

Using learning analytics to understand peer assessment:

The importance of context data

Kamila Misiejuk

Thesis for the degree of Philosophiae Doctor (PhD)
University of Bergen, Norway
2023

UNIVERSITY OF BERGEN



Using learning analytics to understand peer assessment:

The importance of context data

Kamila Misiejuk



Thesis for the degree of Philosophiae Doctor (PhD)
at the University of Bergen

Date of defense: 10.02.2023

© Copyright Kamila Misiejuk

The material in this publication is covered by the provisions of the Copyright Act.

Year: 2023

Title: Using learning analytics to understand peer assessment:

Name: Kamila Misiejuk

Print: Skipnes Kommunikasjon / University of Bergen

Scientific environment

The research presented in this thesis was conducted in the scientific milieu at The Centre for the Science of Learning & Technology (SLATE), the national centre for learning analytics. SLATE is funded by the Ministry of Education and the University of Bergen. SLATE is an interdisciplinary centre of researchers with backgrounds in information science, education, pedagogy, sociology, informatics, psychology, music, fine arts and law.

During my research, I was employed by the University of Bergen at the Department of Information Science and Media Studies.

Acknowledgements

This research would never have been possible without the people who supported me during my PhD studies.

First of all, my deep gratitude goes to my supervisor, Professor Barbara Wasson. You are the reason why I embarked on the academic path by sparking my interest in the topic of technology-enhanced learning in 2015. Thank you for believing in me, for your guidance, and for giving me the freedom to explore during my PhD journey.

My gratitude is extended to my co-supervisor, Professor Ingeborg Krange. Thank you for your advice and genuine concern during my PhD journey.

I would like to thank Dr. David Kofoed Wind, CEO of Peergrade and Eduflow, for providing research data and technical support to enable this research, and Dr. Charlie Negri for helping me with data analysis.

Thanks to my co-authors in the two papers in this research: Dr. Kjetil Egelandstal, who contributed to the first paper, and Dr. Jarle Bastesen, who offered an opportunity to analyze the data from his class and collaborated on the research and writing of the fourth paper.

My appreciation extends to all current and former Slate colleagues who made this journey more fun.

Rosaline Barendregt, thank you not only for being the most amazing office mate ever and a great friend but also for supporting me through all ups and downs of my PhD journey.

Dana Kvietkute, thank you for all the interesting chats during coffee breaks. Your hard work was always an inspiration to me!

Dr. Ingunn Johanne Ness, thank you for our morning conversations in the Slate jazz lounge and for helping me with the philosophical and epistemological parts of this dissertation.

Christina Gkini, thank you for your in-depth feedback on this dissertation and for our chats about the academic life.

I would also like to thank those who took the time to read this dissertation and comment on it: André Rabello Mestre, Qinyi Liu, Fredrik Breien, Geerte Koster, Jeanette Samuelsen, Yağmur Çisem Yılmaz, and Cecilie Hansen.

Special thanks to Jeanette Samuelsen and Dr. Ingunn Johanne Ness for helping with the Norwegian translation of the abstract.

Finally, I would like to thank Dr. Mohammad Khalil for advocating for the amazing coffee machine that was the Slate life support for all these years.

In addition, I need to thank the Quantitative Ethnography research community, which was a big part of my academic development during my PhD journey. In particular, I would like to thank my colleagues in the ISQE resources committee, Dr. Brendan Eagan, Dr. Szilvia Zörgő, Yuanru Tan, and Cesar Hinojosa, for being supportive and giving me space to finish writing this dissertation.

I would like to thank collaborators in different endeavours during this PhD journey: Professor David W. Shaffer, Professor Morten Misfeldt, Professor Sølvi Lillejord, Professor Rebecca Ferguson, Dr. Golnaz Arastoopour Irgens, Dr. Robert Gray, Dr. Danielle Espino, Dr. Erik Ruud, Dr. Rogers Kaliisa, Nina Morlandstø, Jennifer J. Scianna, Clare Porter, and Karl Vachuska.

I am thankful to my family for all your love and encouragement. Special thanks to all my wonderful friends who were there for me on this journey: Bogusia, Natalka, Ewa, Laura, Jay, and Marijn.

I would love to thank my daughter, Tamara. You were my main motivation to stay on track and my personal alarm clock in the mornings. Thank you for not getting sick too often and sleeping through the night (most of the nights). Bardzo Cię kocham, mój mały rozbójniku!

Last but not least, I would like to thank my husband, Fakhr, who endured this journey with me. Thank you for being patient and taking long vacations together with Tamara. Looking forward to new adventures ahead of us. بحبك كثير حبيبي.

Bergen, October 2022

Kamila Misiejuk

Abstract

Learning analytics is a field concerned with collecting, analyzing, and sense-making of educational data to optimize and better understand student learning. The research presented in this dissertation addresses the challenges and opportunities of using learning analytics in a collaborative activity where students grade and/or evaluate each other's work, namely peer assessment. The research was guided by the following research question: How can we use learning analytics to gain new insights into peer assessment?

This article-based dissertation comprises an extended abstract that offers a meta-perspective on two empirical studies and two scoping reviews. The empirical studies adopt a mixed-methods approach and follow a structured process of data collection, preparation, exploration, and analysis to conduct learning analytics. The analysis in the studies was carried out on a dataset without context data and one with a dataset with context data, respectively. Context data is a central concept in this research and refers to non-platform data collected to provide context and meaning to the platform data.

Study 1 focused on the analysis of backward evaluation (i.e., students' perception of feedback) in a context-free dataset collected over two years of operations of an online peer assessment platform. The main finding of this study was that feedback was perceived as useful when students agreed with it or were positive toward it. In addition, the relevance of the feedback was more important than its kindness or justification. Study 1 also highlighted the limitations of working with data generated by a third-party platform without context data—this limited the impact of learning analytics and the analysis possibilities despite the dataset size. As a result, Study 2 analyzed a context-enriched dataset from a college course peer assessment activity that used the same online peer assessment platform and aimed to identify features that influence feedback implementation and essay revision. The findings showed that Boolean feedback, writing a backward evaluation comment, and mitigating praise in the feedback comments had the most significant positive influence on feedback

implementation. The size of the feedback provider group, the draft grade, and previous peer assessment experience were the most important features that negatively influenced feedback implementation and essay revision. Close collaboration with the course instructor aided in understanding and interpreting the data. In addition, the results informed the design of the peer assessment activity in upcoming course offerings.

Scoping review 1 is the first to focus on backward evaluation in online peer assessment platforms and was inspired by the main focus of Study 1. This scoping review gave an overview of the terminology used in the literature for backward evaluation, organized the knowledge about how backward evaluation data is used in research and presented findings about peer assessment activity (related to backward evaluation). Scoping Review 2 is the first review to summarize current research on using learning analytics to either 1) analyze peer assessment data or 2) improve peer assessment activity. This scoping review focused on peer assessment challenges addressed using learning analytics and what new insights were discovered.

Overall, the papers included in this dissertation make theoretical, methodological, and empirical contributions to the fields of learning analytics and peer assessment. The two scoping reviews represent the theoretical contributions of this research. The methodological contributions include the method transparency in empirical studies that tackled two different datasets—context-free and context-rich—from the same online peer assessment platform. Furthermore, the implications of the lack of context data for learning analytics are explored in-depth. Finally, empirically, this research provides new insights into several aspects of peer assessment: rubric design, feedback implementation and essay revision, and backward evaluation.

Sammendrag

Læringsanalyse er et felt som er opptatt av innsamling, analyse og meningsdannelse av utdanningsdata for å optimalisere og bedre forstå studenters læring. Forskningen som presenteres i denne avhandlingen tar for seg utfordringene og mulighetene ved å bruke læringsanalyse i en samarbeidsaktivitet hvor studentene vurderer og/eller vurderer hverandres arbeid, altså medstudentvurdering (peer assessment). Forskningen ble styrt av følgende forskningsspørsmål: Hvordan kan vi bruke læringsanalyse for å få ny innsikt i medstudentvurdering?

Denne artikkelbaserte avhandlingen består av en kappe (extended abstract) som gir et metaperspektiv på to empiriske studier og to "scoping reviews". De empiriske studiene bruker en tilnærming med "mixed methods" og følger en strukturert prosess med datainnsamling, forberedelse, utforskning og analyse for å utføre læringsanalyse. Analysen i studiene ble utført på henholdsvis ett datasett uten kontekstdata og ett datasett med kontekstdata. Kontekstdata er et sentralt begrep i denne forskningen og refererer til ikke-plattformdata (non-platform data) som samles inn for å gi kontekst og mening til plattformdataene. Studie 1 fokuserte på analysen av baklengsvurdering (backward evaluation) (dvs. elevenes oppfatning av tilbakemelding) i et kontekstfritt datasett samlet inn over to års drift av en nettbasert medstudentvurderingsplattform. Hovedfunnet i denne studien var at tilbakemeldinger ble oppfattet som nyttige når studentene var enige i dem eller var positive til dem. I tillegg var relevansen av tilbakemeldingen viktigere enn dens vennlighet eller begrunnelse.

Studie 1 fremhevet også begrensningene ved å jobbe med data generert av en tredjepartsplattform uten kontekstdata – dette begrenset påvirkningskraften (impact) av læringsanalyse og analysemulighetene til tross for datasettstørrelsen. Som et resultat analyserte studie 2 et kontekstberiket datasett fra én medstudentvurderingsaktivitet på høyskolekurs som brukte den samme nettbaserte medstudentvurderingsplattformen, og studien hadde som mål å identifisere funksjoner som påvirker implementering av tilbakemeldinger og omskriving av essay. Funnene viste at boolske tilbakemeldinger,

skrivning av en baklengsvurderingskommentarer og formildende ros i tilbakemeldingskommentarene, hadde den mest signifikante positive innflytelsen på implementeringen av tilbakemeldingene. Størrelsen på gruppen som gir tilbakemeldinger, utkastet til vurderingen og tidligere erfaring med medstudentvurdering var de viktigste egenskapene som påvirket negativt implementering av tilbakemeldinger og omskriving av essay. Tett samarbeid med kursholderen hjalp til med å forstå og tolke dataene. I tillegg informerte resultatene utformingen av medstudentvurderingsaktiviteten i kommende kurstilbud.

Scoping review 1 er den første som fokuserer på baklengsvurdering i nettbaserte medstudentvurderingsplattformer og var inspirert av hovedfokuset i studie 1. Denne scoping reviewen ga en oversikt over terminologien som ble brukt i litteraturen for baklengsvurdering, organiserte kunnskapen om hvordan baklengsvurderingsdata brukes i forskning, samt presenterte funn om medstudentvurderingsaktivitet (relatert til baklengsvurdering). Scoping review 2 er den første som oppsummerer gjeldende forskning om bruk av læringsanalyse til enten 1) å analysere medstudentvurderingsdata eller 2) å forbedre medstudentvurderingsaktiviteten. Denne scoping reviewen fokuserte på utfordringer med medstudentvurdering som ble adressert ved hjelp av læringsanalyse og hvilken ny innsikt som ble oppdaget.

Samlet sett gir artiklene som er inkludert i denne avhandlingen teoretiske, metodiske og empiriske bidrag til feltene læringsanalyse og medstudentvurdering. De to scoping-reviewene representerer de teoretiske bidragene til denne forskningen. De metodiske bidragene inkluderer metodisk transparens i empiriske studier som omhandlet to forskjellige datasett – kontekstfrie og kontekstrike – fra samme nettbaserte medstudentvurderingsplattform. Videre blir implikasjonene av mangelen på kontekstdata for læringsanalyse utforsket i dybden. Tilslutt, gir denne forskningen ny empirisk innsikt i flere aspekter ved medstudentvurdering: rubrikkdesign, implementering av tilbakemeldinger og omskriving av essay, samt baklengsvurdering.

List of publications

Paper 1:

Misiejuk, K., Wasson B. & Egelandstal K. (2021). Using learning analytics to understand student perceptions of peer feedback. *Computers in Human Behavior*, 117. DOI: 10.1016/j.chb.2020.106658.

Paper 2:

Misiejuk, K. & Wasson, B. (2021). Backward evaluation in peer assessment: A scoping review. *Computers & Education*, 175. DOI: 10.1016/j.compedu.2021.104319.

Paper 3:

Misiejuk, K. & Wasson, B. (in press). Learning analytics for peer assessment: A scoping review. In O. Noroozi & B. De Wever (Eds.) *The Power of Peer Learning*. Springer.

Paper 4:

Misiejuk, K., Bastesen, J., Wasson, B. & Krange, I. (submitted). Educational data for learning analytics: Increasing insights into peer assessment with context data. *Assessment & Evaluation in Higher Education*.

Contents

Scientific environment	i
Acknowledgements	iii
Abstract	v
Sammendrag	vii
List of publications	ix
List of figures	xiii
List of tables	xiii
List of abbreviations	xv
Part I The extended abstract	1
1 Introduction	3
1.1 Motivation, aims, and research questions	4
1.2 Research design	5
1.3 Dissertation structure	8
2 Background	9
2.1 Learning analytics	9
2.1.1 Educational data for learning analytics	10
2.1.2 Constructivist approaches to learning analytics	12
2.2 Peer assessment	13
2.2.1 Peer assessment rubrics	15
2.2.2 Feedback implementation	16
2.2.3 Backward evaluation	18
3 Methods	21
3.1 Research paradigm	21
3.2 Scoping reviews	21

3.3	Empirical studies	25
3.3.1	Research methodology	25
3.3.2	Data collection	26
3.3.3	Data preprocessing	30
3.3.4	Data analysis	34
3.3.5	Data interpretation	36
3.3.6	Research ethics	36
4	Results	39
5	Conclusion	47
5.1	Theoretical contributions	47
5.2	Methodological contributions	48
5.3	Empirical contributions	48
5.4	Evaluation of the research approach	51
5.5	Limitations	53
5.6	Conclusions and future work	53
	References	55
	Appendices	71
A	Co-authorship declarations	73
B	Search strings	77
C	Variables and data pre-processing used in empirical studies	79
D	Additional data charting for scoping reviews	85
E	Study 2 approval by the Norwegian Centre for Research Data (NSD)	91
Part II	The papers	93
Paper 1		95
Paper 2		109
Paper 3		123
Paper 4		155

List of figures

Figure 1	Screenshots of the Peergrade interface	6
Figure 2	Overview of the papers included in this research	7
Figure 3	Feedback model	17
Figure 4	Learning Analytics - Principles and Constraints Framework . . .	26
Figure 5	Model of the peer assessment process	28

List of tables

Table 1	Overview of the scoping reviews	23
Table 2	Overview of the empirical studies	27
Table 3	Overview of methods used in the empirical studies	31
Table 4	Short overview of the variables in regression analyses	40
Table 5	Variables and data preprocessing in Study 1	79
Table 6	Variables and data preprocessing in Study 2	82
Table 7	Overview of papers included in the Scoping review 1	85
Table 8	Overview of selected papers included in the Scoping review 2 . .	88

List of abbreviations

BE	Backward Evaluation
ENA	Epistemic Network Analysis
LA	Learning Analytics
LA-PCF	Learning Analytics - Principles and Constraints Framework
NLP	Natural Language Processing
PA	Peer Assessment

Part I

The extended abstract

1. Introduction

Peer assessment (PA) refers to students grading or giving/receiving feedback on each other's work (Topping, 1998). The digitalization of education has resulted in an increase of digital spaces to conduct PA. Online PA platforms automate time-consuming tasks, such as student allocation, which can be especially advantageous in large classrooms (Formanek, Wenger, Buxner, Impey, & Sonam, 2017; Liu & Carless, 2006). Further, they offer additional features such as anonymization, close monitoring of student grading and feedback, or other innovative additions to the activity such as backward evaluation (i.e., students evaluating feedback that they received), gamification elements, or prompts (Gamage, Staubitz, & Whiting, 2021; Babik, Gehringer, Kidd, Pramudianto, & Tinapple, 2016).

Technological developments have opened new opportunities to collect student digital traces. The analysis and sense-making of educational data are foundational in the field of *learning analytics* (LA) (Siemens, 2013; Wilson & Scalise, 2016). Early LA research identified the potential of using LA techniques for constructionist learning activities, such as PA (Berland, Baker, & Blikstein, 2014). Some possible implementations included automatically classifying feedback given by students based on chosen criteria (e.g., a reviewer's reputation), using predictive analytics to indicate peer feedback accuracy, or developing visualizations indicating which peer feedback needs instructor's involvement (Wahid, Chatti, & Schroeder, 2016). Although LA promises new insights from educational data beyond counting clicks (Fincham et al., 2019), constructivist learning activities, such as PA, pose a challenge for LA research as they focus on the learning process rather than the outcome (Berland et al., 2014; Buckingham Shum & Ferguson, 2012). Learning analytics was identified to support feedback processes such as supporting students' reflective processes or increasing their feedback satisfaction (Ryan, Gašević, & Henderson, 2019). However, bridging the gap between raw data and learning constructs (Wise, Knight, & Shum, 2021) and developing actionable insights from LA to improve learning (Clow, 2012) is

challenging. In addition, log data from learning platforms may not include information about the context of a learning activity or access to context information may be restricted due to privacy issues. Therefore, in this research *context data* refers to non-platform data collected to provide context and meaning to the *platform data*.

This research comprises two scoping reviews and two empirical studies and contributes to the fields of LA and PA in several ways. First, it provides a mapping of the field of application of LA in PA research. Second, it presents a subfield of PA, backward evaluation, a promising new direction for LA research. Third, it describes in detail the different methods of working with PA datasets, with or without additional context data using LA. Fourth, it discusses the importance of context data for LA research. Fifth, it expands knowledge of student behaviour during a PA activity, including several aspects of student reaction to feedback and the factors contributing to peer feedback implementation and essay revision. Sixth, it examines the effect of rubric learning design on backward evaluation, feedback implementation, and essay revision.

1.1 Motivation, aims, and research questions

My interest in LA began during the work on my master's thesis, in which I mapped the field of educational data sciences, using papers published at three conferences: the International Conference on Educational Data Mining, the International Learning Analytics and Knowledge Conference, and the ACM Conference on Learning at Scale (Misiejuk, 2017). This work was followed by the "State-of-the-Field Report on Learning Analytics and Knowledge" (Misiejuk & Wasson, 2017) in which I analyzed relevant articles published from 2011–2015 to map the main research themes within the field of LA, the data and methods used, and the characteristics of the LA studies. This early work sparked my interest in educational data and the kinds of insights that we can gain from its analysis.

The main research question of this research is as follows:

How can we use learning analytics to gain new insights into peer assessment?

The following sub-questions were developed to guide this research, and are addressed in the papers enclosed in this dissertation:

Q1: How does the design of a feedback rubric influence backward evaluation, feedback implementation, and essay revision in a peer assessment activity? (Study 1, Scoping review 2)

Q2: What influences essay revision and peer feedback implementation? (Study 2)

Q3: What can we learn about students' perceptions of peer feedback from backward evaluation data? (Study 1, Study 2, Scoping review 2)

Q4: How does context data change the learning analytics analysis of peer assessment data? (Study 1, Study 2, Scoping review 1, Scoping review 2)

The sub-questions capture the different aspects of the main research question that were explored in this research. Four aspects of a PA activity are in focus in this research: rubric design, backward evaluation, essay revision, and feedback implementation. The selection of interest areas was dictated by both the available data and previous research, as well as the potential to use the insights to improve PA in future work. Exploring the effect of rubric design on learners can help instructors develop rubrics that facilitate better quality feedback. Innovative ways to integrate backward evaluation on online platforms have the potential for LA research. It is an important approach to ensure that students engage with the feedback that they receive, and develop their own feedback skills, and hence, improve the PA activity using LA. Gaining more insights into the factors contributing to students revising their own work and implementing peer feedback can help instructors adjust their recommendations about good quality feedback. Finally, this research explored the overarching theme of ways to conduct LA research in the application area of PA, with a special emphasis on context data.

1.2 Research design

Educational data availability and accessibility shaped the LA analysis in this study. Typically, PhD research starts with a literature review on a topic to discover the research gaps, followed by data collection. This research followed a different path, as an opportunity to analyze a big dataset from an online PA platform, Peergrade, presented itself at the beginning of the PhD research (see Figure 1). Peergrade (peergrade.io, now *Eduflow*, eduflow.com) can be integrated with the most popular learning management systems and is used internationally. In addition, supporting students in giving and receiving feedback, the platform enables backward evaluation and calculates peer grade agreement among the students making it possible for teacher intervention in the case of high discrepancies. Knowing that access to data from online platforms is usually limited for research purposes, I agreed to collaborate with this particular platform

and analyze all data collected from 2015 to 2017 (Kitchin & Lauriault, 2015). This collaboration opened another opportunity later in the project; close cooperation with an instructor from a Norwegian university college who used Peergrade in his teaching. This enabled wider data collection that included not only the Peergrade data, but also context data.

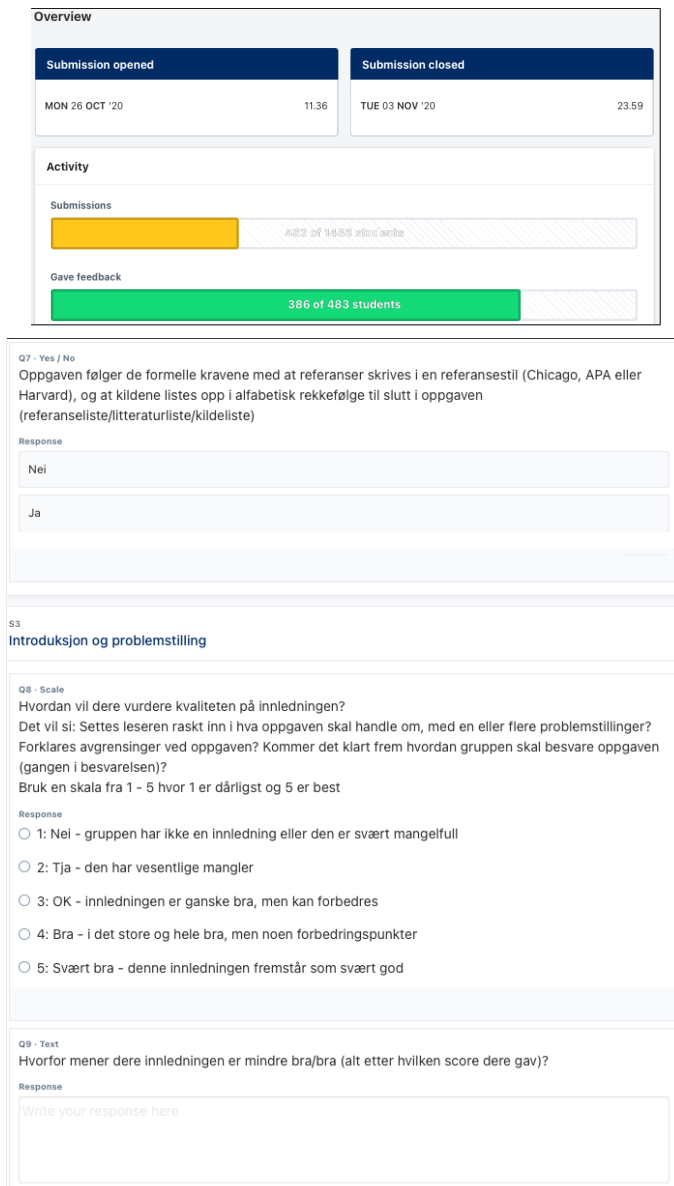


Figure 1: Screenshots of the Peergrade interface

To answer the main research question, two empirical studies analyzing respectively a context-free large dataset and a context-rich dataset from a commercial PA platform, Peergrade, and two scoping reviews were conducted (see Figure 2).

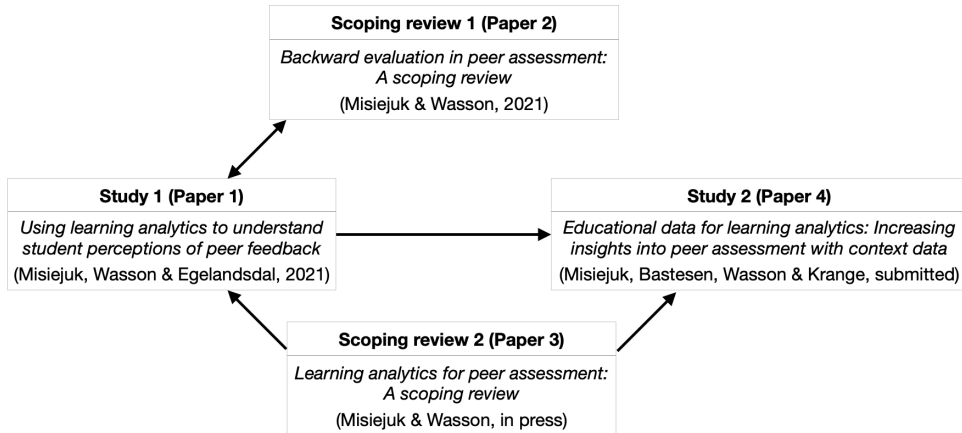


Figure 2: Overview of the papers included in this research

Scoping review 2 (Paper 3, Misiejuk & Wasson, in press) provided the background to this research by mapping previous studies using LA to understand and optimize PA. Study 1 (Paper 1, Misiejuk, Wasson, & Egelanddsdal, 2021) focused on backward evaluation due to the limitations of the context-free dataset provided by Peergrade. The goal of this study was to explore the power of LA used on a typical dataset from a platform not specifically designed for LA. Scoping review 1 (Paper 2, Misiejuk & Wasson, 2021) helped contextualize the findings from Study 1 by mapping the previous backward evaluation research and contributed to an exploration of potential future uses of LA on backward evaluation data. Study 2 (Paper 4, Misiejuk, Bastesen, Wasson, & Krange, submitted) was set out to mitigate the challenges encountered in Study 1 by enriching the Peergrade platform data with context data and focused on feedback implementation and essay revision.

This research applied a type of triangulation mixed-methods design, specifically a data transformation model (Tashakkori & Teddlie, 1998; Creswell & Plano Clark, 2011). Both empirical studies used the *Learning Analytics - Principles and Constraints Framework* (LA-PCF), which presents a structured way of conducting LA research (Khalil & Ebner, 2015). The complementing literature reviews followed the scoping review approach by Levac, Colquhoun, and O'Brien (2010).

1.3 Dissertation structure

The doctoral dissertation is divided into the Extended Abstract (Part I) and the Papers (Part II). The purpose of the extended abstract is to provide more information about the process and decision-making in this research. Furthermore, it links the studies with the main research question and provides a broader view of the findings and contributions to both LA and PA fields.

Part I consists of five chapters, including the introduction. Chapter 2 discusses previous research on the topics relevant to this research and identifies the research gaps. It starts with a short description of the LA field. Next, the challenges of analyzing educational data and the issue of context data are presented, followed by a discussion of constructivist approaches to LA. The chapter ends with a short introduction to the PA research, in particular rubric design, feedback implementation, and backward evaluation. Chapter 3 introduces the methods used in this research. First, the research paradigm is described. This is followed by details of the methods used in the scoping reviews and the empirical studies. Chapter 4 answers the main research questions using the sub-questions posed in the introduction by presenting the results from two empirical studies and two literature reviews conducted in this research. Chapter 5 discusses the theoretical, methodological, and empirical contributions of this research. Furthermore, it reflects on the research limitations and presents the evaluation of the research approach. The chapter ends with future research and conclusions.

Part II consists of the four papers on which the research is based (Paper 1, Misiejuk, Wasson, & Egelandstad, 2021; Paper 2, Misiejuk & Wasson, 2021; Paper 3, Misiejuk & Wasson, in press; Paper 4, Misiejuk et al., submitted). All papers included in this research were co-authored, with me as the first author implying that I contributed the most (see Appendix A for co-author declarations).

During my doctoral scholarship, I published a total of eight papers, and the four papers included as part of this thesis are a selection from these papers (papers not included in this research: Ferguson et al., 2019a; Misiejuk, Scianna, Kaliisa, Vachuska, & Shaffer, 2021; Kaliisa, Misiejuk, Irgens, & Misfeldt, 2021; Misiejuk, Ness, Gray, & Wasson, submitted).

2. Background

In this chapter, the theoretical underpinnings of this research are presented. The chapter begins with an introduction to the learning analytics field, followed by a description of the challenges and opportunities of working with educational data, with a special emphasis on context data. Next, different approaches to learning analytics are presented, and the positioning of this research is described. The chapter continues with a short presentation of peer assessment. Then, three interest areas within peer assessment—feedback rubrics, feedback implementation, and backward evaluation—are presented, and a particular feedback model used in this research is described.

2.1 Learning analytics

The digitization of learning has enabled new ways of collecting and analyzing educational data. This process led to the emergence of the field of *learning analytics* (LA), commonly defined as “the measurement, collection, analysis, and reporting of data about learners and their contexts, for purposes of understanding and optimizing learning and environments in which it occurs” (Siemens, 2013, p. 1382). This definition was expanded by Wilson and Scalise (2016), who highlighted the importance of result interpretation rather than mere reporting, as data is not self-explanatory. As a field, LA is positioned at the intersection of learning, analytics, and human-centred design (*What is Learning Analytics?*, n.d.). Bartimote, Pardo, and Reimann (2018) defined three purposes of analytics in a classroom: description, diagnosis, or prediction. Although scholars include prescription as another purpose of LA (Ifenthaler & Yau, 2020; Du, Yang, Shelton, Hung, & Zhang, 2021); however, Bartimote et al. (2018) argue that prescription should be instructor-driven rather than technology-driven.

Some current mainstream LA research topics include dashboards for personalized learning, multimodal analytics for affect detection, orchestration support through collaborative analytics, and analytics for self-regulated learning in flipped classrooms

(Chen, Zou, & Xie, 2022). Quantitative research methods with a focus on predictive methods are prevalent in the field; however, there has been an increase in the use of qualitative or mixed methods in recent years (Viberg, Hatakka, Bälter, & Mavroudi, 2018).

The primary value of LA is its interventionist nature (Rogers, 2015; Atkisson & Wiley, 2011). The LA process includes data collection, metrics development, data analysis, and finally, an intervention to close the analytic loop (Clow, 2012). The iterative process is set up to refine both analytics and learning processes (Bartimote et al., 2018). However, it is also seen as potentially dangerous, as LA can be used to support student surveillance or only to serve institutional interests (Selwyn, 2019; Williamson, 2017). There are also ethical, privacy, and social responsibility concerns about LA implementations at educational institutions (Drachler et al., 2015; Wise, Sarmiento, & Boothe Jr, 2021; Ferguson, 2019). This research contributes to expanding knowledge about the opportunities and challenges in developing interventions based on working with two different LA datasets, given the privacy and consent constraints.

2.1.1 Educational data for learning analytics

The collection of digital traces that learners leave using online learning platforms is cheaper and less intrusive than other forms of data collection, such as self-reports, interviews, or external tests (Greller & Drachler, 2012). Early research closely connected LA with big data research (Klašnja-Milićević, Ivanović, & Budimac, 2017; Reyes, 2015; Macfadyen, Dawson, Pardo, & Gašević, 2014). Some scholars have even defined LA as “the application of these Big Data techniques to improve learning” (Clow, 2013, p. 684). Big data is commonly characterized by the size of the datasets, the difficulty of analyzing them in a traditional way (volume), the speed at which data are collected (velocity), and the diverse data types (variety) (Kitchin & Lauriault, 2015; Baig, Shuib, & Yadegaridehkordi, 2020). In addition, big data research is often associated with the fourth research paradigm in science, which focuses on exploration and pattern discovery in data rather than theory (Kitchin, 2014).

Although promising, digital data may be messy and a by-product of everyday operations on learning platforms (Kitchin, 2014; Krumm, Means, & Bienkowski, 2018). In addition, educational big data may be unreliable. For example, learners may create fake accounts to abuse the system to obtain certification, or instructors may create student profiles or course sites to test a platform. This data may be included in a dataset as actual learners/courses (Greller & Drachler, 2012). Sophisticated algorithms

to detect such cases are necessary but rarely applied. When not excluded from the analysis, these data points create noise in the dataset and can influence the results (Alexandron, Yoo, Ruipérez-Valiente, Lee, & Pritchard, 2019).

There is also the issue of the accessibility of data for research, which is often connected to data ownership. Student data collected on commercial products are not protected and accessible to the same degree as student data collected at educational institutions. Data ownership is often simply transferred to vendors by registering on a platform (Greller & Drachsler, 2012). Commercial educational data are mainly inaccessible for research purposes, and vendors lack incentives to make their data available for research. If they do grant access, the data provided may be limited (Fischer et al., 2020; Kitchin & Lauriault, 2015). Even in the case of institutional data, such as learning management system data, there may be many administrative and technological hurdles to gaining access, even if students have granted consent to their use.

Not only the right data is necessary to develop effective LA, but also knowledge about what the data mean and how they should be analyzed (Dringus, 2012; Daniel, 2019). Traditional educational researchers are actively involved in data collection and are familiar with the context of a learning activity (Daniel, 2019). In comparison, researchers working with educational big data may not know the context in which their dataset was collected. As LA researchers are usually not involved in the design and development of educational tools adopted at educational institutions, this means that tool developers and not researchers decide which data are generated and collected by the tools. Decisions regarding data collection are not neutral and are prone to human bias (Guzmán-Valenzuela, Gómez-González, Rojas-Murphy Tagle, & Lorca-Vyhmeister, 2021). In addition, some researchers have noticed that these tools may not log meaningful or useful data about learners (Roschelle & Krumm, 2015; Krumm et al., 2018).

Understanding a context is crucial for adequately interpreting LA study results and identifying any meaningful implications and accurate inferences (Song, 2018; Fischer et al., 2020). *Context* can be defined as “any information that can be used to characterize the situation of an entity” (Dey, 2001, p. 5). Previous research on context data in computing environments examined its usefulness for personalized learning (Bicans & Grundspenkis, 2017), data interoperability (Samuelsen, Chen, & Wasson, 2021), and context-aware computing (Dey, 2001). Context data can refer to a variety of data that serve “as a reminder that any pattern takes on its meaning in relation to larger configurations” (Rogers, Gašević, & Dawson, 2016, p. 235), such as temporal and

linguistic metadata, information about the physical activity, student artifacts, learning design, or information about learner knowledge and their social or technical context (Knight, Buckingham Shum, & Littleton, 2014; Rogers et al., 2016; Zheng, Ruan, & Li, 2015). If context data is not collected by a platform or a researcher is not allowed to access the captured context data, other data sources to enrich the dataset may be necessary. In that case, *context data* be defined as non-platform data captured to provide context and meaning to *platform data*. However, obtaining context data can be challenging due to privacy restrictions and a lack of direct access to course instructors or students (Daniel, 2019; Buchanan, Gesher, & Hammer, 2015).

All the above-described issues were encountered during this research and influenced the study design and methodology (see Chapter 3). Platform data used in both empirical studies was a by-product of a commercial platform that was not influenced by the researcher. Context data was missing in Study 1 (Paper 1, Misiejuk, Wasson, & Egelandstad, 2021), leading to difficulties in understanding and making sense of the data and resulting in limited analysis possibilities. In contrast, Study 2 (Paper 4, Misiejuk et al., submitted) began by combining platform data with context data collected by the participating educational institution. Access to context data widened the analysis scope and helped understand and interpret the results.

2.1.2 Constructivist approaches to learning analytics

Learning analytics is often criticized for being positivist, insensitive to contextual factors and shaping behaviourist interventions (Atkisson & Wiley, 2011). The positivist approach assumes that the analysis of data can “establish the objective existence of patterns that lead to novel insights and revolutionise practice by upending conventional understandings” (Rogers, 2015, p. 223) and ignores the non-objective nature of data, the influence of human decisions on data collection, and algorithmic bias (Carter & Egliston, 2021). As an alternative to a positivist approach to LA, a constructivist approach to LA has been suggested by several scholars (Banihashem & Macfadyen, 2021; Knight et al., 2014; Berland et al., 2014; Dietrichson, 2013). Constructivism is a learning theory originating from the works of Jean Piaget and Lev Vygotsky (Fosnot & Perry, 1996), and accounts for learners constructing their own meaning, learning building on prior knowledge, learning being enhanced by social interaction and learning developing through authentic tasks (Cooperstein & Kocevar-Weidinger, 2004). A constructivist approach to LA should avoid classifying students as good or bad but rather aim to improve the learning environments to facilitate learning and focus

on the quality of student constructs (Berland et al., 2014; Knight et al., 2014). Such approaches are deeply connected to the learning design tradition, with a focus on the learning process rather than learning outcomes (Banihashem & Macfadyen, 2021).

Peer assessment, an application area for LA in this research, is a collaborative learning activity facilitating skill development in line with constructivist principles (Gielen, Peeters, Dochy, Onghena, & Struyven, 2010; Tsai, 2001; Yurdabakan, 2011). Thus, the research described in this doctoral thesis applied the principles of the constructivist approach to LA. For example, the analysis did not focus on the student's final grades but rather on the learning process. In addition, this research aimed to identify ways to improve the learning activity and examined the learning design elements, such as feedback rubrics and student constructs, such as feedback comments and essays.

2.2 Peer assessment

Feedback is essential to the learning process, but students do not have many opportunities to receive feedback on their work, especially in large classrooms (Patchan, Schunn, & Correnti, 2016). One solution is to engage the students in *peer assessment* (PA), 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). Peer assessment is an umbrella term that includes different forms of PA activity, such as peer feedback, peer grading, or peer review. Formative PA focused on helping to improve other peers' work rather than grading the quality of peers' work (summative PA) was adopted in this research (Patchan, Schunn, & Clark, 2018; Topping, 1998).

During a PA activity, students engage in observational learning in which they are exposed to their peers' ideas and strategies to tackle an assignment (Chen, 2017; Patchan & Schunn, 2015; Ching & Hsu, 2016). In addition, they engage with their peers' work and use their skills in an authentic context. During a PA activity, students usually perform the roles of both feedback receivers and feedback providers. As feedback providers, students not only assess and reflect on the work of their peers but also find a way to convey their feedback. Previous research has found that providing feedback leads to more benefits than receiving feedback (Lundstrom & Baker, 2009; Patchan & Schunn, 2015). As feedback receivers, students need to reflect on the feedback received and decide which feedback to implement (Li, Liu, & Zhou, 2012). Peer assessment can help improve students' work and identify their own strengths

and weaknesses (Cho & Cho, 2011). Peer feedback is given in a language close to the feedback receiver's and with a level of complexity adapted to their subject understanding (Topping, 2009), which may help with feedback engagement since students typically do not feel intimidated by peer feedback (Liu, Lu, Wu, & Tsai, 2016).

Four meta-analyses reported a positive effect of PA activity on learning achievement in comparison to no PA (Zheng, Zhang, & Cui, 2020; Double, McGrane, & Hopfenbeck, 2020; Sanchez, Atkinson, Koenka, Moshontz, & Cooper, 2017), and to teacher assessment (Li, Xiong, Hunter, Guo, & Tywoniw, 2020). In addition, some aspects of the design of a PA activity can result in a larger effect, such as PA training before the activity, anonymity, and a combination of grading and commenting (Zheng et al., 2020). Peer assessment can also be helpful in the development of non-cognitive skills, such as problem detection and problem diagnosis (Patchan & Schunn, 2015), critical thinking skills (Lynch, McNamara, & Seery, 2012), learning strategies and academic mindsets (Li, Bialo, Xiong, Hunter, & Guo, 2021) or metacognitive skills, such as self-monitoring, planning, and self-efficacy skills (Tsai, Lin, & Yuan, 2002; Boud & Molloy, 2013; Double et al., 2020; Li et al., 2020; Baleghizadeh & Mortazavi, 2014; Ertmer et al., 2010; Liu et al., 2016). Furthermore, PA helps students develop their evaluative judgment, i.e., the ability to evaluate the quality of their own or others' work (Tai, Ajjawi, Boud, Dawson, & Panadero, 2018). They learn to recognize good quality work by better understanding the assessment criteria within a specific domain (Adachi, Tai, & Dawson, 2018a). In a broader context, the adoption of PA in authentic, real-life tasks was found to facilitate the development of essential and transferable skills for students' future employment, such as communication, collaboration, problem-solving, and reflection skills (Sokhanvar, Salehi, & Sokhanvar, 2021; Klucsevsek, 2016).

Although previous research has found that peer grading is just as valid as teacher grading (Double et al., 2020; Falchikov & Goldfinch, 2000; Li et al., 2016), students and instructors may be skeptical of students' abilities to provide accurate peer grading (Planas-Lladó et al., 2021; Wu & Schunn, 2021; Wu, 2019). Insufficient diagnosticity can reduce learning opportunities during a PA activity through peer over-marking, evaluating the work more favourably than an instructor would, or peer under-marking, evaluating peer work less favourably than an instructor would (Heyman & Sailors, 2011; Falchikov & Goldfinch, 2000). As a solution, one artifact is typically evaluated by multiple peers. The validity of PA can also be improved by providing students with clear instructions and well-designed rubrics (Song, Hu, Guo, & Gehringer, 2016).

Since the emergence of the Internet, many online PA platforms with different functionalities have been developed, and many learning management systems have adopted PA functionalities (Babik et al., 2016). Online PA platforms make a PA activity less time-consuming and more accessible to the instructors (Badea & Popescu, 2022) while the digital traces generated by learners on a PA platform have the potential to be used in an LA analysis. However, the effective use of online PA platforms is not trivial, as was highlighted during the emergency remote teaching caused by the COVID-19 pandemic, during which many instructors had to adopt online tools quickly. Research reported a decrease in the use of PA after the switch to emergency remote teaching and that students have a negative perception of online assessment activities (Şenel & Şenel, 2021; Panadero, Fraile, Pinedo, Rodríguez-Hernández, & Díez, 2022). This research contributes to a deeper understanding of online PA by using LA in aspects such as rubric design, feedback implementation, and backward evaluation.

2.2.1 Peer assessment rubrics

To improve PA effectiveness, a proper design of a rubric, “a simple assessment tool that describes levels of performance on a particular task” (Hafner & Hafner, 2003, p. 1509), is crucial. Online PA platforms open new possibilities for creating rubrics where additional prompts or calibrations can be added to the activity (Babik et al., 2016). In a PA activity, a rubric is used to familiarize students with quality criteria and to guide their evaluation.

The quality of the feedback that students receive is closely connected to the rubric design (Nilson, 2003). Previous research found that a PA activity without a rubric led to lower validity (Panadero, Romero, & Strijbos, 2013), a rubric with specific guidelines reduced the differences in commenting styles between high- and low-ability students (Patchan, Charney, & Schunn, 2009), the types of questions in rubrics, task structure, or artifact presentation resulted in different feedback quality (Hicks, Pandey, Fraser, & Klemmer, 2016) and a rubric can be a powerful tool to direct the focus of peer comments (Wallace et al., 1996).

There are still open questions regarding the effective design of a PA rubric. Previous research has found that combining commenting and grading has the best results in terms of increased learning achievement and the development of non-cognitive skills (Li et al., 2021 Zheng et al., 2020). The Peergrade platform offers both options and Boolean questions (i.e. yes/no questions), which were rarely explored in previous research. The relationship between the rubric design and backward evaluation was examined in Study

1 (Paper 1, Misiejuk, Wasson, & Egelandstal, 2021), while Study 2 (Paper 4, Misiejuk et al., submitted) explored the influence of rubric design on feedback implementation and revision.

2.2.2 Feedback implementation

When feedback is given for formative purposes, it is generally agreed that it should not only be passively received but also acted upon (Cartney, 2014). Revising of own work is challenging for learners, as they do not yet possess the skills to judge their work from an outside perspective (Wichmann, Funk, & Rummel, 2015). A goal of PA is to aid the revision process through peer feedback. *Feedback implementation* (also feedback uptake) refers to “changes made to the assessee’s product during revision that are clearly based on and related to received feedback” (Funk, Wichmann, & Rummel, 2013, p. 253). It is a complex process involving feedback sense-making and a reflective process to plan which feedback to implement and how (Wichmann et al., 2015). However, many students lack the literacy skills to understand the feedback they receive or how to implement it, and they may reject the feedback even before engaging with it (Funk et al., 2013). Another reason for not engaging with feedback could be a lack of feedback skills, confidence, time or domain knowledge on the feedback giver’s side, resulting in low-quality feedback (Dressler, Chu, Crossman, & Hilman, 2019; Patchan & Schunn, 2015; Fertalj, Brkić, & Mekterović, 2022).

Previous research has attempted to determine the specific feedback characteristics that constitute good-quality feedback and lead to its implementation. Many peer feedback models have emerged over the years (e.g., Hattie & Timperley, 2007; Espasa, Guasch, Mayordomo, Martinez-Melo, & Carless, 2018; van den Berg, Admiraal, & Pilot, 2006). The model used in Study 2 (Paper 4, Misiejuk et al., submitted) was inspired by the feedback model developed by Nelson and Schunn (2009) to examine how different types of feedback affect writing performance in an undergraduate course (see Figure 3 for the adapted model used in Study 2). This model was developed using previous empirical research and theories regarding potentially influential feedback features on feedback implementation, and was adapted in a variety of PA studies examining different aspects of PA (e.g., Wu & Schunn, 2020; Patchan et al., 2016; Sun, Lavoué, Aritajati, Tabard, & Rosson, 2019). Since feedback implementation was the focus of this study, this model was applied to identify relevant feedback features and code the feedback comments.

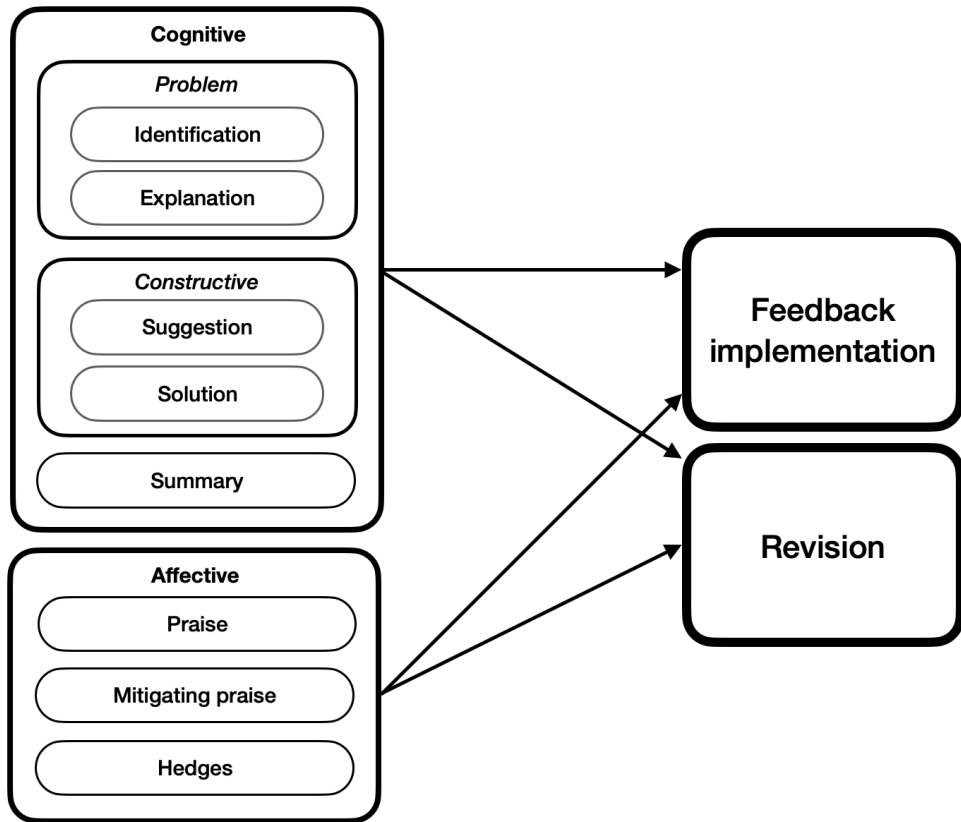


Figure 3: Feedback model (adapted from Nelson & Schunn, 2009)

The model distinguishes between two types of feedback. *Cognitive feedback* focuses on the content of the work and includes 1) problem detection through the identification of a problem (*identification*) and elaboration of the reasons why it is a problem (*explanation*), and 2) constructive comments, such as general advice (*suggestion*) and specific advice (*solution*) to improve the work, as well as 3) description of what students have done in their artifacts without detecting a problem (*summary*). *Affective feedback* is form-oriented and focuses on how students give feedback. It includes positive comments (*praise*), positive comments used to soften criticism (*mitigating praise*), and vague language (*hedges*).

The adapted model did not use two feedback features from the original model, *scope* (if a problem or a solution targets a local or a global issue) and *localization* (if a problem or a solution could be easily found). The structure of the PA rubric in Study 2 was divided into sections corresponding to sections in an essay and overall feedback, thus

implying feedback scope and localization. In addition, two feedback mediators were absent in the adapted model: *understanding* (students expressing that they understood the feedback received in a BE comment) and *agreement* (students expressing that they agreed with the feedback received in a BE comment). Initially, these feedback mediators were supposed to be used to code BE comments. However, due to the content and a low number of BE comments, agreement and understanding were not applicable in this study. Finally, the original model focused only on the *feedback implementation*; however, this study also included the *revision* to examine the extent to which students changed their original artifact, regardless of whether it was a result of the feedback that they received or not. This decision was made in line with the distinction made in previous research by Patchan et al. (2016).

Feedback implementation is one of the best-established benchmarks of effective PA. It was used in Study 2 (Paper 4, Misiejuk et al., submitted) to contribute to the body of knowledge regarding the factors influencing peer feedback implementation.

2.2.3 Backward evaluation

Over the years, students have been asked about their feedback preferences in surveys deployed after a PA activity. The research has found that students and instructors had similar perceptions regarding feedback, i.e. it should be timely, constructive, encouraging, based on an assessment rubric, and detailed about the shortcomings and future directions (Mulliner & Tucker, 2017). Online PA platforms have opened the possibility of a new way to gain insights into students' perceptions of feedback in a way that is integrated into the PA process. *Backward evaluation* (BE) allows students to evaluate the feedback they have received through grading, commenting, or multiple-choice options, depending on the platform (Luxton-Reilly, 2009). Through BE, students develop their feedback skills as a way to ensure that students actively engage with the feedback, reflect on it, and perhaps give advice to the feedback provider on how to improve it (Cook, 2019; Winstone, Nash, Parker, & Rowntree, 2017; Yuan & Kim, 2015). In addition, it helps students recognize the characteristics of useful feedback through exposure (Luxton-Reilly, 2009; Patchan et al., 2018). If they disagree with the feedback, they must utilize skills to defend their work and reflect on their decision-making process during the artifact development.

In a PA activity, BE can be implemented as an accountability method to incentivize students to provide higher-quality feedback and have a higher commitment to the PA task (Luxton-Reilly, 2009; Patchan et al., 2018). There are, however, some

disadvantages to BE, such as an increased workload for students, retaliation, or bias, where students only react positively to positive feedback (Patchan et al., 2018).

Backward evaluation is still an under-researched topic within the PA field. The online PA platform used in this research offers multiple BE features, such as BE grading, BE comments, and multiple-choice improvement suggestions. This data was used to expand the knowledge on BE in Study 1 (Paper 1, Misiejuk, Wasson, & Egelandstad, 2021), Study 2 (Paper 4, Misiejuk et al., submitted), and Scoping review 1 (Paper 2, Misiejuk & Wasson, 2021).

3. Methods

This chapter presents the methods used in this research. First, the adopted research paradigm is described. Next, the methods used to conduct the scoping reviews are detailed. The chapter continues with a description of the research methodology used in the empirical studies. The data sources and data collection are then presented, followed by the steps in the learning analytics process: data preprocessing, data analysis, and data interpretation. The chapter ends with a section on research ethics.

3.1 Research paradigm

This research adopts a pragmatic paradigm. A pragmatic approach aims not to present reality accurately but to be useful and effective in a particular context. The pragmatic paradigm emphasizes the importance of context in knowledge construction and that knowledge discovered through research is relative and not absolute (Eickhoff & Wieneke, 2018; Cherryholmes, 1992). It assumes that researchers are to be “flexible and open to the emergence of unexpected data” (Feilzer, 2010, p. 14), and it strives to abandon the split between qualitative and quantitative methods, encouraging the use of methods that work best for a particular problem (Tashakkori & Teddlie, 2003). A landmark of a pragmatic approach is the use of abductive reasoning, which seeks to infer the best explanation of a phenomenon by moving “back and forth between induction and deduction through a process of inquiry” (Doyle, Brady, & Byrne, 2009, p. 178).

3.2 Scoping reviews

Literature reviews are crucial for synthesizing knowledge in a field, revealing trends and patterns, and identifying research gaps (Paré & Kitsiou, 2017). The *scoping review* is a relatively new approach to conducting literature reviews that is best suited to “determine the scope or coverage of a body of literature on a given topic and give

clear indication of the volume of literature and studies available as well as an overview (broad or detailed) of its focus” (Munn et al., 2018, p. 2).

The scoping review approach was chosen as a method to conduct the literature reviews in this research due to the subject novelty and the lack of previous reviews on the relevant topics (see Table 1 for an overview). Scoping reviews are typically used to map and describe a broad range of literature without critically evaluating individual studies, which may lead to bias (Pham et al., 2014). More systematic methods for conducting a scoping review emerged to mitigate this issue. One approach described by Levac et al. (2010), built upon previous work by Arksey and O’Malley (2005), was used in this research to conduct both scoping reviews. This methodological framework requires transparency and a detailed description of each review step to improve the methodological validity of a scoping review and includes the following five main steps:

1. Identifying the research question
2. Identifying relevant studies
3. Selecting studies
4. Charting the data
5. Collating, summarizing, and reporting the results
6. Involving stakeholders (optional)

The optional step 6 calls for stakeholder involvement to add insights on the topic outside the literature; however, this step was not relevant in the two scoping reviews.

Table 1: Overview of the scoping reviews

	Scoping review 1 (Paper 2) Misiejuk & Wasson, 2021	Scoping review 2 (Paper 3) Misiejuk & Wasson, in press
Research questions	<p>1) What are the characteristics of the studies employing backward evaluation in peer assessment?</p> <p>2) How is backward evaluation conducted?</p> <p>3) What did the analyses of the backward evaluation data reveal?</p>	<p>1) Where in the peer assessment process are the analytics employed? What is the role of learning analytics in peer assessment research?</p> <p>2) What are the reported peer assessment challenges the research addressed with learning analytics? How are they addressed?</p> <p>3) What insights into peer assessment can we gain from learning analytics?</p>
Databases	ProQuest, ERIC, Web of Science, Science Direct, Google Scholar	ProQuest, ERIC, Web of Science, Science Direct, ACM DL, SAGE
Search string	See Appendix B	
# of papers identified	Search 1: 1,262 papers Search 2: 293 papers	1,569 papers
Inclusion criteria	<p>Timeframe: 2000-2021</p> <p>Peer-reviewed</p> <p>Language: English</p> <p>Discipline: Education</p> <p>Relevant to answering of the research questions</p> <p>Data from a PA platform with BE features</p> <p>BE data used in the PA analysis</p>	<p>Timeframe: 2011-2022</p> <p>Peer-reviewed</p> <p>Language: English</p> <p>Discipline: Education</p> <p>Relevant to answering of the research questions</p> <p>“Learning analytics” mentioned in full text or keywords OR paper published at the Learning Analytics and Knowledge Conference or in the Journal of Learning Analytics</p> <p>PA described in the methods section of the paper</p>
# of papers included	10 papers	27 papers
Coding process	Iterative and inductive	
Coding validity	Social moderation	
Methods	Descriptive statistics, thematic analysis	

The aim of the Scoping review 1 (Paper 2, Misiejuk & Wasson, 2021), the first review on the subject, was 1) to position findings from Study 1 (Paper 1, Misiejuk, Wasson, & Egelandstad, 2021) in the broader backward evaluation (BE) research, and 2) contribute to answering the main research question by possibly detecting other learning analytics (LA) methods dealing with a specific type of peer assessment (PA) data, namely, BE data. In particular, this scoping review mapped how previous research dealt with BE data and what insights were gained. The search string development (see Appendix B for search string) was not trivial because of the many terms used for both BE and PA found in the literature. In addition, due to the low number of papers identified in trial searches, the ultimate search was not limited to only LA papers. Overall, two searches were conducted to find relevant papers.

Scoping review 2 (Paper 3, Misiejuk & Wasson, in press) on using LA in PA research was carried out to fill a gap in the field and was the first review on the subject. This review aimed to answer the main research question by positioning this research in a broader context and identifying current developments in the field. The main challenge was determining whether the research reported in the paper was using LA. Many LA methods may be used in other fields; therefore, inclusion criteria were established to select only papers that explicitly positioned themselves within the LA field. Due to a large number of papers in the initial screening stages, a word-matching search was conducted in papers using phrases such as *learning analytics* to determine if they matched the review topic. Later, a closer reading of the full text focused on the method section determined if a paper was focused on PA.

EPPI-Reviewer Web (eppi.ioe.ac.uk), an online tool that supports conducting literature reviews, was used at the beginning stages of both scoping reviews and helped to manage references and facilitate the selection process. The in-depth data coding during the full-text reading of the papers was conducted in Excel. Thematic analysis was applied to categorize the codes into themes in both reviews, implying an iterative and inductive coding process (Braun & Clarke, 2006). The coding categories emerged from the data rather than from theory and were defined in multiple iterations (Thomas, 2006). Both reviews were conducted with a supervisor using social moderation to resolve disagreements about paper inclusion or paper coding and to ensure the validity of the process. Simple descriptive statistics complemented the qualitative analysis.

3.3 Empirical studies

3.3.1 Research methodology

The mixed-methods approach aims to produce a richer understanding of data. In addition, its goal is to confirm or challenge the findings (Parks & Peters, 2022) and to take advantage of the strengths of both qualitative and quantitative methods (Eickhoff & Wieneke, 2018). The empirical studies in this research used a triangulation mixed-methods design (Creswell & Plano Clark, 2011). The motivation to utilize mixed methods was the exploratory and interdisciplinary nature of LA research, which combines many research traditions, both qualitative and quantitative. As a relatively new field, LA does not have well-established guidelines about the method types, nor the amount and kind of data needed to answer a particular research question. A triangulation design implies a concurrent data analysis in which both qualitative and quantitative methods have equal weight. In this research, a variant of a triangulation model, a data transformation model, was chosen. Qualitative data was transformed into quantitative data, and the data was mixed during the analysis stage (Tashakkori & Teddlie, 1998; Creswell & Plano Clark, 2011).

All data analysis and visualizations were conducted using the R programming language, except the epistemic network analysis, which was developed using the ENA web tool (epistemicnetwork.org). The data was preprocessed using both R and Python programming languages.

Learning Analytics - Principles and Constraints Framework

Both empirical studies followed the *Learning Analytics - Principles and Constraints Framework* (LA-PCF) developed by Khalil and Ebner (2015) (See Figure 4). The LA-PCF combines and expands previous LA frameworks proposed by Clow (2012), Chatti, Dyckhoff, Schroeder, and Thüs (2012), and Greller and Drachsler (2012). The central part of the framework is the *LA Life Cycle* which describes four “proceeding steps, starting from the learning environment and ending with the appropriate intervention” (Khalil & Ebner, 2015, p. 1327). The LA-PCF steps were renamed for better readability (see new names in brackets):

1. *Learning environment* (data capture) shows where and how the stakeholders produce data.
2. *Big data* (data preprocessing) indicates the collected data and its preprocessing.

3. *Analytics* (data analysis) describes various LA techniques applied to analyze the data.
4. *Act* (data interpretation) clarifies how the analysis results are interpreted and used to intervene in or optimize LA.

The other part of the LA-PCF is *LA Constraints* which represents aspects of LA implementation that should be considered, such as *privacy*, *transparency*, or *ownership*. Every LA study is subject to all the limitations listed in the LA-PCF; however, only the constraints that impacted the empirical studies the most are described in the 3.3.6 section.

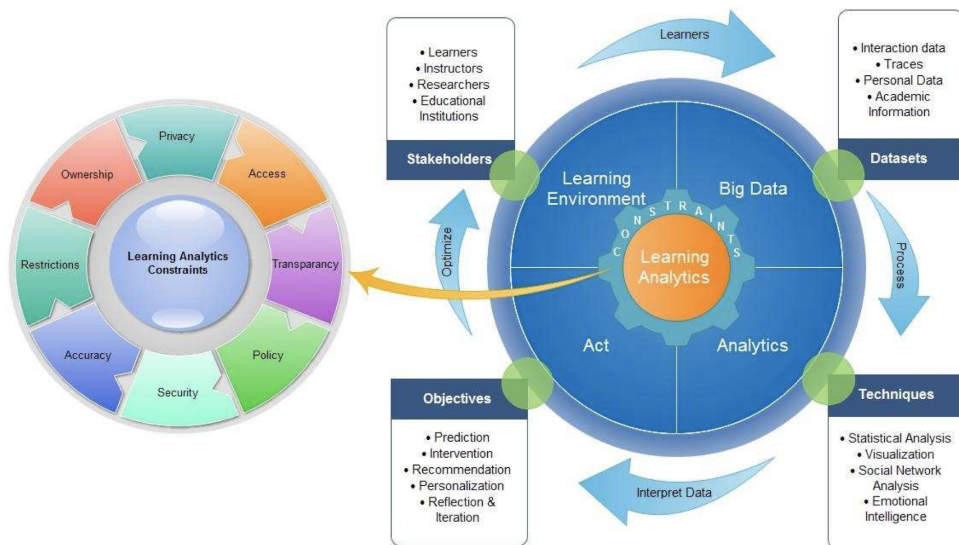


Figure 4: Learning Analytics - Principles and Constraints Framework (LA-PCF) (Khalil & Ebner, 2015, p. 1333).

3.3.2 Data collection

Data collection involves the recording and capturing of data. In Study 1 (Paper 1, Misiejuk, Wasson, & Egelandstad, 2021), Peergrade AsP shared three years of log data that they captured on their online PA platform. In Study 2 (Paper 4, Misiejuk & Wasson, in press), data about one PA activity conducted in an undergraduate course was collected 1) on the Peergrade (platform log data), 2) from the student administrative systems (essays, final grades), and 3) from the instructor (discipline, PA learning design, course design) (see Table 2 for study comparison).

Table 2: Overview of the empirical studies

	Study 1 (Paper 1) (Misiejuk & Wasson, 2021)	Study 2 (Paper 4) (Misiejuk et al., submitted)
Research questions	1) What is the relationship between student's perception of the usefulness of feedback, improvement suggestions, and comments on the feedback? 2) What is the relationship between rubric characteristics and student's perception of the usefulness of feedback?	1) How does context data change the LA analysis of peer assessment data 2) What influences student revision and feedback implementation in peer assessment?
Learning environment	Courses at multiple high schools and higher education institutions	Undergraduate course at a university college
PA typology	Unknown	Subject: Organization theory Objectives: Cognitive gains Training: Short pre-activity training Focus: Quantitative/qualitative/formative Product: Essay Relation to staff assessment: Substitutional Directionality: One-way assessment Privacy: Anonymous Contact: Distance Constellation: Groups Allocation: Random Place: Out of class Time: Free time Requirement: Compulsory Reward: Prerequisite for taking exam
Stakeholders	Learners, instructors	Learners, instructors
Datasets	Platform data: Peergrade log data	Platform data: Peergrade log data Context data: student artifacts (i.e. essays), final grades, discipline, PA learning design, course design
Sample	7,660 backward evaluations	863 implementable feedbacks

Both studies analyzed log data from the online PA platform, Peergrade AsP. Peer assessment can be implemented using a variety of design elements (Gielen, Dochy, & Onghena, 2011; Adachi, Tai, & Dawson, 2018b; Topping, 2021). A typical PA activity on the Peergrade platform starts with an instructor developing an assignment with instructions about an artifact (e.g., an essay or a critique) that a student should develop (see Figure 5). At the same time, the instructor develops a rubric using free-text, multiple-choice, or Boolean questions that should be used to evaluate the student’s artifact. In Peergrade, the instructor configures the details of the activity by entering the rubric and choosing, for example, submission dates and artifact formats (e.g., PDF, mp4 files). In addition, the PA activity information needs to be specified, such as whether students should give feedback in groups or individually, whether peer feedback should be given anonymously, or how many artifacts a student or a group should evaluate. Once the assignment is configured, students create their artifacts and upload them to Peergrade.

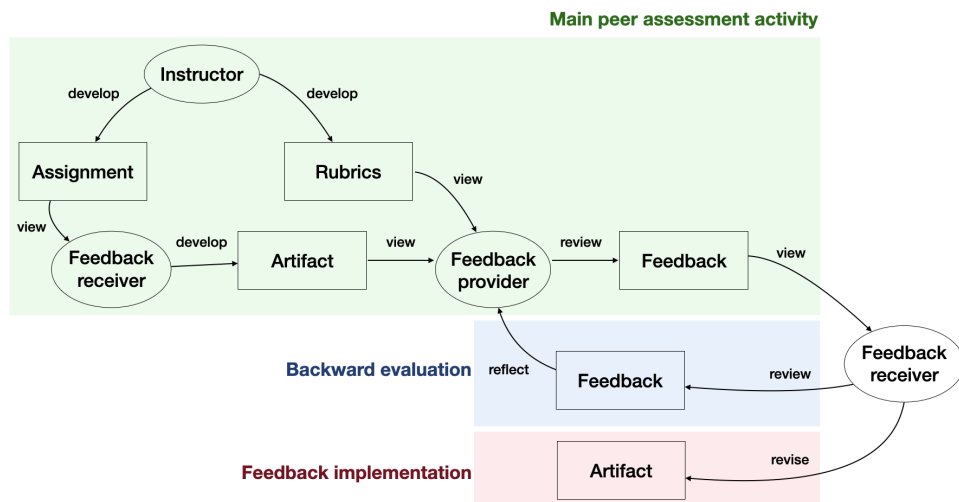


Figure 5: Model of the peer assessment process (adapted from Indriasari et al., 2020)

Students review their peers’ artifacts by answering the rubric questions. This feedback should be reflected on and (possibly) used to revise and improve the original artifact (*feedback implementation*). Feedback receivers can evaluate the feedback on their artifacts by grading it, writing a comment, or selecting one or more improvement suggestions that the feedback provider can later reflect on (*backward evaluation*). If students disagree with the feedback they received, they can also flag it for the instructor to intervene. Furthermore, students can “like” the feedback they particularly

appreciated. Instructors have access to a dashboard with analytics displaying, for example, the proportion of students who have already given feedback and grader agreement.

The Peergrade platform does not store the click stream data (i.e. every click) but rather does store the final text inputs (i.e. student answers to the rubrics or texts of the rubrics) and metadata (e.g., submission dates or time spent on giving feedback). Instructors can download a simplified and formatted version of the Peergrade log data from their PA activity.

The anonymized dataset provided by Peergrade ApS and used in Study 1 included 13 JSON files capturing all platform data logged between 2015 and 2017. Peergrade is a platform used internationally in various high schools and higher education institutions in different disciplines. Understanding the data structure was challenging, and connecting the files through a common ID was not possible in all cases. Due to privacy restrictions, no student artifacts were included in the provided dataset. The data was limited to the assignment texts, rubrics, student answers to the rubrics, backward evaluation, and some metadata, for example, time spent giving feedback and submission dates. The dataset did not include context information such as languages, educational level, geographical location, or discipline. Peer assessment activities were conducted in various languages, including cases of multiple languages appearing in one PA session, presumably indicating the use of Peergrade in foreign language courses.

Without the student artifacts (i.e. what was being evaluated and given feedback on), the feedback analysis was limited, as the feedback could not be compared with the content of the artifacts. In addition, analyzing the feedback itself would have been difficult since many assumptions would have to be made about the educational level and the subject area of an activity to make sense of the feedback. These limitations and challenges influenced the development of research questions that focused on students' perceptions of feedback (backward evaluation) and rubric design. In summary, the data in Study 1 was captured independently by a commercial company and not for research purposes, the learning environments (i.e., education level, subject area, and learning design) of the data collection were unknown, and the data included only the students' and instructors' input on the PA platform and limited metadata.

The data collection for the research reported in Study 2 mitigated the research limitations in Study 1. Rather than focusing on a large sample, the aim was to collect a richer dataset with more context information. To achieve this, collaboration was

initiated with an instructor of a large introductory class at a Norwegian university college who was already using Peergrade for peer feedback on an essay assignment. The PA activity examined was a part of an undergraduate course in the spring semester of 2019. My involvement in the PA design was limited to ensure an authentic learning setting. Two additional questions were added to the PA rubric about: 1) consent to use the data for research purposes, and 2) the student's previous experience with PA.

The data was generated by the students and instructor activity during an obligatory group PA activity in an undergraduate course, where the students had to complete the PA activity to be able to take the 3-hour written exam at the end of the semester (one-third of the final exam was directly related to the PA assignment; however, students did not know this beforehand). The PA activity started with a short training on PA and an introduction to the Peergrade platform and was then followed by a typical PA process. There was an additional step in which students could resubmit their revised essays to the platform. Only the essay resubmission and the BE part of the PA activity were voluntary.

Similar to Study 1, Peergrade data was collected in Study 2. Initially, the Peergrade data for this research was transferred directly from Peergrade ApS in 13 JSON files. However, similar issues with connecting the different files (as in Study 1) were encountered. To mitigate this, the simplified version of PA data available for instructors was downloaded from the Peergrade platform. As the BE data was not a part of downloadable data but could be found in the JSON files, direct support from Peergrade ApS technicians helped connect the two datasets. Guided by the research question, which focused on factors influencing feedback implementation, student artifacts (the draft essays and the revised essays) were also captured from a student administrative system. Moreover, final student grades gleaned from a student administrative system were added to the dataset, while interactions with the course instructor provided information about the course's learning design and aims.

3.3.3 Data preprocessing

Data preprocessing is essential in preparing raw data for actual data analysis (Romero, Romero, & Ventura, 2014). The data in both studies was preprocessed using data cleansing, data coding for textual analysis, and data transformation (see Table 3). The variable selection was motivated by the relevance to the research questions or by the amount of data available for a variable (see Tables in Appendix C for a detailed overview of the variables). For example, Peergrade allows students to flag feedback

they disagree with; however, very few students used this feature in both studies. Thus, it was not included in the analysis.

Table 3: Overview of methods, objectives and main constrains in the empirical studies

	Study 1 (Paper 1) Misiejuk & Wasson, 2021	Study 2 (Paper 4) Misiejuk et al., submitted
Data preprocessing	Inductive data coding Natural language processing Variable construction	Deductive data coding Variable construction
Data analysis	Descriptive statistics Spearman correlation analysis Ordinal logistic regression Epistemic network analysis	Descriptive statistics Pearson correlation analysis Logistic regression Quasi-binomial regression
Objectives	Prediction Exploration	Prediction Intervention
Main constraints	Privacy Accuracy	Access Ownership

Data cleansing

Data cleansing is a process of removing incorrect or incomplete data from a dataset; skipping this step can lead to inaccurate or unreliable analytics results (Chu, Ilyas, Krishnan, & Wang, 2016). The Peergrade dataset in Study 1 was first filtered for English entries, while missing/incomplete data points were deleted. A strict consent policy was applied in Study 2, ensuring that both the feedback-giving and feedback-receiving groups consented to the use of their data for research purposes. Data from non-consenting groups was removed, as well as data from groups with missing values.

Coding for textual analysis

Qualitative coding is a “structured analytical process of organizing qualitative data, primarily text data” (Eickhoff & Wieneke, 2018, p. 904). There are two main types of qualitative coding of text data: *inductive coding*, a bottom-up approach, where codes emerge from the data, and *deductive coding*, a top-down approach using a set of predetermined codes to classify the data (Thomas, 2006). There are also quantitative methods to code the data using natural language processing (NLP) algorithms, such as

sentiment analysis or topic modelling. Quantitative methods are faster than qualitative methods. However, the results of quantitative coding can be challenging to interpret and typically miss their textual context (Eickhoff & Wieneke, 2018).

The first attempts to code BE comments in Study 1 included several unsupervised NLP algorithms, such as topic modelling that uses statistical analyses to determine a set of words, a topic. However, the obtained results were not meaningful and difficult to interpret. Next, sentiment analysis, a method to identify valence in text data by assigning a numerical value to text (Medhat, Hassan, & Korashy, 2014), was conducted using a lexicon and rule-based algorithm, Vader sentiment analyzer (Hutto & Gilbert, 2014). Since this algorithm was specifically designed for analyzing social media, it was suited for BE comments, as they were relatively short, and the language used by students was mostly colloquial. The sentiment scores ranging from -1 (negative) to 1 (positive) were included in the data analysis, but there was a need for richer data coding. Hence, an inductive coding approach was adopted. A random sample of BE comments was used to determine main themes in the data and develop the codes. Due to a large number of BE comments, manual coding would have been too time- and resource-consuming. Therefore, the whole dataset was coded automatically using a simple string-matching algorithm, where specific phrases indicated a presence of a code in a BE comment. Cohen's kappa inter-rater reliability, a standard measurement to determine the level of agreement among raters controlling for agreement due to chance (Burla et al., 2008), was calculated by comparing the coding with a random sample of 10% of the whole dataset coded by a senior researcher.

Given the limitations of quantitative coding methods in Study 1 and the higher complexity of written feedback comments in Study 2, another approach was selected to code the data. First, the data was coded manually by a group of researchers instead of using automated coding techniques. Second, a deductive coding approach was adopted using an adapted version of the feedback feature coding scheme developed by Nelson and Schunn (2009). The coding scheme was refined in an iterative process to build a shared understanding of the meaning of the codes among the raters and to determine additional coding rules. This was followed by coding the entire dataset and calculating Cohen's kappa to ensure the validity of the coding. Initially, BE comments were supposed to be coded according to Nelson and Schunn's BE coding scheme (2009). However, only a few groups left a BE comment; thus, BE comments were excluded from the analysis. Instead, a binary variable was constructed to indicate whether a group left a BE comment.

Data transformation

The goal of data transformation is to convert cleaned data into measurable indicators through variable construction techniques, where new variables are created from a set of existing variables by aggregating, standardizing, applying external algorithms to add new features, or changing the format of the existing data (Vijayarani, Ilamathi, & Nithya, 2015; Fancsali, 2011).

In this research, some numerical variables were aggregated; for example, instead of using individual student grades, the average group grade was used in Study 2. Another way of transforming the numerical data was to calculate proportions, such as in the case of the proportion of specific question types in a rubric in Study 1 or the proportion of students with previous PA experience in Study 2. In addition, text data could be transformed into numerical data. For example, for the *improvement suggestions* in Study 1, a text variable listing improvement suggestions selected by a student, was transformed into the *number of improvement suggestions*, a numerical variable ranging from 0 to 5. Another example would be variables indicating the number of words in a BE comment in Study 1 or in an essay or written feedback comments in Study 2, which were calculated automatically from text data and used in the analysis as numerical data. Some variables were derived from already transformed variables. For example, the *number of improvement suggestions* in Study 1 was used to construct the *number of improvement suggestions category*, a two-level categorical variable categorizing 1-3 suggestions as *fewCat*, and 4-5 suggestions as *manyCat*. Other variables were constructed using external algorithms, such as part-of-speech tagging to extract grammatical properties of BE comments in Study 1 or Jaccard similarity (Mullen, 2015) used to determine the similarity between the draft and final essay in Study 2. Finally, some variables were constructed manually. For example, the *implementation level* in Study 2 was established by comparing the first and revised draft with implementable feedback, that is, feedback that may trigger implementation, to determine the extent of feedback implementation on a 3-level scale. The coding process was iterative and included the coding of a sample by four researchers to find a common understanding of the codes before proceeding to code the whole dataset. Cohen's kappa inter-rater reliability was used to validate the coding.

3.3.4 Data analysis

Descriptive statistics, including correlation analysis, was used to explore the data and the relationships between the variables in both studies. In addition, epistemic network analysis was utilized as an additional data exploration tool in Study 1. Finally, regression analysis determined which features predict students' perceptions of the feedback they received (backward evaluation) in Study 1 and feedback implementation in Study 2.

Descriptive statistics

Both studies used descriptive statistics and visualizations to summarize the preprocessed data by detecting patterns and providing an overview of the dataset (Fisher & Marshall, 2009). In addition, two types of correlation analyses were used in this research. Spearman's rank correlation, the most appropriate method to measure associations for ordinal variables, was used in Study 1 to examine the relationship between BE grades (an ordinal variable with five levels) and other variables (Mukaka, 2012). Since most variables were either Boolean, dichotomous, or continuous in Study 2, Pearson's correlation, a method to determine the strength of linear association between variables, was applied (Prion & Haerling, 2014). The correlation used in Study 2 also calculated the phi coefficient for correlations between two Boolean variables (Harrell Jr & Harrell Jr, 2019). The strength of association was determined as follows: 0–0.19 was regarded as very weak, 0.2–0.39 as weak, 0.40–0.59 as moderate, 0.6–0.79 as strong, and 0.8–1 as very strong (Swinscow & Campbell, 1997).

Regression analysis

Regression analysis is one of the most popular LA methods for determining a casual relationship between variables and is often used for prediction purposes (Charitopoulos, Rangoussi, & Koulouriotis, 2020; Sykes, 1993). One ordinal logistic regression model was developed in Study 1, while three logistic regression models and one quasi-binomial regression model were developed in Study 2.

The regression model in Study 1 was applied to examine which variables contributed to student perceptions of feedback usefulness. Ordinal logistic regression was selected to model BE grades, an ordinal variable, as the dependent variable. Ordinal logistic regression considers the ordering of the levels in the outcome variable, estimates odds ratios and assumes the same effect on the odds regardless of the cut point (Kleinbaum

& Klein, 2010). Stepwise regression using backward selection was conducted to select the variables for the analysis (Venables & Ripley, 2002). This method begins with a full model, and every iteration removes the least significant variable (Liddell & Kruschke, 2018; Healy, 1995). The regression assumptions were checked successfully by calculating the generalized variance inflation factor, a measurement to estimate multicollinearity, and the Brant test for parallel regression assumption (Brant, 1990).

Four statistical models were developed in Study 2: Model 1 predicting no feedback implementation; Model 2 predicting partial feedback implementation; Model 3 predicting full feedback implementation; and, Model 4 predicting the rate of essay revision. The outcome variables of the first three models were Boolean variables indicating different levels of feedback implementation. Thus, logistic regression, the most appropriate method for modelling binary dependent variables, was chosen to analyze the data (DeMaris, 1995). Logistic regression assumes a nonlinear association between the outcome variable and predictors and observations to be independent of each other (French, Immekus, & Yen, 2013). The dependent variable in Model 4 was the revision rate, a proportion. Although proportion variables are continuous, their range is limited to 0–1, and they often violate the normal error term and constant variance assumptions in statistical analysis (Douma & Weedon, 2019). One method for modelling under-dispersed data, where the observed variability is lower than the expected variability, is quasi-binomial regression, (Xekalaki, 2014; Shoukri & Aleid, 2022). Best-fitting models were selected using both-direction stepwise selection, which optimizes the Akaike information criterion value with the help of both forward and backward selection (Venables & Ripley, 2002). The independent variables in the models were successfully examined for the tolerance and variance inflation factor applied to detect multicollinearity. At the same time, Cook’s distance analysis showed no influential outliers in any of the models, and the Hosmer-Lemeshow test indicated a good fit of all models to data (Peng, Lee, & Ingersoll, 2002).

Epistemic network analysis

Epistemic network analysis (ENA) is a novel method used to model the connections between different codes to project them onto a two-dimensional space as a non-directional network (Shaffer & Ruis, 2017). A typical application of ENA would examine the strength of connections among codes for group comparison and project network means. In addition, ENA enables a statistical analysis of the differences between the groups. In Study 1, ENA was utilized to visualize the co-concurrences

of different types of improvement suggestions and the total number of improvement suggestions. The comparison groups were the BE grades. ENA was used to gain a deeper understanding about the number and kinds of suggestions selected by students depending on their feedback perceptions.

3.3.5 Data interpretation

Data interpretation in LA research is closely connected to the closing of the LA loop (Clow, 2012). The effectiveness of LA is measured by its ability to provide meaningful information to stakeholders to improve the learning environment (Rogers, 2015; Atkisson & Wiley, 2011). Since there was no contact with any of the course instructors in Study 1, the practical impact of LA was limited. The main goal of Study 1 was to understand what new insights could be gained by using LA on Peergrade data. The challenges of working with context-free data in Study 1 limited the analysis and resulted in informing the design of the subsequent empirical study, Study 2. In comparison, as the course instructor participated in the data coding and interpretation in Study 2, the analysis results could inform his design decisions regarding future PA activities, thus closing the LA cycle.

3.3.6 Research ethics

There are many concerns about privacy and consent issues in the LA field (Slade & Prinsloo, 2013; Pardo & Siemens, 2014). Study 1 used anonymous data that were limited to platform data only, owned by Peergrade ApS. Since the data transferred from Peergrade ApS were already anonymized and did not include personal data, no student consent was required. The main constraints encountered in this study were connected to privacy and accuracy, as described by Khalil and Ebner (2015). The privacy laws and/or an inability to gather student consent to share their artifacts (as a consequence, these were not shared) resulted in a limited analysis of context-free platform data. In addition, big datasets tend to include data from test users or users who misuse the system. Therefore, sophisticated algorithms are needed to filter them out, which would have been too difficult to develop for this study.

The design of Study 2 was approved by the NSD, the Norwegian Centre for Research Data (see Appendix E). Students were asked for their consent, and their data were anonymized by a third party, a collaborator at the NORCE Norwegian Research Centre. Any data from a non-consenting group was removed from the analysis. This study's two main LA constraints were access and ownership (Khalil & Ebner, 2015). The strict

consent policy adopted in this study led to a significant reduction in the size of the final dataset, which may have influenced the results. This led to some consideration to soften this policy and, for example, still include the feedback data from non-consenting groups and only exclude their essays from the analysis, but this was not conducted. Further, the results from this study influenced the learning design of future course work—the essay was changed to a graded exam—and an additional voluntary round of peer feedback was added to the PA activity. The data analysis process in this study was slow and resource-consuming, but if sped up, it might have caused more ethical dilemmas. Only 40% of all groups consented to share their data. If the data analysis had been conducted in real-time, the instructor would have gained insights based on the activities of a minority. This raises methodological and ethical questions about the potential actionability and validity of the findings if generalized to an entire classroom (Mathrani, Susnjak, Ramaswami, & Barczak, 2021; Cormack, 2016).

4. Results

This chapter describes the main results from each study in this research and how these findings contribute to answering the main research question regarding the insights into peer assessment that can be gained using learning analytics. First, the papers included in this research are summarized. Next, four sub-questions presented in the introduction are used to structure the presentation in this chapter.

The work described in this research answers the main research question by examining different implementations of learning analytics (LA) to expand knowledge of peer assessment (PA). Research using LA to examine PA is emerging and scattered; therefore, this research adopted both literature review and empirical approaches. Four interrelated papers that build on insights from one another are included in this research. The two empirical studies answer the main research question using different datasets, while two scoping reviews position the results of the empirical analysis within the broader research landscape. Study 1 (Paper 1, Misiejuk, Wasson, & Egelanddal, 2021) focused on exploring what new insights about PA can be gained from a context-free dataset captured by an online PA platform. The dataset limitations led the analysis towards backward evaluation (BE) topics and the rubric design in a PA activity. Study 2 (Paper 4, Misiejuk et al., submitted) was an empirical study analyzing a context-rich dataset using LA. The analysis focused on both PA aspects such as feedback implementation and essay revision, and LA aspects such as the potential and challenges in adding context data to platform data. Table 4 presents an overview of dependent and independent variables used in regression analysis in the empirical studies (see Tables in Appendix C for more details). Scoping review 1 (Paper 2, Misiejuk & Wasson, 2021) examined how previous research used BE data and what insights about PA could be discovered when including this data. Scoping review 2 (Paper 3, Misiejuk & Wasson, in press) mapped the landscape of LA applications in PA research, including PA challenges addressed with LA methods and new findings about PA gained using LA.

Table 4: Short overview of the variables in regression analyses by research question

	Study	Dependent variables	Independent variables
Q1	1	BE grade	Rubric questions: <ul style="list-style-type: none"> • Total number per rubric • Type: Boolean, scale, free-text
	2	No feedback implementation Partial feedback implementation Full feedback implementation Revision rate	Implementable feedback: <ul style="list-style-type: none"> • Boolean • Scale • Written: solution or suggestion
Q2	2	No feedback implementation Partial feedback implementation Full feedback implementation Revision rate	Group: <ul style="list-style-type: none"> • Size • Average final grade • Previous PA experience Draft: <ul style="list-style-type: none"> • Length • Grade Feedback: <ul style="list-style-type: none"> • Comment length • Total number of implementable feedback • Total number of praise-only comments • Total number of praise+summary comments
Q3	1	BE grade	Improvement suggestions: <ul style="list-style-type: none"> • Type: specificity, constructivity, relevance, kindness, justification • Total number of suggestions BE comments: <ul style="list-style-type: none"> • Sentiment score • Length • Part-of-speech tagging: verbs, nouns, adjectives • Coding: acceptance, defence, gratitude
	2	No feedback implementation Partial feedback implementation Full feedback implementation Revision rate	BE grade BE comment exists

Q1: *How does the design of a feedback rubric influence backward evaluation, feedback implementation, and essay revision in a peer assessment activity?*

Instructors can choose the number and the type (Boolean, scale or free-text) of questions in Peergrade to design their feedback rubrics.

In Study 1, the design of the feedback rubric was characterized by the type and number of questions asked, as the dataset included rubrics from multiple courses. One variable counted the total number of questions per rubric, and the proportion of each type of question (Boolean, scale, and free-text) in a rubric was calculated. The effect of the rubric was measured against student perception expressed by the backward evaluation (BE) grade using ordinal logistic regression. The only statistically significant result indicated that increasing from one level of *BE grade* to the next multiplies the odds of *free-text questions* by 1.011.

Study 2 analyzed the data from only one PA activity; hence only one rubric that included a mix of Boolean, free-text, and scale questions, was examined. The analysis of the rubric design focused on comparing the implementable feedback given using different question types with the levels of feedback implementation and revision levels. The latter two represent the outcome variables of regression modelling performed in this study. First, implementable feedback was defined for every question type (see Table 6 in Appendix C for details). Next, the answers to the rubric questions were classified as one of four variables: Boolean implementable feedback; scale implementable feedback; and two types of written implementable feedback (answers to free-text questions): solutions; and, suggestions. Descriptive statistics indicated that different feedback types influenced the implementation rate: 90% of all *Boolean implementable feedback* was either partially or fully implemented, and in comparison, 71% of all *scale implementable feedback* and 65% of all *written implementable feedback (suggestions+solutions)* was partially or fully implemented. The regression analysis identified two variables as statistically significant. *Boolean implementable feedback* doubled the odds of *partial feedback implementation* and was associated with a 72% reduction in the odds of *no feedback implementation*. The *solution* in a feedback comment decreased the odds of *full feedback implementation* by 32%.

In summary, both Study 1 and Study 2 revealed few insights into the effects of the question types used in the rubric. The effect of free-text questions on student perception in Study 1 was positive, but relatively small. Study 2 indicated that Boolean feedback was more manageable for students to implement than solutions. In contrast, the solutions—that is, specific suggestions in the written feedback comments—had a negative effect on full feedback implementation.

Q2: What influences essay revision and peer feedback implementation?

The aim of Study 2 was to examine the factors contributing to feedback implementation and essay revision. Four outcome variables were constructed: 1) no implementation, a binary variable indicating feedback that was not implemented at all; 2) partial implementation, a binary variable indicating partial implementation of feedback; 3) full implementation, a binary variable indicating full feedback implementation; and 4) revision rate, a proportion indicating the extent of the essay revision. The first three variables were used in three logistic regression models, while the fourth variable was modelled in a quasi-binomial regression. The predictors in this study included: student characteristics (group size, group average final grade, and previous PA experience), essay characteristics (draft length and grade), and feedback characteristics (feedback comment length, total number of implementable feedback, total number of praise-only comments, total number of praise+summary comments).

For feedback provider groups, each additional increase in the *average grade* was associated with an 18% increase in the odds of a higher *revision rate*, while each additional increase in *group size* was associated with a 46% decrease in the odds of *full feedback implementation*, a 42% increase in the odds of *no feedback implementation* and a decrease in the odds of *revision rates*. For feedback receiver groups, each additional increase in the *group size* was associated with a 39% decrease in the odds of *partial feedback implementation* and a 37% increase in the odds of *full feedback implementation*. In addition, each additional increase in the *average grade* was associated with a 28% decrease in the odds of *no feedback implementation*, a 25% increase in the odds of *full feedback implementation*, and an increase in the odds of *revision rates*. Finally, each increase in *previous PA experience* in feedback receiver groups was associated with a 55% increase in the odds of *full feedback implementation* and a decrease in the odds of *revision rates*.

Furthermore, an increase in the *draft grade*, i.e. peer grade on the draft, was associated with a 23% increase in the odds of *no feedback implementation* and a decrease in the odds of *revision rates*. In addition, the correlation analysis detected a strong negative relationship between *draft length* and *revision rate*, $r(861) = -.72, p < .001$.

The analysis of feedback characteristics showed that an increase in *praise-only* comments was associated with a 28% increase in the odds of *partial feedback implementation*. Further, an increase in the number of *praise-only* or *summary+praise* feedback comments was associated with an over 20% decrease in the odds of *full feedback implementation*. In contrast, *mitigating praise* used in constructive comments was associated with a 93% increase in the odds of *full feedback implementation*. Finally, there was a strong positive correlation between *explanation* and *solution*, $r(861) = .64, p < .001$.

In summary, the results highlight the importance of group size in PA activity. Bigger groups gave feedback that tended not to be implemented, but tackled received feedback more effectively. In addition, previous PA experience of feedback-giving groups did not predict full feedback implementation and revisions. If the first draft was graded high by peers, feedback receivers were less likely to revise the draft and implement the feedback. Mitigating praise had a positive effect on feedback implementation. A high total amount of praise-only comments was motivating to partially implement feedback but negatively affected full feedback implementation. In addition, praise+summary had a negative effect on full feedback implementation.

Q3: *What can we learn about students' perceptions of peer feedback from backward evaluation data?*

The broader view of the use of BE data captured from online PA platforms was presented in Scoping review 1. The review included 9 papers and did not focus only on LA papers, as only two papers explicitly stated their affiliation with the field (Misiejuk & Wasson, 2021; Tsivitanidou & Ioannou, 2019), while four papers were published before 2011, the year of the first International Learning Analytics and Knowledge Conference (Cho & Kim, 2007; Cho & Schunn, 2007; van der Pol, van den Berg, Admiraal, & Simons, 2008; Nelson & Schunn, 2009). The review found that scales measuring feedback helpfulness and agreement were the most popular, while BE comments were typically coded based on agreement and/or understanding of the feedback. In addition, BE data was mostly used to help detect if students were engaged in tit-for-tat strategies and to measure student feedback implementation.

The Peergrade platform offers three ways to conduct a BE activity: by grading the feedback usefulness on a scale of 1 to 5 (BE grade); by commenting on the feedback received (BE comment); and, by selecting improvement suggestions to improve feedback quality in a multiple-choice question including five options: specificity, constructivity, relevance, kindness, and justification (improvement suggestions).

Study 1 analyzed student feedback perception by modelling a BE grade as the outcome variable of the ordinal regression. The following independent variables were used in the analysis (see Table 5 in Appendix C for more details): selected improvement suggestions, total number of improvement suggestions selected; sentiment score of the written BE comments; length of the BE comment; part-of-speech tagging of the BE comment, including verbs, nouns, and adjectives; and, coding of the written BE comment (*acceptance*: expressing praise, error acknowledgment, or intention of revision; *defence*: expressing confusion, criticism, or disagreement; *gratitude*: thanking for the feedback).

The analysis in Study 1 found that expressing *defence* in the feedback comments was associated with a 91% decrease in odds of higher *BE grade* in comparison with the baseline, that is, BE comments coded only with *acceptance*. *Defence+gratitude* was associated with a 72% decrease; *acceptance+defence* was associated with a 71% decrease; and, *acceptance+defence+gratitude* was associated with a 53% decrease. The results showed that selecting any improvement suggestions by a student predicted a higher likelihood that the student would find feedback less useful than more useful. *Relevance* had the strongest effect and was associated with a 71% decrease in odds, while *justification* had the weakest effect, with a decrease of 45% in odds. Backward evaluation comments coded only with *gratitude* increased the odds that the students were more likely to find feedback more useful compared to the baseline (i.e., *acceptance* by 20%), while BE comments coded with *acceptance+gratitude* were more likely by 96%. In addition, for every one-unit increase in *sentiment score*, the odds of feedback being more useful rather than less useful was multiplied by 2.54 times, when holding all other variables constant. Finally, the results of the epistemic network analysis examining which improvement suggestions were selected depending on student perception of feedback usefulness, determined that students who graded the feedback as *not useful at all* selected mostly a combination of *specificity*, *constructivity*, and *relevance*, and rarely selected *kindness* or *relevance*. Also, students who graded the feedback as *extremely useful* typically suggested that feedback should be more *specific* and *constructive*.

Two variables were used to examine the role of BE in feedback implementation and essay revision in Study 2: 1) BE grade, and 2) a binary variable indicating whether a group left a BE comment. Improvement suggestions were rarely used; therefore, they were excluded from the analysis. A BE grade was found to have the opposite effect of a BE comment on no feedback implementation, full feedback implementation, and essay revision rate. For each increase in *BE grade*, the odds of *no feedback implementation* increased by 21%, while the odds of *full feedback implementation* and *revision rates* decreased. In contrast, writing a *BE comment* decreased the odds of *no feedback implementation* by 51%, doubled the odds of *full feedback implementation*, and was associated with an increase in *revision rates*.

In summary, students' perception of feedback usefulness as expressed by BE grade in Study 1 was strictly connected with their level of agreement with the feedback. The feedback that was not useful was perceived as such because it was not relevant, rather than unkind or unjustified. In Study 2, only recognizing the usefulness of feedback by grading had a negative effect on feedback implementation while writing a BE comment had positively affected feedback implementation.

Q4: How does context data change the learning analytics analysis of peer assessment data?

This research defines *context data* as data that helps understand *platform data* and usually has to be obtained from additional data sources such as student administrative systems or course instructors. Two different datasets from the same online PA platform were investigated in this research. The first dataset used in Study 1 included only platform data. In contrast, the second dataset analyzed in Study 2 included both platform and context data (student artifacts, student grades, discipline, PA learning design, and course information).

Study 1 showed that it is possible to gain new insights into PA using LA techniques, even from a context-free dataset. However, understanding the meaning of the data without context was challenging. Furthermore, the lack of context data restricted the research scope and analysis possibilities. Finally, the findings had a limited impact on improving PA activity. In contrast, the inclusion of context data in Study 2 helped to discover more profound insights into the PA activity during the authentic offering of a course. The collaboration with the course instructor added rich information about the course structure, the rationale behind using PA in the class, and domain knowledge on the topic. In addition, the inclusion of the instructor in the data coding and analysis

led to new actionable insights that were reflected in changes to the new offerings of the class. This study showed how LA could be used to gain new insights and motivate the improvement of a learning activity.

Paper 1 (Study 1) appeared in the search results in both Scoping review 1 and Scoping review 2 (see Tables in Appendix D for an overview of included papers). All papers in Scoping review 1 used various context data to describe and complement the platform data in the analysis, except for Paper 1. Similarly, most papers focused on applying LA to gain new insights from PA data in Scoping review 2 collected and used extensive context data. Two other papers besides Paper 1 in Scoping review 2 had little or no context data. The context data was irrelevant in the case of the study by Babik et al. (2016), who worked with simulated datasets. The analysis was conducted only with platform data in Djelil et al. (2021), while limited context data, including educational level and discipline, was used to describe the dataset. This study is an interesting example of analysis with little context data. The focus of the study was narrowed to exploring the role transitions of students and teachers in PA activities over time. However, it did not examine the importance of these transitions in light of other PA aspects. Possibly, additional data could be captured directly from the online platform, depending on the platform owner's design decisions. Both scoping reviews showed that the inclusion of context data, even if limited, made the analysis more powerful and highlighted the limitations of analyzing only platform data in Paper 1.

In summary, including context data in LA analysis improves the quality and depth of the insights that can be discovered. Access to context data prerequisites some connection to the stakeholders producing the data, leading to more impactful and actionable findings and closing of the LA cycle. Also, a close collaboration with stakeholders can help understand and contextualize the data collected, which is crucial if the researcher has to use a third-party platform to capture the data.

5. Conclusion

The first part of this chapter describes the theoretical, methodological, and empirical contributions of this research guided by the following main research question:

How can we use learning analytics to gain new insights into peer assessment?

Next, the evaluation of the research approach and the research limitations are presented. The chapter concludes with possible future research and conclusions.

5.1 Theoretical contributions

The first theoretical contribution is the mapping of the use of learning analytics (LA) to gain insights into the peer assessment (PA) process and how to optimize it in Scoping review 2 (Paper 3, Misiejuk & Wasson, in press). This scoping review aimed to show the broad landscape of current research, identify potential research gaps, and recommend potential future research directions. It focused on the PA challenges addressed using LA, how they were solved, and which insights were gained. This review showed that LA has the potential to better understand and improve PA activities through new insights into student behaviour and the artifacts that they produce, interpersonal and intergroup interactions, or tool improvement. However, research is still emerging and scattered. Twenty of the 27 papers included in this review were published in the last four years, indicating a growing interest in supporting constructivist learning activities such as PA using LA.

The second theoretical contribution showed the use of backward evaluation (BE) data in research in Scoping review 1 (Paper 2, Misiejuk & Wasson, 2021). This scoping review mapped different terms used for BE, compiled the coding schemes used for BE comments, and summarized the main results of utilizing this accountability method in PA activities. Analysis of BE data could be a new promising avenue for LA research, as BE data has the potential to answer new research questions and gain new insights into student feedback perception and processing.

5.2 Methodological contributions

The transparency of the methods used in Study 1 (Paper 1, Misiejuk, Wasson, & Egelandsdal, 2021) and Study 2 (Paper 4, Misiejuk et al., submitted) can be considered a methodological contribution. Detailed descriptions of the data, analysis, and decision-making process regarding two very different datasets, a context-free dataset and a context-rich dataset, are presented in the papers and Chapter 3. Further, a description of the influence of data availability on the analysis are provided. Learning analytics research is often criticized for collecting, measuring, and analyzing what is easiest to obtain (Guzmán-Valenzuela et al., 2021; Selwyn, 2019). Study 1 and Study 2 show the possibilities and challenges 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. In addition, this research is part of a broader LA landscape that experiments by applying different methods to various datasets to improve or better understand PA, as mapped in Scoping Review 2.

Another methodological contribution is the exploration of the importance of context data to improve the understanding of and complement platform data in LA research. Context-free platform data in Study 1 limited the analysis significantly, whereas adding context data to platform data in Study 2 enabled a more in-depth investigation of the PA phenomenon. This research strengthens the previous recommendations about the necessity of context data for LA analysis highlighted by scholars working at the intersection of LA and learning design (Mangaroska & Giannakos, 2018). In addition, this research contributes to the methodological considerations around the importance of meaningful and impactful data rather than big data in LA research (Merceron, Blikstein, & Siemens, 2015; Kitchin & Lauriault, 2015; Yudelson et al., 2014).

5.3 Empirical contributions

This research provides empirical contributions by expanding the knowledge in three aspects of peer assessment (PA): rubric design (Study 1, Study 2); feedback implementation and revision (Study 2); and backward evaluation (BE) (Study 1, Study 2). The findings provide several practical implications for instructors to design PA activities.

Study 1 showed that using free-text questions in the feedback rubric had a positive, but small effect on students' perceptions of the usefulness of feedback. However, this finding is difficult to interpret without context data. Study 2 indicated that Boolean feedback was less challenging to implement than solutions in written feedback comments (answers to free-text questions in a rubric).

Boolean feedback was used to ensure that the essays followed formal requirements, such as the right font type or citation style, so it could have been easier to implement. In contrast, solutions—the specific suggestions in the feedback comments—were written by students in response to general questions about the content and possible improvements of the essay. Feedback written by students could have been more difficult to understand, or feedback receivers may have lacked feedback literacy skills to implement it. This finding confirms previous research on the effectiveness of a structured way to provide peer feedback (Ashton & Davies, 2015). In addition, although the students determined if an essay followed the formal requirements, the actual feedback was provided by the instructor in the form of Boolean questions. Since the solutions were written by peers, they did not carry the same weight as feedback mediated by the instructor in the Boolean feedback. Previous studies have found that some students prefer teacher feedback over peer feedback (Motlagh, 2015; Zhang, 1995). Another explanation could be that feedback receivers did not agree with the solution suggested by their peers, in line with the findings by Wu and Schunn (2020). Also, the solutions may not have been provided in a language motivating students to implement them, for example, by using praise, which was a significant indicator of full implementation found in this study. The results of Study 2 can inform the design of PA rubrics. For example, instructors could use Boolean questions for important domain concepts and free-text questions for less complex aspects of the assignment. This strategy may be helpful in courses where students are unfamiliar with the domain and need more guidance. Another option would be to add more details about the assessment criteria in the free-text questions to help students provide higher-quality feedback (Gielen & De Wever, 2015). Finally, additional grading exercises and intragroup rubric discussions during PA training could help students better understand the rubric and improve feedback implementation (Liu, Li, & Zhang, 2018).

The influence of the size of feedback provider and feedback receiver groups was analyzed separately in Study 2. Previous research reported the advantage of smaller groups in a PA activity (van den Berg et al., 2006; Pelati, Grion, Li, & Serbati, 2020). This finding was confirmed for feedback-giving groups, as feedback from larger

groups was less likely to be implemented. The feedback-giving process may have been more challenging in bigger groups resulting in lower-quality feedback. However, bigger feedback-receiving groups were advantageous for feedback implementation. Discussions about the meaning of feedback and strategies to implement it were probably better facilitated with more students in a group. These results have practical implications for PA design. Smaller groups would be more appropriate if the group PA activity's primary goal is to develop feedback-giving skills. However, if artifact improvement is the aim of the activity, bigger group size should be considered. Another strategy would be integrating externally prompted regulation to improve collaboration in feedback-giving groups (Cho & Lim, 2017).

Previous PA experience in the feedback provider group did not positively influence feedback implementation or revision, as reported in Study 2. Although previous research found that students in any PA condition improved their skills with repetition (Gielen & De Wever, 2015), negative experiences with PA can affect the extent of future PA participation (Zong, Schunn, & Wang, 2022). Details of previous PA activities were not available in this study; students may not have enjoyed PA in the past, or previous PA activities did not help them develop the feedback skills necessary to give appropriate feedback.

The overall amount of praise-only or praise with summary had a negative effect on full feedback implementation and revision in Study 2, in line with the results reported by Wu and Schunn (2021) and Wu and Schunn (2020), but contradicting Patchan et al. (2016). In addition, mitigating praise was found to affect implementation and revision positively. This finding contradicts previous research, where mitigating praise reduced feedback implementation (Patchan et al., 2016) or did not predict feedback implementation (Wu & Schunn, 2021). An explanation for the positive influence of mitigating praise could be sociocultural conventions of giving and receiving feedback (Ramani, Könings, Ginsburg, & van der Vleuten, 2019), as studies mentioned above took place at US universities in English. In contrast, this study was conducted at a Norwegian university college in Norwegian. Based on the results in Study 2, students should use praise-only or praise with summary moderately and try to incorporate praise in constructive comments. However, this recommendation should be adapted with caution due to conflicting results on praise.

The negative effect of praise on feedback implementation could have been connected to another finding in Study 2: that high-quality drafts received a lot of praise-only or praise and summary. Groups with high-quality drafts tended to not revise their essays and did

not implement much feedback. Possibly, groups with high-quality drafts disagreed with peer feedback, as also reported by Wu and Schunn (2020); however, this should not affect draft revisions, and some essay improvements should be expected. This poses a challenge for instructors to motivate students with high-quality drafts to engage more in a PA activity. One strategy would be to focus on developing feedback literacy skills during the PA training (Winstone, Mathlin, & Nash, 2019). Another option would be to integrate regulation scripts with pre-structured dialogue and reflection questions to support groups struggling to trust peer feedback (Cheng, Li, Su, & Gao, 2022).

Study 1 found that the perceived peer feedback usefulness was more connected to feedback relevance rather than a lack of kindness or justification. In addition, students perceived feedback as more useful if they agreed with it. Although Mulliner and Tucker (2017) reported that students' perception of feedback is similar to the instructor's, Huisman, Saab, van Driel, and van den Broek (2018) determined that student feedback perception did not influence writing performance. The lack of context data in Study 1 limited the ability to investigate the effect of students' perception on other aspects of PA activity, such as implementation. Study 2 indicated that grading feedback usefulness did not increase the odds of students implementing feedback in contrast to writing a BE comment. This result contradicts (van der Pol et al., 2008), who found that high feedback implementation was associated with high BE grades. Developing feedback skills is integral for PA and can be facilitated by integrating a BE activity into the PA activity. Study 2 suggests the importance of encouraging students to engage with the feedback received by writing a BE comment rather than just grading. In addition, BE activity should be emphasized during PA training or become an obligatory part of a PA activity to increase its effect (Patchan et al., 2018).

5.4 Evaluation of the research approach

This research applied a mixed-method approach, which means that the research *reliability*, “the consistency of collecting, analyzing, and interpreting the data” (Zohrabi, 2013, p. 260), and *validity*, “the extent to which a concept is accurately measured” (Heale & Twycross, 2015, p. 66), must be addressed for both quantitative and qualitative methods. To ensure credibility, a detailed description of the methods used in this research is presented in Chapter 3.

Scoping review 1 and Scoping review 2 followed a structured approach, documenting every step in detail. In addition, the coding schemes and inclusion/exclusion of several papers were discussed with a senior researcher.

As a relatively new field, LA is still considered to be in the “proof of concept” phase (Ferguson, Clow, Griffiths, & Brasher, 2019b). As a consequence, appropriate evaluations of reliability and validity are still emerging. The nature of digital traces at the LA field’s foundation can be unreliable since log data is considered second-order proxies (Atkisson & Wiley, 2011). The intentionality of students’ actions is assumed, and many student behaviours are not recorded in the log data if they happen outside an online platform (Guzmán-Valenzuela et al., 2021). If data is not reliable, it cannot be considered valid and should not guide any interventions (Winne, 2020). The data limitations and the challenges with the analysis were clearly described for Study 1 and Study 2 to ensure transparency. The reliability of text data coding was achieved through an iterative process involving multiple researchers to negotiate the codes’ meaning and to develop additional coding rules, if necessary. At the same time, Cohen’s kappa was calculated to ensure coding validity. In addition, the development of indicators was well-documented, including preprocessing of raw data. Finally, the statistical models were evaluated to check if they fit the data well and if they did not validate any statistical assumptions.

Predictive modelling poses several challenges for LA research. Low data quality or a shallow understanding of the data can lead to a lack of trustworthiness. Learning processes are dynamic, and applying results from historic data to current learners can be inaccurate and lead to the over-fitting or under-fitting of statistical models (Mathrani et al., 2021). To mitigate these challenges, reliability and validity claims should be approached from a point-in-time view, implying a temporality of the findings’ accuracy (Winne, 2020). Finally, there is a potential that new technological developments, such as bootstrapping, will help validate LA results (Koedinger, McLaughlin, Jia, & Bier, 2016; Winne, 2020). Considering all these issues, the findings from Study 1 and Study 2 were situated within the previous PA research and are to be understood as accurate in a point-in-time context. Thus, the results can be considered to improve PA activity, but due to the dynamic nature of LA, new modelling may be needed for future contexts. At the same time, detailed descriptions of the analysis process can be used to replicate this research in other contexts.

5.5 Limitations

There are several limitations to this research that must be considered.

First, the lack of context data in Study 1 narrowed the analysis options. It did not allow for a deeper analysis of student perception of feedback and its relation to other aspects of PA.

Second, the marginal influence of the researcher on the design of the PA activity in Study 2 resulted in limited data on BE, which may have been an interesting comparison between Study 1 and Study 2. This could have been mitigated by suggesting to the instructor to emphasize the BE activity during the PA training or to make this part of the PA activity obligatory.

Third, no additional study implementing LA insights from Study 2 to influence PA activity in future iterations was conducted. The iterative process of LA is crucial; however, the in-depth manual coding of text data in Study 2 was resource- and time-consuming and limited the possibility of a follow-up study due to time constraints.

Finally, this research focuses on PA, an underresearched topic in the field of LA; hence, two first-ever scoping reviews on the topics on LA in PA and BE were part of this research. This contributed to the exploratory nature of the research presented, as work in this area is still emerging.

5.6 Conclusions and future work

Digital data have the potential to contribute new insights into learning and assessment processes, as well as to improve them. Learning analytics can help find hidden patterns and insights in data from PA activities that are not easily accessible to humans without the help of algorithms. However, such potential has to be considered in light of the data quality from which these insights come. This research showed the difficulties of working with context-free and context-rich datasets from a commercial PA platform and how these challenges were mitigated in the data analysis and study design.

Future research in LA should focus on longitudinal studies, where LA insights could be tested in new iterations and/or new contexts. Online PA platforms open new possibilities to examine the influence of various features, such as integration of BE activity or changes in the rubric, to improve the PA activity. Learning analytics could be an appropriate technique to build complex models examining the effects of different PA

learning designs in online spaces. However, future research should consider providing descriptions of how data availability influenced the decision-making process of the analysis. This would ensure transparency and be an essential resource as the LA field matures.

References

- Adachi, C., Tai, J., & Dawson, P. (2018a). Academics' perceptions of the benefits and challenges of self and peer assessment in higher education. *Assessment & Evaluation in Higher Education*, 43(2), 294–306.
- Adachi, C., Tai, J., & Dawson, P. (2018b). A framework for designing, implementing, communicating and researching peer assessment. *Higher Education Research & Development*, 37(3), 453–467.
- 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).
- Alexandron, G., Yoo, L. Y., Ruipérez-Valiente, J. A., Lee, S., & Pritchard, D. E. (2019). Are MOOC learning analytics results trustworthy? With fake learners, they might not be! *International Journal of Artificial Intelligence in Education*, 29(4), 484–506.
- Arksey, H., & O'Malley, L. (2005). Scoping studies: Towards a methodological framework. *International Journal of Social Research Methodology*, 8(1), 19–32.
- Ashton, S., & Davies, R. S. (2015). Using scaffolded rubrics to improve peer assessment in a mooc writing course. *Distance Education*, 36(3), 312–334.
- Atkisson, M., & Wiley, D. (2011). Learning analytics as interpretive practice: Applying Westerman to educational intervention. In *Proceedings of the 1st International Conference on Learning Analytics and Knowledge* (pp. 117–121).
- Babik, D., Gehringer, E. F., Kidd, J., Pramudianto, F., & Tinapple, D. (2016). Probing the landscape: Toward a systematic taxonomy of online peer assessment systems in education. *Teaching & Learning Faculty Publications*, 2.
- Badea, G., & Popescu, E. (2022). A dynamic review allocation approach for peer assessment in technology enhanced learning. *Education and Information Technologies*, 1–32.
- Baig, M. I., Shuib, L., & Yadegaridehkordi, E. (2020). Big data in education: a state of the art, limitations, and future research directions. *International Journal of Educational Technology in Higher Education*, 17(1), 1–23.
- 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.
- Banihashem, K., & Macfadyen, L. P. (2021). Pedagogical design: Bridging learning theory and learning analytics. *Canadian Journal of Learning and Technology*, 47(1).
- Bartimote, K., Pardo, A., & Reimann, P. (2018). The perspective realism brings to learning analytics in the classroom. In R. J. Lodge, J. Horvath, & L. Corrin (Eds.), *Learning Analytics in the Classroom* (pp. 22–42). Routledge.
- Berland, M., Baker, R. S., & Blikstein, P. (2014). Educational data mining and learning analytics: Applications to constructionist research. *Technology, Knowledge and*

- Learning*, 19(1), 205–220.
- Bicans, J., & Grundspenkis, J. (2017). Student learning style extraction from on-campus learning context data. *Procedia Computer Science*, 104, 272–278.
- Boud, D., & Molloy, E. (2013). Rethinking models of feedback for learning: The challenge of design. *Assessment & Evaluation in Higher Education*, 38(6), 698–712.
- Brant, R. (1990). Assessing proportionality in the proportional odds model for ordinal logistic regression. *Biometrics*, 1171–1178.
- Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2), 77–101.
- Bridges, S. M., Chan, L. K., Chen, J. Y., Tsang, J. P., & Ganotice, F. A. (2020). Learning environments for interprofessional education: A micro-ethnography of sociomaterial assemblages in team-based learning. *Nurse Education Today*, 94, 104569.
- Buchanan, E., Geshler, A., & Hammer, P. (2015). Privacy, security, and ethics. *Data-intensive Research in Education: Current Work and Next steps*, 89–98.
- Buckingham Shum, S., & Ferguson, R. (2012). Social learning analytics. *Journal of Educational Technology & Society*, 15(3), 3–26.
- Burla, L., Knierim, B., Barth, J., Liewald, K., Duetz, M., & Abel, T. (2008). From text to codings: Intercoder reliability assessment in qualitative content analysis. *Nursing Research*, 57(2), 113–117.
- Carter, M., & Egliston, B. (2021). What are the risks of virtual reality data? Learning analytics, algorithmic bias and a fantasy of perfect data. *New Media & Society*, 1–20.
- Cartney, P. (2014). Exploring the use of peer assessment as a vehicle for closing the gap between feedback given and feedback used. In S. Hatzipanagos & R. Rochon (Eds.), *Approaches to Assessment that Enhance Learning in Higher Education* (pp. 71–84). Routledge.
- Charitopoulos, A., Rangoussi, M., & Koulouriotis, D. (2020). On the use of soft computing methods in educational data mining and learning analytics research: A review of years 2010–2018. *International Journal of Artificial Intelligence in Education*, 30(3), 371–430.
- Chatti, M. A., Dyckhoff, A. L., Schroeder, U., & Thüs, H. (2012). A Reference Model for Learning Analytics. *International Journal of Technology Enhanced Learning*, 4(5-6), 318–331.
- Chen, W. (2017). Knowledge convergence among pre-service mathematics teachers through online reciprocal peer feedback. *Knowledge Management & E-Learning: An International Journal*, 9(1), 1–18.
- Chen, X., Zou, D., & Xie, H. (2022). A decade of learning analytics: Structural topic modeling based bibliometric analysis. *Education and Information Technologies*, 1–45.
- Cheng, J., & Lei, J. (2021). A description of students' commenting behaviours in an online blogging activity. *E-Learning and Digital Media*, 18(2), 209–225.

- Cheng, L., Li, Y., Su, Y., & Gao, L. (2022). Effect of regulation scripts for dialogic peer assessment on feedback quality, critical thinking and climate of trust. *Assessment & Evaluation in Higher Education*, 1–13.
- Cherryholmes, C. H. (1992). Notes on pragmatism and scientific realism. *Educational Researcher*, 21(6), 13–17.
- Ching, Y.-H., & Hsu, Y.-C. (2016). Learners' interpersonal beliefs and generated feedback in an online role-playing peer-feedback activity: An exploratory study. *International Review of Research in Open and Distributed Learning*, 17(2), 105–122.
- Chiu, H.-Y., Kang, Y.-N., Wang, W.-L., Chen, C.-C., Hsu, W., Tseng, M.-F., & Wei, P.-L. (2019). The role of active engagement of peer observation in the acquisition of surgical skills in virtual reality tasks for novices. *Journal of Surgical Education*, 76(6), 1655–1662.
- 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).
- 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.
- Cho, M.-H., & Lim, S. (2017). Using regulation activities to improve undergraduate collaborative writing on wikis. *Innovations in Education and Teaching International*, 54(1), 53–61.
- Cho, Y. H., & Cho, K. (2011). Peer reviewers learn from giving comments. *Instructional Science*, 39(5), 629–643.
- Choi, H., Dowell, N., Brooks, C., & Teasley, S. (2019). Social comparison in moocs: Perceived ses, opinion, and message formality. In *Proceedings of the 9th International Conference on Learning Analytics and Knowledge* (pp. 160–169).
- Chu, X., Ilyas, I. F., Krishnan, S., & Wang, J. (2016). Data cleaning: Overview and emerging challenges. In *Proceedings of the 2016 International Conference on Management of Data* (pp. 2201–2206).
- Clow, D. (2012). The learning analytics cycle: Closing the loop effectively. In *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge* (pp. 134–138).
- Clow, D. (2013). An overview of learning analytics. *Teaching in Higher Education*, 18(6), 683–695.
- Cook, A. (2019). *Using interactive learning activities to address challenges of peer feedback systems* (Unpublished doctoral dissertation). University of California, San Diego.
- Cooperstein, S. E., & Kocevar-Weidinger, E. (2004). Beyond active learning: A constructivist approach to learning. *Reference Services Review*, 32(2), 141–148.
- Cormack, A. N. (2016). A data protection framework for learning analytics. *Journal of Learning Analytics*, 3(1), 91–106.
- Creswell, J. W., & Plano Clark, V. L. (2011). Choosing a mixed methods design. *Designing*

- and Conducting Mixed Methods Research*, 2, 53–106.
- Daniel, B. K. (2019). Big data and data science: A critical review of issues for educational research. *British Journal of Educational Technology*, 50(1), 101–113.
- 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).
- DeMaris, A. (1995). A tutorial in logistic regression. *Journal of Marriage and the Family*, 57(4), 956–968.
- Dey, A. K. (2001). Understanding and using context. *Personal and Ubiquitous Computing*, 5(1), 4–7.
- Dietrichson, A. (2013). Beyond clickometry: Analytics for constructivist pedagogies. *International Journal on E-Learning*, 12(4), 333–351.
- Divjak, B., & Maretic, M. (2015). Learning analytics for e-assessment: The state of the art and one case study. In *Proceedings of the 2015 Central European Conference on Information and Intelligent Systems*.
- Djelil, F., Brisson, L., Charbey, R., Bothorel, C., Gilliot, J.-M., & Ruffieux, P. (2021). Analysing peer assessment interactions and their temporal dynamics using a graphlet-based method. In *Proceedings of the 16th European Conference on Technology Enhanced Learning* (pp. 82–95).
- 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(2), 481–509.
- Douma, J. C., & Weedon, J. T. (2019). Analysing continuous proportions in ecology and evolution: A practical introduction to beta and dirichlet regression. *Methods in Ecology and Evolution*, 10(9), 1412–1430.
- Doyle, L., Brady, A.-M., & Byrne, G. (2009). An overview of mixed methods research. *Journal of Research in Nursing*, 14(2), 175–185.
- Drachsler, H., Hoel, T., Scheffel, M., Kismihók, G., Berg, A., Ferguson, R., ... Manderveld, J. (2015). Ethical and privacy issues in the application of learning analytics. In *Proceedings of the 5th International Conference on Learning Analytics and Knowledge* (pp. 390–391).
- Dressler, R., Chu, M.-W., Crossman, K., & Hilman, B. (2019). Quantity and quality of uptake: Examining surface and meaning-level feedback provided by peers and an instructor in a graduate research course. *Assessing Writing*, 39, 14–24.
- Dringus, L. P. (2012). Learning analytics considered harmful. *Journal of Asynchronous Learning Networks*, 16(3), 87–100.
- Du, X., Yang, J., Shelton, B. E., Hung, J.-L., & Zhang, M. (2021). A systematic meta-review and analysis of learning analytics research. *Behaviour & Information Technology*, 40(1), 49–62.

- Eickhoff, M., & Wieneke, R. (2018). Understanding topic models in context: a mixed-methods approach to the meaningful analysis of large document collections. In *Proceedings of the 51st Hawaii International Conference on System Sciences* (pp. 903–912).
- Er, E., Villa-Torrano, C., Dimitriadis, Y., Gasevic, D., Bote-Lorenzo, M. L., Asensio-Pérez, J. I., ... Martínez Monés, A. (2021). Theory-based learning analytics to explore student engagement patterns in a peer review activity. In *Proceedings of the 11th International Learning Analytics and Knowledge Conference* (pp. 196–206).
- Ertmer, P. A., Richardson, J. C., Lehman, J. D., Newby, T. J., Cheng, X., Mong, C., & Sadaf, A. (2010). Peer feedback in a large undergraduate blended course: Perceptions of value and learning. *Journal of Educational Computing Research*, 43(1), 67–88.
- Espasa, A., Guasch, T., Mayordomo, R., Martinez-Melo, M., & Carless, D. (2018). A dialogic feedback index measuring key aspects of feedback processes in online learning environments. *Higher Education Research & Development*, 37(3), 499–513.
- 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.
- Fancsali, S. E. (2011). Variable construction for predictive and causal modeling of online education data. In *Proceedings of the 1st International Learning Analytics and Knowledge Conference* (pp. 54–63).
- Feilzer, M. Y. (2010). Doing mixed methods research pragmatically: Implications for the rediscovery of pragmatism as a research paradigm. *Journal of Mixed Methods Research*, 4(1), 6–16.
- Ferguson, R. (2019). Ethical challenges for learning analytics. *Journal of Learning Analytics*, 6(3), 25–30.
- Ferguson, R., Clow, D., Griffiths, D., & Brasher, A. (2019b). Moving forward with learning analytics: Expert views. *Journal of Learning Analytics*, 6(3), 43–59.
- Ferguson, R., Coughlan, T., Egelandsdal, K., Gaved, M., Herodotou, C., Hillaire, G., ... Whitelock, D. (2019a). *Innovating Pedagogy 2019 (Open University Innovation Report 7)*. The Open University.
- Fertalj, M., Brkić, L. J., & Mekterović, I. (2022). A systematic review of peer assessment approaches to evaluation of open-ended student assignments. In *Proceedings of the 45th Jubilee International Convention on Information, Communication and Electronic Technology* (pp. 1076–1081).
- Fincham, E., Whitelock-Wainwright, A., Kovanović, V., Joksimović, S., van Staaldin, J.-P., & Gašević, D. (2019). Counting clicks is not enough: Validating a theorized model of engagement in learning analytics. In *Proceedings of the 9th International Conference on Learning Analytics and Knowledge* (pp. 501–510).
- Fischer, C., Pardos, Z. A., Baker, R. S., Williams, J. J., Smyth, P., Yu, R., ... Warschauer, M. (2020). Mining big data in education: Affordances and challenges. *Review of Research*

- in Education*, 44(1), 130–160.
- Fisher, M. J., & Marshall, A. P. (2009). Understanding descriptive statistics. *Australian Critical Care*, 22(2), 93–97.
- Formanek, M., Wenger, M. C., Buxner, S. R., Impey, C. D., & Sonam, T. (2017). Insights about large-scale online peer assessment from an analysis of an astronomy mooc. *Computers & Education*, 113, 243–262.
- Fosnot, C. T., & Perry, R. S. (1996). Constructivism: A psychological theory of learning. *Constructivism: Theory, Perspectives, and Practice*, 2(1), 8–33.
- French, B. F., Immekus, J. C., & Yen, H.-J. (2013). Logistic regression. In T. Teo (Ed.), *Handbook of Quantitative Methods for Educational Research* (pp. 145–165). Brill.
- Funk, A. L., Wichmann, A., & Rummel, N. (2013). Supporting feedback uptake in online peer assessment. In *Proceedings of the 10th International Conference on Computer-Supported Collaborative Learning. Volume 2: Short Papers, Panels, Posters, Demos, & Community Events* (pp. 253–254). International Society of the Learning Sciences.
- Gamage, D., Staubitz, T., & Whiting, M. (2021). Peer assessment in MOOCs: Systematic literature review. *Distance Education*, 42(2), 268–289.
- Gielen, M., & De Wever, B. (2015). Structuring peer assessment: Comparing the impact of the degree of structure on peer feedback content. *Computers in Human Behavior*, 52, 315–325.
- Gielen, S., Dochy, F., & Onghena, P. (2011). An inventory of peer assessment diversity. *Assessment & Evaluation in Higher Education*, 36(2), 137–155.
- Gielen, S., Peeters, E., Dochy, F., Onghena, P., & Struyven, K. (2010). Improving the effectiveness of peer feedback for learning. *Learning and Instruction*, 20(4), 304–315.
- Greller, W., & Drachslar, H. (2012). Translating learning into numbers: A generic framework for learning analytics. *Journal of Educational Technology & Society*, 15(3), 42.
- Gunnarsson, B. L., & Alterman, R. (2014). Peer promotions as a method to identify quality content. *Journal of Learning Analytics*, 1(2), 126–150.
- Guzmán-Valenzuela, C., Gómez-González, C., Rojas-Murphy Tagle, A., & Lorca-Vyhmeister, A. (2021). Learning analytics in higher education: a preponderance of analytics but very little learning? *International Journal of Educational Technology in Higher Education*, 18(1), 1–19.
- 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.
- Harrell Jr, F. E., & Harrell Jr, M. F. E. (2019). Package ‘hmisc’. *CRAN2018, 2019*, 235–236.
- Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, 77(1), 81–112.
- Heale, R., & Twycross, A. (2015). Validity and reliability in quantitative studies. *Evidence-based Nursing*, 18(3), 66–67.

- Healy, M. (1995). Statistics from the inside. 15. Multiple regression (1). *Archives of Disease in Childhood*, 73(2), 177.
- Heyman, J. E., & Sailors, J. J. (2011). Peer assessment of class participation: Applying peer nomination to overcome rating inflation. *Assessment & Evaluation in Higher Education*, 36(5), 605–618.
- 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 Conference on Human Factors in Computing Systems* (pp. 458–469).
- 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.
- Huisman, B., Saab, N., van Driel, J., & van den Broek, P. (2018). Peer feedback on academic writing: Undergraduate students' peer feedback role, peer feedback perceptions and essay performance. *Assessment & Evaluation in Higher Education*, 43(6), 955–968.
- Hutto, C., & Gilbert, E. (2014). Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the 8th International AAAI Conference on Web and Social Media* (pp. 216–225).
- Ifenthaler, D., & Yau, J. Y.-K. (2020). Utilising learning analytics to support study success in higher education: a systematic review. *Educational Technology Research and Development*, 68(4), 1961–1990.
- 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(6), 5205–5234.
- Kaliisa, R., Misiejuk, K., Irgens, G. A., & Misfeldt, M. (2021). Scoping the emerging field of quantitative ethnography: Opportunities, challenges and future directions. In *Proceedings of the 2nd International Conference on Quantitative Ethnography* (pp. 3–17).
- Khalil, M., & Ebner, M. (2015). Learning analytics: Principles and constraints. In *Proceedings of the 2015 World Conference on Educational Multimedia, Hypermedia and Telecommunications* (pp. 1326–1336).
- Khosravi, H., Gyamfi, G., Hanna, B. E., & Lodge, J. (2020). Fostering and supporting empirical research on evaluative judgement via a crowdsourced adaptive learning system. In *Proceedings of the 10th international conference on Learning Analytics and Knowledge* (pp. 83–88).
- Kitchin, R. (2014). Big data, new epistemologies and paradigm shifts. *Big Data & Society*, 1(1), 1–12.
- Kitchin, R., & Lauriault, T. P. (2015). Small data in the era of big data. *GeoJournal*, 80(4), 463–475.
- Klašnja-Milićević, A., Ivanović, M., & Budimac, Z. (2017). Data science in education: Big

- data and learning analytics. *Computer Applications in Engineering Education*, 25(6), 1066–1078.
- Kleinbaum, D. G., & Klein, M. (2010). *Ordinal Logistic Regression*. Springer.
- Klucsevsek, K. M. (2016). Transferring skills from classroom to professional writing: Student-faculty peer review as an extension of cognitive apprenticeship. *Journal of the Scholarship of Teaching and Learning*, 16(6), 106–123.
- Knight, S., Buckingham Shum, S., & Littleton, K. (2014). Epistemology, assessment, pedagogy: Where learning meets analytics in the middle space. *Journal of Learning Analytics*, 1(2), 23–47.
- Koedinger, K. R., McLaughlin, E. A., Jia, J. Z., & Bier, N. L. (2016). Is the doer effect a causal relationship? How can we tell and why it's important. In *Proceedings of the 6th International Conference on Learning Analytics and Knowledge* (pp. 388–397).
- Krumm, A., Means, B., & Bienkowski, M. (2018). *Learning Analytics Goes to School: A Collaborative Approach to Improving Education*. Routledge.
- Levac, D., Colquhoun, H., & O'Brien, K. K. (2010). Scoping studies: Advancing the methodology. *Implementation Science*, 5(1), 1–9.
- Li, H., Bialo, J. A., Xiong, Y., Hunter, C. V., & Guo, X. (2021). The effect of peer assessment on non-cognitive outcomes: A meta-analysis. *Applied Measurement in Education*, 34(3), 179–203.
- 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., L. Kornhaber, M., Lyu, Y., Chung, K. S., & K. Suen, H. (2016). Peer assessment in the digital age: A meta-analysis comparing peer and teacher ratings. *Assessment & Evaluation in Higher Education*, 41(2), 245–264.
- Li, L., Liu, X., & Zhou, Y. (2012). Give and take: A re-analysis of assessor and assessee's roles in technology-facilitated peer assessment. *British Journal of Educational Technology*, 43(3), 376–384.
- 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.
- Lin, C.-J. (2019). An online peer assessment approach to supporting mind-mapping flipped learning activities for college english writing courses. *Journal of Computers in Education*, 6(3), 385–415.
- 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), 286–297.
- Liu, N.-F., & Carless, D. (2006). Peer feedback: the learning element of peer assessment. *Teaching in Higher Education*, 11(3), 279–290.
- Liu, X., Li, L., & Zhang, Z. (2018). Small group discussion as a key component in

- online assessment training for enhanced student learning in web-based peer assessment. *Assessment & Evaluation in Higher Education*, 43(2), 207–222.
- Lundstrom, K., & Baker, 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(1), 30–43.
- Luxton-Reilly, A. (2009). A systematic review of tools that support peer assessment. *Computer Science Education*, 19(4), 209–232.
- Lynch, R., McNamara, P. M., & Seery, N. (2012). Promoting deep learning in a teacher education programme through self-and peer-assessment and feedback. *European Journal of Teacher Education*, 35(2), 179–197.
- Macfadyen, L. P., Dawson, S., Pardo, A., & Gašević, D. (2014). Embracing big data in complex educational systems: The learning analytics imperative and the policy challenge. *Research & Practice in Assessment*, 9, 17–28.
- Mangaroska, K., & Giannakos, M. (2018). Learning analytics for learning design: A systematic literature review of analytics-driven design to enhance learning. *IEEE Transactions on Learning Technologies*, 12(4), 516–534.
- Mathrani, A., Susnjak, T., Ramaswami, G., & Barczak, A. (2021). Perspectives on the challenges of generalizability, transparency and ethics in predictive learning analytics. *Computers and Education Open*, 2, 1–9.
- Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, 5(4), 1093–1113.
- Merceron, A., Blikstein, P., & Siemens, G. (2015). Learning analytics: From big data to meaningful data. *Journal of Learning Analytics*, 2(3), 4–8.
- Misiejuk, K. (2017). *Mapping the field of educational data sciences. analysis of the proceedings from educational data mining, learning analytics and knowledge, and learning at scale* (Unpublished master's thesis). Humboldt University of Berlin, Germany.
- Misiejuk, K., Bastesen, J., Wasson, B., & Krange, I. (submitted). Student implementation of peer feedback: Increasing insights with context data. *Assessment & Evaluation in Higher Education*.
- Misiejuk, K., Ness, I. J., Gray, R., & Wasson, B. (submitted). Changes in online course designs: Before, during, and after the pandemic. *Frontiers in Education*.
- Misiejuk, K., Scianna, J., Kaliisa, R., Vachuska, K., & Shaffer, D. W. (2021). Incorporating sentiment analysis with epistemic network analysis to enhance discourse analysis of twitter data. In *Proceedings of the 2nd International Conference on Quantitative Ethnography* (pp. 375–389).
- Misiejuk, K., & Wasson, B. (2017). *State of the Field Report on Learning Analytics. SLATE Report 2017-2*. Centre for the Science of Learning & Technology (SLATE), University of Bergen.
- Misiejuk, K., & Wasson, B. (2021). Backward evaluation in peer assessment: A scoping

- review. *Computers & Education*, 175, 1-12.
- Misiejuk, K., & Wasson, B. (in press). Learning analytics for peer assessment - A scoping review. In O. Noroozi & B. De Wever (Eds.), *The Power of Peer Learning*. Springer.
- Misiejuk, K., Wasson, B., & Egelandstal, K. (2021). Using learning analytics to understand student perceptions of peer feedback. *Computers in Human Behavior*, 117, 1-13.
- Mørch, A. I., Engeness, I., Cheng, V. C., Cheung, W. K., & Wong, K. C. (2017). Essaycritic: Writing to learn with a knowledge-based design critiquing system. *Journal of Educational Technology & Society*, 20(2), 213–223.
- Motlagh, L. N. (2015). Who do learners prefer to be corrected by? Teachers or classmates? *Procedia-Social and Behavioral Sciences*, 199, 381–386.
- Mukaka, M. M. (2012). A guide to appropriate use of correlation coefficient in medical research. *Malawi Medical Journal*, 24(3), 69–71.
- Mullen, L. (2015). *Textreuse: Detect text reuse and document similarity*. rOpenSci.
- Mulliner, E., & Tucker, M. (2017). Feedback on feedback practice: perceptions of students and academics. *Assessment & Evaluation in Higher Education*, 42(2), 266–288.
- Munn, Z., Peters, M. D., Stern, C., Tufanaru, C., McArthur, A., & Aromataris, E. (2018). Systematic review or scoping review? Guidance for authors when choosing between a systematic or scoping review approach. *BMC Medical Research Methodology*, 18(1), 1–7.
- Nelson, M., & Schunn, C. D. (2009). The nature of feedback: How different types of peer feedback affect writing performance. *Instructional Science*, 37(4), 375–401.
- Nilson, L. B. (2003). Improving student peer feedback. *College Teaching*, 51(1), 34–38.
- Panadero, E., Fraile, J., Pinedo, L., Rodríguez-Hernández, C., & Díez, F. (2022). Changes in classroom assessment practices during emergency remote teaching due to COVID-19. *Assessment in Education: Principles, Policy & Practice*, 1–22.
- Panadero, E., Romero, M., & Strijbos, J.-W. (2013). The impact of a rubric and friendship on peer assessment: Effects on construct validity, performance, and perceptions of fairness and comfort. *Studies in Educational Evaluation*, 39(4), 195–203.
- Pardo, A., & Siemens, G. (2014). Ethical and privacy principles for learning analytics. *British Journal of Educational Technology*, 45(3), 438–450.
- Paré, G., & Kitsiou, S. (2017). Methods for literature reviews. In F. Lau & C. Kuziemsky (Eds.), *Handbook of eHealth Evaluation: An Evidence-based Approach*. University of Victoria.
- Parks, L., & Peters, W. (2022). Natural language processing in mixed-methods text analysis: A workflow approach. *International Journal of Social Research Methodology*, 1–13.
- Patchan, M. M., Charney, D., & Schunn, C. D. (2009). A validation study of students' end comments: Comparing comments by students, a writing instructor, and a content instructor. *Journal of Writing Research*, 1(2), 124–152.
- Patchan, M. M., & Schunn, C. D. (2015). Understanding the benefits of providing peer

- feedback: How students respond to peers' texts of varying quality. *Instructional Science*, 43(5), 591–614.
- 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.
- Patchan, M. M., Schunn, C. D., & Correnti, R. J. (2016). The nature of feedback: How peer feedback features affect students' implementation rate and quality of revisions. *Journal of Educational Psychology*, 108(8), 1098.
- Pelati, C., Grion, V., Li, L., & Serbati, A. (2020). Peer assessment practices in an online context: does the group size matter? *Form@re - Open Journal per la Formazione in Rete*, 20(1), 143–153.
- Peng, C.-Y. J., Lee, K. L., & Ingersoll, G. M. (2002). An introduction to logistic regression analysis and reporting. *The Journal of Educational Research*, 96(1), 3–14.
- Pham, M. T., Rajić, A., Greig, J. D., Sargeant, J. M., Papadopoulos, A., & McEwen, S. A. (2014). A scoping review of scoping reviews: advancing the approach and enhancing the consistency. *Research Synthesis Methods*, 5(4), 371–385.
- Planas-Lladó, A., Feliu, L., Arbat, G., Pujol, J., Suñol, J. J., Castro, F., & Martí, C. (2021). An analysis of teamwork based on self and peer evaluation in higher education. *Assessment & Evaluation in Higher Education*, 46(2), 191–207.
- Prion, S., & Haerling, K. A. (2014). Making sense of methods and measurement: Pearson product-moment correlation coefficient. *Clinical Simulation in Nursing*, 10(11), 587–588.
- Ramani, S., Könings, K. D., Ginsburg, S., & van der Vleuten, C. P. (2019). Feedback redefined: Principles and practice. *Journal of General Internal Medicine*, 34(5), 744–749.
- Reyes, J. A. (2015). The skinny on big data in education: Learning analytics simplified. *TechTrends*, 59(2), 75–80.
- Rogers, T. (2015). Critical realism and learning analytics research: epistemological implications of an ontological foundation. In *Proceedings of the 5th International Conference on Learning Analytics and Knowledge* (pp. 223–230).
- Rogers, T., Gašević, D., & Dawson, S. (2016). Learning analytics and the imperative for theory driven research. In C. Haythornthwaite, R. Andrews, J. Fransman, & E. M. Meyers (Eds.), *The SAGE Handbook of E-Learning Research* (pp. 232–250).
- Romero, C., Romero, J. R., & Ventura, S. (2014). A survey on pre-processing educational data. In A. Peña-Ayala (Ed.), *Educational Data Mining* (pp. 29–64). Springer.
- Roschelle, J., & Krumm, A. (2015). Infrastructures for improving learning in information-rich classrooms. In P. Reimann, S. Bull, M. Kickmeier-Rust, R. Vatrupu, & B. Wasson (Eds.), *Measuring and Visualizing Learning in the Information-Rich Classroom* (pp. 19–26). Routledge.

- Ryan, T., Gašević, D., & Henderson, M. (2019). Identifying the impact of feedback over time and at scale: Opportunities for learning analytics. In M. Henderson, R. Ajjawi, D. Boud, & E. Molloy (Eds.), *The Impact of Feedback in Higher Education* (pp. 207–223). Springer.
- Samuelsen, J., Chen, W., & Wasson, B. (2021). Enriching context descriptions for enhanced LA scalability: A case study. *Research and Practice in Technology Enhanced Learning*, 16(1), 1–26.
- Sanchez, C. E., Atkinson, K. M., Koenka, A. C., Moshontz, H., & Cooper, H. (2017). Self-grading and peer-grading for formative and summative assessments in 3rd through 12th grade classrooms: A meta-analysis. *Journal of Educational Psychology*, 109(8), 1049.
- Sedrakyan, G., Snoeck, M., & De Weerd, J. (2014). Process mining analysis of conceptual modeling behavior of novices—empirical study using jmermaid modeling and experimental logging environment. *Computers in Human Behavior*, 41, 486–503.
- Selwyn, N. (2019). What’s the problem with learning analytics? *Journal of Learning Analytics*, 6(3), 11–19.
- Şenel, S., & Şenel, H. C. (2021). Remote assessment in higher education during COVID-19 pandemic. *International Journal of Assessment Tools in Education*, 181–199.
- 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).
- Shoukri, M. M., & Aleid, M. M. (2022). Quasi-binomial regression model for the analysis of data with extra-binomial variation. *Open Journal of Statistics*, 12(1), 1–14.
- Siemens, G. (2013). Learning analytics: The emergence of a discipline. *American Behavioral Scientist*, 57(10), 1380–1400.
- Slade, S., & Prinsloo, P. (2013). Learning analytics: Ethical issues and dilemmas. *American Behavioral Scientist*, 57(10), 1510–1529.
- Sokhanvar, Z., Salehi, K., & Sokhanvar, F. (2021). Advantages of authentic assessment for improving the learning experience and employability skills of higher education students: A systematic literature review. *Studies in Educational Evaluation*, 70, 1–10.
- Song, D. (2018). Learning analytics as an educational research approach. *International Journal of Multiple Research Approaches*, 10(1), 102–111.
- Song, Y., Hu, Z., Guo, Y., & Gehringer, E. F. (2016). An experiment with separate formative and summative rubrics in educational peer assessment. In *Proceedings of the 2016 IEEE Frontiers in Education Conference* (pp. 1–7).
- Sun, N., Lavoué, E., Aritajati, C., Tabard, A., & Rosson, M. B. (2019). Using and perceiving emoji in design peer feedback. In *Proceedings of the 13th International Conference on Computer-Supported Collaborative Learning* (pp. 296–303).
- Swinscow, T. D. V., & Campbell, M. J. (1997). *Statistics at square one*. BMJ.
- Sykes, A. O. (1993). *An Introduction to Regression Analysis*. Coase-Sandor Institute for Law

- & *Economics Working Paper No. 20*. University of Chicago Law School.
- 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.
- Tashakkori, A., & Teddlie, C. (1998). *Mixed Methodology: Combining Qualitative and Quantitative Approaches* (Vol. 46). SAGE.
- Tashakkori, A., & Teddlie, C. (2003). Issues and dilemmas in teaching research methods courses in social and behavioural sciences: US perspective. *International Journal of Social Research Methodology*, 6(1), 61–77.
- Thomas, D. R. (2006). A general inductive approach for analyzing qualitative evaluation data. *American Journal of Evaluation*, 27(2), 237–246.
- 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.
- Topping, K. (2021). Peer assessment: Channels of operation. *Education Sciences*, 11(3), 91.
- Tsai, C.-C. (2001). The interpretation construction design model for teaching science and its applications to internet-based instruction in taiwan. *International Journal of Educational Development*, 21(5), 401–415.
- Tsai, C.-C., Lin, S. S., & Yuan, S.-M. (2002). Developing science activities through a networked peer assessment system. *Computers & Education*, 38(1-3), 241–252.
- 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).
- van den Berg, I., Admiraal, W., & Pilot, A. (2006). Peer assessment in university teaching: evaluating seven course designs. *Assessment & Evaluation in Higher Education*, 31(1), 19–36.
- van der Pol, J., van den Berg, B., 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*. Springer.
- Viberg, O., Hatakka, M., Bälter, O., & Mavroudi, A. (2018). The current landscape of learning analytics in higher education. *Computers in Human Behavior*, 89, 98–110.
- Vijayarani, S., Ilamathi, M. J., & Nithya, M. (2015). Preprocessing techniques for text mining: An overview. *International Journal of Computer Science & Communication Networks*, 5(1), 7–16.
- Vogelsang, T., & Ruppertz, L. (2015). On the validity of peer grading and a cloud teaching assistant system. In *Proceedings of the 5th International Conference on Learning Analytics And Knowledge* (pp. 41–50).
- Wahid, U., Chatti, M. A., & Schroeder, U. (2016). Improving peer assessment by using

- learning analytics. In *Proceedings of DeLFI Workshops* (pp. 52–54).
- Wallace, D. L., Hayes, J. R., Hatch, J. A., Miller, W., Moser, G., & Silk, C. M. (1996). Better revision in eight minutes? prompting first-year college writers to revise globally. *Journal of Educational Psychology*, 88(4), 682.
- What is learning analytics?* (n.d.). <https://www.solaresearch.org/about/what-is-learning-analytics/>. Society for Learning Analytics Research (SoLAR). (Accessed: 03.10.2022)
- Wichmann, A., Funk, A. L., & Rummel, N. (2015). Maximizing benefit of peer-feedback to increase feedback uptake in academic writing. In *Proceedings of the 11th International Conference on Computer-Supported Collaborative Learning* (pp. 411–418).
- Williamson, B. (2017). *Big data in education: The digital future of learning, policy and practice*. SAGE.
- Wilson, M., & Scalise, K. (2016). Learning analytics: Negotiating the intersection of measurement technology and information technology. *Learning, Design, and Technology*, 1–23.
- Winne, P. H. (2020). Construct and consequential validity for learning analytics based on trace data. *Computers in Human Behavior*, 112, 1–5.
- Winstone, N. E., Mathlin, G., & Nash, R. A. (2019). Building feedback literacy: Students' perceptions of the developing engagement with feedback toolkit. *Frontiers in Education*, 4, 1–11.
- 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.
- Wise, A. F., Knight, S., & Shum, S. B. (2021). Collaborative learning analytics. In U. Cress, C. Rosé, A. Friend Wise, & J. Oshima (Eds.), *International Handbook of Computer-Supported Collaborative Learning* (pp. 425–443). Springer.
- Wise, A. F., Sarmiento, J. P., & Boothe Jr, M. (2021). Subversive learning analytics. In *Proceedings of the 11th International Learning Analytics and Knowledge Conference* (pp. 639–645).
- Wu, Y., & Schunn, C. D. (2020). 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), 492–526.
- Wu, Z. (2019). Lower english proficiency means poorer feedback performance? A mixed-methods study. *Assessing Writing*, 41, 14–24.
- Xekalaki, E. (2014). Under-and overdispersion. *Wiley StatsRef: Statistics Reference Online*, 1–9.
- Yuan, J., & Kim, C. (2015). Effective feedback design using free technologies. *Journal of*

Educational Computing Research, 52(3), 408–434.

- Yudelson, M., Fancsali, S., Ritter, S., Berman, S., Nixon, T., & Joshi, A. (2014). Better data beats big data. In *Proceedings of the 7th International Conference on Educational Data Mining* (pp. 205–208).
- Yurdabakan, I. (2011). The investigation of peer assessment in primary school cooperative learning groups with respect to gender. *International Journal of Primary, Elementary and Early Years Education*, 39(2), 153–169.
- Zhang, S. (1995). Reexamining the affective advantage of peer feedback in the esl writing class. *Journal of Second Language Writing*, 4(3), 209–222.
- Zheng, L., Zhang, X., & Cui, P. (2020). The role of technology-facilitated peer assessment and supporting strategies: A meta-analysis. *Assessment & Evaluation in Higher Education*, 45(3), 372–386.
- Zheng, Y., Ruan, S., & Li, L. (2015). A conceptual design framework for big data based learning analysis. In *Proceedings of the 2nd International Conference on Education, Management and Computing Technology* (pp. 1476–1479).
- Zohrabi, M. (2013). Mixed method research: Instruments, validity, reliability and reporting findings. *Theory and Practice in Language Studies*, 3(2), 254.
- Zong, Z., Schunn, C. D., & Wang, Y. (2022). Do experiences of interactional inequality predict lower depth of future student participation in peer review? *Computers in Human Behavior*, 127, 1–15.

Appendices

A. Co-authorship declarations



UNIVERSITY OF BERGEN
Faculty of Social Sciences

Statement of co-authorship describing the independent research contribution of the candidate

Article no: 1

Authors: Kamila Misiejuk, Barbara Wasson, Kjetil Egelandstal

Stipulated total contribution of the candidate (%): 70%

Title: Using learning analytics to understand student perceptions of peer feedback

The independent contribution of the candidate
Problem formulation and research design: <ul style="list-style-type: none">• <i>Conceptualization</i>• <i>Methodology</i>
Data collection: <ul style="list-style-type: none">• <i>Data curation</i>
Analysis and interpretation: <ul style="list-style-type: none">• <i>Software</i>• <i>Formal analysis</i>• <i>Visualization</i>
Writing/presentation: <ul style="list-style-type: none">• <i>Writing – original draft</i>• <i>Writing – review & editing</i>

 Signature of the candidate Name (bold letters): KAMILA MISIEJUK	 Signature of co-author 1 Name (bold letters): BARBARA WASSON
Any Comments:	 Signature of co-author 2 Name (bold letters): KJETIL EGELANDSDAL



UNIVERSITY OF BERGEN
Faculty of Social Sciences

**Statement of co-authorship
describing the independent research contribution of the candidate**



Article no: 2

Authors: Kamila Misiejuk, Barbara Wasson

Stipulated total contribution of the candidate (%): 80%

Title: Backward evaluation in peer assessment: A scoping review

The independent contribution of the candidate
Problem formulation and research design: <ul style="list-style-type: none"> • <i>Conceptualization</i> • <i>Methodology</i>
Data collection: <ul style="list-style-type: none"> • <i>Data curation</i>
Analysis and interpretation: <ul style="list-style-type: none"> • <i>Software</i> • <i>Formal analysis</i> • <i>Visualization</i>
Writing/presentation: <ul style="list-style-type: none"> • <i>Writing – original draft</i> • <i>Writing – review & editing</i>

 Signature of the candidate Name (bold letters): KAMILA MISIEJUK	 Signature of co-author 1 Name (bold letters): BARBARA WASSON
Any Comments:	



UNIVERSITY OF BERGEN
Faculty of Social Sciences

**Statement of co-authorship
describing the independent research contribution of the candidate**



Article no: 3

Authors: Kamila Misiejuk, Barbara Wasson

Stipulated total contribution of the candidate (%): 80%

Title: Learning analytics for peer assessment: A scoping review

The independent contribution of the candidate:
Problem formulation and research design: <ul style="list-style-type: none"> • <i>Conceptualization</i> • <i>Methodology</i>
Data collection: <ul style="list-style-type: none"> • <i>Data curation</i>
Analysis and interpretation: <ul style="list-style-type: none"> • <i>Software</i> • <i>Formal analysis</i>
Writing/presentation: <ul style="list-style-type: none"> • <i>Writing – original draft</i> • <i>Writing – review & editing</i>

 Signature of the candidate	 Signature of co-author 1
Name (bold letters): KAMILA MISIEJUK	Name (bold letters): BARBARA WASSON
Any Comments:	



UNIVERSITY OF BERGEN
Faculty of Social Sciences

**Statement of co-authorship
describing the independent research contribution of the candidate**




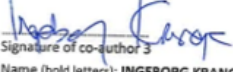
Article no: 4

Authors: Kamila Misiejuk, Jarle Bastesen, Barbara Wasson, Ingeborg Krange

Stipulated total contribution of the candidate (%): 70%

Title: Student implementation of peer feedback: Increasing insights with context data

The independent contribution of the candidate
Problem formulation and research design: <ul style="list-style-type: none"> • Conceptualization • Methodology
Data collection: <ul style="list-style-type: none"> • Data curation
Analysis and interpretation: <ul style="list-style-type: none"> • Software • Formal analysis • Visualization
Writing/presentation: <ul style="list-style-type: none"> • Writing – original draft • Writing – review & editing

 Signature of the candidate Name (bold letters): KAMILA MISIEJUK	 Signature of co-author 1 Name (bold letters): JARLE BASTESEN
Any Comments:	 Signature of co-author 2 Name (bold letters): BARBARA WASSON
	 Signature of co-author 3 Name (bold letters): INGEBORG KRANGE

B. Search strings

Scoping review 1

("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 "backreview 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 "metagrading" OR "meta-grading" OR "meta-assessment")

Scoping review 2

"learning analytics" AND ("peer feedback" OR "peer review" OR "peer grading" OR "peer evaluation" OR "peer assessment" OR "peer rating")

C. Variables and data pre-processing used in empirical studies

Table 5: Variables and data preprocessing in Study 1

Variable	Preprocessing	Description
<i>Backward evaluation grade</i>		
Backward evaluation grade (ordinal variable)	-	Range: 1-5 1: Not useful at all 2: Not very useful 3: Somewhat useful, although it could have been more elaborate 4: Very useful, although minor things could have been better 5: Extremely useful, constructive and justified
<i>Improvement suggestions</i>		
Kindness (binary variable)	-	Feedback is too harsh and uses harsh language
Justification (binary variable)	-	Feedback should be more justified and give more arguments for the decisions
Constructivity (binary variable)	-	Feedback should be more constructive and propose things to improve
Relevance (binary variable)	-	Feedback does not feel relevant to my hand-in or addresses the wrong things
Specificity (binary variable)	-	Feedback should be more specific and point to concrete things that can be improved

Table 5: Variables and data preprocessing in Study 1 (cont.)

Variable	Preprocessing	Description
# of improvement suggestions (discrete variable)	Variable construction	Range: 0-5
fewCat (binary variable)	Variable construction	1-3 improvement suggestions selected
manyCat (binary variable)	Variable construction	4-5 improvement suggestions selected
<i>Backward evaluation comment</i>		
Backward evaluation code (categorical variable)	Manual inductive coding String matching using Pandas Python package Variable construction	7 levels: only <i>acceptance</i> : Backward evaluation comments expressing praise, error acknowledgment, or intention of revision only <i>defence</i> : Backward evaluation comments expressing confusion, criticism, or disagreement only <i>gratitude</i> : Backward evaluation comments expressing gratitude <i>acceptance + gratitude</i> <i>acceptance + defence</i> <i>defence + gratitude</i> <i>acceptance + defence + gratitude</i>
Sentiment score (continuous variable)	Sentiment analysis using the Vader sentiment analyser	Range: -1 (negative) - 1 (positive)
Backward evaluation comment length (continuous variable)	Variable construction using stringr R package	# of characters per backward evaluation comment normalized to a 0-1 range
Part-of-speech tags (continuous variable)	Part-of-speech tagging using the spaCy Python package	Proportion of <i>verbs</i> , <i>nouns</i> and <i>adjectives</i> per all words in a backward evaluation comment

Table 5: Variables and data preprocessing in Study 1 (cont.)

Variable	Preprocessing	Description
<i>Rubric design</i>		
Rubric question type (continuous variable)	Variable construction	Proportion of <i>boolean</i> , <i>numerical</i> and <i>text</i> questions per rubric.
# of questions per rubric (continuous variable)	Variable construction	# of questions per rubric normalized to a 0-1 range

Table 6: Variables and data preprocessing in Study 2

Variable	Preprocessing	Description
<i>Student characteristics (platform data)</i>		
Group size (discrete variable)	-	Range: 1-3
Peer assessment experience (continuous variable)	Variable construction	Proportion of students in a group with previous peer assessment experience
<i>Student characteristics (context data)</i>		
Group average grade (continuous variable)	Variable construction	Mean of all individual final grades in a group
<i>Essay characteristics (context data)</i>		
Draft length (discrete variable)	Variable construction	# of words in a draft
<i>Essay characteristics (platform data)</i>		
Draft grade (ordinal variable)	-	5 levels: 1: Unacceptable, does not fulfil the minimum requirements 2: Poor, fulfils only the minimum requirements <i>Satisfactory</i> : has significant shortcomings 3: Good, a satisfactory performance in most areas 4: Very good, demonstrates sound judgement and a very good degree of independent thinking 5: Excellent, demonstrates excellent judgement and a high degree of independent thinking
<i>Feedback characteristics (platform data)</i>		
Written feedback length (discrete variable)	Variable construction	Total # of words in a feedback response
Praise + summary (discrete variable)	Deductive coding Variable construction	Total # of praise or positive comments that also summarise what a group has done and refers to contents of the essay in one feedback response

Table 6: Variables and data preprocessing in Study 2 (cont.)

Variable	Preprocessing	Description
Praise only (discrete variable)	Deductive coding Variable construction	Total # of praise or positive comments without summary in one feedback response
Mitigating praise (binary variable)	Deductive coding	Feedback comment includes a positive comment in a negative feedback to soften criticism
Hedges (binary variable)	Deductive coding	Feedback comment includes an indirect advice (using words like "maybe" or "can")
Identification (binary variable)	Deductive coding	Feedback comment includes a problem identification without an explanation
Explanation (binary variable)	Deductive coding	Feedback comment includes a problem explanation and refers to contents of the essay
Suggestion (binary variable)	Deductive coding	Feedback comment includes a general advice or an indirect solution in written feedback
Solution (binary variable)	Deductive coding	Feedback comment includes a specific advice or a direct solution
Boolean implementable feedback (binary variable)	Variable construction	Negative answer to a boolean feedback question
Scale implementable feedback (binary variable)	Variable construction	Lower than highest rating on a scale question
Implementable feedback (discrete variable)	Variable construction	Total # of implementable feedback in one feedback response

Table 6: Variables and data preprocessing in Study 2 (cont.)

Variable	Preprocessing	Description
<i>Backward evaluation characteristics (platform data)</i>		
Backward evaluation grade (ordinal variable)	-	Range: 1-5 1: Not useful at all 2: Not very useful 3: Somewhat useful, although it could have been more elaborate 4: Very useful, although minor things could have been better 5: Extremely useful, constructive and justified
Backward evaluation comment (binary variable)	Variable construction	Indicates, if a group wrote a backward evaluation comment
<i>Feedback implementation (platform + context data)</i>		
Implementation level 0 (binary variable)	Deductive coding Variable construction	No feedback implementation
Implementation level 1 (binary variable)	Deductive coding Variable construction	Partial feedback implementation
Implementation level 2 (binary variable)	Deductive coding Variable construction	Full feedback implementation
<i>Essay revision (context data)</i>		
Revision rate (continuous variable)	Variable construction using Jaccard similarity index from the textreus R package	Range: 0 (no revision) - 1 (the whole essay was revised)

D. Additional data charting for scoping reviews

Table 7: Overview of papers included in the Scoping review 1

Paper	Sample size	Data*	Methods
<i>Focus: Feedback uptake</i>			
Van der Pol et al. (2008)	Study 1: 27 Study 2: 38	<i>PA platform data</i> Discipline PA learning design Course characteristics <i>Student artefacts</i>	Text data coding Descriptive statistics Correlation analysis Regression analysis
Nelson and Schunn (2009)	24	<i>PA platform data</i> Discipline PA learning design Demographic data Course characteristics <i>Student artefacts</i>	Text data coding Descriptive statistics Correlation analysis Hypothesis testing
Wu and Schunn (2020)	185	<i>PA platform data</i> Discipline Demographic data PA learning design <i>Course characteristics</i> <i>Student artefacts</i>	Text data coding Descriptive statistics Correlation analysis Regression analysis

* Data used in the analysis is in italics, data used to describe the dataset is in regular font.

Table 7: Overview of papers included in the Scoping review 1 (cont.)

Paper	Sample size	Data*	Methods
<i>Focus: Tit-for-tat strategy</i>			
de Alfaro and Shavlovsky (2016)	23,762	<i>PA platform data</i> <i>Discipline</i> Student artefacts	Descriptive statistics Correlation analysis Parameterized probabilistic model
Adewoyin et al. (2016)	284	<i>PA platform data</i> <i>Discipline</i> PA learning design Course characteristics	Descriptive statistics Correlation analysis
Cho and Kim (2007)	617	<i>PA platform data</i> <i>Discipline</i> PA learning design <i>Interface type</i>	Descriptive statistics Correlation analysis
<i>Focus: Peer feedback quality</i>			
Patchan et al. (2018)	287	<i>PA platform data</i> <i>Discipline</i> Demographic data PA learning design Course characteristics <i>Pre-activity survey</i>	Text data coding Descriptive statistics Correlation analysis Hypothesis testing
<i>Focus: Improvement of writing skills</i>			
Cho and Schunn (2007)	87	<i>PA platform data</i> <i>Discipline</i> Demographic data PA learning design Course characteristics <i>Pre-activity skill test</i>	Descriptive statistics Hypothesis testing

* Data used in the analysis is in italics, data used to describe the dataset is in regular font.

Table 7: Overview of papers included in the Scoping review 1 (cont.)

Paper	Sample size	Data*	Methods
<i>Focus: LA insights into BE</i>			
Misiejuk and Wasson (2021)	7,660	<i>PA platform data</i>	Descriptive statistics Correlation analysis Regression analysis Epistemic network analysis
Tsivitanidou and Ioannou (2019)	21	<i>PA platform data</i> Discipline PA learning design <i>Pre-instructional questionnaire</i> <i>Think aloud protocols</i>	Text data coding Descriptive statistics Correlation analysis Hypothesis testing

* Data used in the analysis is in italics, data used to describe the dataset is in regular font.

Table 8: Overview of papers included in the Scoping review 2 that used learning analytics to gain new insights from peer assessment data

Paper	Sample size	Data *	Methods
Babik et al. (2019)	Two networks N = {12, 120}	<i>Simulated datasets</i>	Descriptive statistics Correlation analysis Hypothesis testing Monte-Carlo simulation
Bridges et al. (2020)	13	<i>VL platform data</i> <i>Discipline</i> <i>Demographic data</i> <i>PA learning design</i> <i>Survey</i> <i>Audio recordings</i> <i>Video recordings</i>	Descriptive statistics Discourse analysis Spatial analysis
Chiu et al. (2019)	50	<i>PA data</i> <i>Demographic data</i> <i>PA learning design</i> <i>Survey</i> <i>Simulator data</i>	Descriptive statistics Correlation analysis Hypothesis testing
Choi et al. (2019)	456 (2 studies)	<i>MOOC platform data</i> <i>Discipline</i> <i>Demographic data</i> <i>PA learning design</i>	Descriptive statistics Correlation analysis Hypothesis testing Natural Language Processing
Divjak and Maretic (2015)	62 (2 studies)	<i>LMS platform data</i> <i>Discipline</i> <i>PA learning design</i> <i>Course design</i> <i>Survey</i>	Descriptive statistics Taxicab of Manhattan distance
Djelil et al. (2021)	422 (7 courses)	<i>PA platform data</i> <i>Discipline</i>	Descriptive statistics Hypothesis testing Cluster analysis Social network analysis

* Data used in the analysis is in italics, data used to describe the dataset is in regular font.

Table 8: Overview of papers included in the Scoping review 2 that used learning analytics to gain new insights from peer assessment data (cont.)

Paper	Sample size	Data*	Methods
Er et al. (2021)	30	<i>PA platform data</i> Discipline PA learning design <i>Student grades</i>	Data coding Process mining
Gunnarsson and Alterman (2014)	157 (2 studies)	<i>Blogging data</i> Discipline PA learning design Course learning design <i>Student grades</i>	Descriptive statistics Correlation analysis Regression analysis
Huang et al. (2019)	96	<i>LMS platform data</i> Discipline Demographic data PA learning design Course learning design <i>Pre- and post-test</i>	Data coding Descriptive statistics Hypothesis testing Social network analysis
Khosravi et al. (2020)	384	<i>Adaptive learning platform data</i> Discipline PA learning design <i>Student grades</i>	Descriptive statistics Correlation analysis Hypothesis testing Root mean squared error
Lin (2019)	57	<i>LMS platform data</i> Discipline PA learning design <i>Course learning design</i> <i>Pre- and post-test</i> <i>Survey</i>	Data coding Descriptive statistics Hypothesis testing Regression analysis

* Data used in the analysis is in italics, data used to describe the dataset is in regular font.

Table 8: Overview of papers included in the Scoping review 2 that used learning analytics to gain new insights from peer assessment data (cont.)

Paper	Sample size	Data *	Methods
Misiejuk et al. (2021)	7,660	<i>PA platform data</i>	Descriptive statistics Correlation analysis Regression analysis Epistemic network analysis
Mørch et al. (2017)	125	Discipline PA learning design <i>Student artefact</i> <i>Student grades</i> <i>Pre- and post-test</i> <i>Video recordings</i>	Data coding Descriptive statistics Hypothesis testing Discourse analysis
Sedrakyan et al. (2014)	86	<i>Modeling tool data</i> Discipline Demographic data PA learning design Course characteristics <i>Student grades</i>	Descriptive statistics Process mining
Vogelsang and Ruppertz (2015)	467	<i>MOOC platform data</i> Discipline PA learning design Course characteristics Student artefacts <i>Student grades</i>	Descriptive statistics Correlation analysis
Cheng and Lei (2021)	24	<i>Bloggng data</i> Discipline PA learning design <i>Course characteristics</i>	Descriptive statistics Regression analysis Social network analysis

* Data used in the analysis is in italics, data used to describe the dataset is in regular font.

E. Study 2 approval by the Norwegian Centre for Research Data (NSD)

[Meldeskjema](#) / [Evaluering av bruk av Peergrade i faget Organisasjon og ledelse](#) / Vurdering

Vurdering

Dato	Type
09.03.2021	Standard

Referansenummer
875275

Prosjekttittel
Evaluering av bruk av Peergrade i faget Organisasjon og ledelse

Behandlingsansvarlig institusjon
Høgskolen Kristiania – Ernst G. Mortensens Stiftelse / School of Communication, Leadership, and Marketing / institutt for ledelse og organisasjon

Prosjektansvarlig
Jarle Bastesen

Prosjektperiode
01.11.2020 - 31.12.2022

[Meldeskjema](#) 

Kommentar

Det er vår vurdering at behandlingen av personopplysninger i prosjektet vil være i samsvar med personvernlovgivningen så fremt den gjennomføres i tråd med det som er dokumentert i meldeskjemaet 09.03.2021 med vedlegg, samt i meldingsdialogen mellom innmelder og NSD. Behandlingen kan starte.

MELD VESENTLIGE ENDRINGER

Dersom det skjer vesentlige endringer i behandlingen av personopplysninger, kan det være nødvendig å melde dette til NSD ved å oppdatere meldeskjemaet. Før du melder inn en endring, oppfordrer vi deg til å lese om hvilke type endringer det er nødvendig å melde: nsd.no/personverntjenester/fyll-ut-meldeskjema-for-personopplysninger/melde-endringer-i-meldeskjema

Du må vente på svar fra NSD før endringen gjennomføres.

TYPE OPPLYSNINGER OG VARIGHET

Prosjektet vil behandle alminnelige kategorier av personopplysninger frem til 31.12.2022.

LOVLIG GRUNNLAG

Prosjektet vil innhente samtykke fra de registrerte til behandlingen av personopplysninger. Vår vurdering er at prosjektet legger opp til et samtykke i samsvar med kravene i art. 4 og 7, ved at det er en frivillig, spesifikk, informert og utvetydig bekreftelse som kan dokumenteres, og som den registrerte kan trekke tilbake. Lovlig grunnlag for behandlingen vil dermed være den registrertes samtykke, jf. personvernforordningen art. 6 nr. 1 bokstav a.

PERSONVERNPRINSIPPER

NSD vurderer at den planlagte behandlingen av personopplysninger vil følge prinsippene i personvernforordningen om:

- lovlighet, rettferdighet og åpenhet (art. 5.1 a), ved at de registrerte får tilfredsstillende informasjon om og samtykker til behandlingen
- formålsbegrensning (art. 5.1 b), ved at personopplysninger samles inn for spesifikke, uttrykkelig angitte og berettigede formål, og ikke viderebehandles til nye uforenlige formål
- dataminimering (art. 5.1 c), ved at det kun behandles opplysninger som er adekvate, relevante og nødvendige for formålet med prosjektet
- lagringsbegrensning (art. 5.1 e), ved at personopplysningene ikke lagres lengre enn nødvendig for å oppfylle formålet

DE REGISTRERTES RETTIGHETER

NSD vurderer at informasjonen om behandlingen som de registrerte vil motta oppfyller lovens krav til form og innhold, jf. art. 12.1 og art. 13.

Så lenge de registrerte kan identifiseres i datamaterialet vil de ha følgende rettigheter: innsyn (art. 15), retting (art. 16), sletting (art. 17), begrensning (art. 18) og dataportabilitet (art. 20).

Vi minner om at hvis en registrert tar kontakt om sine rettigheter, har behandlingsansvarlig institusjon plikt til å svare innen en måned.

FØLG DIN INSTITUSJONS RETNINGSLINJER

NSD legger til grunn at behandlingen oppfyller kravene i personvernforordningen om riktighet (art. 5.1 d), integritet og konfidensialitet (art. 5.1 f) og sikkerhet (art. 32).

For å forsikre dere om at kravene oppfylles, må dere følge interne retningslinjer og eventuelt rådføre dere med behandlingsansvarlig institusjon.

OPPFØLGING AV PROSJEKTET

NSD vil følge opp ved planlagt avslutning for å avklare om behandlingen av personopplysningene er avsluttet.

Lykke til med prosjektet!

Kontaktperson hos NSD: Lene Chr. M. Brandt
Tlf. Personverntjenester: 55 58 21 17 (tast 1)

Part II

The papers

Paper 1

Misiejuk, K., Wasson B. & Egelandstal K. (2021). Using learning analytics to understand student perceptions of peer feedback. *Computers in Human Behavior*, 117. DOI: 10.1016/j.chb.2020.106658.



ELSEVIER

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 Egelandstad^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 & William, 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 (William, 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 & William, 1998; Black & William, 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

0747-5632/© 2020 The Author(s). Published by Elsevier Ltd. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

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, Geshler, & 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 (Hrepić, 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

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.

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.

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;

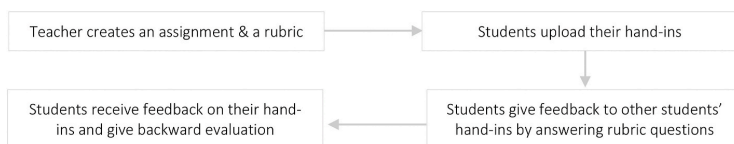


Fig. 1. Peer Assessment activity in Peergrade.

Wickham, 2016), the *sjPlot* package for Spearman Rank Correlation (v2.7.0; Lüdtke, 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 pre-processed 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., “tjlkldfjsldkdfj”, “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

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.”

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).

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

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		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	q_type	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			

(*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

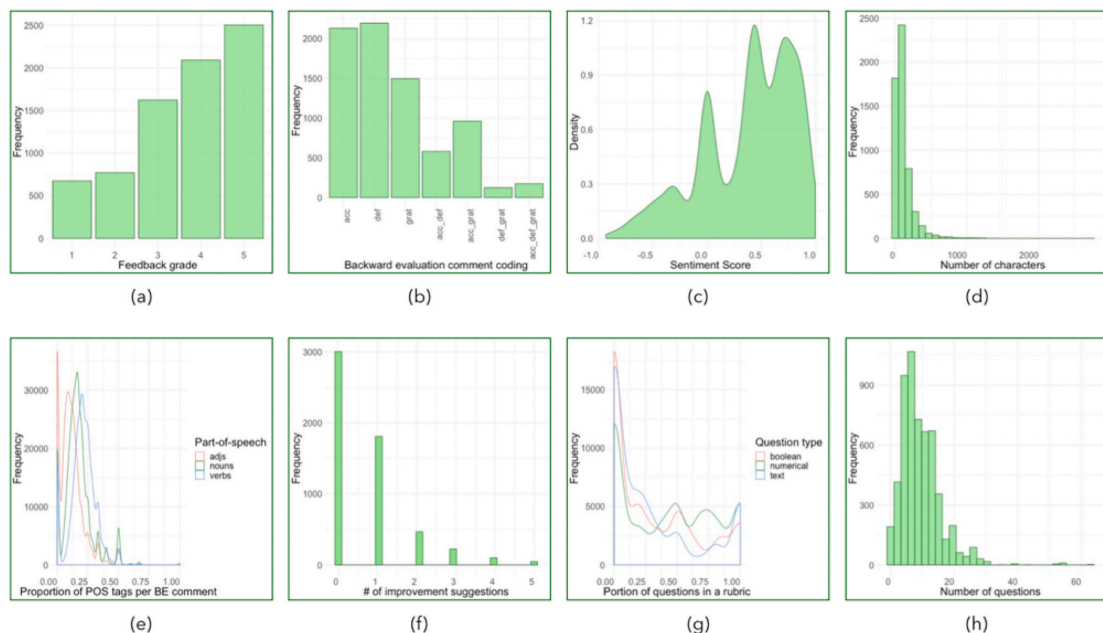


Fig. 2. Distributions of dependent and independent variables: (a) feedback grade, (b) BE comment codes, (c) sentiment score, (d) BE comment length, (e) part-of-speech tags, (f) number of selected improvement suggestions, (g) question type, (h) number of questions.

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 ($\beta =$

$0.001, p \leq .05$), and more *#_of_qs* per rubric makes students more likely to perceive feedback as more useful rather than less useful (*#_of Impr_suggs*, $\beta = 0.39$), 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

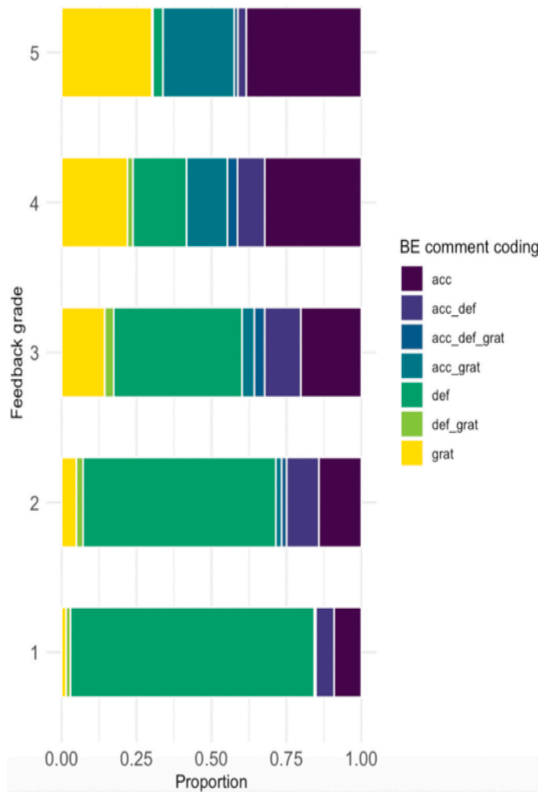


Fig. 3. The proportion of BE comment codes per feedback grade.

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).

variable		f_grade				
		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	sentiment	-0.016	0.083***	0.077***	0.027**	-0.136***
	sentiment	-0.268***	-0.168***	-0.093***	0.080***	0.275***
#_of_impr_suggs	#_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			

Abbreviations: 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

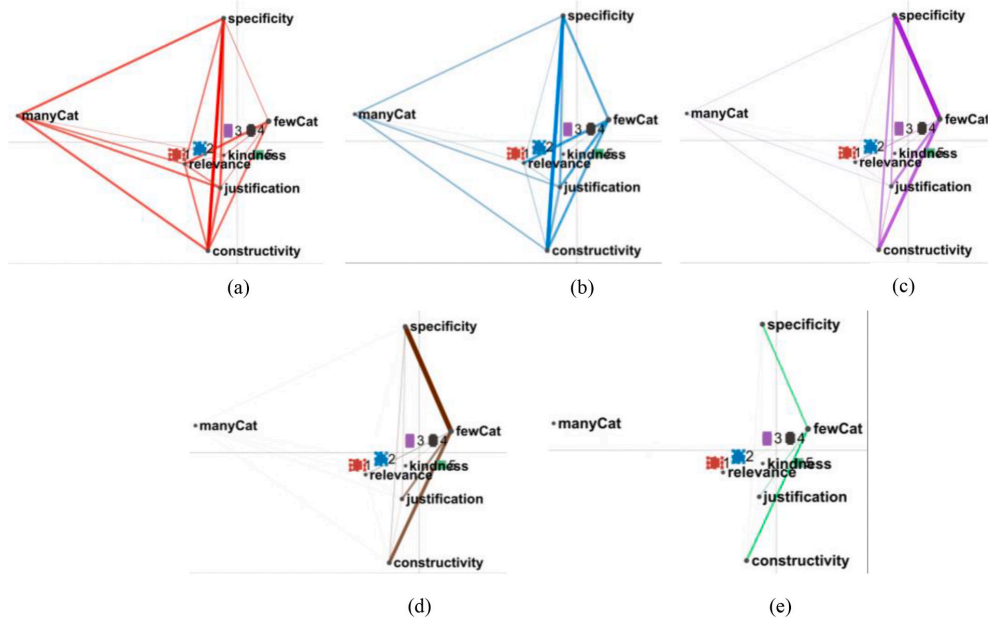


Fig. 4. ENA model plotting improvement suggestions with the number of improvement suggestions: (a) Plot for feedback grade 1 (FG1), (b) Plot for feedback grade 1 (FG2), (c) Plot for feedback grade 3 (FG3), (d) Plot for feedback grade 4 (FG4), (e) Plot for feedback grade 5 (FG5) NOTE: FG=Feedback Grade; 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 (feed-forward). 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 Ege-landsdal: 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., & Maretic, 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., & Gasevic, D. (2019). Synergy: An online platform for dialogic peer feedback at scale. In K. Lund, G. P. Nicolai, 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](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. *Educuse Review*, 46(5), 31–40.
- Lüdtke, D. (2019). *sjPlot: Data visualization for statistics in social science*. <https://doi.org/10.5281/zenodo.3308157>
- Lundstrom, K., & Baker Smemo, 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). *eNA: 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](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. Vatrappu, & B. Wasson (Eds.), *Measuring and visualizing learning in the information-rich classroom* (pp. 19–26). New York: Routledge.
- Ryan, T., Gasević, 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. Gasević (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](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.
- William, 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 reciprocity 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., Litman, 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](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).

Paper 2

Misiejuk, K. & Wasson, B. (2021). Backward evaluation in peer assessment: A scoping review. *Computers & Education*, 175. DOI: [10.1016/j.compedu.2021.104319](https://doi.org/10.1016/j.compedu.2021.104319).



Contents lists available at ScienceDirect

Computers & Education

journal homepage: www.elsevier.com/locate/compedu

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

ARTICLE INFO

Keywords:
Peer assessment
Backward evaluation
Scoping review

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

* Corresponding author. Department of Information Science & Media Studies, 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.compedu.2021.104319>

Received 25 November 2020; Received in revised form 26 July 2021; Accepted 28 August 2021

Available online 28 August 2021

0360-1315/© 2021 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY license

(<http://creativecommons.org/licenses/by/4.0/>).

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 Massive 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, & Egelanddsdal, 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

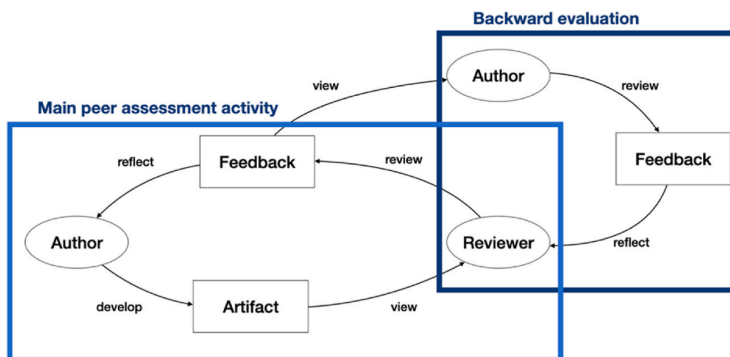


Fig. 1. A model of the peer assessment process including backward evaluation (adapted from Indriasari, Luxton-Reilly, & Denny, 2020).

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.

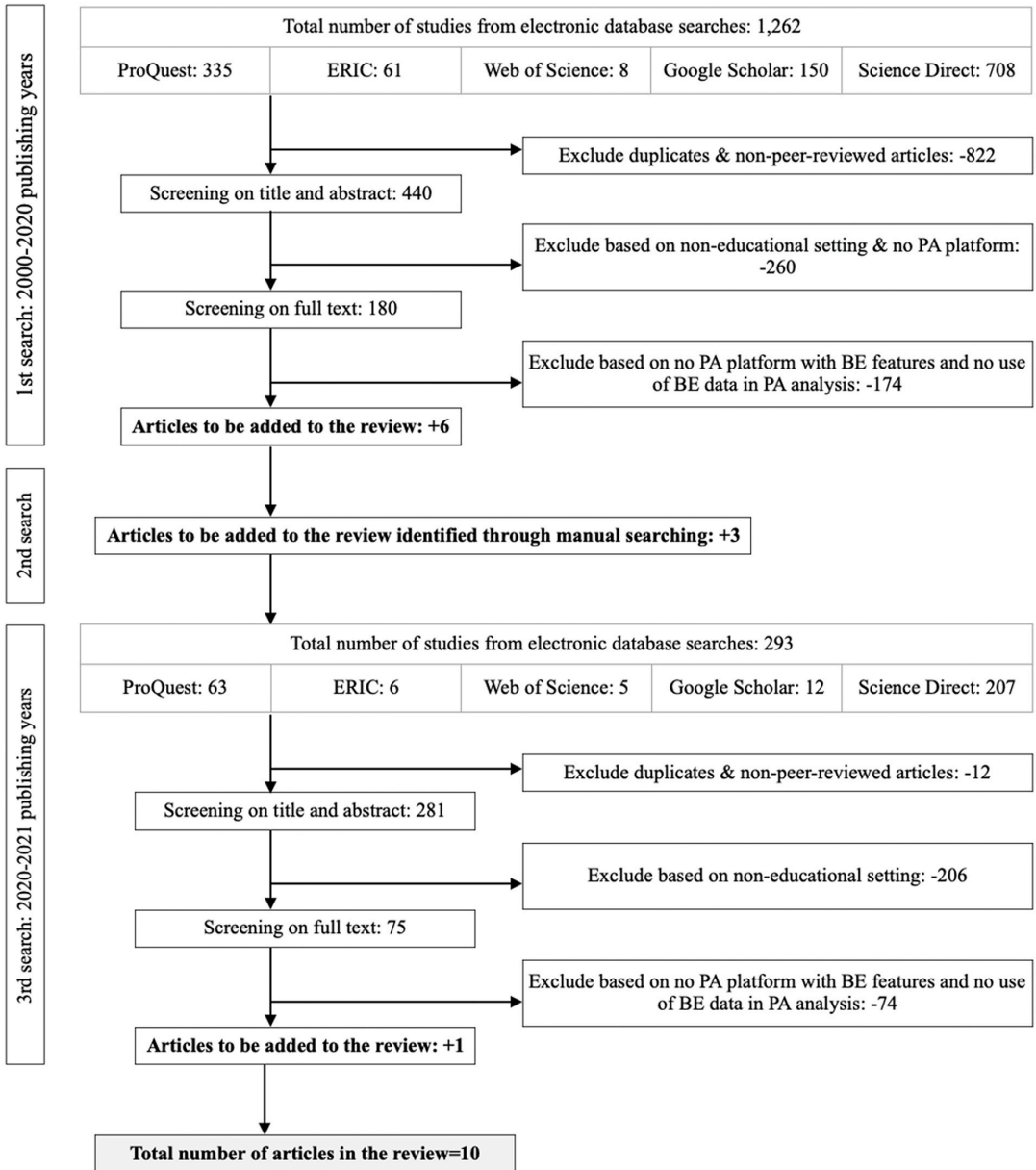


Fig. 2. Inclusion/exclusion process.

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, Broekmans, 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, Haşlamam, & 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’. Tsvitanidou 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 & Schumm, 2007	SWoRD/Peerceptiv	Higher education	Physics	87	Scale, comment	Yes	7-point helpfulness scale	Not coded	–
Study 1 by Vander 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 Vander 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 Schumm (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 Schumm (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.
	Wu and Schunn (2020a)	back-review, back-evaluation	Low feedback understanding if feedback included problem explanation. 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.
improvement of writing skills	Cho & Schunn, 2007	helpfulness rating	High problem agreement if feedback comments included mitigating praise. Longer feedback comments predict higher helpfulness ratings.
learning analytics insights into BE	Tsvitanidou and Ioannou (2019)	react phase	High writing performance in the final draft if students gave more helpful feedback.
	Misiejuk et al. (2021)	backward evaluation	More backward evaluation comments if low agreement with feedback. 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) de Alfano and Shavlowsky (2016) Adewoyin et al. (2016)	back-review review feedback back-evaluation	Low chance of tit-for-tat strategy if cognitive interface design is implemented. Evidence of tit-for-tat strategy based on helpfulness ratings independent of subject area. 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. Tsvitanidou 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

This research is supported by a PhD research grant from the University of Bergen, Norway. The authors would like to thank the reviewers for their excellent comments that have improved the quality of this article.

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., Haşlamani, 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.
- Gamege, 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., Dřopcová, 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.
- Tsitivanidou, 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, L., 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 reciprocity 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.

Paper 3

Misiejuk, K. & Wasson, B. (accepted). Learning analytics for peer assessment - A scoping review. In O. Noroozi & B. De Wever (Eds.) *The Power of Peer Learning*. Springer.

Paper 4

Misiejuk, K., Bastesen, J., Wasson, B. & Krange, I. (submitted). Educational data for learning analytics: Increasing insights into peer assessment with context data. *Assessment & Evaluation in Higher Education*.



Graphic design: Communication Division, UIB / Print: Skjipes Kommunikasjon AS



uib.no

ISBN: 9788230868485 (print)
9788230867204 (PDF)