, how would such a conclusion influence the teacher's development of PA rubrics and preparation of the students for the PA activity? These aspects require further investigation, and probably more fine-grained coding of the BE comments.

That student's sensemaking of the feedback correlates with their experience of its usefulness is known from previous studies and relates to the problem of the feedback gap (Jonsson, 2013; Nelson & Schunn, 2009). Students who experience feedback as less useful and responded with criticism and disagreement, however, might also be affected by their motivation and educational beliefs, as well as the actual comments from their peers. However, analysing the motivation or educational beliefs of students was outside of the scope of this study.

That students used the improvement category *kindness* to a lesser extent than the improvement suggestions *specificity, justification,* and *constructivity*, resonates well with studies that have found that the affective features of feedback has less impact on student improvement than cognitive features (Hattie & Timperley, 2007; Nelson & Schunn, 2009). This should not be interpreted as "feedback should not be kind", but rather that kindness itself does not provide students with information on how to improve.

The results of our exploratory study suggest that most feedback was not *specific* or *constructive* enough, even in cases when students graded the feedback as *extremely useful*, as indicated by the improvement suggestions that they have chosen. This suggests that students did not receive sufficient preparation for the PA activity, or they did not take the task seriously. Patchan et al. (2018) found that students who believed that their peer feedback was graded based on the perceived helpfulness by feedback receivers, gave better quality feedback. These two results show that there is an interdependency between the feedback giver and the feedback receiver. Thus, including BE as a part of the PA activity might help students develop their evaluative judgment of what is good quality feedback, in particular, if guided by the instructor. However, this would require implementing PA more than once in the course design in order to develop these skills.

The use of LA to give insight into student perceptions of PA moves us beyond what has been studied before through the use of questionnaires. The literature review by Ashenafi (2017) found that most PA activities are non-iterative, and not fully integrated into the whole educational program, which makes it hard to measure the impact of PA on long-term learning (Ashenafi, 2017). This could be addressed by LA. The automation of tasks, such as coding of the text data comes with new opportunities and challenges. It can speed up the data analysis process and enables an analysis of larger datasets, however, it might come at the cost of simplification of the content of the feedback. The regression analysis gave us general insights into the patterns in the data, while the correlation analysis revealed more details about student's behaviour depending on their feedback perception. Finally, ENA helped us develop a visual representation of the connections among different variables, and thus, revealed more detailed information about aspects of the data.

This current exploratory study shows that the insights from LA depend significantly on the availability of data and context information, and the quality of the available data. Without the student hand-ins, it is not possible to assess the quality of students' feedback, since we do not know to what the students are referring. Without the context data, the analysis is limited to basic measures, such as comment length and sentiment analysis, and limited our ability to "go back to the data" and close the interpretative cycle. Furthermore, the mixed quality of the feedback comments prevented a more sophisticated feedback coding. These challenges have to be taken into consideration while conducting LA research with big datasets.

Moreover, this study is an example of working with data collected by an educational platform that has not been developed to provide data specifically for LA, but rather to run smoothly. This is a common issue in LA, and we tried to mitigate it by matching variables and results from previous research. This study confirms a larger question about the meaningfulness of this kind of analysis of big data without the possibility to connect this data with external context information and when our data making-sense capabilities are restricted. The addition of contextual data could strengthen the results and help the data sense-making process (Mangaroska & Giannakos, 2018). On the other hand, the automatization of data coding is a clear advantage of LA methods over traditional research methods where hand-coding is the default, as this takes more time and resources.

6.1. Limitations

The current study has some important limitations. The first is lack of control variables due to weaknesses in the Peergrade dataset including 1) the absence of background information about the students, 2) the context of the PA activity, such as discipline (e.g., history or art), educational level (e.g., K-12 or college), pedagogical approach, or course structure and 3) assignment mark and/or the final course mark for the students. This indicates that the results might be caused by other variables that are absent from our dataset.

Second, the coding of the BE comments is quite broad as a result of the heterogeneous dataset (i.e., there is a wide variety of types of feedback characteristics (length; quality, full sentences, phrases, etc.), and lack of context (e.g., domain information such as are the students writing in their mother tongue, was the feedback assignment obligatory, etc.). Conducting the analysis on a more homogenous dataset, or a dataset with control variables, would allow for a more detailed analysis, such as examining the relationship between feedback characteristics and perceived feedback usefulness, if perceived feedback usefulness led to revision of the hand-in, or if student characteristics, such as previous experience with PA, influences their perception of feedback usefulness. Third, the dataset did not include the original work—the "hand-in" or item on which the feedback was being given. This lack of essential data makes it impossible to analyse if the feedback was used to improve their work.

6.2. Future work

We are embarking on a series of studies with higher education institutions in Norway that are focused on PA supported by the Peergrade tool. Future work will use the findings and experience from this exploratory analysis of the big dataset coming from a variety of institutions and disciplines when analysing a big dataset coming from a single course at a higher education institution (we currently have 2 such datasets from two different institutions and more information about the students, the PA activity, hand-ins, their final grades, etc.). This will allow the inclusion of more control variables about the students and the PA activity, as well as more opportunities for more specific coding of the text data (e.g., to include domain terms into coding). It will also add a new challenge in that the written language in the hand-ins and feedback comments is not English.

Regarding the analysis of rubrics and its relationship to student perception of feedback, it would be interesting with more fine-grained coding (i.e. boolean, text, or numerical) of the questions, e.g., using Bloom's Taxonomy (Bloom, Engelhart, Furst, Hill, & Krathwohl, 1956), for more in-depth analysis.

To address the risk of multicollinearity influencing the data results (Perez, 2017), other methods of data analysis will be applied in future research, such as Principal Component Analysis.

We are also interested in applying more text analysis that will allow us to analyse the content of the feedback text in a more sophisticated way. Previous educational research on peer feedback can give us insight into what we might look for, and other additional information we should collect. For example, in a study of peer feedback using 1,073 feedback segments from an online peer review system (SWoRD), Nelson and Schunn (2009) found that student's comprehension of the feedback was the only significant mediator for student implementation. That is, if the students understood the problem that was addressed, they were more likely to implement the suggestions that the feedback provided. In particular, they found that students were more likely to understand the feedback if it offered concrete solutions, a location of the problem(s), or if the feedback included a summary. Student perception of feedback, however, is not only affected by the feedback message itself, but also by their ability to interpret the feedback. So, while the clarity and form of the feedback might lead to confusion in some cases, this might also be caused by differences in the students' conceptual understanding. In a review, Jonsson (2013) found that a lack of understanding of academic terminology and assessment criteria was a common problem across many studies on student perception and use of feedback.

Finally, it will be possible to map the behaviour of a student during the entire PA process and identify patterns in the relationship between a student's own hand-in, the feedback they give to other students, and how they react the feedback that they receive.

7. Conclusion

Finally, we have shown that LA has the potential to show new insights into the BE aspect of PA, although there are many challenges as highlighted above. Furthermore, the research community needs to evolve theories about what various types of data reveal about learning, and therefore what to collect; the problem space is too large to simply gather all available data and attempt to mine it for patterns that might reveal generalizable insights. In addition, in collecting and analysing student data, issues of privacy, safety, and security pose new challenges not found in most scientific disciplines.

Credit author statement

Kamila Misiejuk: Conceptualisation, Visualisation, Writing – original draft, Writing – review & editing, Data curation, Formal analysis, Methodology, Software Barbara Wasson: Conceptualisation, Supervision, Writing – review & editing, Writing – original draft Kjetil Egelandsdal: Writing – original draft, Writing – review & editing.

Acknowledgment

The authors would like to thank Charlie Negri, NORCE Teknologi for collaboration on the data analysis and thank Peergrade CEO (David Wind) for access to the anonymised dataset. This research is funded by a PhD stipend from the University of Bergen, Norway.

References

Adewoyin, O., Araya, R., & Vassileva, J. (2016). Peer review in mentorship: Perception of the helpfulness of review and reciprocal ratings. *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 *Proceedings of the 9th international conference on educational data mining* (pp. 62–69).

Ashenafi, M. M. (2017). Peer-assessment in higher education–twenty-first century practices, challenges and the way forward. *Assessment & Evaluation in Higher Education, 42*(2), 226–251.

Ashton, S., & Davies, R. S. (2015). Using scaffolded rubrics to improve peer assessment in a MOOC writing course. *Distance Education, 36*(3), 312–334.

Misiejuk, K., & Wasson, B. (2017). *State of the Field report on Learning Analytics. SLATE Report 2017-2*. Bergen, Norway: Centre for the Science of Learning & Technology (SLATE).

Baleghizadeh, S., & Mortazavi, M. (2014). The impact of different types of journaling techniques on EFL learners' self-efficacy. *Profile - Issues in Teachers' Professional Development, 16*(1), 77–88.

Beckmann, N., Beckmann, J. F., & Elliott, J. G. (2009). Self-confidence and performance goal orientation interactively predict performance in a reasoning test with accuracy feedback. *Learning and Individual Differences, 19*(2), 277–282.

Birney, D. P., Beckmann, J. F., Beckmann, N., & Double, K. S. (2017). Beyond the intellect: Complexity and learning trajectories in Raven's Progressive Matrices depend on self-regulatory processes and conative dispositions. *Intelligence, 61*, 63–77. https://doi.org/10.1016/j.intell.2017.01.005

Black, P., & Wiliam, D. (1998). Inside the black box: Raising standards through classroom Assessment. *Phi Delta Kappan, 80*(2), 139–144.

Black, P., & Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability, 21*(1), 5–31.

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

Bloxham, S., & Campbell, L. (2010). Generating dialogue in assessment feedback: Exploring the use of interactive cover sheets. *Assessment & Evaluation in Higher Education, 35*(3), 291–300. https://doi.org/10.1080/02602931003650045

Brant, R. (1990). Assessing proportionality in the proportional odds model for ordinal logistic regression. *Biometrics, 46*, 1171–1178.

Buchanan, E., Gesher, A., & Hammer, P. (2015). Privacy, security, and ethics. In C. Dede (Ed.), *Data-intensive research in education: Current work and next steps* (pp. 89–98). Washington, DC: Computing Research Association.

Carless, D., Salter, D., Yang, M., & Lam, J. (2010). Developing sustainable feedback practices. *Studies in Higher Education, 36*(4), 395–407. https://doi.org/10.1080/03075071003642449

Cheng, K. H., Liang, J. C., & Tsai, C. C. (2015). Examining the role of feedback messages in undergraduate students' writing performance during an online peer assessment activity. *The Internet and Higher Education, 25*, 78–84.

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., & Kwon, K. (2007). Learning writing by reviewing. In *Proceedings of the 8th international conference on computer-supported collaborative learning* (pp. 141–143).

Cook, A. (2019). *Using interactive learning activities to address challenges of peer feedback systems*. Doctoral dissertation).

Dawson, P., Henderson, M., Mahoney, P., Phillips, M., Ryan, T., Boud, D., et al. (2019). What makes for effective feedback: Staff and student perspectives. *Assessment & Evaluation in Higher Education, 44*(1), 25–36.

De Wever, B., Van Keer, H., Schellens, T., & Valcke, M. (2011). Assessing collaboration in a wiki: The reliability of university students' peer assessment. *The Internet and Higher Education, 14*(4), 201–206.

Divjak, B., & Maretić, M. (2017). Learning analytics for peer-assessment:(dis) advantages, reliability and implementation. *Journal of Information and Organizational Sciences, 41*(1), 21–34.

Double, K., McGrane, J., & Hopfenbeck, T. (2018). *The impact of peer assessment on academic performance: A meta-analysis of (quasi) experimental peer assessment studies*.

Er, E., Dimitriadis, Y., & Gašević, D. (2019). Synergy: An online platform for dialogic peer feedback at scale. In K. Lund, G. P. Niccolai, E. Lavoué, C. Hmelo-Silver, G. Gweon, & M. Baker (Eds.), *13th international conference on computer supported collaborative learning (CSCL) 2019: Vol. 2. A wide lens: Combining embodied, enactive, extended, and embedded learning in collaborative settings* (pp. 1005–1008). Lyon, France: International Society of the Learning Sciences.

Ertmer, P. A., Richardson, J. C., Lehman, J. D., Newby, T. J., Cheng, X., Mong, C., et al. (2010). Peer feedback in a large undergraduate blended course: Perceptions of value and learning. *Journal of Educational Computing Research, 43*(1), 67–88.

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

Fox, J., & Weisberg, S. (2011). *An R companion to applied regression* (2nd ed.). Thousand Oakes, CA: Sage Publishing.

Gielen, M., & De Wever, B. (2015). Structuring the peer assessment process: A multilevel approach for the impact on product improvement and peer feedback quality. *Journal of Computer Assisted Learning, 31*(5), 435–449.

Graner, M. H. (1987). Revision workshops: An alternative to peer editing groups. *English Journal, 76*(3), 40–45. https://doi.org/10.2307/818540

Hafner, J., & Hafner, P. (2003). Quantitative analysis of the rubric as an assessment tool: An empirical study of student peer-group rating. *International Journal of Science Education, 25*(12), 1509–1528.

Hattie, J., & Gan, M. (2011). Instruction based on feedback. In R. E. Mayer, & P. A. Alexander (Eds.), *Handbook of research on learning and instruction* (pp. 249–271). New York, NY: Routledge.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*(1), 81–112.

Healy, M. J. (1995). Statistics from the Inside. 16. Multiple regression (2). *Archives of Disease in Childhood, 73*(3), 270–274.

Higgins, R., Hartley, P., & Skelton, A. (2001). Getting the Message Across: The problem of communicating assessment feedback. *Teaching in Higher Education, 6*(2), 269–274. https://doi.org/10.1080/13562510120045230

Hrepic, Z., Zollman, D. A., & Rebello, N. S. (2007). Comparing students' and experts' understanding of the content of a lecture. *Journal of Science Education and Technology, 16*(3), 213–224. https://doi.org/10.1007/s10956-007-9048-4

Huang, B., Hwang, G. J., Hew, K. F., & Warning, P. (2019). Effects of gamification on students' online interactive patterns and peer-feedback. *Distance Education, 40*(3), 350–379.

Hutto, C. J., & Gilbert, E. E. (2014). Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Eighth international conference on weblogs and social media* (ICWSM-14).

Jacobs, G. M., Curtis, A., Braine, G., & Huang, S.-Y. (1998). Feedback on student writing: Taking the middle path. *Journal of Second Language Writing, 7*(3), 307–317. https://doi.org/10.1016/S1060-3743(98)90019-4

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

Kaufman, J. H., & Schunn, C. D. (2011). Students' perceptions about peer assessment for writing: Their origin and impact on revision work. *Instructional Science, 39*(3), 387–406. https://doi.org/10.1007/s11251-010-9133-6

Kitchin, R. (2013). Big data and human geography: Opportunities, challenges and risks. *Dialogues in Human Geography, 3*(3), 262–267. https://doi.org/10.1177/2043820613513388

Kitchin, R. (2014). *The data revolution: Big data, open data, data infrastructures and their consequences*. London: SAGE. https://doi.org/10.4135/9781473909472

Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin, 119*(2), 254–284. https://doi.org/10.1037/0033-2909.119.2.254

Krumm, A., Means, B., & Bienkowski, M. (2018). Data used in educational data-intensive research. *Learning analytics goes to school*. New York: Routledge.

Liddell, T. M., & Kruschke, J. K. (2018). Analyzing ordinal data with metric models: What could possibly go wrong? *Journal of Experimental Social Psychology, 79*, 328–348.

Liu, C.-C., Lu, K.-H., Wu, L. Y., & Tsai, C.-C. (2016). The impact of peer review on creative self-efficacy and learning performance in Web 2.0 learning activities. *Journal of Educational Technology & Society, 19*(2).

Long, P., & Siemens, G. (2011). Penetrating the fog: Analytics in learning and education. *Educause Review, 46*(5), 31–40.

Lüdecke, D. (2019). *_sjPlot: Data visualization for statistics in social science_*. https://doi.org/10.5281/zenodo.1308157

Lundstrom, K., & Baker Smemoe, W. (2009). To give is better than to receive: The benefits of peer review to the reviewer's own writing. *Journal of Second Language Writing, 18*, 30–43. https://doi.org/10.1016/j.jslw.2008.06.002

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

Mangaroska, K., & Giannakos, M. N. (2018). *Learning analytics for learning design: A systematic literature review of analytics-driven design to enhance learning*. IEEE Transactions on Learning Technologies.

Marquart, L. C., Swiecki, Z., Collier, W., Eagan, B., Woodward, R., & Shaffer, D. W. (2019). *rENA: Epistemic network analysis*.

Mukaka, M. M. (2012). A guide to appropriate use of correlation coefficient in medical research. *Malawi Medical Journal, 24*(3), 69–71.

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

Nelson, G. L., & Carson, J. G. (1998). ESL students' perceptions of effectiveness in peer response groups. *Journal of Second Language Writing, 7*(2), 113–131. https://doi.org/10.1016/S1060-3743(98)90010-8

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. V., & Litman, D. J. (2014). Improving peer feedback prediction: The sentence level is right. In *Proceedings 9th workshop on innovative use of NLP for building educational applications* (pp. 99–108).

Nicol, D. (2009). Assessment for learner self-regulation: Enhancing achievement in the first year using learning technologies. *Assessment & Evaluation in Higher Education, 34* (3), 335–352. https://doi.org/10.1080/02602930802255139

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

Nilson, L. B. (2003). Improving student peer feedback. *College Teaching, 51*(1), 34–38.

Pardo, A. (2018). A feedback model for data-rich learning experiences. *Assessment & Evaluation in Higher Education, 43*(3), 428–438.

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.

Perez, L. V. (2017). *Principal component analysis to address multicollinearity*. Retrieved from https://www.whitman.edu/Documents/Academics/Mathematics/2017/Perez.pdf.

Piech, C., Huang, J., Chen, Z., Do, C., Ng, A., & Koller, D. (2013). *Tuned models of peer assessment in MOOCs*. arXiv preprint arXiv:1307.2579.

Prinsloo, P., & Slade, S. (2017). Big data, higher education and learning analytics: Beyond justice, towards an ethics of care. In B. K. Daniel (Ed.), *Big data and learning analytics in higher education* (pp. 109–124). Cham: Springer. https://doi.org/10.1007/978-3-319-06520-5_8.

Roschelle, J., & Krumm, A. (2016). Infrastructures for improving learning in information-rich classrooms. In P. Reimann, S. Bull, M. Kickmeier-Rust, R. Vatrapu, & B. Wasson (Eds.), *Measuring and visualizing learning in the information-rich classroom* (pp. 19–26). New York: Routledge.

Ryan, T., Gašević, D., & Henderson, M. (2019). Identifying the impact of feedback over time and at scale: Opportunities for learning analytics. *The impact of feedback in higher education* (pp. 207–223). Cham: Palgrave Macmillan.

Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional Science, 18*(2), 119–144. https://doi.org/10.2307/23369143

Sadler, D. R. (2009). Indeterminacy in the use of preset criteria for assessment and grading. *Assessment & Evaluation in Higher Education, 34*(2), 159–179. https://doi.org/10.1080/02602930801956059

Sadler, D. R. (2010). Beyond feedback: Developing student capability in complex appraisal. *Assessment & Evaluation in Higher Education, 35*(5), 535–550. https://doi.org/10.1080/02602930903541015

Shaffer, D. W. (2017). *Quantitative ethnography*. Madison, WI: Cathcart Press.

Shaffer, D., & Ruis, A. (2017). Epistemic network analysis: A worked example of theory-based learning analytics. In C. Lang, G. Siemens, A. Wise, & D. Gašević (Eds.), *Handbook of learning analytics* (pp. 175–187). Society for Learning Analytics and Research.

Shibani, A. (2017). Combining automated and peer feedback for effective learning design in writing practices. In *25th international conference on computers in education: Technology and innovation, doctoral student consortia proceedings*.

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

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*, 249–276.

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

Tsui, A. B. M., & Ng, M. (2000). Do secondary L2 writers benefit from peer comments? *Journal of Second Language Writing, 9*(2), 147–170. https://doi.org/10.1016/S1060-3743(00)00022-9

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.

Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with S*. New York: Springer.

Wahid, U., Chatti, M. A., & Schroeder, U. (2016). Improving peer assessment by using learning analytics. In R. Zender (Ed.), *Proceedings of DeLFI workshops 2016* (pp. 52–54).

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. New York: Springer.

Wiliam, D. (2011). What is assessment for learning? *Studies In Educational Evaluation, 37*, 3–14. https://doi.org/10.1016/j.stueduc.2011.03.001

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. (2020). From feedback to revisions: Effects of feedback features and perceptions. *Contemporary Educational Psychology, 60*, 1–17.

Xiong, W., Litmaan, D., & Schunn, C. (2012). Natural language processing techniques for researching and improving peer feedback. *Journal of Writing Research, 4*(2), 155–176.

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

Zhang, S. (1995). Reexamining the affective advantage of peer feedback in the ESL writing class. *Journal of Second Language Writing, 4*(3), 209–222. https://doi.org/10.1016/1060-3743(95)90010-1

Zingle, G., Radhakrishnan, B., Xiao, Y., Gehringer, E., Xiao, Z., Pramudianto, F., et al. (2019). Detecting suggestions in peer assessments. In *Proceedings of the 12th international conference on educational data mining* (pp. 474–479).