Received: 20 November 2022 | Accepted: 1 May 2023

DOI: 10.1111/bjet.13336

ORIGINAL ARTICLE

British Journal of Educational Technology BERA

Using natural language processing to support peer-feedback in the age of artificial intelligence: A cross-disciplinary framework and a research agenda

Elisabeth Bauer[1,2] | Martin Greisel[3] | Ilia Kuznetsov[4] | Markus Berndt[5] | Ingo Kollar[3] | Markus Dresel[3] | Martin R. Fischer[5] | Frank Fischer[1]

[1]Department of Psychology, Ludwig-Maximilians-Universität München, Munich, Germany

[2]School of Social Sciences and Technology, Technical University of Munich, Munich, Germany

[3]Faculty for Philosophy and Social Sciences, University of Augsburg, Augsburg, Germany

[4]Ubiquitous Knowledge Processing Lab, Department of Computer Science and Hessian Center for AI (hessian.AI), Technical University of Darmstadt, Darmstadt, Germany

[5]Institute of Medical Education, University Hospital of LMU Munich, Munich, Germany

Correspondence
Elisabeth Bauer, School of Social Sciences and Technology, Technical University of Munich, Friedl Schöller Endowed Chair for Educational Psychology, Arcisstr. 21, 80333 Munich, Germany.
Email: eli.bauer@tum.de

Martin Greisel, Faculty for Philosophy and Social Sciences, University of Augsburg, Universitätsstr. 10, 86159 Augsburg, Germany.
Email: martin.greisel@uni-a.de

Advancements in artificial intelligence are rapidly increasing. The new-generation large language models, such as ChatGPT and GPT-4, bear the potential to transform educational approaches, such as peer-feedback. To investigate peer-feedback at the intersection of natural language processing (NLP) and educational research, this paper suggests a cross-disciplinary framework that aims to facilitate the development of NLP-based adaptive measures for supporting peer-feedback processes in digital learning environments. To conceptualize this process, we introduce a peer-feedback process model, which describes learners' activities and textual products. Further, we introduce a terminological and procedural scheme that facilitates systematically deriving measures to foster the peer-feedback process and how NLP may enhance the adaptivity of such learning support. Building on prior research on education and NLP, we apply this scheme to all learner activities of the peer-feedback process model to exemplify a range of NLP-based adaptive support measures. We also discuss the current challenges and suggest

Elisabeth Bauer and Martin Greisel contributed equally to this study.

directions for future cross-disciplinary research on the effectiveness and other dimensions of NLP-based adaptive support for peer-feedback. Building on our suggested framework, future research and collaborations at the intersection of education and NLP can innovate peer-feedback in digital learning environments.

KEYWORDS

adaptivity, artificial intelligence, digital learning, large language models, learner support, natural language processing, peer-feedback

Funding information

Bundesministerium für Bildung und Forschung, Grant/Award Number: 16DHL1040; Deutsche Forschungsgemeinschaft, Grant/Award Number: DFG FOR 2385; Stiftung Innovation in der Hochschullehre, Grant/Award Number: FMM120-2020

Practitioner notes

What is already known about this topic

- There is considerable research in educational science on peer-feedback processes.
- Natural language processing facilitates the analysis of students' textual data.
- There is a lack of systematic orientation regarding which NLP techniques can be applied to which data to effectively support the peer-feedback process.

What this paper adds

- A comprehensive overview model that describes the relevant activities and products in the peer-feedback process.
- A terminological and procedural scheme for designing NLP-based adaptive support measures.
- An application of this scheme to the peer-feedback process results in exemplifying the use cases of how NLP may be employed to support each learner activity during peer-feedback.

Implications for practice and/or policy

- To boost the effectiveness of their peer-feedback scenarios, instructors and instructional designers should identify relevant leverage points, corresponding support measures, adaptation targets and automation goals based on theory and empirical findings.
- Management and IT departments of higher education institutions should strive to provide digital tools based on modern NLP models and integrate them into the respective learning management systems; those tools should help in translating the automation goals requested by their instructors into prediction targets, take relevant data as input and allow for evaluating the predictions.

INTRODUCTION

Feedback, one of the most powerful means to support learning (Hattie, 2008), refers to “all post-response information that is provided to a learner to inform the learner about his or her actual state of learning or performance” (Narciss, 2008, p. 127). In formal education, feedback is often provided by instructors, such as teachers or lecturers, or peer learners, especially the latter offering high potential for learning (eg, Double et al., 2020). However, learners' often insufficient competencies (eg, knowledge, skills, and attitudes; see Blömeke et al., 2015) concerning a learning task and feedback production can diminish the effectiveness of peer-feedback. To exploit peer-feedback's potential benefits, learners need support. Theoretically, a support measure might be any aspect of instructional design that helps learners pursue a task goal or develop their competencies. Owing to technological advancements in artificial intelligence (AI), these support measures can be adapted and automated (Ninaus & Sailer, 2022).

Natural language processing (NLP)—a subfield of AI dedicated to text and speech processing—has attained great progress in recent years, with the latest generation of large language models (LLMs)—ChatGPT and GPT-4 (OpenAI, 2023)—attracting wide public attention and enabling novel applications of AI in many areas of human activity, such as education (Kasneci et al., 2023). In digital learning environments, where communication during peer-feedback often involves an exchange of written text (eg, Patchan et al., 2016), automating adaptive support measures via NLP bears great promise. Prior research at the intersection of education and NLP has investigated, for example, using NLP to automate computer-supported adaptive feedback on students' reasoning about case vignettes in digital learning environments and found positive effects on students' performance during and after the learning process (Sailer et al., 2023). However, compared to designing adaptive support for peers giving each other feedback, the design process underlying the use case of NLP-based adaptive feedback, which is automatically provided as part of digital learning environments, appears far less complex because peer-feedback involves several more options for leveraging students' learning processes to maximize their potential learning benefits.

This paper aims to initiate and facilitate the development of NLP-based adaptive measures to support peer-feedback in digital learning environments. To this end, (a) we propose a model to describe the activities that learners engage in and the textual products they generate during a peer-feedback process. On this basis, we systematically derive potential support measures. However, to make these measures adaptive and automatize their employment, we need to cross the boundaries of educational research and NLP. For this reason, (b) we introduce a terminological and procedural scheme for designing NLP-based adaptive support measures. Then, (c) we apply this scheme to the peer-feedback process and provide examples of potential NLP-based adaptive measures for supporting peer-feedback. Finally, we discuss current challenges and future research directions.

THE PEER-FEEDBACK PROCESS

Peer-feedback involves at least two learners performing a series of activities (see Figure 1) and thereby engaging in cognitive, meta-cognitive, motivational-affective and social processes (Aben et al., 2019; Narciss et al., 2014). This paper refers to peer-feedback as a process organized around solving a specific task, guided by instruction and realized within a digital learning environment. Due to peer-feedback's dialogic nature, learners repeatedly engage in social activities, such as sharing information, negotiating meaning and regulating the learning process (Liu et al., 2016), and they often generate several textual products as well as cognitive, meta-cognitive and affective learning outcomes.

In the following section, we briefly illustrate a peer-feedback scenario in a digital learning environment in teacher education. The scenario aims to facilitate pre-service teachers' skills in evidence-informed reasoning about classroom-related teaching problems (Hornstein et al., 2023). The pre-service teachers are confronted with written case vignettes about a school lesson. In the design phase (see Figure 1), an instructor specifies a rule for peer

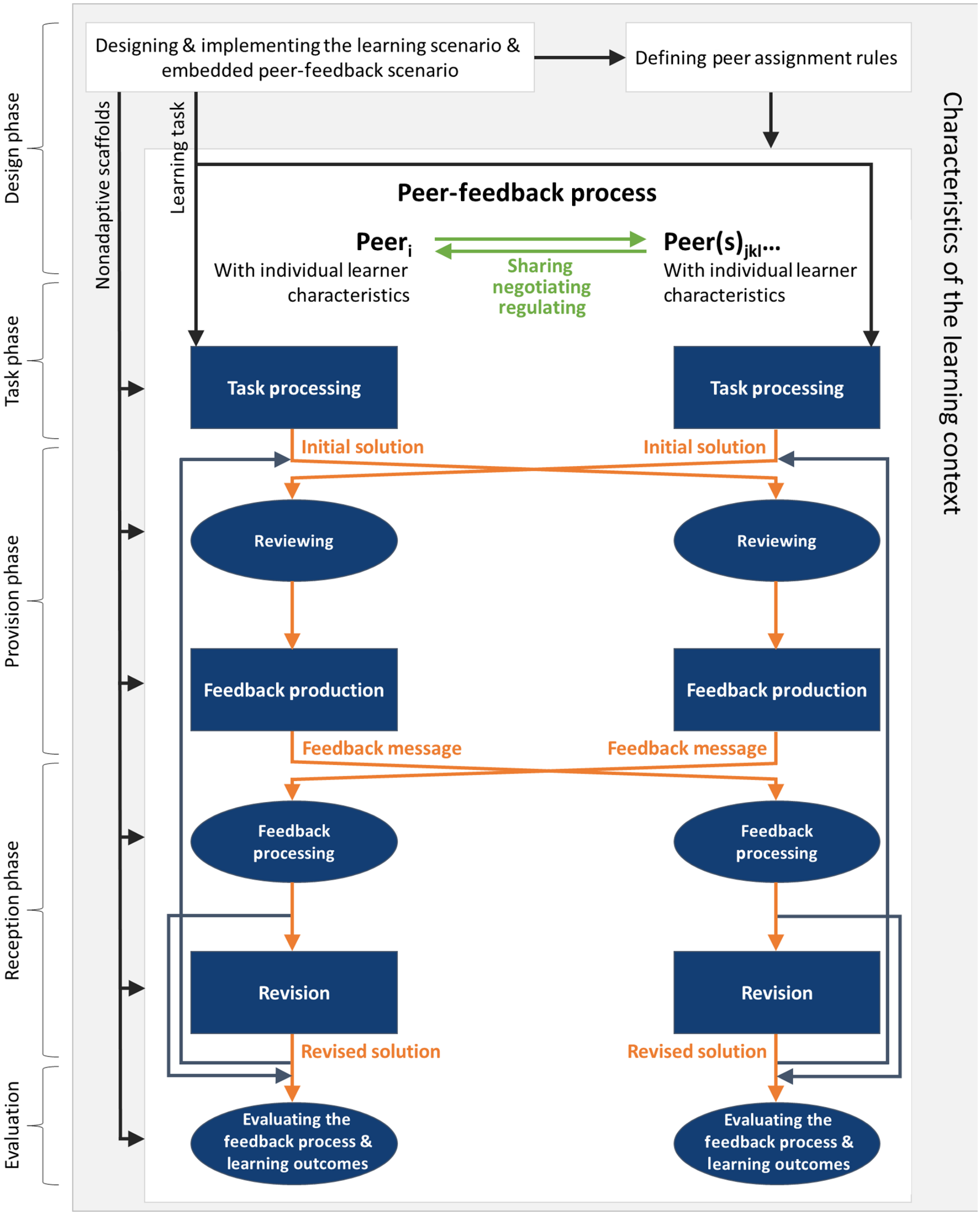

FIGURE 1 A peer-feedback process model.

assignment. Each student will review the initial solutions of two randomly assigned peer learners and receive feedback from two randomly assigned peer reviewers. In the task phase of the peer-feedback process, the learners receive a learning task: to identify, analyse and solve the teaching problems in the case vignettes. During *task processing*, the learners produce a written *initial solution* comprising a short essay for each identified teaching problem. In the provision phase, peers are prompted to engage in *reviewing* and *feedback production*. The learners compose *feedback messages* for the initial solutions of two peers. In the reception phase, each learner receives two feedback messages on their own initial solutions. The learners subsequently engage in *feedback processing* and *revision* and produce a *revised solution*. During evaluation, the learners *evaluate the feedback process and the learning outcomes*. All these activities feed back into the individual learner characteristics (eg, increasing task knowledge).

The peer-feedback process is influenced by the characteristics of individual learners, the learning context, and the task characteristics. *Learner characteristics* include prior task knowledge, feedback skills, digital skills and meta-cognitive skills, as well as motivation, beliefs and attitudes toward the learning task and peer-feedback (Aben et al., 2019; Berndt et al., 2022; Narciss et al., 2014; Panadero, 2016; Sailer et al., 2021; Strijbos et al., 2021). *Contextual characteristics* include the usability of the learning environment and the instructional design (eg, the motivational design of the peer-feedback scenario; Brewer & Klein, 2006; Narciss et al., 2014; Patchan et al., 2018). *Task characteristics* relate to the features of the learning content and tasks, such as the functional complexity and comprehensiveness of the required learning activities (Fischer et al., 2022; Van Merriënboer & Kirschner, 2018).

To facilitate the peer-feedback process, learners usually need instructional support. This is most often realized as static, *non-adaptive scaffolds* which are the same for every learner, such as prompts with stepwise instructions or assessment rubrics. However, ideally, the support considers the individual, contextual, and task characteristics (Double et al., 2020; Li et al., 2020). Such *adaptive support measures* adjust to the observed changes in learners' characteristics and performance. The type and degree of the provided support is thus tailored to the learners' current needs and hence capable of further enhancing the learning benefits (Plass & Pawar, 2020; Tetzlaff et al., 2021).

EDUCATIONAL RESEARCH AND NLP FOR ADAPTIVE PEER-FEEDBACK SUPPORT

The peer-feedback process model (see Figure 1) describes the activities and textual products that can be identified during learning with peer-feedback. To develop adaptive support measures that benefit from advancements in NLP, we need to bridge the boundaries between educational research and NLP. We thus suggest a terminological and procedural scheme (see Figure 2) that links concepts from educational research with concepts from NLP to better describe, investigate and design NLP-based adaptive measures for supporting peer-feedback. In the subsequent section, we will apply this scheme to exemplify several NLP-based measures that target the individual activities depicted in the peer-feedback process model.

Adaptively supporting the peer-feedback process

To be effective, support measures need to target relevant *leverage points* in the peer-feedback process. Leverage points are the factors and instances in the peer-feedback process that can either facilitate or impair the learning process and outcomes. Leverage points can be identified at the cognitive, meta-cognitive, affective-motivational and social-dialogical levels of the peer-feedback process (eg, Aben et al., 2019; Narciss et al., 2014). We will primarily focus on the cognitive level in the examples throughout this paper, such as the correctness of the initial solution in the task phase; however, we will, at times, also illustrate meta-cognitive leverage points (eg, reflection processes or awareness of important aspects of the peer-feedback process during the phase of evaluation) and motivational leverage points (eg, when processing multiple feedback messages).

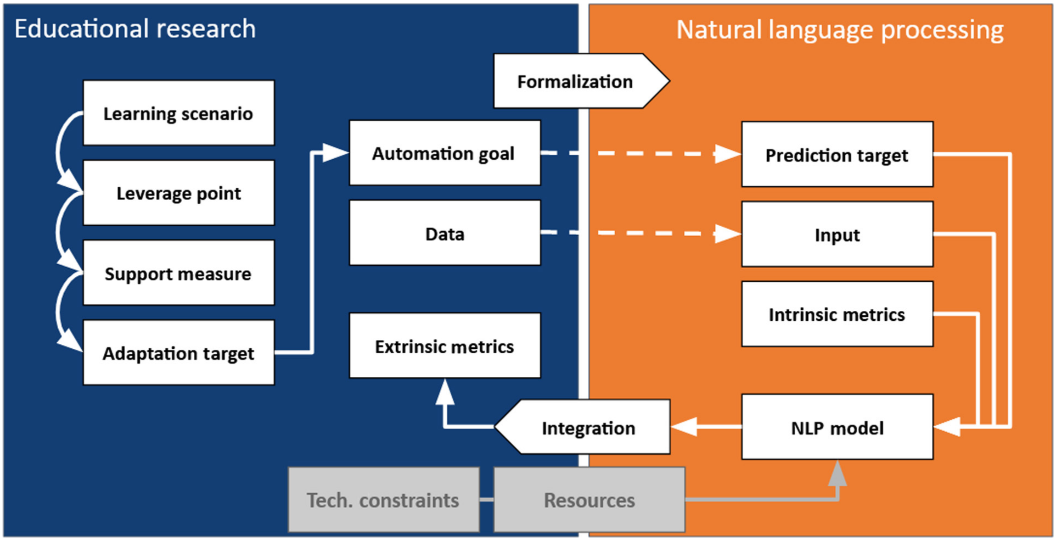

FIGURE 2 Terminological and procedural scheme for designing NLP-based adaptive support measures.

Adaptive *support measures* target leverage points to increase the effectiveness of the peer-feedback process by providing additional task-relevant information or explanations or by exerting direct or indirect *regulation*. While scaffolds—such as prompts and hints—regulate peer learning directly by offering additional instructions and explanations (Bannert & Mengelkamp, 2013; Quintana et al., 2004), awareness tools regulate indirectly by offering performance indicators to inform learners' self-regulation (Lin & Tsai, 2016; Michel et al., 2017). Moreover, some support measures offer learners options for individualization (Kucirkova et al., 2021): Compared to adaptivity, *adaptability* offers learners themselves (rather than the technical learning environment with which they work) the option to adjust the overall use, timing or degree of learner support (Fischer, 2001; Wang et al., 2017). Consequently, learners can practice self-regulation (eg, Vogel et al., 2022).

Adaptive support measures can vary across use cases, depending on the *adaptation targets*, such as learner characteristics, processes or outcomes to which the support is adapted. To adjust the learner support to dynamically changing adaptation targets (eg, advancing skills; Kalyuga, 2007; Tetzlaff et al., 2021), observable activities and products can be used for performance assessment. For example, as a reviewing support, the to-be-reviewed initial solution of a peer might be pre-processed in such a way that its important flaws are automatically highlighted. In addition, the reviewing support could be personalized for the reviewer by adapting to their task performance so that those reviewers with low prior performance receive hints with additional instructions.

To automate adaptive learner support in digital learning environments, technological advancements in AI can augment learner activities and automate routine operations and actions (eg, Ninaus & Sailer, 2022). For the peer-feedback process, such *automation goals* can be classified as supporting individual activities (eg, prompting to further structure insufficiently structured feedback messages), modifying individual products (eg, highlighting important aspects in feedback messages), and modifying the sequence of activities in the peer-feedback process (eg, omitting the revision step if feedback does not include suggestions for improvement; see Figure 1).

Enhancing adaptive support via NLP

The products exchanged during peer-feedback—such as the feedback message—often constitute textual data. NLP is thus a primary candidate for automating and augmenting adaptive support measures for peer-feedback scenarios. NLP's general objective is to create computational *models* that make accurate predictions about a *target* based on textual *input* according to one or multiple *evaluation metrics*. This objective encompasses a wide range of task types (Wang et al., 2022), ranging from surface-level tasks (such as text categorization or span and relation extraction) to deep semantics-driven tasks (such as summarization and text generation; see Figure 3). Modern deep-learning-based NLP builds upon pre-trained LLMs (Brown et al., 2020; Devlin et al., 2019). Through self-supervised training on large, unlabeled textual collections, these models create generally applicable neural representations of text that can be tailored to particular end tasks. NLP's applications range from predicting the sentiment of tweets (Agarwal et al., 2011) to automatic question answering for scientific literature (Dasigi et al., 2021). The latest generation of LLMs, such as GPT-4 and Tk-Instruct (Wang et al., 2022), blends the boundaries between task types by offering a unified text-to-text interface for model querying, which allows one to apply the models to unfamiliar tasks solely based on a brief task definition and a few examples.

Previously, NLP has been applied to support scenarios similar or related to peer-feedback, such as scholarly peer review (Cheng et al., 2020; Hua et al., 2019; Kennard et al., 2022; Kuznetsov et al., 2022), case-study analysis (Pfeiffer et al., 2019; Schulz et al., 2019), and argumentative writing and essay grading (eg, Burstein et al., 1998; Zhang & Litman, 2021). However, these existing applications of NLP have been limited to the isolated scenarios for which they were originally designed; no holistic framework for NLP-based peer-feedback support has been proposed to date. The potential of the new generation of LLMs to support peer-feedback remains equally underexplored. Moreover, the idiosyncrasies of peer-feedback data

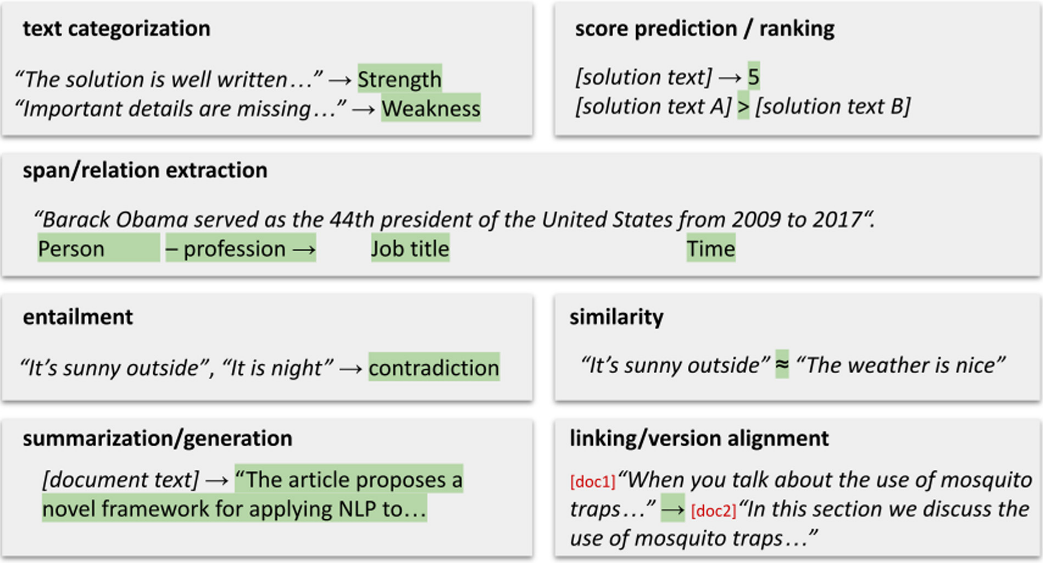
text categorization
"The solution is well written…" → Strength
"Important details are missing…" → Weakness
score prediction / ranking
[solution text] → 5
[solution text A] > *[solution text B]*
span/relation extraction
"Barack Obama served as the 44th president of the United States from 2009 to 2017".
Person – profession → Job title Time
entailment
"It's sunny outside", "It is night" → contradiction
similarity
"It's sunny outside" ≈ *"The weather is nice"*
summarization/generation
[document text] → "The article proposes a novel framework for applying NLP to…
linking/version alignment
[doc1]*"When you talk about the use of mosquito traps…"* → [doc2]*"In this section we discuss the use of mosquito traps…"*

FIGURE 3 Examples of the NLP task types discussed in this paper and their inputs (italic) and prediction targets (in colour). Each task type can accommodate a wide range of tasks: For example, the sentiment analysis of blog posts and stance prediction for peer-feedback texts can be both cast as score prediction tasks or text categorization tasks.

and associated tasks in the educational domain pose a range of methodological challenges, such as data scarcity, language and domain shift, interpretability, sensitivity to bias and increased privacy demands. We revisit these challenges in the final section of the paper.

A generic approach to NLP model development for peer-feedback can be derived following the recently proposed framework of Translational NLP (Newman-Griffis et al., 2021). Based on the automation goal and available data, an appropriate NLP task type is chosen, along with the formal definitions of the prediction target, textual input, and evaluation metric. For example, *solution grading* (goal) based on *task solution texts* (data) can be cast as a *score prediction task* (NLP task type). That is, given *plain text* (input), the model predicts a *numerical score* in a given range (target) with the aim of minimizing the *deviation between the true and the predicted score* (evaluation metric). Based on the data's size and availability, technical constraints (eg, hardware limitations), and existing NLP resources (eg, pre-trained language models and corpora), NLP practitioners collaborate with educational researchers to select the appropriate NLP methodology for the task. Moreover, the nature of the support measure must be considered: while some support measures, such as essay grading (eg, to identify learners' current performance level), can be performed offline, others, such as real-time feedback analysis, place additional requirements on the model's throughput. The trade-off between the processing speed and the acceptable error rate of NLP-based assistance must be established in a collaborative cross-disciplinary process.

Based on the chosen approach, an NLP model is created, which, in most cases, involves obtaining a set of inputs paired with target values for the chosen task and using this data to develop and evaluate an NLP model. Once the model reaches adequate performance on a held-out test portion of the data, it can be integrated into the learning environment to support the automation goal, either by *fully automating* the target procedure (eg, predicting the solution score) or by *augmenting* the learner experience via continuous real-time support or via an adaptive scaffold (eg, notifying the learner that the current task solution text is insufficient prior to submission). A model might perform well *intrinsically*—that is, on the test data (eg, solution score deviation is low)—but fail to adequately support the chosen support measure, with reasons ranging from a poor selection of the intrinsic evaluation metrics and data to the particularities of implementation and overall user experience. Hence, *extrinsic* evaluation is the final and crucial step in measuring the adequacy of NLP-based support. Unlike intrinsic evaluation, which is usually performed by the NLP development team itself, extrinsic evaluation requires the NLP model to be deployed in a learning setting and incorporates both user feedback and formal measurements of intervention success. For example, the effectiveness of an NLP-enhanced adaptive support measure can be investigated in a pretest–posttest control group study regarding extrinsic outcomes, such as learners' skill improvement. If applying the newly developed support measure results in the desired extrinsic outcomes, the implementation can be considered successful.

NLP ADAPTIVE MEASURES FOR SUPPORTING THE PEER-FEEDBACK PROCESS

Below, we exemplify a range of (a) leverage points for effective peer-feedback, (b) several options for support measures that target these leverage points, (c) adaptation targets for these measures, (d) corresponding automation goals, (e) the necessary textual data and (f) the NLP prediction targets that may be used for doing so.

Design phase: Preparing the peer-feedback process

In the context of formal education, a peer-feedback scenario in a digital learning environment is usually designed by an instructor (ie, teacher, lecturer or tutor), who specifies the learning task(s), learning materials, instructions, and nonadaptive scaffolds (eg, reflection prompts; Bannert & Mengelkamp, 2013). The instructor also needs to define the rules for the peer assignment—that is, to whom each learner gives feedback and from whom each learner receives feedback. The peer assignment can be random or based on learner characteristics, such as learners' task-related skills. Research found that learners who delivered better initial solutions also provided more hints on self-regulation aspects instead of providing only corrective feedback (Alqassab et al., 2018b). Their feedback comments were also more critical (Cho & Cho, 2011). The same was true for learners with a higher general writing ability (Patchan & Schunn, 2015). In addition, Alqassab et al. (2018a) found that the feedback accuracy decreased when the initial solution contained more flaws. However, Patchan et al. (2013) showed that writers with high general verbal ability also received similar feedback from low-ability peers, whereas low-ability writers of initial solutions received more valuable feedback from high-ability peers. Therefore, as a *leverage point*, it might be recommendable to assign initial solutions of more competent learners to less competent learners and initial solutions of less competent learners to more competent learners (comparable to positive vs. erroneous worked examples; Große & Renkl, 2007; Tsovaltzi et al., 2012).

To optimize the effectiveness of the peer-feedback processes, a *support measure* for the peer assignment might focus on the learners' level of domain knowledge and initial solution performance as an *adaptation target*. The *automation goal* for NLP would consist of automatically identifying high-performing and low-performing learners.

The design phase requires *textual data* to determine learners' performance. The peer assignment can be scheduled to take place *after* the initial processing of the learning task, making it possible to estimate learners' performance within the current learning scenario. In cases when learners are assigned to dyads or groups prior to task processing (eg, in collaborative or otherwise dialogical learning situations), learners' *prior* textual data can be obtained from prior learning tasks. The indicators and automation targets for distinguishing high-performing and low-performing learners based on their textual products (ie, initial solution, feedback message, and revised solution) are detailed in the subsequent sections. From an NLP perspective, the quality of a task solution can be cast as a score prediction or ranking task and approached in an end-to-end fashion. Several lines of research in NLP are dedicated to the automatic evaluation of texts, including automatic essay scoring (Burstein, 2003; Dasgupta et al., 2018; Ke & Ng, 2019; Page, 1968; Zhang & Litman, 2021), quantifying readability (Deutsch et al., 2020), factuality (Maynez et al., 2020) or specificity (Lugini & Litman, 2017).

Task processing and producing an initial solution

Through initial task processing, learners are familiarized with the content and instructions of the learning task and build a mental model of the approaches needed to process and solve the task, which can facilitate the quality of the subsequently produced feedback message and increase the effectiveness of the peer-feedback process (eg, Alemdag & Yildirim, 2022; Greene & Azevedo, 2009). However, learner characteristics, such as prior knowledge, impact task processing and the quality of the initial solution. Therefore, a *leverage point* involves facilitating learners' task processing toward producing a sufficient initial solution.

Support measures to facilitate task processing should serve one or more of the following mechanisms (Belland, 2014, p. 507): “enlisting student interest, controlling frustration, providing feedback, indicating important task/problem elements to consider, modeling expert processes, and questioning.” For example, indicating some important elements to consider could be realized by highlighting or annotating the relevant aspects in learners' initial solutions. The *adaptation target* for adjusting such support measures is the learners' task performance in the present learning task or prior learning tasks. Current task performance can be assessed by letting learners submit a draft of their initial solution. An *automation goal* for such adaptive support measures is the automatic detection of the relevant structural aspects or content aspects in the initial solution texts that distinguish adequate initial solutions from inadequate ones. The relevant aspects depend on the task at hand: argumentation structures (eg, claims, premises and their connections; Rapanta & Walton, 2016; Wambsganss et al., 2020) are a key requirement for argumentative essay writing; pre-service teachers' initial solutions about a case vignette can be analysed, for example, concerning epistemic activities, such as identifying the problem, generating hypotheses, generating evidence, evaluating evidence, and drawing conclusions (Bauer et al., 2020; Fischer et al., 2014).

NLP automation at this stage can be based on the current and prior initial solution texts as well as the sample or expert solution texts available to the instructor. Detecting the structural aspects of initial solutions can be naturally cast as a span and relation extraction or as a text categorization task, backed by the existing body of work in NLP: automatic argumentation mining is an active research area (Lippi & Torroni, 2016; Lugini & Litman, 2020); NLP-based automatic analysis of epistemic activities in diagnostic essays written by pre-service teachers and medical students (Schulz et al., 2019) has been used to automatically provide adaptive feedback (Pfeiffer et al., 2019; Sailer et al., 2023). Moreover, state of the art approaches for recognizing textual entailment (Bowman et al., 2015) can be used to judge the validity of the claims made in the solution cast as a text pair classification task; the recent advances in text similarity measurement (Reimers & Gurevych, 2019), text generation quality metrics (Zhang et al., 2020), and linking (Kuznetsov et al., 2022) can be used to compare learner's solutions to sample solutions and determine the differences to be communicated to the learner to facilitate task processing.

Reviewing peer solutions

After submitting the initial solutions, learners adopt the role of peer reviewers and receive the initial solutions of their peer learners for review. The peer reviewers subsequently analyse the assigned initial solutions and compare them to the assumed performance goal. The reviewers thus prepare for producing a feedback message (Gielen & De Wever, 2015) and learn from the alternative solutions (Cho & Cho, 2011; Lundstrom & Baker, 2009; Nicol et al., 2014). However, learners' task-relevant knowledge may limit the effectiveness of reviewing. Therefore, a *leverage point* for reviewing is to facilitate learners' understanding of the performance goal—that is, the characteristics that separate adequate and inadequate initial solutions (Berndt et al., 2018; Huisman et al., 2018; Peters et al., 2018).

Prior studies have provided *support measures*, such as evaluation rubrics, worked examples, and templates with assessment criteria (Alemdag & Yildirim, 2022; Alqassab et al., 2018b; Gielen & De Wever, 2015; Peters et al., 2018; Rotsaert et al., 2018; Voet et al., 2018). Adaptive support measures might highlight or annotate the structural or content aspects of initial solutions—essentially analogous to the support measures for task processing. In addition, the relationships between a learner's initial solution and the reviewed initial solution can be highlighted and communicated to the learner. The reviewing support can be adjusted to two *adaptation targets*: (a) reviewers may receive support adapted to the reviewed initial solution, which might be implemented as a default support or an adaptable support; (b) the degree of reviewing support might be adapted to the reviewer's level of task-relevant competencies, indicated by the performance in their own initial solution. Similar to the initial task-processing support, the *automation goals* for adaptive reviewing support entail detecting the relevant structural aspects or content aspects in the initial solution texts.

Given the similarity of goals, NLP automation developed for the task phase can be reused to provide adaptive support in the reviewing phase. The availability of one's own and other learners' initial solution texts at this stage allows the application of the previously developed NLP models to compare initial solutions to sample solutions. The solution under review can be compared to one's own solution text, which highlights the differences and establishes cross-document links.

Feedback production

After reviewing a peer's initial solution, the reviewers craft a feedback message, and the quality of the message determines its effectiveness in facilitating learning processes and outcomes (Hattie & Timperley, 2007; Narciss et al., 2014). Several models for conceptualizing feedback (Panadero & Lipnevich, 2022) and peer-feedback (eg, Patchan et al., 2016; Wu & Schunn, 2020) exist; the most well-known model, which addresses feedback quality, was introduced by Hattie and Timperley (2007), according to whom high-quality feedback is structured based on three feedback questions that address (1) the performance goal (feed-up), (2) the current performance (feed-back), and (3) the possible improvements (feed-forward). They offer recommendations concerning four *content levels* of feedback: Novice learners require guidance on the task processes (process level); intermediate learners can already benefit from guidance for self-monitoring (self-regulation level); for advanced learners, corrective performance-information can suffice (task level); and personal evaluations serve social-affective purposes instead of improving learning outcomes (person level). Therefore, a central *leverage point* for increasing the effectiveness of peer-feedback processes entails advancing the quality of the peer-feedback messages regarding the degree of elaboration, structuring, and the choice of content level for the feedback.

The *support measures* in prior research addressed the three feedback questions and four content levels by providing instructions and prompts, assessment scripts with structural scaffolds, structural guidelines, and structural templates (Alqassab et al., 2018b; Gielen & De Wever, 2015). Possible *adaptation targets* for adjusting feedback support would be: (1) the reviewer's feedback skills, indicated by a submitted draft of the current feedback message (eg, for prompts concerning the feedback questions) or feedback messages from prior peer-feedback scenarios (eg, for prompts on avoiding common structural flaws); (2) the peer's task performance in the reviewed initial solution (eg, for prompts concerning the content level for the feedback); (3) the reviewer's task performance in the initial solution (for personalizing the degree of feedback support). The *automation goals* for adaptive feedback support focus on automatically detecting relevant aspects of feedback messages (in addition to analyses of the prior initial solution texts).

NLP automation for the feedback production phase can utilize a wide spectrum of textual data as input, including one's own and other learners' initial solutions and the feedback message draft. Feedback messages are subject to structural analysis cast as span and relation extraction or as a text categorization task. Prior NLP studies successfully applied discourse parsing to analyse feedback messages in scholarly peer review. For example, Kuznetsov et al. (2022) proposed a corpus of scholarly feedback texts in which each sentence is labelled with pragmatic categories, such as recap (feed-up), strength and weakness (feedback), todo (feed-forward) and so on. Alternatively, Hua et al. (2019) focused on extracting argumentation structures, labelling parts of peer review reports as evaluation, fact, request and so on. Similar efforts exist in the domain of English argumentative essay writing; for example, Nguyen et al. (2016) proposed, implemented, and deployed an instant feedback system to detect solutions in peer review texts. The automatic analysis of the content level addressed in feedback texts has not been widely studied, but it might be approached similarly as a text classification or a span extraction NLP task.

Feedback processing

To process the feedback, learners need to read the received message(s) and try to comprehend it and evaluate its relevance. Their mindful processing of the feedback—that is, paying attention to the relevant information and investing mental effort for comprehension—can increase their recall performance (Alemdag & Yildirim, 2022; Bolzer et al., 2015). Mindful processing is particularly important for dealing with complex feedback (Berndt et al., 2022). However, if learners lack motivation or are overwhelmed by feedback (Aben et al., 2022; Huisman et al., 2018, 2020; Strijbos et al., 2021), they can react maladaptively, for example, by ignoring the feedback (Butler & Winne, 1995). In addition, to minimize the impact of varying feedback quality, receiving feedback from multiple peers can average the feedback information, thus increasing the objectivity of the peer-feedback (Cho & Schunn, 2007; De Wever et al., 2011). This advantage, however, must be balanced with the downside of increased effort for processing multiple feedback messages (Rouet & Britt, 2011). Therefore, the *leverage points* for supporting learners' feedback processing entail facilitating the mindful processing and information integration of multiple feedback messages.

As a *support measure* for facilitating feedback processing and comprehension, prior research explored the use of rubrics (Wichmann et al., 2018). Such scaffolds might also be used to support the processing and integration of multiple feedback messages. The *adaptation targets* for adaptive processing support primarily comprise the received feedback message(s). In addition, the amount of processing support can be personalized to the feedback receiver based on their initial solution and their previous feedback messages. The *automation goals* for such adaptive support measures include, for example, the automatic detection of the relevant aspects in feedback messages (eg, for prompting or highlighting) or summarizing the information common in multiple feedback messages.

NLP for assisting peer-feedback processing largely follows the tasks required for feedback production, including the structural and content level analysis of peer-feedback texts. However, the feedback processing stage offers new opportunities for NLP automation: since each learner receives more than one feedback message from their peers in many scenarios, NLP can be applied to aggregate the information from these messages, either by summarizing them (text generation) or by connecting multiple feedback texts to the initial solution text (linking). The former goal can be supported by existing NLP research on multi-document summarization (Fabbri et al., 2019); the latter can be addressed using the latest developments in NLP for scholarly peer review analysis (Kennard et al., 2022; Kuznetsov et al., 2022) that study the problem of connecting paper drafts to peer review texts and author responses as well as emerging general-purpose NLP approaches to multi-document language modelling (Caciularu et al., 2021).

Revision

After processing the received feedback, learners will ideally revise their initial solution, which offers them a range of benefits for learning (Linn et al., 2013): recognizing new ideas in the feedback messages and adding them to their understanding of the subject matter, generating connections between ideas, and monitoring their progress while working these new ideas into their initial solution. However, while revising, learners (a) must avoid making new mistakes or adopting incorrect feedback comments (Wichmann et al., 2018); moreover,

(b) they might experience difficulties making fundamental changes to their initial solution, instead focusing on micro-level changes, such as clarifying or elaborating a sentence or paragraph (Cho & MacArthur, 2010). Therefore, the *leverage points* for revision support include avoiding new mistakes and facilitating making fundamental changes when revising initial solutions.

As a *support measure* for adopting only adequate ideas, Wichmann et al. (2018) used a rubric, which prompted strategic behaviours, such as reflecting on the feedback information and engaging in planning, monitoring, and evaluating throughout revisions. Directly regulating scaffolds or awareness support, such as highlighting, can be used to identify the parts in the initial solution that must be revised according to the received feedback. In particular, if suggested in the feedback, the need to make fundamental changes can be emphasized. The *adaptation targets* for revision support are primarily the received feedback messages. In addition, the degree of revision support might be personalized for the learner by adjusting it to their prior performance (eg, initial solution and quality of feedback message). The *automation goals* for revision support measures might entail detecting revision-critical information in the feedback messages and connecting it to the parts in the initial solution and comparing the revised solution draft to the initial solution by considering the received feedback.

In addition to the NLP automation for feedback production and processing assistance, the availability of the revised solution draft enables new NLP automation scenarios to support the revision automation goals. The revised solution can be compared to the initial solution via version alignment, and the differences introduced during revision can be attributed to the received feedback messages: a linking task type was recently proposed by Kuznetsov et al. (2022) in the context of scholarly peer review; related efforts also exist in the domain of English argumentative writing (Afrin et al., 2020; Afrin & Litman, 2018; Zhang et al., 2017). The automatic classification of edit purposes and intentions (eg, as text pair classification) was explored in a research on Wikipedia (Yang et al., 2017). In contrast, judging whether the edits indeed address the feedback is an underexplored NLP area related to recent studies in edit- and change-aware NLP and language modelling (Logan IV et al., 2022; Schick et al., 2022).

Evaluating the feedback process and learning outcomes

Experiences with activities in the peer-feedback process feed into the individual learners' internal evaluation of the peer-feedback process and learning outcomes (Nicol, 2021). For example, learners who evaluate the received feedback as low-quality might conclude that peer-feedback is not a beneficial instructional approach (Kaufman & Schunn, 2011). With such an attitude, they will likely invest less effort into future peer-feedback scenarios and offer low-quality peer-feedback themselves. This makes learners' evaluation a relevant condition for the benefits of future learning with peer-feedback. However, without external guidance, learners' final evaluations often remain rather superficial (Anderson, 2012). An important *leverage point* might thus entail provoking learners' in-depth reflection about the peer-feedback process, its benefits and options for future improvements.

As *support measures* for in-depth evaluation, learners' awareness of the various aspects of the peer-feedback process might be increased. This could be achieved by using reflection prompts that address aspects of the peer-feedback process, such as the intended behavioural changes for future peer-feedback activities or the learning outcomes gained from receiving the feedback (Anderson, 2012; Raković et al., 2022). The summary statistics and visuals of the tracked changes or rebutted feedback information might provide further options for supporting individual learners in their final evaluation. These options could be employed for learners in the roles of both feedback receiver and reviewer. In addition, such summarizing information could be provided at a group level (eg, a whole class or course) to illustrate the benefits and challenges of collaborative knowledge construction—that is, how information (eg, advances but also misconceptions) is distributed and "travelled" between learners (Huisman et al., 2020). The *adaptation targets* might primarily entail distinguishing learners' successes and difficulties in the different activities of peer-feedback (eg, incorporating vs. not incorporating the feedback information) and offering personalized instructions.

The *automation goals* for these support measures comprise comparing the revised and initial solution, determining the extent to which the changes reflect the feedback and how similar the revised solution is to a sample solution (if the task allows the creation of such a sample solution), and calculating the group-level indicators based on the information from individual peer-feedback loops (eg, the extent of revisions as well as the most frequently implemented feedback targets and structural elements). At this stage, the results of prior NLP automation can be used to produce aggregate statistics and insights from the peer-feedback process. Additionally, unsupervised NLP approaches can be applied to aggregate and cluster parts of the task solution and feedback message texts to determine the most commonly mentioned topics and improvement suggestions, for example, by using neural sentence representations (Reimers & Gurevych, 2019).

CURRENT CHALLENGES AND A RESEARCH AGENDA

In this paper, we proposed a peer-feedback process model that offers an overview of peer learners' activities and products, developed a terminological and procedural scheme to bridge educational research and NLP perspectives, and applied the latter to the former to exemplify how each peer-feedback activity might be adaptively supported using NLP. As our examples show, peer-feedback offers a wide ground for applying NLP support, from the pre-scoring of initial solutions during peer assignment, to judging the sufficiency of changes resulting from peer-feedback, to extracting new insights about the overall peer-feedback process. NLP models, we note, can often be repurposed between phases as part of different support measures—yet, the extrinsic evaluation measures remain specific to each phase. For example, while a feedback provider might be interested in writing a comprehensive feedback message, a feedback receiver might be interested in investing reduced feedback processing time. We also note that, as the peer-feedback scenario proceeds, more information becomes available for potential support measures, and more data becomes accessible for developing NLP automation models to enhance those measures.

In the following, we describe current challenges and opportunities that arise when applying NLP to peer-feedback and propose an agenda of seven major themes for future research directions:

1. The *effectiveness* of NLP-based adaptive support measures for all phases and activities of the peer-feedback process constitutes the primary desideratum to investigate in future research. NLP bears great potential to amplify the effectiveness and efficiency of peer-feedback by augmenting and automating adaptive support measures that target the relevant leverage points of an effective peer-feedback process. Unlike many other applications of NLP that are primarily concerned with improving the efficiency of text work per se, the extrinsic evaluation of NLP applications in the educational domain is additionally concerned with facilitating learners' skills or other latent learner variables (eg, affective-motivational states). Such variables are challenging to define and measure, requiring close cross-disciplinary collaboration between NLP practitioners and educational scientists. Future research might investigate the relative

effectiveness of different NLP-based adaptive support measures—that is, for which phases and activities NLP-based adaptive support can bring the most merit according to the chosen extrinsic evaluation metrics. In addition, within the phases and activities, it remains unclear which support measures are most effective for facilitating peer-feedback processes and increasing learning outcomes (eg, indirectly vs. directly regulating support, adaptive vs. adaptable support, and so on) and which adaptive support measures are best to be automated or augmented by NLP. Multiple support measures might also create synergies for learning when used in combination (Tabak & Kyza, 2018), which suggests potential for further research on the combinations of NLP-based learner support. Such research questions might be investigated in studies using pretest-posttest control group designs. The initial investigations might be conducted in laboratory settings that allow detailed investigations of the intervention effects on learning and peer-feedback processes; subsequently, it seems important to conduct field studies in actual classrooms to further investigate the practical feasibility of the NLP adaptive support measures for different peer-feedback scenarios (eg, smaller classes vs. lectures) and to investigate the stability of effects found in the laboratory under field conditions.

2. The *generalizability of peer-feedback processes* as described in the suggested peer-feedback process model across peer-feedback scenarios with varying contextual characteristics is another open question. Our examples mostly presume a linear and nicely pre-structured peer-feedback scenario in which learners follow a more or less fixed sequence without synchronously communicating much during the individual activities. Peer-feedback scenarios that require a higher degree of self-regulation and co-regulation might impose additional challenges (Greisel et al., 2021; Koivuniemi et al., 2017) but also potential for NLP-based adaptive support measures. Moreover, the characteristics of the learning task—its complexity or dynamics (see Fischer et al., 2022) or the non-existence of a correct solution (see Fischer & Wolf, 2015)—might be relevant to both the generalizability of the model as well as the effectiveness of NLP-based adaptive support measures. Educational research on peer-feedback needs to continue with detailed process analyses to investigate the learning processes during peer-feedback as well as the potential influences of varying learner characteristics, context characteristics and task characteristics, which might also affect the leverage points for optimal learner support.
3. In addition, future research must further investigate the *potential leverage points and support measures for facilitating peer-feedback processes*. In this paper, we exemplified a limited number of potential leverage points and support measures; they do not represent a comprehensive list and must be complemented. We mostly focused on the cognitive level of the peer-feedback process. However, the peer-feedback process also involves affective-motivational and social-dialogical levels (eg, Aben et al., 2019; Narciss et al., 2014), which involve a variety of further options for leveraging the effectiveness of the peer-feedback process and increasing learners' benefits by providing suitable adaptive support measures. The leverage points and support measures targeting these levels of the peer-feedback process must be identified and systematically investigated. For the cognitive leverage points and support measures, one can distinguish between structural dimensions (eg, argumentation structures or the three feedback questions of high-quality feedback) and content dimensions (eg, the correctness of an initial solution or the feedback message concerning the content of the learning task) in the products of the peer-feedback process. The content aspects facilitate specifying the support needed for the content of the learning task, which might be more comprehensible and, thus, more beneficial for the learners; however, structural aspects might facilitate learning transfer across different learning tasks (Hetmanek et al., 2018). The comparative effectiveness of targeting these

two dimensions with NLP-based adaptive learner support requires further investigation as well.

4. The *generalizability of collaborative processes between NLP and education*, as described by the proposed terminological and procedural scheme for designing NLP-based adaptive support measures, is another open question. We developed this scheme motivated by the idea of researching peer-feedback at the intersection of NLP and education. The use case of supporting peer-feedback involves some characteristics that seem relevant to consider for designing automatic adaptive learner support; for example, the peer-feedback process implies early access to text data (eg, initial solution) to which learner support can be adapted. Another important characteristic is that NLP is used to augment the peer-feedback process to give students further opportunities for learning, whereas other use cases would strive for complete automatization (eg, fully automating adaptive feedback; eg, Sailer et al., 2023). Despite these specificities of using NLP for supporting peer-feedback, the developed scheme might as well be beneficial for facilitating other cross-disciplinary collaboration at the intersection of NLP and education, such as automating adaptive learner support for educational contexts other than peer-feedback (eg, supporting essay writing) and for developing teaching support for instructors (eg, dashboards for teachers or lecturers). However, while a joint terminology and framework can facilitate cross-disciplinary research (Heitzmann et al., 2021), the generalizability of the suggested scheme for further cross-disciplinary collaborations between NLP and education remains to be explored.
5. A major challenge and research direction concerning NLP entails handling *data scarcity*. Modern NLP is driven by the availability of data; however, compared to other application areas of NLP, such as news and social networks, peer-feedback data is scarce. Although many of the NLP approaches mentioned in our overview come bundled with labelled datasets and pre-trained models, performance degradation due to domain and language shifts can prevent their wide reuse. For instance, an argumentative essay scoring model for English will underperform on German medical case-study essays. Even the most advanced NLP systems suffer from substantial performance decrease when applied to previously unseen languages and domains. The application of NLP to peer-feedback thus demands *advancements in domain and language adaptation technology* (Chronopoulou et al., 2022; Pfeiffer et al., 2020), as well as *quantifying and collecting data from pre-existing peer-feedback workflows* (Dycke et al., 2022, 2023). Simultaneously, peer-feedback has the potential to generate great amounts of diverse textual data for NLP research and to provide excellent testing grounds for the study of language- and domain-adaptation in NLP.
6. Handling important *properties of peer-feedback data*—such as handling biases and personal data—is also crucial. NLP models exhibit a wide range of biases, including reproducing *pre-existing biases* in the data (eg, non-native speakers' solution texts being graded lower), *technical bias* due to algorithm behaviour, as well as *emergent bias* from applying NLP models outside of their intended task and data distribution (Bender & Friedman, 2018). While some application areas are less sensitive to bias, it is clearly an undesirable property for peer-feedback support, as even small and undetectable biases can be amplified by repeated exposure and thus put the affected learner groups at a disadvantage. Bias can be addressed by carefully documenting and curating NLP datasets and models (Bender & Friedman, 2018; Mohammad, 2022), applying specialized debiasing techniques (Utama et al., 2020), and choosing automation goals less vulnerable to known biases. At the same time, the diversity of the participants involved in peer-feedback and the resulting data form an excellent basis for the study of bias mitigation in NLP. In addition, when applying NLP to peer-feedback, the ethical and legal challenges related to learners' *personal data management* and *data licensing* must be addressed. While responsible research is already gaining traction in NLP (Dycke et al., 2022, 2023; Rogers et al., 2021), we envision that

cross-disciplinary collaborations between the fields would greatly enrich the overall data and participant handling practice in NLP.

7. The *interpretability* of NLP models often seems to be a crucial requirement for peer-feedback support: simply informing a learner that their peer-feedback is insufficient is likely not as useful as being able to highlight the problematic instances, for example, in the feedback message. Yet, despite their predictive power, modern NLP models are notoriously opaque: while state-of-the-art neural models, such as GPT-4, might offer correct predictions for a wide range of tasks, they lack the ability to reliably explain *why* a prediction is made. While potentially acceptable in many other scenarios, such as predicting the sentiment of tweets, this is clearly problematic for NLP application in the context of learner support. There is an active line of research on NLP that addresses the interpretability of NLP models (Tenney et al., 2020), and further research on their interpretability can be conducted in the context of supporting peer-feedback. Applying NLP to peer-feedback offers an excellent opportunity to study related questions of human–AI interaction, such as what kind of explanations are most helpful and required to increase interpretability and support the learners' understanding of their learning processes. Such research could benefit from combining different research approaches: qualitative interview studies could explore learners' perceptions and acceptance of technology enhancement; moreover, large-scale studies could use learning analytics to understand how learners adopt the behavioural changes suggested by automated learning support.

Future research and design work in the fields of education and NLP might build on the proposed cross-disciplinary framework and research agenda to collaborate more systematically and innovate peer-feedback in digital learning environments.

ACKNOWLEDGEMENTS

This research was supported by the German Federal Ministry of Research and Education (FAMULUS-Project 16DHL1040), Deutsche Forschungsgemeinschaft (DFG FOR 2385) and Stiftung Innovation in der Hochschullehre (FMM120-2020). Open Access funding enabled and organized by Projekt DEAL.

CONFLICT OF INTEREST STATEMENT

We have no known conflict of interest to disclose.

DATA AVAILABILITY STATEMENT

Data sharing not applicable to this article as no datasets were generated or analysed during the current study.

ETHICS STATEMENT

This article does not present primary data or primary studies involving human participants performed by any of the authors.

ORCID

Elisabeth Bauer https://orcid.org/0000-0003-4078-0999
Martin Greisel https://orcid.org/0000-0002-9586-5714
Ilia Kuznetsov https://orcid.org/0000-0002-6359-2774
Markus Berndt https://orcid.org/0000-0002-4467-5355
Ingo Kollar https://orcid.org/0000-0001-9257-5028
Markus Dresel https://orcid.org/0000-0002-2131-3749
Martin R. Fischer https://orcid.org/0000-0002-5299-5025
Frank Fischer https://orcid.org/0000-0003-0253-659X

REFERENCES

Aben, J. E., Dingyloudi, F., Timmermans, A. C., & Strijbos, J. W. (2019). Embracing errors for learning: Intrapersonal and interpersonal factors in feedback provision and processing in dyadic interactions. In M. Henderson, R. Ajjawi, D. Boud, & E. Molloy (Eds.), *The impact of feedback in higher education* (pp. 107–125). Palgrave Macmillan. https://doi.org/10.1007/978-3-030-25112-3_7

Aben, J. E. J., Timmermans, A. C., Dingyloudi, F., Lara, M. M., & Strijbos, J.-W. (2022). What influences students' peer-feedback uptake? Relations between error tolerance, feedback tolerance, writing self-efficacy, perceived language skills and peer-feedback processing. *Learning and Individual Differences*, *97*, 102175. https://doi.org/10.1016/j.lindif.2022.102175

Afrin, T., & Litman, D. (2018). Annotation and classification of sentence-level revision improvement. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 240–246). New Orleans, LA. Association for Computational Linguistics. https://doi.org/10.18653/v1/W18-0528

Afrin, T., Wang, E., Litman, D., Matsumura, L. C., & Correnti, R. (2020). Annotation and classification of evidence and reasoning revisions in argumentative writing. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 75–84). Seattle, WA, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.bea-1.7

Agarwal, A., Xie, B., Vovsha, I., Rambow, O., & Passonneau, R. J. (2011). Sentiment analysis of Twitter data. In *Proceedings of the Workshop on Language in Social Media* (pp. 30–38). Portland, OR. Association for Computational Linguistics. https://aclanthology.org/W11-0705.pdf

Alemdag, E., & Yildirim, Z. (2022). Effectiveness of online regulation scaffolds on peer feedback provision and uptake: A mixed methods study. *Computers & Education*, *188*, 104574. https://doi.org/10.1016/j.compedu.2022.104574

Alqassab, M., Strijbos, J. W., & Ufer, S. (2018a). The impact of peer solution quality on peer-feedback provision on geometry proofs: Evidence from eye-movement analysis. *Learning and Instruction*, *58*, 182–192. https://doi.org/10.1016/j.learninstruc.2018.07.003

Alqassab, M., Strijbos, J. W., & Ufer, S. (2018b). Training peer-feedback skills on geometric construction tasks: Role of domain knowledge and peer-feedback levels. *European Journal of Psychology of Education*, *33*(1), 11–30. https://doi.org/10.1007/s10212-017-0342-0

Anderson, N. J. (2012). Student involvement in assessment: Healthy self-assessment and effective peer assessment. In C. Coombe, P. Davidson, B. O'Sullivan, & S. Stoynoff (Eds.), *The Cambridge guide to second language assessment* (pp. 187–197). Cambridge University Press.

Bannert, M., & Mengelkamp, C. (2013). Scaffolding hypermedia learning through metacognitive prompts. In R. Azevedo & V. Aleven (Eds.), *International handbook of metacognition and learning technologies*. Springer. https://doi.org/10.1007/978-1-4419-5546-3_12

Bauer, E., Fischer, F., Kiesewetter, J., Shaffer, D. W., Fischer, M. R., Zottmann, J. M., & Sailer, M. (2020). Diagnostic activities and diagnostic practices in medical education and teacher education: An interdisciplinary comparison. *Frontiers in Psychology*, *11*, 562665. https://doi.org/10.3389/fpsyg.2020.562665

Belland, B. R. (2014). Scaffolding: Definition, current debates, and future directions. In J. Spector, M. Merrill, J. Elen, & M. Bishop (Eds.), *Handbook of research on educational communications and technology* (pp. 505–518). Springer. https://doi.org/10.1007/978-1-4614-3185-5_39

Bender, E. M., & Friedman, B. (2018). Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, *6*, 587–604. https://doi.org/10.1162/tacl_a_00041

Berndt, M., Strijbos, J. W., & Fischer, F. (2018). Effects of written peer-feedback content and sender's competence on perceptions, performance, and mindful cognitive processing. *European Journal of Psychology of Education*, *33*(1), 31–49. https://doi.org/10.1007/s10212-017-0343-z

Berndt, M., Strijbos, J. W., & Fischer, F. (2022). Impact of sender and peer-feedback characteristics on performance, cognitive load, and mindful cognitive processing. *Studies in Educational Evaluation*, *75*, 101197. https://doi.org/10.1016/j.stueduc.2022.101197

Blömeke, S., Gustafsson, J.-E., & Shavelson, R. J. (2015). Beyond dichotomies: Competence viewed as a continuum. *Zeitschrift für Psychologie*, *223*, 3–13. https://doi.org/10.1027/2151-2604/a000194

Bolzer, M., Strijbos, J. W., & Fischer, F. (2015). Inferring mindful cognitive-processing of peer-feedback via eye-tracking: Role of feedback-characteristics, fixation-durations and transitions. *Journal of Computer Assisted Learning*, *31*(5), 422–434. https://doi.org/10.1111/jcal.12091

Bowman, S. R., Angeli, G., Potts, C., & Manning, C. D. (2015). A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing* (pp. 632–642). Lisbon, Portugal. Association for Computational Linguistics. https://doi.org/10.48550/arXiv.1508.05326

Brewer, S., & Klein, J. D. (2006). Type of positive interdependence and affiliation motive in an asynchronous, collaborative learning environment. *Educational Technology Research and Development*, *54*(4), 331–354. https://doi.org/10.1007/s11423-006-9603-3

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., … Amodei, D. (2020). Language models are few-shot learners. In *Advances in neural information processing systems* (pp. 1877–1901). Curran Associates. https://doi.org/10.18653/v1/2021.mrl-1.1

Burstein, J. (2003). The E-rater® scoring engine: Automated essay scoring with natural language processing. In M. D. Shermis & J. C. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 113–121). Routledge. https://doi.org/10.4324/9781410606860

Burstein, J., Kukich, K., Wolff, S., Lu, C., Chodorow, M., Braden-Harder, L., & Harris, M. D. (1998). Automated scoring using a hybrid feature identification technique. In *COLING 1998 Volume 1: The 17th International Conference on Computational Linguistics*. https://aclanthology.org/C98-1032

Butler, D. L., & Winne, P. H. (1995). Feedback and self-regulated learning: A theoretical synthesis. *Review of Educational Research*, *65*(3), 245–281. https://doi.org/10.3102/00346543065003245

Caciularu, A., Cohan, A., Beltagy, I., Peters, M. E., Cattan, A., & Dagan, I. (2021). CDLM: Cross-document language modeling. In *Findings of the Association for Computational Linguistics: EMNLP 2021* (pp. 2648–2662). Punta Cana, Dominican Republic. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.findings-emnlp.225

Cheng, L., Bing, L., Yu, Q., Lu, W., & Si, L. (2020). APE: Argument pair extraction from peer review and rebuttal via multi-task learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing* (pp. 7000–7011). Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.569

Cho, K., & MacArthur, C. (2010). Student revision with peer and expert reviewing. *Learning and Instruction*, *20*(4), 328–338. https://doi.org/10.1016/j.learninstruc.2009.08.006

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. https://doi.org/10.1016/j.compedu.2005.02.004

Cho, Y. H., & Cho, K. (2011). Peer reviewers learn from giving comments. *Instructional Science*, *39*(5), 629–643. https://doi.org/10.1007/s11251-010-9146-1

Chronopoulou, A., Peters, M. E., & Dodge, J. (2022). Efficient hierarchical domain adaptation for pretrained language models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 1336–1351). Seattle, WA. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.naacl-main.96

Dasgupta, T., Naskar, A., Dey, L., & Saha, R. (2018, July). Augmenting textual qualitative features in deep convolution recurrent neural network for automatic essay scoring. In *Proceedings of the 5th Workshop on Natural Language Processing Techniques for Educational Applications* (pp. 93–102). https://doi.org/10.18653/v1/W18-3713

Dasigi, P., Lo, K., Beltagy, I., Cohan, A., Smith, N. A., & Gardner, M. (2021). A dataset of information-seeking questions and answers anchored in research papers. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 4599–4610). Online. Association for Computational Linguistics. https://aclanthology.org/2021.naacl-main.365.pdf

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. https://doi.org/10.1016/j.iheduc.2011.07.003

Deutsch, T., Jasbi, M., & Shieber, S. (2020). Linguistic features for readability assessment. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 1–17). Seattle, WA, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.bea-1.1

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 4171–4186). Minneapolis, MN. Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423

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. https://doi.org/10.1007/s10648-019-09510-3

Dycke, N., Kuznetsov, I., & Gurevych, I. (2022). Yes-yes-yes: Proactive data collection for ACL rolling review and beyond. In *Findings of the Association for Computational Linguistics: EMNLP 2022* (pp. 300–318). Abu Dhabi, United Arab Emirates. Association for Computational Linguistics. https://aclanthology.org/2022.findings-emnlp.23.pdf

Dycke, N., Kuznetsov, I., & Gurevych, I. (2023). NLPeer: A unified resource for the computational study of peer review. *arXiv Preprints*. https://doi.org/10.48550/arXiv.2211.06651

Fabbri, A. R., Li, I., She, T., Li, S., & Radev, D. R. (2019). Multi-news: A large-scale multi-document summarization dataset and abstractive hierarchical model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 1074–1084). Florence, Italy. Association for Computational Linguistics. https://doi.org/10.48550/arXiv.1906.01749

Fischer, F., Bauer, E., Seidel, T., Schmidmaier, R., Radkowitsch, A., Neuhaus, B. J., Hofer, S. I., Sommerhoff, D., Ufer, S., Kuhn, J., Küchemann, S., Sailer, M., Koenen, J., Gartmeier, M., Berberat, P., Frenzel, A., Heitzmann, N., Holzberger, D., Pfeffer, J., … Fischer, M. R. (2022). Representational scaffolding in digital simulations—Learning professional practices in higher education. *Information and Learning Sciences*, *123*(11/12), 645–665. https://doi.org/10.1108/ILS-06-2022-0076

Fischer, F., Kollar, I., Ufer, S., Sodian, B., Hussmann, H., Pekrun, R., Neuhaus, B., Dorner, B., Pankofer, S., Fischer, M., Strijbos, J.-W., Heene, M., & Eberle, J. (2014). Scientific reasoning and argumentation: Advancing an interdisciplinary research agenda in education. *Frontline Learning Research*, *2*(3), 28–45. https://doi.org/10.14786/flr.v2i2.96

Fischer, G. (2001). User modeling in human–computer interaction. *User Modeling and User-Adapted Interaction*, *11*(1), 65–86. https://doi.org/10.1023/A:1011145532042

Fischer, G., & Wolf, K. D. (2015). What can residential, research-based universities learn about their core competencies from MOOCs (massive open online courses)? In H. Schelhowe, M. Schaumburg, & J. Jasper (Eds.), *Teaching is touching the future. Academic teaching within and across disciplines* (pp. 65–75). UniversitätsVerlagWebler.

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. https://doi.org/10.1111/jcal.12096

Greene, J. A., & Azevedo, R. (2009). A macro-level analysis of SRL processes and their relations to the acquisition of a sophisticated mental model of a complex system. *Contemporary Educational Psychology*, *34*(1), 18–29. https://doi.org/10.1016/j.cedpsych.2008.05.006

Greisel, M., Spang, L., Fett, K., Melzner, N., Dresel, M., & Kollar, I. (2021). "Houston, we have a problem!" Homogeneous problem perception, and immediacy and intensity of strategy use in online collaborative learning. In C. E. Hmelo-Silver, B. De Wever, & J. Oshima (Eds.), *Proceedings of the 14th International Conference on Computer-Supported Collaborative Learning—CSCL 2021* (pp. 99–106). International Society of the Learning Sciences. https://repository.isls.org//handle/1/7365

Große, C. S., & Renkl, A. (2007). Finding and fixing errors in worked examples: Can this foster learning outcomes? *Learning and Instruction*, *17*(6), 612–634. https://doi.org/10.1016/j.learninstruc.2007.09.008

Hattie, J. (2008). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. Routledge. https://doi.org/10.4324/9780203887332

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, *77*(1), 81–112. https://doi.org/10.3102/003465430298487

Heitzmann, N., Opitz, A., Stadler, M., Sommerhoff, D., Fink, M. C., Obersteiner, A., Schmidmaier, R., Neuhaus, B. J., Ufer, S., Seidel, T., Fischer, M. R., & Fischer, F. (2021). Cross-disciplinary research on learning and instruction—Coming to terms. *Frontiers in Psychology*, *1539*, 562658. https://doi.org/10.3389/fpsyg.2021.562658

Hetmanek, A., Engelmann, K., Opitz, A., & Fischer, F. (2018). Beyond intelligence and domain knowledge: Scientific reasoning and argumentation as a set of cross-domain skills. In F. Fischer, A. C. Clark, K. Engelmann, & J. Osborne (Eds.), *Scientific reasoning and argumentation: The roles of domain-specific and domain-general knowledge* (pp. 203–226). Routledge. https://doi.org/10.4324/9780203731826

Hornstein, J., Greisel, M., Ott, J., Weidenbacher, A., & Kollar, I. (2023). Promoting evidence-informed reasoning in student teachers through peer feedback [Paper presentation]. In *20th Biennial EARLI Conference Thessaloniki*, Greece.

Hua, X., Nikolov, M., Badugu, N., & Wang, L. (2019). Argument mining for understanding peer reviews. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 2131–2137). Minneapolis, MN. Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1219

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. https://doi.org/10.1080/02602938.2018.1424318

Huisman, B., Saab, N., Van Driel, J., & Van Den Broek, P. (2020). A questionnaire to assess students' beliefs about peer-feedback. *Innovations in Education and Teaching International*, *57*(3), 328–338. https://doi.org/10.1080/14703297.2019.1630294

Kalyuga, S. (2007). Expertise reversal effect and its implications for learner-tailored instruction. *Educational Psychology Review*, *19*(4), 509–539. https://doi.org/10.1007/s10648-007-9054-3

Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., … Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, *103*, 102274. https://doi.org/10.1016/j.lindif.2023.102274

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

Ke, Z., & Ng, V. (2019). Automated essay scoring: A survey of the state of the art. In *Proceedings of IJCAI-19* (pp. 6300–6308). https://www.ijcai.org/proceedings/2019/0879.pdf

Kennard, N., O'Gorman, T., Das, R., Sharma, A., Bagchi, C., Clinton, M., Yelugam, P. K., Zamani, H., & McCallum, A. (2022). DISAPERE: A dataset for discourse structure in peer review discussions. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 1234–1249). Seattle, WA. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.naacl-main.89

Koivuniemi, M., Panadero, E., Malmberg, J., & Järvelä, S. (2017). Higher education students' learning challenges and regulatory skills in different learning situations/Desafíos de aprendizaje y habilidades de regulación en distintas situaciones de aprendizaje en estudiantes de educación superior. *Infancia y Aprendizaje*, *40*(1), 19–55. https://doi.org/10.1080/02103702.2016.1272874

Kucirkova, N., Gerard, L., & Linn, M. C. (2021). Designing personalised instruction: A research and design framework. *British Journal of Educational Technology*, *52*(5), 1839–1861. https://doi.org/10.1111/bjet.13119

Kuznetsov, I., Buchmann, J., Eichler, M., & Gurevych, I. (2022). Revise and resubmit: An intertextual model of text-based collaboration in peer review. *Computational Linguistics*, *48*(4), 1–38. https://doi.org/10.1162/coli_a_00455

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. https://doi.org/10.1080/02602938.2019.1620679

Lin, J. W., & Tsai, C. W. (2016). The impact of an online project-based learning environment with group awareness support on students with different self-regulation levels: An extended-period experiment. *Computers & Education*, *99*, 28–38. https://doi.org/10.1016/j.compedu.2016.04.005

Linn, M. C., Eylon, B.-S., & Davis, E. A. (2013). The knowledge integration perspective on learning. In M. C. Linn, E. A. Davis, & P. Bell (Eds.), *Internet environments for science education* (pp. 29–46). Routledge. https://doi.org/10.4324/9781410610393

Lippi, M., & Torroni, P. (2016). Argumentation mining: State of the art and emerging trends. *ACM Transactions on Internet Technology (TOIT)*, *16*(2), 1–25. https://doi.org/10.1145/2850417

Liu, L., Hao, J., von Davier, A. A., Kyllonen, P., & Zapata-Rivera, J. (2016). A tough nut to crack: Measuring collaborative problem solving. In Y. Rosen, S. Ferrara, & M. Mosharraf (Eds.), *Handbook of research on technology tools for real-world skill development* (pp. 344–359). IGI Global. https://doi.org/10.4018/978-1-4666-9441-5.ch013

Logan, R. L., IV, Passos, A., Singh, S., & Chang, M. W. (2022). FRUIT: Faithfully reflecting updated information in text. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 3670–3686). Seattle, WA. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.naacl-main.269

Lugini, L., & Litman, D. (2017). Predicting specificity in classroom discussion. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 52–61). Copenhagen, Denmark. Association for Computational Linguistics. https://doi.org/10.18653/v1/W17-5006

Lugini, L., & Litman, D. (2020). Contextual argument component classification for class discussions. In *Proceedings of the 28th International Conference on Computational Linguistics* (pp. 1475–1480). Barcelona, Spain. International Committee on Computational Linguistics. https://doi.org/10.18653/v1/2020.coling-main.128

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*, 30–43. https://doi.org/10.1016/j.jslw.2008.06.002

Maynez, J., Narayan, S., Bohnet, B., & McDonald, R. (2020). On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 1906–1919). Online. Association for Computational Linguistics. https://aclanthology.org/2020.acl-main.173.pdf

Michel, C., Lavoué, E., George, S., & Ji, M. (2017). Supporting awareness and self-regulation in project-based learning through personalized dashboards. *International Journal of Technology Enhanced Learning*, *9*(2/3), 204–226. https://doi.org/10.1504/IJTEL.2017.084500

Mohammad, S. (2022). Ethics sheets for AI tasks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics* (pp. 8368–8379). Dublin, Ireland. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.acl-long.573

Narciss, S. (2008). Feedback strategies for interactive learning tasks. In D. Jonassen, M. J. Spector, M. Driscoll, M. D. Merrill, J. van Merrienboer, & M. P. Driscoll (Eds.), *Handbook of research on educational communications and technology* (3rd ed., pp. 125–144). Routledge. https://doi.org/10.4324/9780203880869

Narciss, S., Sosnovsky, S., Schnaubert, L., Andrès, E., Eichelmann, A., Goguadze, G., & Melis, E. (2014). Exploring feedback and student characteristics relevant for personalizing feedback strategies. *Computers & Education*, *71*, 56–76. https://doi.org/10.1016/j.compedu.2013.09.011

Newman-Griffis, D., Lehman, J. F., Rosé, C., & Hochheiser, H. (2021). Translational NLP: A new paradigm and general principles for natural language processing research. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 4125–4138). Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.naacl-main.325

Nguyen, H., Xiong, W., & Litman, D. (2016). Instant feedback for increasing the presence of solutions in peer reviews. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations* (pp. 6–10). San Diego, CA. Association for Computational Linguistics. https://aclanthology.org/N16-3002.pdf

Nicol, D. (2021). The power of internal feedback: Exploiting natural comparison processes. *Assessment & Evaluation in Higher Education*, *46*(5), 756–778. https://doi.org/10.1080/02602938.2020.1823314

Nicol, D., Thomson, A., & Breslin, C. (2014). Rethinking feedback practices in higher education: A peer review perspective. *Assessment & Evaluation in Higher Education*, *39*(1), 102–122. https://doi.org/10.1080/02602938.2013.795518

Ninaus, M., & Sailer, M. (2022). Closing the loop—The human role in artificial intelligence for education. *Frontiers in Psychology*, *13*, Article 956798. https://doi.org/10.3389/fpsyg.2022.956798

OpenAI. (2023). GPT-4 Technical Report. *arXiv Preprints*. https://arxiv.org/abs/2303.08774

Page, E. B. (1968). The use of the computer in analyzing student essays. *International Review of Education*, *14*(2), 210–225. https://www.jstor.org/stable/3442515

Panadero, E. (2016). Is it safe? Social, interpersonal, and human effects of peer assessment. In G. Brown & L. Harris (Eds.), *Handbook of human and social conditions in assessment* (pp. 247–266). Routledge. https://doi.org/10.4324/9781315749136

Panadero, E., & Lipnevich, A. A. (2022). A review of feedback models and typologies: Towards an integrative model of feedback elements. *Educational Research Review*, *35*, 100416. https://doi.org/10.1016/j.edurev.2021.100416

Patchan, M. M., Hawk, B., Stevens, C. A., & Schunn, C. D. (2013). The effects of skill diversity on commenting and revisions. *Instructional Science*, *41*(2), 381–405. https://doi.org/10.1007/s11251-012-9236-3

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. https://doi.org/10.1007/s11251-015-9353-x

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. https://doi.org/10.1080/03075079.2017.1320374

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), 1098–1120. https://doi.org/10.1037/edu0000103

Peters, O., Körndle, H., & Narciss, S. (2018). Effects of a formative assessment script on how vocational students generate formative feedback to a peer's or their own performance. *European Journal of Psychology of Education*, *33*(1), 117–143. https://doi.org/10.1007/s10212-017-0344-y

Pfeiffer, J., Meyer, C. M., Schulz, C., Kiesewetter, J., Zottmann, J., Sailer, M., Bauer, E., Fischer, F., Fischer, M. R., & Gurevych, I. (2019). Famulus: Interactive annotation and feedback generation for teaching diagnostic reasoning. In S. Padó (Ed.), *The 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing—Proceedings of system demonstrations: Emnlp-IJCNLP 2019* (pp. 73–78). Association for Computational Linguistics (ACL). https://aclanthology.org/D19-3013.pdf

Pfeiffer, J., Vulić, I., Gurevych, I., & Ruder, S. (2020). MAD-X: An adapter-based framework for multi-task cross-lingual transfer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing* (pp. 7654–7673). Online. Association for Computational Linguistics. https://doi.org/10.48550/arXiv.2005.00052

Plass, J. L., & Pawar, S. (2020). Toward a taxonomy of adaptivity for learning. *Journal of Research on Technology in Education*, *52*(3), 275–300. https://doi.org/10.1080/15391523.2020.1719943

Quintana, C., Reiser, B. J., Davis, E. A., Krajcik, J., Fretz, E., Duncan, R. G., Kyza, E., Edelson, D., & Soloway, E. (2004). A scaffolding design framework for software to support science inquiry. *Journal of the Learning Sciences*, *13*(3), 337–386. https://doi.org/10.1207/s15327809jls1303_4

Raković, M., Bernacki, M. L., Greene, J. A., Plumley, R. D., Hogan, K. A., Gates, K. M., & Panter, A. T. (2022). Examining the critical role of evaluation and adaptation in self-regulated learning. *Contemporary Educational Psychology*, *68*, 102027. https://doi.org/10.1016/j.cedpsych.2021.102027

Rapanta, C., & Walton, D. (2016). The use of argument maps as an assessment tool in higher education. *International Journal of Educational Research*, *79*, 211–221. https://doi.org/10.1016/j.ijer.2016.03.002

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing* (pp. 3982–3992). Hong Kong, China. Association for Computational Linguistics. https://doi.org/10.48550/arXiv.1908.10084

Rogers, A., Baldwin, T., & Leins, K. (2021). 'Just what do you think you're doing, Dave?' A checklist for responsible data use in NLP. In *Findings of the Association for Computational Linguistics: EMNLP 2021* (pp. 4821–4833). Punta Cana, Dominican Republic. Association for Computational Linguistics. https://doi.org/10.48550/arXiv.2109.06598

Rotsaert, T., Panadero, E., Schellens, T., & Raes, A. (2018). "Now you know what you're doing right and wrong!" Peer feedback quality in synchronous peer assessment in secondary education. *European Journal of Psychology of Education*, *33*(2), 255–275. https://doi.org/10.1007/s10212-017-0329-x

Rouet, J.-F., & Britt, M. A. (2011). Relevance processes in multiple document comprehension. In M. T. McCrudden, J. P. Magliano, & G. Schraw (Eds.), *Text relevance and learning from text* (pp. 19–52). IAP Information Age Publishing.

Sailer, M., Bauer, E., Hofmann, R., Kiesewetter, J., Glas, J., Gurevych, I., & Fischer, F. (2023). Adaptive feedback from artificial neural networks facilitates pre-service teachers' diagnostic reasoning in simulation-based learning. *Learning and Instruction*, *83*, 101620. https://doi.org/10.1016/j.learninstruc.2022.101620

Sailer, M., Schultz-Pernice, F., & Fischer, F. (2021). Contextual facilitators for learning activities involving technology in higher education: The C♭-model. *Computers in Human Behavior*, *121*, 106794. https://doi.org/10.1016/j.chb.2021.106794

Schick, T., Dwivedi-Yu, J., Jiang, Z., Petroni, F., Lewis, P., Izacard, G., You, Q., Nalmpantis, C., Grave, E., & Riedel, S. (2022). PEER: A collaborative language model. *arXiv Preprints*. https://doi.org/10.48550/arXiv.2208.11663

Schulz, C., Meyer, C. M., & Gurevych, I. (2019). Challenges in the automatic analysis of students' Ddiagnostic reasoning. In A. Korhonen, D. Traum, & L. Màrquez (Eds.), *The 57th Annual Meeting of the Association for Computational Linguistics—Proceedings of the Conference* (pp. 6974–6981). Florence, Italy. Association for Computational Linguistics. https://doi.org/10.1609/aaai.v33i01.33016974

Strijbos, J.-W., Pat-El, R., & Narciss, S. (2021). Structural validity and invariance of the feedback perceptions questionnaire. *Studies in Educational Evaluation*, *68*, 100980. https://doi.org/10.1016/j.stueduc.2021.100980

Tabak, I., & Kyza, E. A. (2018). Research on scaffolding in the learning sciences: A methodological perspective. In F. Fischer, C. E. Hmelo-Silver, S. R. Goldman, & P. Reimann (Eds.), *International handbook of the learning sciences* (pp. 191–200). Routledge. https://doi.org/10.4324/9781315617572

Tenney, I., Wexler, J., Bastings, J., Bolukbasi, T., Coenen, A., Gehrmann, S., Jiang, E., Pushkarna, M., Radebaugh, C., Reif, E., & Yuan, A. (2020). The language interpretability tool: Extensible, interactive visualizations and analysis for NLP models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* (pp. 107–118). Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-demos.15

Tetzlaff, L., Schmiedek, F., & Brod, G. (2021). Developing personalized education: A dynamic framework. *Educational Psychology Review*, *33*(3), 863–882. https://doi.org/10.1007/s10648-020-09570-w

Tsovaltzi, D., McLaren, B. M., Melis, E., & Meyer, A. K. (2012). Erroneous examples: Effects on learning fractions in a web-based setting. *International Journal of Technology Enhanced Learning*, *4*(3–4), 191–230. https://doi.org/10.1504/IJTEL.2012.051583

Utama, P. A., Moosavi, N. S., & Gurevych, I. (2020). Towards debiasing NLU models from unknown biases. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing* (pp. 7597–7610). Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.613

Van Merriënboer, J. J., & Kirschner, P. A. (2018). 4C/ID in the context of instructional design and the learning sciences. In F. Fischer, C. E. Hmelo-Silver, S. R. Goldman, & P. Reimann (Eds.), *International handbook of the learning sciences* (pp. 169–179). Routledge. https://doi.org/10.4324/9781315617572

Voet, M., Gielen, M., Boelens, R., & De Wever, B. (2018). Using feedback requests to actively involve assessees in peer assessment: Effects on the assessor's feedback content and assessee's agreement with feedback. *European Journal of Psychology of Education*, *33*(1), 145–164. https://doi.org/10.1007/s10212-017-0345-x

Vogel, F., Kollar, I., Fischer, F., Reiss, K., & Ufer, S. (2022). Adaptable scaffolding of mathematical argumentation skills: The role of self-regulation when scaffolded with CSCL scripts and heuristic worked examples. *International Journal of Computer-Supported Collaborative Learning*, *17*, 39–64. https://doi.org/10.1007/s11412-022-09363-z

Wambsganss, T., Niklaus, C., Cetto, M., Söllner, M., Handschuh, S., & Leimeister, J. M. (2020). AL: An adaptive learning support system for argumentation skills. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (pp. 1–14). https://doi.org/10.1145/3313831.3376732

Wang, X., Kollar, I., & Stegmann, K. (2017). Adaptable scripting to foster regulation processes and skills in computer-supported collaborative learning. *International Journal of Computer-Supported Collaborative Learning*, *12*, 153–172. https://doi.org/10.1007/s11412-017-9254-x

Wang, Y., Mishra, S., Alipoormolabashi, P., Kordi, Y., Mirzaei, A., Arunkumar, A., Ashok, A., Dhanasekaran, A. S., Naik, A., Stap, D., Pathak, E., Karamanolakis, G., Lai, H. G., Purohit, I., Mondal, I., Anderson, J., Kuznia, K., Doshi, K., Patel, M., … Khashabi, D. (2022). Super-natural instructions: Generalization via declarative instructions on 1600+ nlp tasks. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing* (pp. 5085–5109). https://aclanthology.org/2022.emnlp-main.340/

Wichmann, A., Funk, A., & Rummel, N. (2018). Leveraging the potential of peer feedback in an academic writing activity through sense-making support. *European Journal of Psychology of Education*, *33*(1), 165–184. https://doi.org/10.1007/s10212-017-0348-7

Wu, Y., & Schunn, C. D. (2020). When peers agree, do students listen? The central role of feedback quality and feedback frequency in determining uptake of feedback. *Contemporary Educational Psychology*, *62*, 101897. https://doi.org/10.1016/j.cedpsych.2020.101897

Yang, D., Halfaker, A., Kraut, R., & Hovy, E. (2017). Identifying semantic edit intentions from revisions in Wikipedia. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing* (pp. 2000–2010). Copenhagen, Denmark. Association for Computational Linguistics. https://doi.org/10.18653/v1/D17-1213

Zhang, F., Hashemi, H. B., Hwa, R., & Litman, D. (2017). A corpus of annotated revisions for studying argumentative writing. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics* (pp. 1568–1578). Vancouver, Canada. Association for Computational Linguistics. https://doi.org/10.18653/v1/P17-1144

Zhang, H., & Litman, D. (2021). Essay quality signals as weak supervision for source-based essay scoring. In *Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 85–96). Online. Association for Computational Linguistics. https://aclanthology.org/2021.bea-1.9

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2020). BERTScore: Evaluating text generation with BERT. *International Conference on Learning Representations*. *arXiv preprint*. https://doi.org/10.48550/arXiv.1904.09675

How to cite this article: Bauer, E., Greisel, M., Kuznetsov, I., Berndt, M., Kollar, I., Dresel, M., Fischer, M. R., & Fischer, F. (2023). Using natural language processing to support peer-feedback in the age of artificial intelligence: A cross-disciplinary framework and a research agenda. *British Journal of Educational Technology*, *54*, 1222–1245. https://doi.org/10.1111/bjet.13336