Learning and Instruction 83 (2023) 101620

Contents lists available at ScienceDirect

Learning and Instruction

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

Adaptive feedback from artificial neural networks facilitates pre-service teachers' diagnostic reasoning in simulation-based learning

Michael Sailer [a,*], Elisabeth Bauer [a], Riikka Hofmann [b], Jan Kiesewetter [c], Julia Glas [a], Iryna Gurevych [d], Frank Fischer [a]

[a] Department of Psychology, Ludwig-Maximilians-Universität München, Munich, Germany
[b] Faculty of Education, University of Cambridge, Cambridge, UK
[c] Institute for Medical Education, University Hospital, Ludwig-Maximilians-Universität München, Munich, Germany
[d] Ubiquitous Knowledge Processing Lab, Department of Computer Science, Technical University of Darmstadt, Darmstadt, Germany

ARTICLE INFO

Keywords:
Simulation-based learning
Teacher education
Artificial intelligence
Adaptive feedback
Natural language processing

ABSTRACT

In simulations, pre-service teachers need sophisticated feedback to develop complex skills such as diagnostic reasoning. In an experimental study with $N = 178$ pre-service teachers about simulated pupils with learning difficulties, we investigated the effects of automatic adaptive feedback, which is based on artificial neural networks, on pre-service teachers' diagnostic reasoning. Diagnostic reasoning was operationalised as diagnostic accuracy and the quality of justifications. We compared automatic adaptive feedback with static feedback, which we provided in form of an expert solution. Further, we experimentally manipulated whether the learners worked individually or in dyads on the computer lab-based simulations. Results show that adaptive feedback facilitates pre-service teachers' quality of justifications in written assignments, but not their diagnostic accuracy. Further, static feedback even had detrimental effects on the learning process in dyads. Automatic adaptive feedback in simulations offers scalable, elaborate, process-oriented feedback in real-time to high numbers of students in higher education.

1. Introduction

Teachers' diagnostic reasoning skills are essential for dealing with increasing diversity and heterogeneity in classrooms: pupils have diverse and changing learning prerequisites that teachers must consider in order to offer individual support (Reinke, Stormont, Herman, Puri, & Goel, 2011). However, there are indications that diagnostic reasoning is often neglected in teacher education and that teachers themselves consider their diagnostic skills insufficient (Poznanski, Hart, & Graziano, 2021). In teacher education as in many other higher education (HE) programmes, it is often not possible to offer extensive real-life practice of specific instances of diagnostic reasoning (Grossman et al., 2009; Heitzmann et al., 2019).

One promising option to overcome this gap between education and practice is to provide pre-service teachers with simulation-based learning opportunities, which are less overwhelming than real-life situations by isolating skills early on in professional learning (Chernikova et al., 2020). However, simulations might not be helpful per se, but need to be accompanied by further instructional guidance like targeted feedback to become effective. Specifically, due to the complexity involved in simulation-based learning of diagnostic reasoning, learners may need specific support and feedback to make full use of their learning opportunities (Kiesewetter et al., 2020). Feedback that is adapted to learners' needs is resource intensive for HE teachers (see Henderson, Ryan, & Phillips, 2019); partially automating the feedback seems promising but also challenging.

Involving collaborative learning scenarios is another pedagogical approach in the context of simulation-based learning. Existing studies found that, compared to individuals, collaborative learners often perform better in solving reasoning problems in simulated scenarios (Csanadi, Kollar, & Fischer, 2021). Moreover, learners seem to be better in critical evaluation of other's arguments than their own arguments (Mercier & Sperber, 2017), suggesting collaborative scenarios may be beneficial for learning complex diagnostic tasks. We conducted a study in which we investigated the effects of automated adaptive feedback on pre-service teachers' simulation-based learning of diagnostic reasoning in individual and collaborative learning settings.

* Corresponding author. Department of Psychology, Ludwig-Maximilians-Universität München, Leopoldstraße 13, 80802, Munich, Germany.
E-mail address: Michael.Sailer@psy.lmu.de (M. Sailer).

https://doi.org/10.1016/j.learninstruc.2022.101620
Received 30 June 2021; Received in revised form 10 January 2022; Accepted 23 March 2022
Available online 11 April 2022

1.1. Teachers' diagnostic reasoning

Facilitating pre-service teachers' learning of diagnostic reasoning seems important to prepare them for diagnostic reasoning in real classroom settings and in school. We consider teachers' diagnostic reasoning broadly as the goal-oriented collection and integration of information, aiming to reduce uncertainty in order to make educational decisions (Heitzmann et al., 2019). Studies have shown that teachers' diagnosis of their pupils' various learning prerequisites is important in order to provide individual pupils with suitable support (Reinke et al., 2011). Part of pupils' learning prerequisites may also be learning difficulties such as dyslexia or behavioural disorders like attention deficit hyperactivity disorder (ADHD). These require early identification to intervene in a timely manner and provide affected pupils with the necessary support. To avoid disadvantaging single pupils due to insufficient or unsuited support, generating greater problems in the future, it is vital that teachers identify cues to learning difficulties early in a pupil's school career. We conceptualise the recognition of these cues as an important part of teachers' effective diagnostic reasoning (Poznanski et al., 2021).

Diagnostic reasoning skills can develop by engaging in practice, i.e., by repeated knowledge application and exposure to various diagnostic problems. Thereby, knowledge becomes increasingly encapsulated into higher-level concepts (Schmidt & Rikers, 2007). With further experience and thus practice, knowledge is integrated into episodic representations of diagnostic problems, which is referred to as script formation (Charlin et al., 2012; Lachner, Jarodzka, & Nückles, 2016). Diagnostic reasoning can be assessed regarding the achievement of target criteria, such as diagnostic accuracy, which indicates the degree of correctness of a teacher's diagnostic judgement (Kolovou, Naumann, Hochweber, & Praetorius, 2021). Beyond achieving diagnostic accuracy, justifying diagnoses and explaining the underlying diagnostic reasoning are helpful and crucial for collaborating with other teachers or school psychologists (Csanadi et al., 2021). Justifying diagnoses by providing supporting evidence facilitates collaborators' understanding of the diagnostic reasoning (see Hitchcock, 2005). Justifications can also facilitate a process of considering and reconciling explanations within collaborative diagnostic reasoning and thus help improve the diagnosing (see Berland & Reiser, 2009). Therefore, we conceptualise the pre-service teachers' learning of diagnostic reasoning with both outcomes of teachers' reasoning, diagnostic accuracy and the quality of justifications.

1.2. Simulation-based learning to foster diagnostic reasoning

To foster diagnostic reasoning, there is evidence that pre-service teachers' practicing diagnostic reasoning in authentic contexts or with authentic cases during HE can be effective (Van Merriënboer, 2013; VanLehn, 1996), which is one reason why simulation-based learning has been identified as an innovative way forward in teacher education. Simulations are partial representations of professional situations, with a set of features that can be manipulated by learners (see Codreanu, Sommerhoff, Huber, Ufer, & Seidel, 2020). This can involve authentic cases of simulated pupils that teachers will deal with as part of their work. Authentic cases of simulated pupils provide learning opportunities to practice diagnostic reasoning, which is needed as an in-service teacher. Simulations are particularly beneficial in terms of practicing critical but infrequent situations and focusing on specific subsets of practices in which learners can repeatedly engage (Grossman et al., 2009). Therefore, simulation-based learning is considered a highly promising instructional approach for learning diagnostic reasoning in teacher education (Codreanu et al., 2020).

However, diagnosing simulated pupils even in simulation-based learning is a complex task for pre-service teachers and might not be effective per se (see Kiesewetter et al., 2020). Research emphasised that especially novice learners, who lack a certain level of prior knowledge and skills, need particular support and feedback to effectively learn complex skills (Cook et al., 2013; Wisniewski, Zierer, & Hattie, 2019). Therefore, pre-service teachers may need specific support and feedback to effectively learn diagnostic reasoning in simulation-based learning.

1.2.1. Adaptive feedback in simulation-based learning

Receiving feedback is considered a necessary condition for harnessing the potentials of simulation-based learning of complex skills, such as diagnostic reasoning (Cook et al., 2013; Scheuer, McLaren, Loll, & Pinkwart, 2012). In order to be effective in supporting the learning of complex skills, feedback needs to elaborate on ways to appropriately process the task, not only provide information about correct task solutions (Narciss et al., 2014; Wisniewski et al., 2019). Elaborating on appropriate or optimal processing of the task is often done by presenting expert solutions, which exemplify the processing of the task following the learner's own efforts to solve the problem (Renkl, 2014). Presenting an expert solution as a form of static feedback (i.e., non-adaptive feedback) is resource-efficient in HE, because all learners receive the same generic feedback; besides, it can easily be provided automatically in digital learning environments. However, learners need to determine their current state of knowledge and performance and figure out options for improvement by themselves, by comparing their own processing and solution with the expert solution. This process can be demanding and difficult for learners, involving being confronted with a large amount of information, possibly exceeding learners' cognitive capacity (Sweller, van Merriënboer, & Paas, 2019). In contrast to such static feedback, adaptive feedback can accommodate learners' specific needs by making appropriate adjustments to the feedback based on learners' performance (see Plass & Pawar, 2020). Such adaptive feedback can highlight and thus facilitate learners' understanding of their current state of knowledge and options for improvement, for example by identifying gaps between a learner's current and desired knowledge state or providing additional explanations if the task processing was flawed (Bimba, Idris, Al-Hunaiyyan, Mahmud, & Shuib, 2017; Narciss et al., 2014; Plass & Pawar, 2020). Adaptive feedback might thus increase germane cognitive load, that is, the cognitive resources invested in actual learning processes (Sweller et al., 2019). Freeing up cognitive resources for learning processes to actually happen might be particularly helpful in learning complex skills like diagnostic reasoning. Therefore, pre-service teachers' simulation-based learning of diagnostic reasoning might be effectively supported by process-oriented, adaptive feedback on their diagnostic reasoning (Wisniewski et al., 2019).

1.2.2. Automation of adaptive feedback in simulation-based learning

Adaptive feedback, however, is resource-intensive for HE teachers if done manually for every learner's task solution. Automating adaptive feedback on the learners' task processing to make process-oriented, adaptive feedback accessible to numerous learners is a potential solution. Research explored various possible applications of automatically assessing closed format questions or log data in cognitive tutors and intelligent tutoring systems (Graesser, Hu, & Sottilare, 2018). However, complex reasoning tasks in simulations require justifications consolidated in written explanations. For open ended explanations, recent advancements in artificial intelligence and machine learning offer new technical capabilities with help of artificial neural networks. In particular, methods of Natural Language Processing (NLP) aim to parse, analyse, and understand human language (Manning & Schütze, 2005) and thus, enable automating a real-time measurement of certain aspects of learners' written solutions without a human corrector (see Plass & Pawar, 2020). In the context of diagnostic reasoning, artificial neural network-based NLP models for sequence tagging can be specialised for the particular context of diagnostic reasoning: they can be trained to automatically detect diagnostic entities (e.g., cues or diagnoses) and epistemic activities (e.g., hypothesis generation or evidence evaluation; see Schulz, Meyer, and Gurevych (2019) in learners' written explanations. Based on predictions provided by the models, pre-service teachers can be automatically offered predefined feedback elements that are

adapted to the corresponding detected diagnostic entities and epistemic activities in written explanations of their diagnostic reasoning (Pfeiffer et al., 2019). However, the use of NLP involves challenges: the predominant learning paradigm in NLP utilises transfer learning strategies where models or word representations are pre-trained on freely available text corpora (Howard & Ruder, 2018; Mikolov, Sutskever, Chen, Corrado, & Dean, 2013). These models are subsequently fine-tuned on the target task, which, depending on the similarity of the source and target domain, can result in a considerable decrease in performance. This is particularly challenging when the target task involves domain-specific terms, which might have been seldomly used in text corpora or even in training data that are used for fine-tuning the target task. Further, as the pre-training approaches rely on unsupervised methods, i.e., trained on unlabelled data, biases (e.g. gender) can be encoded in the pre-trained representations (Bolukbasi, Chang, Zou, Saligrama, & Kalai, 2016), which have to be monitored and considered. Despite these challenges, there is initial evidence that using NLP to automate adaptive feedback on learners' written solutions concerning an online task about climate change leads learners to revise their solutions, which improved the quality of their justifications (Zhu et al., 2017; Zhu, Liu, & Lee, 2020).

In summary, we assume that compared to providing an expert solution as a form of non-adaptive, static feedback, NLP-based automatic adaptive feedback may be employed to support pre-service teachers in simulation-based learning of diagnostic reasoning in terms of the quality of their justifications. To what extent NLP-based adaptive feedback can also advance diagnostic accuracy has hardly been investigated.

1.2.3. Individual and collaborative learning in simulation-based learning

The need for adaptive feedback on pre-service teachers' learning of diagnostic reasoning in simulation-based learning may differ depending on the social form of learning. Learners seem better in the critical evaluation of other's arguments than their own arguments (Mercier & Sperber, 2017). Throughout collaboration, learners can use their partners as resources in negotiating meaning and as an additional source of feedback (Weinberger, Stegmann, & Fischer, 2010), by adaptively correcting each other and filling the collaboration partner's knowledge gaps. Therefore, collaborative learners' need for adaptive feedback in diagnostic reasoning may be lower compared to individual learners. However, there is also evidence that collaborative learners show higher learning gains with adaptive feedback (Chuang & O'Neil, 2012; Hsieh & O'Neil, 2002). As collaborative learners might be affected by transaction costs, because they need cognitive capacity for interacting, expressing thoughts, and monitoring another's understanding, they might be particularly in danger of cognitive overload in complex tasks like diagnostic reasoning (Dillenbourg, 2002; Janssen & Kirschner, 2020). Thus, the effects of adaptive feedback on germane cognitive load might be even more pronounced in collaborative learning contexts where cognitive resources are taxed through additional *collaboration load* (Kirschner, Paas, & Kirschner, 2009).

Regarding diagnostic reasoning, research comparing collaborative processing of diagnostic reasoning tasks with individual processing indicates differences in the processing approaches. These different approaches may affect diagnostic accuracy and the quality of justifications: on the one hand, collaborative learners generate more hypotheses and evaluate more evidence before suggesting a solution (Csanadi et al., 2021) and they apply a more reflective approach by processing the task information under different perspectives (Okada & Simon, 1997). Individuals, on the other hand, seem more determined in proposing solutions (Csanadi et al., 2021).

Based on these contrasting findings regarding the benefits and costs of collaboration – particularly in the field of simulation-based learning (Cook et al., 2012, 2013) – it is difficult to derive directed hypotheses. However, it is plausible to assume that the effects of different kinds of feedback *differ* depending on whether learners learn alone or together. Thus, we hypothesise an interaction of feedback type with the social form of learning (i.e., an undirected hypothesis).

1.3. The present study

In this study, we employ an automatic adaptive feedback algorithm in a simulation setting that is based on NLP methods. The algorithm was implemented to provide feedback on pre-service teachers' written explanations of their diagnostic reasoning about simulated pupils with learning difficulties. In this context, diagnostic reasoning is expressed in diagnostic accuracy (i.e., whether or not diagnoses are correct), and in the quality of diagnostic justifications (i.e., the extent to which relevant supporting pieces of evidence for the diagnoses are presented). We investigate effects of automatic adaptive NLP-based feedback compared to static feedback (i.e., expert solutions). Automatic adaptive feedback provides process-related feedback and can foster learners' understanding of their current state of knowledge and suggest where and how to improve current performance (Narciss et al., 2014; Plass & Pawar, 2020; Wisniewski et al., 2019). We hypothesise that automatic adaptive feedback is more effective than static feedback in fostering learners' diagnostic reasoning in the learning process (see Zhu et al., 2017; Zhu et al., 2020; Hypothesis 1a for diagnostic accuracy; 1b for the quality of justification). We further investigate whether potential effects of automatic adaptive feedback might interact with the social form of learning, that is, whether pre-service teachers learn individually or collaboratively. On the one hand, adaptive feedback might have higher impact for individual learners than for collaborators since collaborative learners can provide each other with adaptive feedback during the task processing (Weinberger et al., 2010). On the other hand, collaborators' needs for adaptive feedback might be higher compared to individual learners because of transaction costs (Janssen & Kirschner, 2020). We hypothesise an interaction of the social form of learning and the type of feedback on diagnostic reasoning in the learning process (Hypothesis 2a for diagnostic accuracy; 2b for the quality of justification). Further, we hypothesise a positive effect of automatic adaptive feedback on learners' diagnostic reasoning skills in a post-test (Hypothesis 3a for diagnostic accuracy; 3b for the quality of justification). We also hypothesise an interaction effect of the social form of learning and the type of feedback on diagnostic reasoning skills in a post-test (Hypothesis 4a for diagnostic accuracy; 4b for the quality of justification).

2. Method

2.1. Sample and design

A total of $N = 178$ pre-service teachers for primary school and higher track secondary school from a German university participated in the study. We recruited the pre-service teachers by advertising in lectures, in online courses, and on campus. The study utilised a 2X2 between-subjects experimental design. Learners were randomly assigned to one of four conditions: they received either static or adaptive feedback and worked either collaboratively or individually. Sixty pre-service teachers learned individually, half of them received static feedback while the other half received adaptive feedback. The remaining 118 pre-service teachers were randomly grouped into dyads. One dyad was excluded, because one of the partners turned out not to be a pre-service teacher. Half of the dyads received static feedback; the other half received adaptive feedback. The adjusted sample size is $N = 118$ units (60 individual learners and 58 dyads).

The 137 (77%) women and 39 (22%) men were on average 23.34 years old ($SD = 3.58$, min = 18, max = 35) and their study semester varied from semester 1 to 16 ($M = 5.84$, $SD = 3.73$). The distribution regarding age and gender in our study is comparable with the population of pre-service teachers in Germany.

2.2. Learning environment and the learners' task

The study was conducted using the computer-based learning platform CASUS (https://www.instruct.eu/casus/), on which pre-service teachers learned with simulated pupil cases. The simulated pupils are constructed building child profiles with various learning difficulties. In the learning phase, the learners worked on six simulated pupil cases, which are available in an open science repository https://osf.io/knfmu/. Three of the cases were concerned with children with specific learning difficulties with impairment in reading and/or writing (dyslexia or isolated reading or spelling disorder). The other three cases dealt with diseases from the spectrum of Attention-Deficit and/or Hyperactivity Disorder (ADD or ADHD). We used document-based simulations (Heitzmann et al., 2019): the learners had access to different types of materials that described the behaviour of the simulated pupil. The material included a transcript of a conversation between the teacher and the parents of the child, the pupil's school assignments and certificates, and a description of the pupil's learning and social behaviour. Learners decided how many information sources they examined and in which order. Following each case, pre-service teachers wrote an explanation of their diagnostic reasoning. After that, the learners received either automated or static feedback on their written explanation. A detailed explanation of the learning environment can be found in Bauer et al. (2022).

2.3. Manipulation of independent variables

Depending on the experimental condition, the learning environment provided either static or adaptive feedback and the pre-service teachers learned either individually or collaboratively.

2.3.1. Static and adaptive feedback

Static feedback: After learners had entered and justified their diagnosis, they received an expert solution of the case and were asked to compare it with their own solution. Two independent domain experts validated the expert solutions prior to their use in the study. An example of static feedback is shown in Fig. 1.

Automatic adaptive feedback: Learners' diagnostic explanation was analysed in real-time using NLP: we applied a sequence labelling approach to identify diagnostic classes, consisting of diagnostic entities (e.g., reading problems, hyperactivity) and diagnostic activities (hypothesis generation, evidence generation, evidence evaluation, and drawing conclusions), in pre-service teachers' written explanations. To automatically and adaptively provide feedback on learners' current diagnostic reasoning, a system consisting of three components (NeuralWeb, INCEpTION, and CASUS) was implemented (for in-depth explanation see Pfeiffer et al., 2019). First, in a "cold-start" phase, domain experts coded explanations written by learners of a prior study with $N = 118$ pre-service teachers, who worked on the same simulations in the learning environment CASUS (see Bauer et al., 2020). The experts used the annotation platform INCEpTION (https://inception-project.github.io/) and coded the data according to diagnostic entities and epistemic activities (for details see Schulz, Meyer, & Gurevych, 2019). Second, the coded data was used to initially train a predictive model in NeuralWeb, a Python-based web service. Third, the written explanations of new learners, who participated in the present study, were processed through the NeuralWeb model to output a label-set of discrete diagnostic classes (diagnostic entities and epistemic activities). In a nutshell, we utilised state-of-the-art artificial neural network-based models for sequence tagging (see Akbik et al., 2019), specialised for the setting of diagnostic reasoning (see Pfeiffer et al., 2019), which have been shown to outperform standard baselines for these types of tasks (for a comparison of alternative models see Schulz, Meyer, & Gurevych, 2019).

Depending on the automatically identified classes, specific paragraphs of predefined feedback text were adaptively activated. The feedback paragraphs were created by domain experts and validated by independent domain experts prior to the study. Experts created approximately 40 feedback paragraphs for every case. For example, these feedback paragraphs informed the learner that a specific symptom was correctly identified in the simulated pupil case. When the corresponding element of a pupil's profile was not detected in a learner's written explanation, the feedback informed the learner that they missed mentioning that symptom. The identified classes and the automatic adaptive feedback, consisting of a range of different feedback paragraphs, were sent back to the learning environment CASUS. CASUS then presented this adaptive feedback to the user.

The automatic adaptive feedback targets two levels of the learner's written explanation of their diagnostic reasoning: diagnostic activities

Fig. 1. Static feedback in CASUS.

(whether appropriate reasoning activities were applied or missing) and diagnostic entities (whether the chosen diagnosis and its justification are correct, incorrect, or missing in terms of the domain-specific and case-specific content). By clicking on feedback paragraphs, the relating text part, which was identified in their answer, was highlighted (see Fig. 2).

After the feedback, learners were presented with the next case. There was no opportunity for a revision of the initial written explanation after the feedback.

2.3.2. Individual and collaborative learning

In the individual condition, every learner worked on their own in a computer lab. In the collaborative condition, two learners worked together as a dyad. The two partners each worked on their own computer in separate computer labs and communicated via headsets. We lightly structured the collaboration of the dyads by a scene level script (see Vogel, Wecker, Kollar, & Fischer, 2017). We assigned two roles to the learners, namely main user and secondary user. The main user actively operated in the CASUS learning environment and shared their screen with the collaborating partner. The secondary user was able to advise and discuss with the main user (for details see Kiesewetter et al., 2022). After three of the six learning cases, the roles changed. In the beginning of the study, collaborative learners were explicitly told to exchange ideas with their learning partner and to write their explanation together.

2.4. Procedure

On average the learners needed $M = 170.17$ min to complete the study ($SD = 31.59$ min). At first, we introduced the participants to the learning platform CASUS by a short video. Second, the pre-service teachers completed a questionnaire containing demographics and a conceptual knowledge test. Third, they watched an 18-min video with theoretical input about specific learning difficulties included in the simulation to compensate for possible differences in prior knowledge. Fourth, the participants worked on the six learning cases (individually or collaboratively and receiving static or adaptive feedback). After three of the six learning cases, participants had a 10-min break. After that, participants completed the remaining three learning cases. Lastly, the participants completed a post-test with two unsupported cases; these were cases similar to the learning cases in the learning phase and also concerned with pupils with learning difficulties (dyslexia and ADD), however, without any feedback, and without a learning partner. For their participation, the participants received 50 Euros as compensation.

2.5. Measures

2.5.1. Prior conceptual knowledge

Conceptual knowledge, which refers to knowledge about concepts and their interrelations in a certain domain, is considered a necessary prerequisite for successful diagnostic reasoning (Heitzmann et al., 2019). We operationalised prior conceptual knowledge about learning disorders with 14 questions about reading and writing difficulties as well as behavioural disorders from the spectrum of ADHD or ADD. The questions were single choice questions with four answer options and one right answer option for each question. Choosing the correct answer was awarded with one point. For dyads, which we used as one unit of analysis for the collaborative learning setting, we calculated the mean score of both collaborators for every item. Further, we calculated the mean score of the 14 items to operationalise it. As the conceptual knowledge represented several potentially independent areas (e.g., ADHD, dyslexia) rather than only one, we assume that the scale reflects a formative instead of a reflective construct and thus we will report the variance inflation factor (VIF). A VIF statistic for formative constructs should be lower than 3.3, meaning that less than 70% of the indicator's variance is explained by the other indicators (Stadler, Sailer, & Fischer, 2021). The items for prior conceptual knowledge showed no variance inflation, indicating an appropriate measurement as a formative construct ($VIF_{min} = 1.20$; $VIF_{max} = 1.78$).

2.5.2. Diagnostic accuracy

We used the written explanations of learners' diagnostic reasoning as the data source to determine leaners' diagnostic accuracy in each case. Based on the expert solutions, we developed a coding scheme to operationalise diagnostic accuracy for each learning case and each post-test

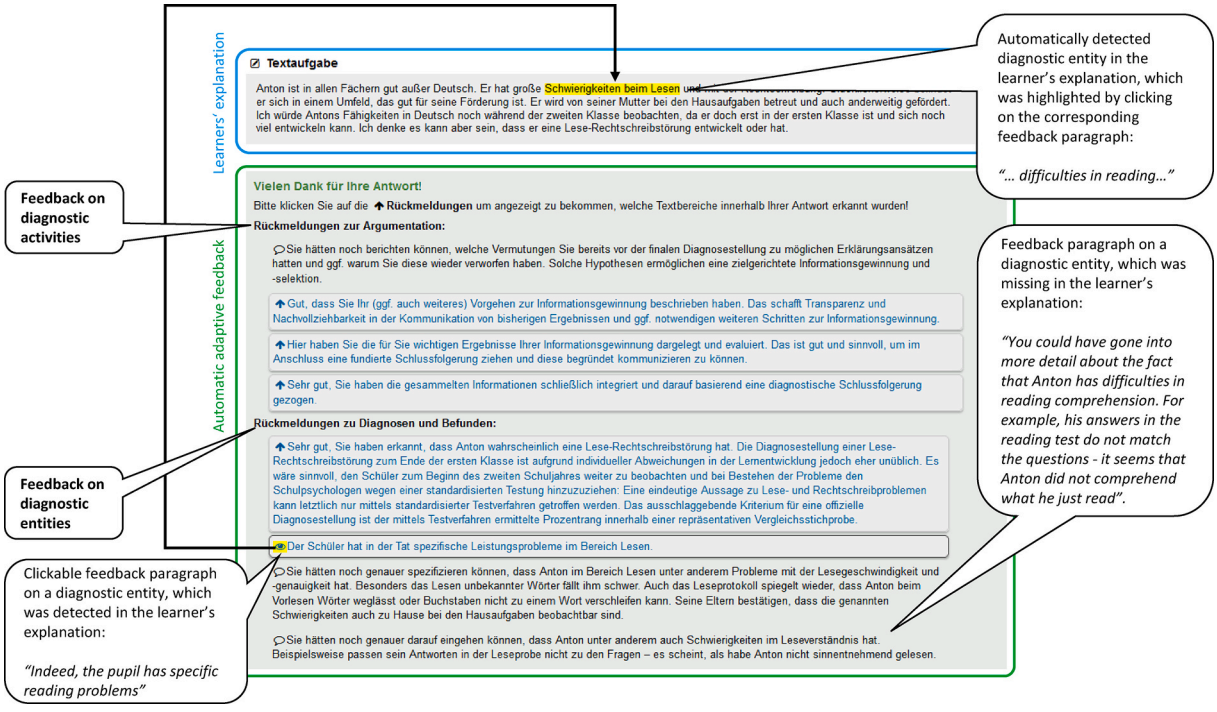

Fig. 2. Automatic adaptive feedback in CASUS.

case. We used one category per case, which represented the presence (coded as 1) or absence (coded as 0) of the correct diagnosis in the learners' explanation. Two trained coders independently coded the written explanations of 11 individual learners and nine dyads (16.9% of the overall data) regarding diagnostic accuracy. The written explanations used in the training were from learners of all four conditions. The inter-rater agreement for diagnostic accuracy, assessed with Cohen's kappa, was high ($\kappa = 0.95$). The remaining material was coded individually.

Diagnostic accuracy in the learning process consisted of five categories, each indicating the presence of the correct diagnoses in the written explanation of learning cases numbered two to six. The first learning case was not included in the learning process measurement as the learners received the first feedback after the completion of the cases. We calculated the mean score of the frequencies of these five categories to operationalise diagnostic accuracy in the learning process. The reliability of this variable was acceptable (McDonald's $\omega = 0.61$).

Post-test diagnostic accuracy consisted of two categories, each indicating the presence of the correct diagnoses in the written explanation of both unsupported post-test cases. All participants solved the two post-test cases individually. To obtain the post-test diagnostic accuracy for the dyads, we calculated the mean value for every category for every dyad. Further, we calculated the mean score of the frequency of the two categories to operationalise post-test diagnostic accuracy. These two categories showed a significant correlation of $r = 0.43$ ($p < .001$), indicating sufficient internal consistency.

2.5.3. Quality of justification

To determine learners' quality of justifications in each case, we used the written explanations of learners' diagnostic reasoning as the data source. We developed a coding scheme, which is based on expert solutions, to operationalise the quality of diagnostic justifications. We used six categories for each case, which indicated the presence (coded as 1) or absence (coded as 0) of the six primary supporting pieces of evidence for the correct diagnosis. Based on experts' solutions, employing all six pieces of evidence in a case is considered a high-quality justification. Again, two trained coders independently coded the written explanations of 11 individual learners and nine dyads (17% of the overall data) regarding the quality of justifications. Cohen's kappa for the quality of justification was high ($\kappa = 0.91$).

We used the described coding of the learners' explanations to operationalise the quality of justification in both the learning process and the post-test quality of justification.

The quality of justification in the learning process was operationalised by 30 categories. We used learners' explanations from five learning cases (learning cases two to six), in which we used six categories representing the presence or absence of the primary supporting pieces of evidence for the correct diagnosis in each learning case. As the learners received the first feedback after completing the first learning cases, we excluded the first case for the measurement of the quality of justification in the learning process. We calculated the mean score of the 30 categories. The reliability of the variable is acceptable (McDonald's $\omega = 0.75$).

Post-test quality of justification consisted of 12 categories, which were assigned in the learners' explanations from two unsupported post-test cases that all students completed individually at the end of the study. In each of these two post-test cases we used six categories, which were coded for the presence or absence of supporting evidence in the corresponding case. To obtain the quality of justification values for the dyads, we calculated the mean value of each of the 12 categories of every dyad. Then, we calculated the mean score for the 12 categories for the quality of justifications in the post-test. The reliability of the variable was rather low (McDonald's $\omega = 0.50$).

2.5.4. Time-on-task

As a control measure, we included time-on-task, which is the time that learners spent with the learning material (e.g., report of grades, description of a pupil's social behaviour), not including the time they spent with writing the explanation and reading the feedback. We measured time-on-task for all six learning cases and computed a sum score.

2.6. Statistical analyses

To investigate Hypotheses 1(a, b) and 2(a, b) we calculated a MANCOVA, using the type of feedback and the social form of learning as independent variables. The diagnostic accuracy and the quality of justifications in the learning process were dependent variables. To control for prior conceptual knowledge, we used this variable as a covariate. To account for individual differences in time-on-task, we included it as a covariate as well. To account for non-normal distribution, time-on-task values were log transformed (see Van der Linden, 2016). Hypotheses 3 (a, b) and 4(a, b) were also tested by a MANCOVA. We included the same independent variables and covariates in this MANCOVA as in the MANCOVA for Hypotheses 1(a, b) and 2(a, b), however this time we used post-test diagnostic accuracy and the post-test quality of justifications as dependent variables. We analysed the data with IBM SPSS Statistics 26 and set the alpha level to $\alpha = 0.05$.

2.7. Ethics clearance

The study was approved by the Medical Faculty's Ethics Committee of Ludwig-Maximilians-Universität München (no.17-249).

3. Results

3.1. Prior conceptual knowledge (Randomisation check)

Descriptive results of the participants' prior conceptual knowledge in the four conditions of the 2X2 design are shown in Table 1. The descriptives indicate comparable levels of prior knowledge across all conditions. An ANOVA with the independent variables type of feedback and social form of learning did not reveal indications for systematic apriori differences (type of feedback, $F(1,114) = 0.09$, $p = .76$, $\eta_p = 0.001$, social form of learning, $F(1,114) = 0.66$, $p = .42$, $\eta_p = 0.006$, and interaction type of feedback X social form of learning, $F(1,114) = 2.00$, $p = .16$, $\eta_p = 0.017$). Thus, the randomisation was successful with respect to prior conceptual knowledge. A correlation matrix of all variables is included in Appendix 1.

3.2. Effects on diagnostic reasoning in the learning process

To analyse effects of automatic adaptive feedback, the social form of learning as well as their interaction on diagnostic reasoning outcomes in the learning process, we conducted a MANCOVA. In this analysis we included log transformed time-on-task, (diagnostic accuracy, $F(1,112) = 11.34$, $p = .001$, $\eta_p = 0.092$, quality of justification, $F(1,112) = 55.34$, $p < .001$, $\eta_p = 0.331$), and prior conceptual knowledge as covariates (diagnostic accuracy, $F(1,112) = 0.22$, $p = .643$, $\eta_p = 0.002$, quality of justification, $F(1,112) = 1.51$, $p = .222$, $\eta_p = 0.013$). The model explains 32% of variance in diagnostic accuracy and 47% of variance in the quality of justifications in the learning process.

Table 1
Means (*M*) and standard deviations (*SD*) for prior conceptual knowledge split by conditions of the 2X2 design.

Social form of learning	Type of feedback	*M*	*SD*
individual	static	.70	.12
	adaptive	.68	.13
collaborative	static	.69	.09
	adaptive	.72	.08

Regarding the *diagnostic accuracy* in the learning process, we found a significant medium-sized interaction effect of the type of feedback and the social form of learning, $F(1,112) = 13.28$, $p < .001$, $\eta_p = 0.106$, supporting Hypothesis 2a. Based on the estimated marginal means in Table 2 and Bonferroni-corrected post hoc comparisons, we conclude that collaborative learners, who have received static feedback scored significantly lower on diagnostic accuracy compared to all other groups. No other post hoc comparisons were significant (see Appendix 2). As the interaction is disordinal the main effects for type of feedback, $F(1,112) = 16.06$, $p < .001$, $\eta_p = 0.125$, and social form of learning, $F(1,112) = 12.41$, $p = .001$, $\eta_p = 0.100$, cannot be meaningfully interpreted. Thus, Hypothesis 1a, in which a main effect of automatic adaptive feedback was hypothesised, was not supported.

Regarding the *quality of justification* in the learning process, we found a large main effect of automatic adaptive feedback, $F(1,112) = 34.07$, $p < .001$, $\eta_p = 0.233$. The main effect of the social form of learning on the quality of justification in the learning process was not significant, $F(1,112) = 0.94$, $p = .336$, $\eta_p = 0.125$, neither was the interaction effect of social form of learning and type of feedback, $F(1,112) = 0.69$, $p = .407$, $\eta_p = 0.006$. These results are in support of Hypothesis 1b as we found a positive effect of adaptive feedback on the quality of justifications. Hypothesis 2b was not supported in our analysis as we found no interaction effect of social form of learning and type of feedback on the quality of justification in the learning process.

3.3. Effects on diagnostic reasoning skills in the post-test

To investigate effects of automatic adaptive feedback, the social form of learning as well as its interaction on diagnostic reasoning skills in the post-test we used a MANCOVA, controlling for log transformed time-on-task, (diagnostic accuracy, $F(1,112) = 13.57$, $p < .001$, $\eta_p = 0.108$, quality of justification, $F(1,112) = 16.42$, $p < .001$, $\eta_p = 0.128$), and prior conceptual knowledge, (diagnostic accuracy, $F(1,112) = 1.95$, $p = .166$, $\eta_p = 0.017$, quality of justification, $F(1,112) = 0.12$, $p = .732$, $\eta_p = 0.001$). The model explains 31% of variance in post-test diagnostic accuracy and 28% of variance in the post-test quality of justifications.

We found a significant medium-sized interaction effect of the type of feedback and the social form of learning on post-test *diagnostic accuracy*, $F(1,112) = 7.96$, $p = .006$, $\eta_p = 0.066$, supporting Hypothesis 4a. As for the learning process, also in the post-test, learners that collaborated in the learning phase and received static feedback reached significantly lower levels of diagnostic accuracy in the post-test than learners in all other conditions (see Table 2). Again, no other post hoc comparisons were significant (see Appendix 2). Because of the disordinal interaction, the main effects for type of feedback, $F(1,112) = 14.60$, $p < .001$, $\eta_p = 0.115$, and social form of learning, $F(1,112) = 17.66$, $p < .001$, $\eta_p = 0.136$, on post-test diagnostic accuracy cannot be meaningfully interpreted. Hypothesis 3a, in which a main effect of automatic adaptive feedback was hypothesised, was not supported.

Results regarding the post-test *quality of justifications* show a large significant main effect of automatic adaptive feedback, $F(1,112) = 21.02$, $p < .001$, $\eta_p = 0.158$. The social form of learning, $F(1,112) = 1.28$, $p = .260$, $\eta_p = 0.011$, and the interaction between type of feedback and social form of learning, $F(1,112) = 0.01$, $p = .923$, $\eta_p < 0.001$, were not significant. These results support Hypothesis 3b based on the main effect of automatic adaptive feedback. As we did not find an interaction for the post-test quality of justifications, Hypothesis 4b was not supported.

Table 2
Estimated marginal means (*M*) and standard errors (*SE*) for all dependent variables split by conditions of the 2X2 design.

Dependent Variable	Social form of learning	Type of feedback	*M*	*SE*
Diagnostic accuracy in the learning process	individual	static	.54	.05
		adaptive	.55	.05
	collaborative	static	.20	.05
		adaptive	.56	.05
Quality of justifications in the learning process	individual	static	.40	.02
		adaptive	.51	.02
	collaborative	static	.36	.02
		adaptive	.50	.02
Post-test diagnostic accuracy	individual	static	.74	.06
		adaptive	.80	.06
	collaborative	static	.32	.06
		adaptive	.71	.06
Post-test quality of justifications	individual	static	.33	.02
		adaptive	.43	.02
	collaborative	static	.35	.02
		adaptive	.45	.02

Note: Estimated means and standard errors with covariates on the following values: log transformed time-on-task = 10.16, prior conceptual knowledge = .70.

4. Discussion

In an experimental study, we investigated the effects of NLP-based adaptive feedback on diagnostic reasoning in individual and collaborative simulation-based learning. Methods of NLP using sophisticated algorithms of artificial neural networks allow for automatic analysis of written texts of learners, which facilitates providing process-oriented automatic feedback in real-time and without the need to involve a human corrector. However, to implement such NLP-based systems, training data is initially required: In this study, data from a prior study of 118 learners was used (see Bauer et al., 2020). These data were fully coded regarding the aspects we provided feedback for. Thus, developing and implementing NLP-based automatic adaptive feedback is a time-consuming and extensive task, which may, however, pay off when it comes to preparing many pre-service teachers for their upcoming challenge of diagnosing learning difficulties in pupils in their daily business as teachers.

4.1. Automatic adaptive feedback fosters learners' quality of justifications

Results showed that adaptive feedback fostered the pre-service teachers' quality of justifications in written assignments for both individual and collaborative learners. Adaptive feedback might have facilitated learners' comparison of their current performance to a desired goal performance in their diagnostic reasoning (e.g., Bimba et al., 2017; Narciss et al., 2014; Plass & Pawar, 2020). Compared with the static feedback, the demand on learners' cognitive resources for processing the feedback might have been reduced by adaptive feedback, which may have helped to improve the quality of justification (see Sweller et al., 2019). Even in simulations, which are designed to be less cognitively taxing than real-life situations, diagnostic reasoning is a complex task requiring a high amount of cognitive resources for information processing. Especially pre-service teachers, who lack prior knowledge and professional experience with respect to learning difficulties (Poznanski et al., 2021), are in danger of cognitive overload. However, for diagnostic reasoning and particularly for elaborated justifications, pre-service teachers need optimal conditions to be able to invest enough cognitive resources into actual learning processes (Sweller et al., 2019).

4.2. Effects of automatic adaptive feedback might Depend on the type of task

The results concerning diagnostic accuracy indicate that adaptive feedback did not outperform static feedback per se; instead, we found an interaction between feedback and the social form of learning on diagnostic accuracy. Collaborative learners receiving static feedback had a significantly lower diagnostic accuracy than learners from the other three experimental conditions. Collaborative learners receiving adaptive feedback and individual learners in both feedback conditions did not differ significantly in their diagnostic accuracy. Individual learners may

not require the same amount of adaptive support to relate their own diagnoses with a correct diagnosis as needed when considering various pieces of evidence to provide high-quality justification. Thus, the information provided in the static feedback may have already been sufficient to foster individual learners' diagnostic accuracy. A reason for this might be found in the complexity of the tasks underlying diagnostic accuracy and the quality of justifications: accuracy requires a diagnostic decision that might be fairly simple to derive in some cases – especially when a limited number of potential diagnoses (here: learning difficulties) is previously introduced to the learners, like in our study. The justification might be more cognitively demanding because of the broad range of evidence, for which learners have to evaluate relationships and interdependencies as relevant or irrelevant in a specific case. In the study, adaptive feedback was more effective for tasks which involved a variety of aspects that have to be considered at a time (such as multiple evidence for a high-quality justification), compared to tasks that require a single decision (such as concluding an accurate diagnosis) – especially when this decision is not too difficult.

4.3. Automatic adaptive feedback is helpful for collaborators' diagnostic decision making

For collaborative learning, the results indicated that compared with static feedback, adaptive feedback particularly facilitated collaborative learners' diagnostic accuracy. In the static condition, learners' cognitive capacities have already been more challenged than in the adaptive condition because learners had to compare their own performance with the expert solution. Additionally, transaction costs like communicating, social regulation, and coordinating might have induced additional collaboration load (see Janssen & Kirschner, 2020; Kirschner et al., 2009). This might especially affect dyads that collaborate for the first time and probably will not do so with the same collaboration partner in the future, such as the pre-service teachers in our laboratory study. In such situations, the chances for making use of the potentials of a collective working memory, which involves also knowledge about the other collaborators, are rather low (Janssen & Kirschner, 2020). Combined with evidence that collaborative learners tend to be less pragmatic in proposing solutions in reasoning tasks compared to individual learners (Csanadi et al., 2021), collaborative learners who received static feedback might have struggled most in proposing a solution in form of a diagnostic decision in their written explanations. Instead, they may have focused more strongly on evaluating evidence without finalising a diagnostic conclusion in explaining their diagnostic reasoning. For evaluating pieces of evidence to construct a high-quality justification, the benefits of collaboration, like the adoption of multiple perspectives (Okada & Simon, 1997), might balance the cognitive capacity disadvantages that occur particularly in the diagnostic accuracy outcomes. Further, compared to the static expert solution, the automatic adaptive feedback may have had a stronger compensating effect, possibly by explicitly scaffolding the learners to determine and explicate a shared diagnosis.

4.4. Limitations and future research

The results may be limited in their generalisability with respect to the diagnostic tasks involved. As there were relatively few possible diagnoses and the associated cues of these diagnoses were fairly similar, the decision task itself might not have been taxing the cognitive resources of the participating pre-service teachers to the extent that more complex decisions would. Further, effects on the two different outcomes, particularly diagnostic accuracy, might also be influenced by the choice of our sample: we included pre-service teachers in our study, who are typically not the ones making formal diagnoses about learning difficulties (i.e. diagnostic accuracy). Instead, they often provide evidence for or against certain differential diagnoses (i.e. quality of justification) to school psychologists or special education teachers. Thus, the pre-service teachers in our study might rather have focussed on the evaluation of different evidences than on a final diagnosis. Another possible limitation to generalisability may be that we used dyads to represent collaborative contexts (Jensen & Wiley, 2006); it will be interesting to investigate the effects of adaptive feedback for bigger groups. With respect to internal validity, there may have been issues because the instruction did not specifically emphasise the necessity to state the diagnosis, but asked the learners to justify their diagnosis, therefore only asking for the diagnosis implicitly. However, independently of inaccuracy or absence of the diagnosis, adaptive feedback – compared with static feedback – helped improve collaborative learners' performance in writing a congruent explanation of their diagnostic reasoning that includes a conclusion regarding the diagnosis. A further limitation relates to the relatively low reliabilities of the measurement of diagnostic accuracy. Potential reasons for that are the different areas of learning difficulties that we included in the study as well as the rather low number of categories.

Future studies might measure outcomes like diagnostic accuracy with a larger amount of codes or items. To do so, a larger number of cases in simulations that require less time for the learners to solve might help to get a more reliable measurement of the diagnostic accuracy outcomes. Future research might further investigate the interaction of the type of feedback and the social form of learning to explore different hypotheses that underlie the interaction effect found in this study. In this regard, the interaction of the type of feedback with other aspects of simulation-based learning may be further investigated, such as the complexity of the cases and tasks that need to be performed while processing the cases, which may also affect the need for specific types of feedback.

To address issues with external validity of the results, field studies with larger samples may be conducted to further specify relevant contextual conditions in which automatic adaptive feedback in simulation-based learning are most effective. Such studies could make use of existing groups, e.g., within university courses, and further investigate our findings of automatic adaptive feedback in collaborative settings as well as the acceptance of automatic adaptive feedback in practice. Acceptance of automatic adaptive feedback might critically depend on the trust of users in the feedback system as well as the transparency of the NLP and feedback system (see Shin, 2021). To ensure end-users' understanding of how our NLP and feedback system works, we implemented a transparent system that highlights detected components (see Fig. 2), making us optimistic about the field use of our simulation with respect to learners' acceptance and trust. In addition, in field studies, learners could be asked to utilise the feedback provided in order to revise their explanations and thus, to take an increasingly active role in their learning (see Zhu et al., 2020).

5. Conclusions

Automatic adaptive feedback in simulation-based learning can be used effectively to foster pre-service teachers' diagnostic reasoning in HE. Using methods of NLP like the algorithms related to artificial neural networks to automate adaptive feedback seems to provide particular benefits in terms of fostering learning more complex reasoning outcomes, such as justifying a diagnosis – even in short term interventions of the length of one single course session. Adaptive feedback is a promising instructional support to help pre-service teachers improve the quality of justifications in written assignments, independent of whether they learn together or alone. In collaborative learning contexts, adaptive feedback rather than static seems to effectively compensate for collaboration costs that lead to performance drops with respect to diagnostic accuracy. However, training and specialisation of artificial neural network-based NLP models is a time-consuming task, as it requires collection of data sets and elaborate manual coding before actual implementation of the automatic adaptive feedback. These efforts might be worthwhile where automatic adaptive feedback is subsequently

implemented in simulations in large programmes, such as teacher education or medical education. In such contexts, automatic adaptive feedback can offer a convenient solution for providing elaborate, process-oriented feedback in real-time to high numbers of students.

Funding

This research was supported by a grant of the German Federal Ministry of Research and Education (Grant No.: 16DHL1040) and by the Elite Network of Bavaria (K-GS-2012-209). We have no conflicts of interest to disclose.

CRediT authorship contribution statement

Michael Sailer: Conceptualization, Formal analysis, Visualization, Investigation, Methodology, Resources, Writing – original draft, Writing – review & editing. **Elisabeth Bauer:** Investigation, Methodology, Resources, Visualization, Software, Writing – original draft, Writing – review & editing. **Riikka Hofmann:** Validation, Writing – review & editing. **Jan Kiesewetter:** Conceptualization, Methodology, Writing – review & editing. **Julia Glas:** Formal analysis, Investigation, Software, Writing – original draft. **Iryna Gurevych:** Conceptualization, Methodology, Funding acquisition, Writing – review & editing. **Frank Fischer:** Conceptualization, Funding acquisition, Project administration, Writing – review & editing.

Acknowledgement

The authors would like to thank Gabrielle Arenge for the proofreading of the manuscript. Further, the authors would like to thank the whole FAMULUS team.

Appendix. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.learninstruc.2022.101620.

References

Akbik, A., Bergmann, T., Blythe, D., Rasul, K., Schweter, S., & Vollgraf, R. (2019). Flair: An easy-to-use framework for state-of-the-art NLP. In *Proceedings of the 2019 conference of the north American chapter of the association for computational linguistics* (pp. 54–59). https://doi.org/10.18653/v1/N19-4010

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*, 2787. https://doi.org/10.3389/fpsyg.2020.562665

Bauer, E., Sailer, M., Kiesewetter, J., Schulz, C., Gurevych, I., Fischer, M. R., & Fischer, F. (2022). Learning to Diagnose Students' Behavioral, Developmental and Learning Disorders in a Simulation-Based Learning Environment for Pre-Service Teachers. In F. Fischer, & A. Opitz (Eds.), *Learning to Diagnose with Simulations* (pp. 97–107). Springer. https://doi.org/10.1007/978-3-030-89147-3_8.

Berland, L. K., & Reiser, B. J. (2009). Making sense of argumentation and explanation. *Science Education, 93*(1), 26–55. https://doi.org/10.1002/sce.20286

Bimba, A. T., Idris, N., Al-Hunaiyyan, A., Mahmud, R. B., & Shuib, N. L. B. M. (2017). Adaptive feedback in computer-based learning environments: A review. *Adaptive Behavior, 25*(5), 217–234. https://doi.org/10.1177/1059712317727590

Bolukbasi, T., Chang, K. W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. *Advances in Neural Information Processing Systems, 29*, 4349–4357.

Charlin, B., Lubarsky, S., Millette, B., Crevier, F., Audétat, M. C., Charbonneau, A., et al. (2012). Clinical reasoning processes: Unravelling complexity through graphical representation. *Medical Education, 46*(5), 454–463. https://doi.org/10.1111/j.1365-2923.2012.04242.x

Chernikova, O., Heitzmann, N., Stadler, M., Holzberger, D., Seidel, T., & Fischer, F. (2020). Simulation-based learning in higher education: A meta-analysis. *Review of Educational Research, 90*(4), 499–541. https://doi.org/10.3102/0034654320933544

Chuang, S. H., & O'Neil, H. F. (2012). Role of task-specific adapted feedback on a computer-based collaborative problem-solving task. In H. F. O'Neil, & R. S. Perez (Eds.), *Web-based learning: Theory, research, and practice* (pp. 239–254). New York: Routledge.

Codreanu, E., Sommerhoff, D., Huber, S., Ufer, S., & Seidel, T. (2020). Between authenticity and cognitive demand: Finding a balance in designing a video-based simulation in the context of mathematics teacher education. *Teaching and Teacher Education, 95*, Article 103146. https://doi.org/10.1016/j.tate.2020.103146

Cook, D. A., Brydges, R., Hamstra, S. J., Zendejas, B., Szostek, J. H., Wang, A. T., … Hatala, R. (2012). Comparative effectiveness of technology-enhanced simulation versus other instructional methods: A systematic review and meta-analysis. *Simulation in Healthcare, 7*(5), 308–320. https://doi.org/10.1097/SIH.0b013e3182614f95

Cook, D. A., Hamstra, S. J., Brydges, R., Zendejas, B., Szostek, J. H., Wang, A. T., … Hatala, R. (2013). Comparative effectiveness of instructional design features in simulation-based education: Systematic review and meta-analysis. *Medical Teacher, 35*(1), 867–898. https://doi.org/10.3109/0142159X.2012.714886

Csanadi, A., Kollar, I., & Fischer, F. (2021). Pre-service teachers' evidence-based reasoning during pedagogical problem-solving: Better together? *European Journal of Psychology of Education, 36*(1), 147–168. https://doi.org/10.1007/s10212-020-00467-4

Dillenbourg, P. (2002). Over-scripting CSCL: The risks of blending collaborative learning with instructional design. In P. A. Kirschner (Ed.), *Three worlds of CSCL. Can we support CSCL?* (pp. 61–91). Heerlen: Open Universiteit Nederland.

Graesser, A. C., Hu, X., & Sottilare, R. (2018). Intelligent tutoring systems. In F. Fischer, C. E. Hmelo-Silver, S. R. Goldman, & P. Reimann (Eds.), *International handbook of the learning sciences* (pp. 246–255). New York, NY: Routledge.

Grossman, P., Compton, C., Igra, D., Ronfeldt, M., Shahan, E., & Williamson, P. (2009). Teaching practice: A cross-professional perspective. *Teachers College Record, 111*(9), 2055–2100.

Heitzmann, N., Seidel, T., Opitz, A., Hetmanek, A., Wecker, C., Fischer, M., … Fischer, F. (2019). Facilitating diagnostic competences in simulations: A conceptual framework and a research agenda for medical and teacher education. *Frontline Learning Research, 7*(4), 1–24. https://doi.org/10.14786/flr.v7i4.384

Henderson, M., Ryan, T., & Phillips, M. (2019). The challenges of feedback in higher education. *Assessment & Evaluation in Higher Education, 44*(8), 1237–1252. https://doi.org/10.1080/02602938.2019.1599815

Hitchcock, D. (2005). Good reasoning on the Toulmin model. *Argumentation, 19*(3), 373–391. https://doi.org/10.1007/s10503-005-4422-y

Howard, J., & Ruder, S. (2018). Universal language model fine-tuning for text classification. In *Proceedings of the 56th annual meeting of the association for computational linguistics* (pp. 328–339).

Hsieh, I.-L. G., & O'Neil, H. F. (2002). Types of feedback in a computer-based collaborative problem-solving group task. *Computers in Human Behavior, 18*(6), 699–715. https://doi.org/10.1016/S0747-5632(02)00025-0

Janssen, J., & Kirschner, P. A. (2020). Applying collaborative cognitive load theory to computer-supported collaborative learning: Towards a research agenda. *Educational Technology Research & Development, 68*(2), 783–805. https://doi.org/10.1007/s11423-019-09729-5

Jensen, M. S., & Wiley, J. (2006). When three heads are better than two. In *Proceedings of the annual meeting of the cognitive science society, 28*. Retrieved from https://escholarship.org/uc/item/9160x87s.

Kiesewetter, J., Hege, I., Sailer, M., Bauer, E., Schulz, C., Platz, M., & Adler, M. (2022). A usability study for implementing remote collaboration in a virtual patient platform. *JMIR Medical Education*. https://doi.org/10.2196/24306

Kiesewetter, J., Sailer, M., Jung, V. M., Schönberger, R., Bauer, E., Zottmann, J. M., … Fischer, M. R. (2020). Learning clinical reasoning: How virtual patient case format and prior knowledge interact. *BMC Medical Education, 20*(1), 73. https://doi.org/10.1186/s12909-020-1987-y

Kirschner, F., Paas, F., & Kirschner, P. A. (2009). A cognitive load approach to collaborative learning: United brains for complex tasks. *Educational Psychology Review, 21*(1), 31–42. https://doi.org/10.1007/s10648-008-9095-2

Kolovou, D., Naumann, A., Hochweber, J., & Praetorius, A.-K. (2021). Content-specificity of teachers' judgment accuracy regarding students' academic achievement. *Teaching and Teacher Education, 100*(4), Article 103298. https://doi.org/10.1016/j.tate.2021.103298

Lachner, A., Jarodzka, H., & Nückles, M. (2016). What makes an expert teacher? Investigating teachers' professional vision and discourse abilities. *Instructional Science, 44*(3), 197–203. https://doi.org/10.1007/s11251-016-9376-y

Manning, C. D., & Schütze, H. (2005). In *Foundations of statistical natural language processing* (8th ed.). Cambridge, MA: MIT Press.

Mercier, H., & Sperber, D. (2017). *The enigma of reason*. Cambridge, Massachusetts: Harvard University Press. https://doi.org/10.4159/9780674977860

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems*, 3111–3119.

Narciss, S., Sosnovsky, S., Schnaubert, L., Andrès, E., Eichelmann, A., Goguadze, G., et al. (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

Okada, T., & Simon, H. A. (1997). Collaborative discovery in a scientific domain. *Cognitive Science, 21*(2), 109–146. https://doi.org/10.1016/S0364-0213(99)80020-2

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* (pp. 73–78). Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-3013.

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

Poznanski, B., Hart, K. C., & Graziano, P. A. (2021). What do preschool teachers know about attention-deficit/hyperactivity disorder (ADHD) and does it impact ratings of

child impairment? *School Mental Health, 13*(1), 114–128. https://doi.org/10.1007/s12310-020-09395-6

Reinke, W. M., Stormont, M., Herman, K. C., Puri, R., & Goel, N. (2011). Supporting children's mental health in schools: Teacher perceptions of needs, roles, and barriers. *School Psychology Quarterly, 26*(1), 1–13. https://doi.org/10.1037/a0022714

Renkl, A. (2014). Toward an instructionally oriented theory of example-based learning. *Cognitive Science, 38*(1), 1–37. https://doi.org/10.1111/cogs.12086

Scheuer, O., McLaren, B. M., Loll, F., & Pinkwart, N. (2012). Automated analysis and feedback techniques to support and teach argumentation: A survey. *Educational Technologies for Teaching Argumentation Skills*, 71–124. https://doi.org/10.2174/978160805015411201010071

Schmidt, H. G., & Rikers, R. M. J. P. (2007). How expertise develops in medicine: Knowledge encapsulation and illness script formation. *Medical Education, 41*(12), 1133–1139. https://doi.org/10.1111/j.1365-2923.2007.02915.x

Schulz, C., Meyer, C. M., & Gurevych, I. (2019). Challenges in the automatic analysis of students' diagnostic reasoning. *Proceedings of the AAAI Conference on Artificial Intelligence, 33*(01), 6974–6981. https://doi.org/10.1609/aaai.v33i01.33016974.

Shin, D. (2021). The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI. *International Journal of Human-Computer Studies, 146*, Article 102551. https://doi.org/10.1016/j.ijhcs.2020.102551

Stadler, M., Sailer, M., & Fischer, F. (2021). Knowledge as a formative construct: A good alpha is not always better. *New Ideas in Psychology, 60*, Article 100832. https://doi.org/10.1016/j.newideapsych.2020.100832

Sweller, J., van Merriënboer, J. J. G., & Paas, F. (2019). Cognitive architecture and instructional design: 20 years later. *Educational Psychology Review, 31*(2), 261–292. https://doi.org/10.1007/s10648-019-09465-5

Van Merriënboer, J. J. G. (2013). Perspectives on problem solving and instruction. *Computers & Education, 64*, 153–160. https://doi.org/10.1016/j.compedu.2012.11.025

VanLehn, K. (1996). Cognitive skill acquisition. *Annual Review of Psychology, 47*, 513–539. https://doi.org/10.1146/annurev.psych.47.1.513

Van der Linden, W. J. (2016). Lognormal response-time model. In W. J. van der Linden (Ed.), *Handbook of item response theory, 1* pp. 289–310). Boca Raton, FL: Chapman & Hall/CRC Press. models.

Vogel, F., Wecker, C., Kollar, I., & Fischer, F. (2017). Socio-cognitive scaffolding with computer-supported collaboration scripts: A meta-analysis. *Educational Psychology Review, 29*(3), 477–511. https://doi.org/10.1007/s10648-016-9361-7

Weinberger, A., Stegmann, K., & Fischer, F. (2010). Learning to argue online: Scripted groups surpass individuals (unscripted groups do not). *Computers in Human Behavior, 26*(4), 506–515. https://doi.org/10.1016/j.chb.2009.08.007

Wisniewski, B., Zierer, K., & Hattie, J. (2019). The power of feedback revisited: A meta-analysis of educational feedback research. *Frontiers in Psychology, 10*, 3087. https://doi.org/10.3389/fpsyg.2019.03087

Zhu, M., Lee, H.-S., Wang, T., Liu, O. L., Belur, V., & Pallant, A. (2017). Investigating the impact of automated feedback on students' scientific argumentation. *International Journal of Science Education, 39*(12), 1648–1668. https://doi.org/10.1080/09500693.2017.1347303

Zhu, M., Liu, O. L., & Lee, H.-S. (2020). The effect of automated feedback on revision behavior and learning gains in formative assessment of scientific argument writing. *Computers & Education, 143*. https://doi.org/10.1016/j.compedu.2019.103668