Please quote as: Wambsganss, T.; Söllner, M.; Koedinger, K.; Leimeister, J. M. (2022): Adaptive Empathy Learning Support in Peer Review Scenarios. ACM CHI Conference on Human Factors in Computing System (CHI). New Orleans, Louisiana, USA.

Adaptive Empathy Learning Support in Peer Review Scenarios

Thiemo Wambsganss
thiemo.wambsganss@unisg.ch
University of St.Gallen
St.Gallen, Switzerland
Carnegie Mellon University
Pittsburgh, United States

Matthias Söllner
soellner@uni-kassel.de
University of Kassel
Kassel, Germany

Kenneth Koedinger
kk1u@andrew.cmu.edu
Carnegie Mellon University
Pittsburgh, United States

Jan Marco Leimeister
janmarco.leimeister@unisg.ch
University of St.Gallen
St.Gallen, Switzerland
University of Kassel
Kassel, Germany

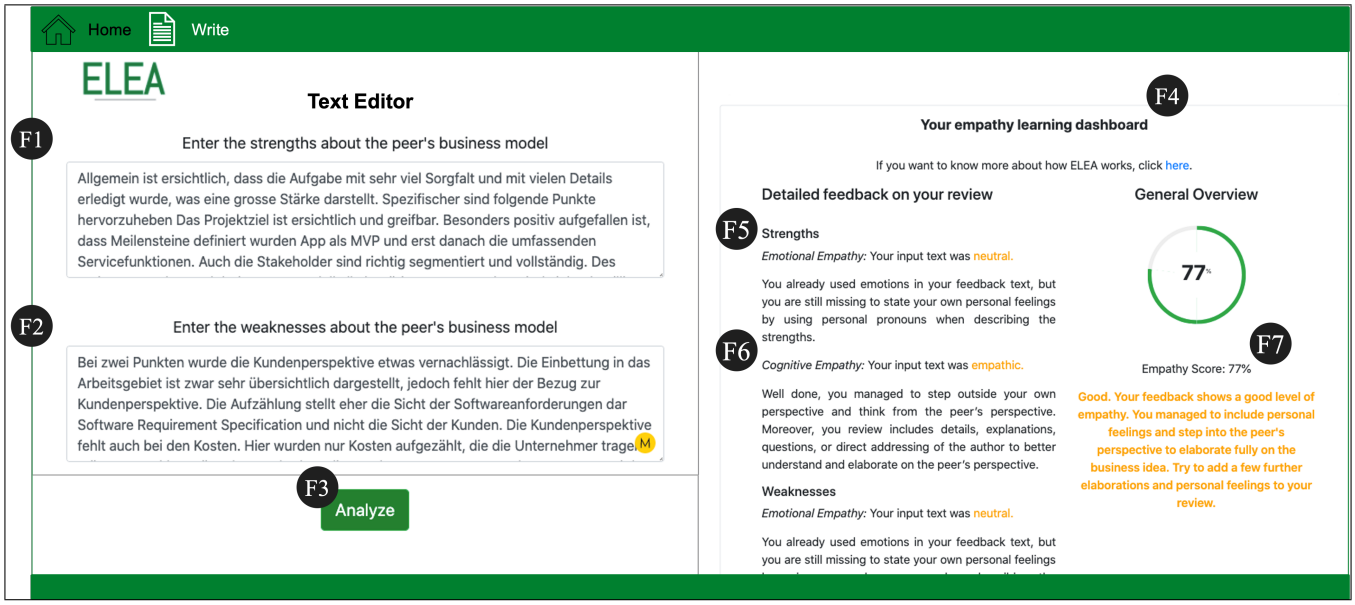

Figure 1: Screenshot of our adaptive empathy learning support system: a user receives feedback on the cognitive and emotional empathy level of her text in a peer review exercise.

ABSTRACT

Advances in Natural Language Processing offer techniques to detect the empathy level in texts. To test if individual feedback on certain students' empathy level in their peer review writing process will help them to write more empathic reviews, we developed ELEA, an adaptive writing support system that provides students with feedback on the cognitive and emotional empathy structures. We compared ELEA to a proven empathy support tool in a peer review setting with 119 students. We found students using ELEA wrote more empathic peer reviews with a higher level of emotional empathy compared to the control group. The high perceived skill learning, the technology acceptance, and the level of enjoyment provide promising results to use such an approach as a feedback application in traditional learning settings. Our results indicate that learning applications based on NLP are able to foster empathic writing skills of students in peer review scenarios.

CHI '22, April 29-May 5, 2022, New Orleans, LA, USA

ACM ISBN 978-1-4503-9157-3/22/04…$15.00
https://doi.org/10.1145/3491102.3517740

CCS CONCEPTS

• **Applied computing** → **Interactive learning environments**; • **Computing methodologies** → **Natural language processing**; • **Human-centered computing** → *Laboratory experiments.*

KEYWORDS

Educational Applications, Writing Support Systems, Automated Feedback, Empathy Learning

ACM Reference Format:

Thiemo Wambsganss, Matthias Söllner, Kenneth Koedinger, and Jan Marco Leimeister. 2022. Adaptive Empathy Learning Support in Peer Review Scenarios. In *CHI Conference on Human Factors in Computing Systems (CHI '22), April 29-May 5, 2022, New Orleans, LA, USA.* ACM, New York, NY, USA, 17 pages. https://doi.org/10.1145/3491102.3517740

1 INTRODUCTION

"The biggest deficit we have in our society, and in the world right now, is an empathy deficit. We are in great need of people being able to stand in somebody else's shoes and see the world through their eyes." (Barack Obama in 2009, talking to Students in Istanbul)

As Barack Obama, former president of the United States, stated, empathy is not only an elementary skill for our society, civil discourse, and daily interaction but also for professional communication as well as successful teamwork, and thus elementary for educational curricula (i.e., OECD Learning Framework 2030 [57]). It is the *"ability to simply understand the other person's perspective [. . .] and to react to the observed experiences of another"* ([24], p.1). Empathy skills not only pave the foundation for successful interaction in digital companies, e.g., in agile work environments [48], but they are also one of the key abilities in the future that distinguish human work force and artificial intelligence agents from one another [67]. However, besides the growing importance of empathy, research has shown that empathy skills of US college students have decreased from 1979 to 2009 by more than thirty percent and even more rapidly from 2000 to 2009 [40]. On these grounds, the Organization for Economic Cooperation and Development (OECD) claims that training empathy skills should receive a more prominent role in today's higher education [57]. To train empathy, institutions, for instance, universities, traditionally rely on experiential learning scenarios, such as shadowing, communication skills training or role-playing, e.g., in medical education [47]. Individual empathy training is therefore only available for a limited number of students, since individual support through a student's learning journey is often hindered due to traditional large-scale lectures or the growing field of distance learning scenarios such as Massive Open Online Classes (MOOCs) [80]. However, to develop skills such as empathy, it is of great importance for the individual student to receive continuous feedback throughout their learning journey [34, 94]. In fact, educational institutions are limited in providing these individual learning conditions especially for empathy skill training.

Due to the interdisciplinary research interest, the term empathy is defined from multiple perspectives in terms of its dimensions or components [26]. Being aware that there are multiple perspectives on empathy, in this paper, we focus on the cognitive and emotional components of empathy as defined by Davis [24] and Lawrence et al. [44]. Therefore, we follow the "Toronto Empathy Scale" [86] as a synthesis of instruments for measuring and validating empathy. Empathy refers to the *"ability to simply understand the other person's perspective [. . .] and to react to the observed experiences of another"* ([24], p.1), and it consists of both emotional and cognitive components [86]. While emotional empathy lets us perceive what other people feel, cognitive empathy is the human ability to recognize and understand other individuals [44].

One avenue to model these empathy constructs in natural language (e.g., to provide adaptive feedback) is offered by recent advances in Natural Language Processing (NLP) and Machine Learning (ML) [74]. Empathy detection is a growing research approach to model empathetic structures and phrases of a given text [17, 18]. This information can be leveraged to score the quality of a text and provide students with individual feedback on their empathy level, e.g., in peer reviews [16, 28]. Researchers especially from the field of educational technology have designed pedagogical scenarios to train the empathy skills of students, such as through virtual reality role-playing for social work education [30], virtual agents to simulate patient treatments for nurses (e.g., [47]) or empathy text feedback on computer-mediated communication platforms to foster empathy for employee–customer relationships [77]. Despite the large number of studies, the existing literature lacks an approach with findings on how to design an adaptive and intelligent learning tool to assist students in developing the ability to simply comprehend another person's point of view and react to their observed experience with intelligent feedback on natural language [69].

Given this potential for leveraging empathy detection to enhance learning, we designed and built ELEA (short for *Empathy Learning Application*), an adaptive learning tool that provides students with feedback on their cognitive and emotional empathy level during their peer review writing process. Being aware that empathy modeling is always biased and wrongly predicted labels (as well as biased correct predicted ones) might harm certain user groups, our aim with this paper was to provide a proof-of-concept study by investigating the hypothesis that individual feedback on certain students' empathy level in their peer review writing process will help them to write more empathic reviews. Hence, we followed two different development approaches to build a user-centered design of ELEA for the "average" learner in a particular learning scenario at our university. First, we used a rigorous theory-driven approach, where we systematically analyzed literature in the field of educational technology and pedagogical theories based on [93] to derive requirements and principles for a first design of ELEA. Second, we followed a user-centered design approach, where we interviewed 28 students from our university to derive user stories and needs for a design of an adaptive empathy learning tool. In this vein, we could ensure and control for specific needs, potential harm, and unintended consequences of a tool that provides students with feedback on their empathic writing skills. To build an individual and adaptive feedback tool, we built on the empathy annotated peer-review corpus of [98], since the data contains 500 student-written texts annotated for their cognitive and emotional empathy level based on a rigorous annotation guideline following [24, 86]. The corpus represents a balanced and representative sample of student peer reviews in German, and thus enables us to train a model which comes with satisfying accuracy not only on the train and test data, but is also able to model the differences and nuances of student-written text in the particular scenario we embedded it into robustly [82]. We trained and tuned a state-of-the-art transfer learning model to detect the cognitive and emotional empathy level of student peer reviews following [27, 43]. This model now serves as the underlying feedback algorithm of ELEA.

To determine the impact of ELEA on students' empathy skills, we evaluated our learning tool in a peer learning scenario in comparison with a proven approach for supporting empathetic writing in technology-mediated communication [77]. In a study with 119 students, we observed that participants who used ELEA wrote more empathic peer reviews with a higher level of emotional empathy compared to the ones using the alternative approach. Moreover, we measured the perceived empathy skill learning [86], the technology acceptance and the perceived level of enjoyment of both tools using key constructs [90, 91]. We found that the perceived empathy skill learning, the perceived usefulness, the intention to use, and the level of enjoyment of ELEA provide promising results. They indicate that ELEA might help students learn how to react to other students' perspectives and motivate them to reflect on their empathetic text writing in peer review scenarios.

Besides the novel contributions, our research comes with several limitations. With ELEA we present the first adaptive learning tool for empathy skills. This is especially novel, since recent advances in NLP and ML have, to the best of our knowledge, not been leveraged to provide adaptive empathy feedback. Past research has mostly built empathy learning tools based on syntactical analysis (e.g., [77]) or embedded emotional modeling in other contexts than education (e.g., [60, 87]). However, our design is only based on learners and on empirical evaluation in the pedagogical scenario of peer reviews in German language. Other languages, domains or pedagogical scenarios might require certain design adaptations (e.g., through a different understanding of empathy due to culture, domain, and user group). Second, we show the effectiveness and usefulness of ELEA by comparing it to the current state of text-based learning tools for empathy skills in an experimental peer learning scenario. The results demonstrate the benefits of leveraging NLP and ML for intelligent feedback on empathy skills on a student's learning journey [74], e.g., in other collaborative learning settings [37, 66]. However, the design of our tool as well as the feedback algorithm is biased by a Western European team of researchers. It also comes with the limitations that the tool is designed for the "average" student, with the drawbacks of possibly excluding certain user groups other than from our research design (e.g., individuals on the autism spectrum). Future research is needed to investigate the effects of our design such as possible harms or bias against these minorities.

2 RELATED WORK AND CONCEPTUAL BACKGROUND

Our work is inspired by previous studies on the concept of empathy, technology-based learning systems for empathy, studies about empathy detection algorithms, and self-regulated learning theory, which serves as an underlying theory for our main hypothesis.

2.1 The Concept of Empathy

Empathy plays an essential role in daily life in many practical situations, such as client communication, leadership, or agile teamwork [57]. Therefore, especially business schools today are increasingly trying to focus on fostering empathy skills (e.g., [63]) to provide students with the right skill set to meet future job profiles (i.e., [57, 92]). Since Titchener's German word "Einfühlung" [88, 105] was related with the term empathy, the construct of empathy has been considered a fundamental component of social cognition that contributes to the human ability to understand and respond adaptively to other people's emotions [86]. Empathy has numerous definitions from various fields corresponding to the different research lenses and perspectives. It was originally translated and understood as "feeling into" [25, 26]. According to several scholars, there is no clear, universal definition of empathy [23, 55]. Table 1 lists a non-exhaustive overview of regularly used empathy definitions from research in chronological sequence starting with the most recent ones.

Definition	Authors
"An emotional process, or an accurate affective insight into the feeling state of another."	Spreng et al. (2009) [86]
"The drive or ability to attribute mental states to another person/animal, and entails an appropriate affective response in the observer to the other person's mental state."	Baron-Cohen and Wheel-wright (2004) [9]
"The ability to experience and understand what others feel without confusion between oneself and others."	Decety and Jackson (2004) [26]
"An affective response that stems from the apprehension or comprehension of another's emotional state or condition, and which is similar to what the other person is feeling or would be expected to feel in the given situation."	Eisenberg (2000) [29]
"An other-oriented emotional response congruent with the other's perceived welfare."	Batsom et al. 1997 [10]
"The ability to put oneself into the mental shoes of another person to understand his or her emotions and feelings."	Goldman 1993 [32]
"Reactions of one individual to the observed experiences of another [...] and simply understanding the other person's perspective."	Davis 1983 [24]

Table 1: Non-exhaustive overview on various definitions of the term empathy ordered by the year of publication.

Aside from defining what empathy is, several studies focus on how to quantify the construct. To measure empathy, most researchers apply questionnaires with self-report measures, although alternative methods exist including neuroscientific or behavioral measurements. Behavioral measurements include the Kids Empathetic Development Scale [68] and neuroscientific measures are, for example, Magnetic Resonance Imaging (e.g., [13]). In our literature review, we found dozens of established empathy scales, ranging from the "hogan empathy scale" [36] including 64-items capturing four different dimensions (social self-confidence, even-temperedness, sensitivity, and nonconformity) or the "Toronto empathy questionnaire" by [86] consisting of 16 items.

In our research, we follow the Toronto empathy scale, since [86] created the scale by investigating the various existing empathy instruments in literature and determining what they have in common. Furthermore, the resulting empathy questionnaire is well cited and based on a measure that represents "the broadest, most general definition of empathy.", where empathy consists of both emotional and cognitive components [86]. In the context of student peer reviews, cognitive empathy (perspective taking) is the writer's ability to use cognitive processes such as role-taking, perspective-taking, or "*decentering*" while evaluating the peers' submitted tasks. The student sets aside her own perspective and recognizes the perspective of the peer. Cognitive empathy can happen purely cognitively in that there is no reference to any affective state [9] but mostly includes understanding the other's emotional state as well. Emotional empathy (emphatic concern) is the writer's emotional response to the peers' affective state. The students can either show the same emotions as read in the peer review or simply state an appropriate feeling towards the peer. Typical examples include sharing excitement with the peer about the content submitted or showing concern over the peer's opinion. The following example depicts high emotional empathy: *"I think your idea is brilliant!"*.

2.2 Technology-Based Learning Systems for Empathy

Besides the importance of empathy in daily life, studies have shown that empathy skills of US college students have decreased from 1979 to 2009 by more than thirty percent and even more rapidly in the last period from 2000 to 2009 [40]. A possible explanation is the growing amount of digital communication in our society [40]. Scientists, therefore, urge that training empathy skills should receive a more prominent role in today's higher education (e.g., [30, 57]). In fact, individual support of empathy learning is missing in most learning scenarios. In some domains, training programs are designed to increase empathy skills through role plays, films, literature or video games (e.g., [11]). Since social professions, in particular, are characterized by interactions, similar training programs that promote empathy or empathetic forms of expression have so far also been successfully implemented for social workers [35], doctors and nurses [6]. In business education, empathy is usually trained through communication scenarios, classroom exercises, role plays or experiential learning (e.g., [63]). In fact, empathy is often regarded as a subcomponent of social competence [102]. Corresponding support measures often take place in extensive programs to promote social development. However, in order to train particular skills such as empathy, it is essential for the individual student to receive continuous feedback, also called formative feedback, throughout the learning process [34, 94, 95]. According to [76], the result of feedback is specific information about the learning task or process that fills a gap between what is understood and what should be understood. Even in areas where empathy is part of the curriculum, such as health or social work, the ability of a teacher to provide tutoring is naturally limited by time and availability constraints.

Especially in more frequent large-scale lectures and distance learning scenarios, the ability to individually support a student's empathy ability is hampered because it always was and still is difficult for educators to provide continuous and individual feedback to a single student. Many scholars, particularly in the field of educational technology, have looked into how technology-based solutions might help students gain empathy and close the gap. When compared to human teachers, the use of information technology in education has several advantages, including consistency, scalability, perceived fairness, widespread use, and better availability. As a result, technology-enhanced empathy learning systems can help relieve some of the burdens on teachers by supporting learners with adaptive empathy feedback. Scientists have mainly used three approaches from educational technology to foster empathy skills of students [41]:

- **Computer-assisted instruction (CAI)** is often embedded in the form of virtual reality (VR) learning tools in pedagogical scenarios to enable students to directly dive into the perspective of a peer, e.g., a client or patient (e.g., [6]).
- **Intelligent tutoring systems (ITS)** are often used in the form of virtual agents built into online tools, e.g., to enable interaction with emotional avatars (e.g., [42]).
- **Computer-supported collaborative learning (CSCL)** tools are, for instance, implemented to enhance empathy in the text communication of learners [77]. In their approach, [77] use a simple library of messengers based on neurolinguistics, psychometrics, and text mining techniques to promote empathy among students' interaction, based on identification and text matching suggestions.

Our approach combines two perspectives: ITS and CSCL. The combination of ITS and CSCL to design adaptive empathy learning tools is scarcely investigated in literature. [87], for example, developed ClientBot, a text-based conversational agent, trained on a library of 2354 psychotherapy transcripts, which provided explicit feedback on the usage of basic interviewing and counseling skills. The results showed that participants using ClientBot used more reflections during practicing feedback. Similarly, our aim is to provide pedagogical feedback on a learner's actions and solutions, hints, and recommendations to encourage and guide future activities in the writing processes or automated evaluation to indicate whether a student's reaction to another person's perspective is emotionally appropriate. We rely on NLP and ML to analyze the given text and provide adaptive feedback in students' peer writing process. [60] investigated that their writing support tool MepsBot, which provides assessment and recommendations about emotional support to users when writing comments to peers in online mental health communities, improves user satisfaction with and confidence in their comments. Building on these literature streams, we hypothesise, that individual feedback on students' empathy levels when writing peer reviews will help them to write more empathic texts. We evaluate our hypothesis by comparing our tool against a CSCL approach, since it is widespread and has been empirically proven to support students' empathy skills [77].

2.3 Empathy Detection in Natural Language

The detection of empathy in texts is a growing research field in NLP and ML. Empathy detection aims to identify and model empathetic structures of a given text [17, 18]. The task is usually regarded as a subset of emotion detection, which in turn is often referred

to as being part of sentiment analysis. The detection of emotions in texts has made major progress, with sentiment analysis being one of the most prominent areas in recent years [46]. However, most scientific studies have been focusing on the prediction of the polarity of words for assessing negative and positive notions. This has been done for various domains such as online forums [1] or twitter postings [75]. Most existing work for empathy detection focuses on spoken dialogue, addressing conversational agents, psychological interventions, or call center applications (e.g., [49], [62], [4], or [81]). Several studies address the detection and prediction of empathy in natural texts [38, 106], e.g., for empathy modelling based on news story reactions [17]. Nevertheless, the potential of empathy detection has been investigated in different domains but barely leveraged for individual feedback in a student's learning progress (i.e., such as [17]). Recently, [98] have annotated a novel corpus to model empathy skills in student-written peer reviews. The annotation scheme is based on constructs in psychological literature [24, 44] and evaluated through a rigorous annotation study. Moreover, the corpus represents a balanced and representative sample of student peer reviews ob business models in German, and thus enables us to train a model which not only comes with a satisfying accuracy on the train and test data, but is also able to model the differences and nuances of student written text in the particular scenario we aim to embed it into [82]. Therefore, we aim to build on that potential to enhance current empathy learning scenarios by providing students with adaptive writing support in a collaborative peer review writing exercise. In fact, the positive impact of NLP and ML algorithms for adaptive skill learning has been demonstrated before [96]. For example, [95] have leveraged a corpus on student-written argumentative peer reviews [97] to provide students adaptive argumentation tutoring in a persuasive writing exercise. However, to the best of our knowledge, no study exists that investigates and evaluates the design of a user-centered empathy learning tool in peer review scenarios based on recent advantages in NLP and ML.

2.4 Individual Differences of User Groups in the Context of Empathy

Literature has highlighted a number of individual differences of user groups in the context of empathy. Even though we aim to "design for the average", we want to outline some prominent individual differences that should be kept in mind to avoid potential undesired negative effects of certain design decisions.

First, we focus on gender differences. [19] provide a compelling overview of gender-based empathy differences. For example, studies conducted by [51, 56] show that gender differences related to empathy already manifest very early in life and can be observed throughout the lifespan, with females typically showing higher levels of empathy compared to males. Further studies even report that the differences appears to grow with age - meaning that the gap between empathy levels of females and males increases [51, 89].

Second, literature has also highlighted certain cultural differences in the context of empathy. For example, a study comparing members of an independent (US) and an interdependent society (Iran) found that participants with interdependent cultural norms (Iran) reported higher empathy than participants with independent cultural norms (US) [107]. A second study shows significant differences in empathy levels comparing Australien caucasian, and mainland Chinese university students [108]. Interestingly, the study combines cultural differences with gender differences. The authors only observed the significant differences among female study participants, but not when comparing male participants. This shows that different forms of individual differences - in this case, gender and culture - can interact with each other to strengthen or weaken the effect of such differences on empathy.

Third, besides gender and cultural differences, literature has also discussed empathy towards different groups of minorities. [52, 53], e.g., observed that perceptions of threat towards a specific group of people (e.g., foreigners) - also called out-group members - impacts the level of empathy towards members of this group. These perceptions can then be mitigated using different mechanisms, such as socialization, which result in higher levels of empathy towards out group members. Another example for out-group members can be neurodiverse people, such as individuals on the autism spectrum [72]. Here, prior research has shown that communication styles that are typically preferred by neurotypical people, such as transformative leadership styles among employees, are less suited for neurodiverse employees, since they prefer a more direct communication that leaves less space for interpretations [58]. When it comes to empathy in specific, a recent study showed that individuals on the autism spectrum showed comparable levels of empathy towards other species compared to the control group [54]. One remarkable outlier in this study, however, was the level of empathy shown towards other human beings. Here, individuals on the autism spectrum showed lower levels than the control group.

To summarize, these individual differences need to be kept in mind when designing systems that relate to the concept of empathy. We want to emphasize that the normative conception of emotional and empathetic language might be understood differently in various contexts from different user groups (e.g., gender, cultural and other minorities). The appropriate design of an empathy learning tool based on empathy detection algorithms from NLP and ML is therefore always biased from a certain perspective. As a result, different recommendations and prompts based on these algorithms need to be always adopted for the specific scenario and user group. In our research, our main goal was to design a tool for the specific context we could control and investigate: the construction of student peer reviews in German language from "average" German master students at a Western European university. As stated, the concept of empathy is very controversial and, thus, scripted as well as predefined support might be confusing or even harm certain user groups from other contexts. In our case, this could be students from other cultural backgrounds, normative conceptions, or neurominorities, such as individuals on the autism spectrum.

2.5 Self-Regulated Learning Theory to Foster Individual Learning

We believe that self-regulated learning theory supports our underlying hypothesis that individual and personal feedback on a student's ability to react to other people's perspectives in a peer review scenario will support her learning activity and engage the student

to train her abilities to write more empathic peer reviews. Self-regulated learning theory reflects that students learn better with formative feedback and goal setting [8]. Especially for students in a learning process, critical reflection through self-monitoring and self-evaluation is an important component for effective learning [110], also reflected in literature on transformation learning (e.g., [50]). It can be an initial trigger for a student's learning process and thus the creation of new knowledge structures [64]. However, the right portion of self-monitoring and self-reflection in combination with a learning goal is important for students to learn effectively [8, 110]. Feedback should specify goals, track progress toward those goals, and identify actions that will help the learner achieve those goals in order to be effective [34]. The level of feedback adaptivity, on the other hand, can vary greatly. As a result, it's critical to consider the impact of different adaptive feedback granularity levels to ensure that our feedback actually aids them in learning. [14] claim that accurate self-evaluation and feedback of one's learning progress are key components for effective self-regulation of learning, in the vein of social cognition theory and self-regulated learning [7]. Humans, on the other hand, have a difficult time monitoring and evaluating their learning and comprehension of complicated content [14]. As a result, the creation of appropriate evaluation and feedback aspects for certain skills may aid learners in learning more successfully over time (i.e., [73, 110]). Combining skill monitoring with formative assessment and performance feedback is one technique to support successful learning monitoring and evolution [15, 34]. Repeated feedback on students' abilities can lead to better results in a certain pedagogical task in short-term and, thus, can increase the overall metacognition skill learning in a long-term intervention [34, 73, 94]. Therefore, we believe that the right level of feedback on a students' skill, such as empathy, could lead to more self-efficacy and thus to motivation to learn how to react to other people's perspectives in an appropriate manner. We test our hypothesis in a short-term intervention scenario to investigate if adaptive user-centered feedback on students' written peer reviews helps them to write more empathetic texts. This lays the foundation to investigate further impacts of adaptive empathy feedback on students' metacognitive empathy skills in longitudinal studies.

3 DESIGN OF LEARNING SYSTEM

ELEA is composed of two main components: an adaptive user interface and an intelligent empathy feedback algorithm in the back end. The basic user interaction concept of ELEA is illustrated in Figure 2. A user conducts a peer writing exercise and receives adaptive feedback on the cognitive and emotional empathy level. The design process is based on self-regulated learning theory as a kernel theory [7, 110]. The feedback mechanism of ELEA is designed with the objective to provide students with self-monitoring and self-evaluation independent of an instructor, time, and location. As stated before, our aim is to provide a proof-of-concept study by investigating the hypothesis that individual feedback on certain students' empathy level in their peer review writing process will help them to take the perspective of the recipient of their review, and thus, to write more empathic reviews. Hence, we designed a learning tool for the "average" students of the specific pedagogical scenario of peer reviews on business models in German language. In this vein, we are aware that the use of our tool in other domains or circumstances, such as, in other cultures might not yield the same results due to bias in the design and different contexts.

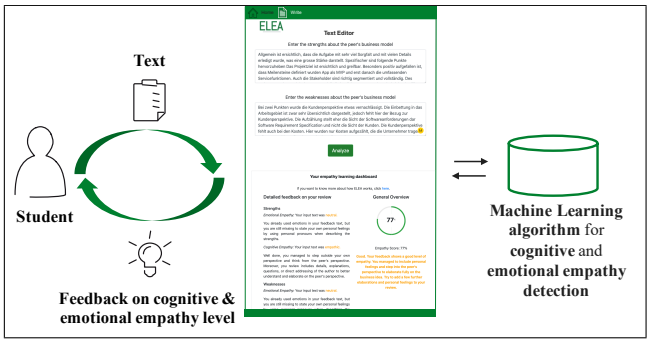

Figure 2: Basic user interaction concept of ELEA: a student receives adaptive empathy feedback on the cognitive and emotional empathy level in a peer writing exercise.

3.1 User Interface of ELEA

3.1.1 Deriving Requirements from Literature and Users. In order to build a novel learning tool, we followed two different approaches: a rigorous theory-driven approach and an agile user-centered approach following the build-measure-learn paradigm [70]. For the rigorous theory-driven approach, we followed the approaches of [22] and [93] to conduct a systematic literature review with the aim to derive a set of theory requirements for the design of an empathy learning system. We focused our research on studies that display the successful implementation of learning tools for empathy skills. Therefore, we identified two research areas to derive requirements: educational technology and learning theories. We first focused on these literature streams since developing a learning tool for empathy skills is a complex endeavor that is studied by psychologists, pedagogues, and computer scientists using various methodologies. Only publications dealing with or contributing to a type of learning instrument in the subject of empathy learning, such as an established learning theory, were included. On this basis, we chose 110 papers for further examination. We grouped related problems in these contributions into five clusters as literary issues, which served as theory criteria for the learning tool for empathy skills. The five clusters were based on the literature streams of 1) formative feedback [15, 34], 2) learner-centered design [85], 3) technology-mediated learning [33], 4) emotional and cognitive empathy support [24, 86], and 5) learner control [78].

Besides the rigorous theory-driven approach, we followed a user-centered design approach at the same time. First, we conducted 28 semi-structured interviews with students to receive an initial understanding of the needs and requirements of learners for a learning tool for empathy skills [31]. Each interview lasted an average of 40.91 minutes (SD = 15.9 minutes) and consisted of 30 questions. The interviewers were a subset of our university's students who may all benefit from an empathy learning tool. The following issues were discussed with the participants: prior experience with

technology-based learning systems, the importance of skills in university education, system needs for learning metacognition skills (e.g., functionality, design), and system requirements for learning empathy (e.g., functionalities, design). Only master's students were selected for the interviews in order to obtain impressions based on many years of learning experience and to ensure that the interviewees had experiences in collaborative learning scenarios. The students interviewed were 24.82 years old on average (SD = 1.98) and all studied economics, law, or psychology; 15 were male and 13 were female. The interviews were examined using a qualitative content analysis after a more precise transcription. The interviews were coded, and abstract categories were created based on the results. During the examination, open coding was used to provide a consistent coding system [31]. We derived user stories and aggregated the most common ones following [21]. The aim of the interviews was to follow a design thinking process to get an understanding of the users' needs and perspectives. The user stories have been derived by one research who conducted the interviews and have been discussed based on the transcriptions with two senior researchers in three workshops to ensure the validity of the findings. The findings provided with an overview of the needs and requirements of users for an adaptive empathy learning system. The results are in line with other design investigations for empathy learning tools such as [99].

Based on those insights, we designed low-fidelity prototypes of ELEA to test different design hypotheses with end-users to learn more about the human-computer interaction of an adaptive learning tool for empathy skills- For example, we hypothesized that students would like to receive a specific numerical empathy score on their different text paragraphs. Therefore, we designed two paper prototypes: one providing text feedback based on three categorical variables *"non-empathic", "neutral"* and *"empathic"*, and one prototype providing feedback based on five numerical variables one to five (1: low, 5 high). The empathy feedback algorithm was simulated by a human. The hypothesis was validated with 12 users. However, we learned, that the majority of students rather like the categorical empathy feedback. Therefore, the final version of ELEA contributes to these findings by providing the users with feedback based on three categorical variables. In total, we conducted three cycles testing several design hypotheses with a total of 65 users (cycle 1: 12 users, cycle 2: 25 users, cycle 3: 28 users). These users were different from the ones recruited for the semi-structured interviews but also students from our university with a similar age and gender distribution. Based on both approaches, we finally came up with several design principles on how to build an adaptive learning system for empathy skills (see Table 2). For example, design principle three described that an effective empathy learning tool should be employed in a theory-based learning scenario in which students can apply and train their empathy skills (e.g., in a peer learning setting) to allow students to receive formative feedback on their skill level or design principle four described that students should receive theory-based explanations and recommendations on different granularity levels for certain empathy feedback categories to allow students to transparently understand and use the feedback to foster their skills. The design principles were followed in the instantiation of our current version of ELEA.

	Design Principles
1)	To design effective learning tools for students to improve their empathy skills, employ a web-based application with a responsive lean and intuitive UX, which includes motivational learning elements (e.g., learning progress indicator) to allow students to use the tool intuitively and stay motivated to learn.
2)	To design effective learning tools for students to improve their empathy skills, employ an individual empathy feedback mechanism that provides instant and individual feedback on different granularity levels based on the learning content to allow students to receive and choose the right amount of needed input.
3)	To design effective learning tools for students to improve their empathy skills, employ a theory-based learning scenario in which students can apply and train their empathy skills (e.g., in a peer learning setting) to allow students to receive formative feedback on their skill level.
4)	To design effective learning tools for students to improve their empathy skills, employ theory-based explanations and recommendations for certain empathy feedback categories to allow students to transparently understand and use the feedback to foster their skills.

Table 2: Derived design principles on how to build an empathy learning tool, which we followed in our design instantiation of ELEA.

3.1.2 User Interaction of ELEA. Based on design principle one, we built ELEA as a responsive web-based application that can be used on all kinds of devices. A screenshot of ELEA and its different functionalities (e.g., *F1 - F7*) can be seen in Figure 1. ELEA provides the user with a rather simple text input field (*F1*) in which they can write or copy a text. In the current version, we embedded ELEA in a peer learning scenario, where students write a business model review to a peer by elaborating on strengths, weaknesses, and improvement suggestions (*F2*). Below the input field, users can click on the analyze button (*F3*) to receive individual feedback on the empathy level of their text through a personal learning dashboard (*F4, F5, F6, F7*). As required in design principle two, the dashboard provides different granularity levels of feedback, which enables the user to control the amount of needed feedback information [78]. A total empathy score and an adaptive recommendation message provide an initial overview of the quality of the text (*F7*). The individual recommendation is based on the empathy feedback level and provides a combination of motivating elements and solution suggestions to improve the individual skill level according to [34]. Moreover, the users can receive detailed feedback on their business model review based on the written strengths, weaknesses, or improvement suggestions. For each review component, the emotional and cognitive empathy level is scored for the variables *"non-empathic", "neutral"* and *"empathic"* (*F5, F6*). Moreover, ELEA provides the user with clear steps on how to improve the respective cognitive and emotional empathy level of the specific business model peer review part. These action steps provide the user with orientation and context to improve their writing quality [34, 85]. Users can implement ELEA's

feedback in the text input and analyze the improved peer review again. ELEA will then adapt the empathy dashboard with a new overall empathy score, which allows the students to detect their empathy progress easily. Moreover, as found in design principles four, best practices and explanations about cognitive and emotional empathy theory are provided by clicking on the explanation button. A pop-up window displays a transparent explanation of how ELEA works and provides a theory-based definition of cognitive and emotional empathy structures (see figure 3).

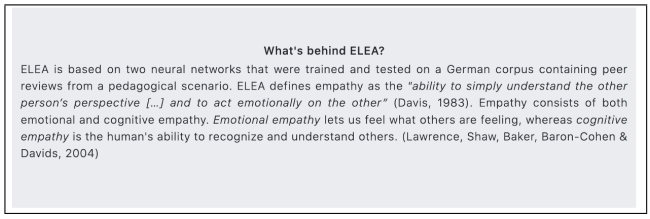

Figure 3: Screenshot of exemplary explanations and details of ELEA.

3.2 Feedback Algorithm of ELEA

In order to fulfill the users' requirements to give instant feedback on their texts, we built on the empathy annotated student-written text corpora of [98]. The corpus serves as the underlying data set to train and tune a state-of-the-art transfer learning model to design and build an adaptive empathy feedback tool. A requirement for developing NLP methods that are able to identify empathetic structures in written texts is the availability of annotated corpora. We searched the literature for a corpus that fulfilled the following criteria: 1) the corpus contains annotated empathic student essays, 2) it has a sufficient corpus size to be able to use the trained model in a real-world scenario that fulfills the user requirements, and 3) the annotations are based on a rigorous annotation guideline for guiding the annotators towards a moderate agreement. The cognitive and emotional empathy annotated student peer reviews corpus published in [98] fulfilled all these requirements. The corpus consists of 500 student-generated peer reviews written in German which are annotated for the cognitive and emotional empathy levels on a 1-5 Likert scale (1: low, 5: high). More information on the text domain, the annotation guidelines, the annotation study as well as on the corpus statistics can be found in [98].

To provide students with feedback on the empathy quality of their texts, we implemented an approach for detecting the cognitive and emotional empathy levels in them [43]. The empathy detection task is considered a paragraph-based multiclass classification task, where each paragraph is either considered to be a *strength, weakness* or *improvement suggestion* and has a "non-empathic", "neutral" or "empathic" cognitive and emotional empathy level. For the current version of ELEA our main objective was to assess the cognitive and emotional empathy level of the specific paragraphs since students enter strengths, weaknesses, and improvement suggestions separately. Therefore, a classification of review components is currently not necessary. Hence, we trained a predictive model on the corpus of [98] following the architecture of Bidirectional Encoder Representations from Transformers (BERT) proposed by [27]. We classified text paragraphs into the cognitive and emotional empathy level based on three labels *"non-empathic", "neutral", and "empathic"*. We used the BERT model from *deepset*[1], since it is available for German and provides a deep pre-trained model that was unsupervised while training on domain-agnostic German corpora (e.g., the German Wikipedia). The novelty of this architecture is the ability to capture semantic information from pre-trained texts, which can then be used for other downstream tasks without the need for retraining, e.g., for identifying the empathy level of text components. We split the data into 70 % training, 20 % validation, and 10 % test data. For applying the model, the corpus texts were split into word tokens to fulfill the preparation requirements for BERT. The special preprocessing for BERT was conducted by utilizing the tokenizer and processor provided by deepset. The goal of our model is to provide accurate predictions to identify and classify the empathy level of review paragraphs that can be used for accessing the skill level of students and thus provide adaptive guidance and feedback on how to improve their empathetic writing. We tested the model with different parameters. The best performing combination incorporated a dropout probability of 10%, a learning rate of $3e^{-5}$ and the number of epochs were 3 [27]. Each combination was evaluated using the f1 score metric. After several iterations, we reached a micro f1 score of 74.96% for the detection of emotional empathy and 69.98% for the detection of the cognitive empathy level of a text paragraph. To ensure the validity of our BERT model, we benchmarked against bidirectional Long-Short-Term-Memory-Conditional-Random-Fields classifiers (BiLSTM-CRF). In combination with the corresponding embeddings vocabulary (GloVe) [61] our LSTM reached an unsatisfying f1 score of 61% for detecting the emotional empathy level and 51% for detecting the cognitive empathy level. Therefore, we embedded our BERT model in the back end of ELEA to provide students with adaptive recommendations based on their empathy level in their peer review writing process.

Moreover, based on user requirements, we provide students a total empathy score as an initial overview of the empathy level in their texts (F7). The total empathy score is calculated by summing up all the scores received throughout the detailed feedback, divided by the maximum score possible, and multiplied by 100 to obtain a percentage score. The higher the score value, the more empathetic is the text of the user. A score of 70 to 80 % is considered good but not perfect, since the user already wrote a neutral or an empathic text, but did not use full empathic language. Based on the overall score, the overall empathy recommendation is adopted. In total, we pre-scripted 23 recommendation messages for adaptive empathy feedback in the domain of student-written peer reviews on business models at a Western European university. The scripts are designed for the "average" student in the pedagogical scenario of German student peer reviews on business models. In our three design cycles with a total of 65 users (see section 3.1.1), we also tested different versions and formulations of the recommendation messages. In this way, we could control for harm, misunderstandings, or unintended consequences of our empathy recommendations by qualitatively asking the students about their perception after they received certain scripts. No objections were reported. However, we still want

[1] https://github.com/deepset-ai/FARM

to note that the empathy constructs modeled with our algorithm as well as the scripts are only valid for the context in which we derived and tested them. The formulation of the recommendations as well as the annotations - and, thus, the model - is biased from a Western European perspective. The predictions might be confusing or even harmful for other contexts (see subsection 2.4). Hence, different scenarios, user groups, or languages might require other empathy concepts.

4 EXPERIMENTAL SETUP

Our goal was to evaluate the hypothesis that individual feedback on students' empathy levels in a peer review exercise will foster their empathy skill level in the written texts. To evaluate our hypothesis, we designed an experiment in which participants were asked to write a peer review based on a provided business model essay, as this is a common large-scale teaching scenario to foster the skill learning of students across different business domains (e.g., [71, 96]).

4.1 Participants

We recruited 119 students from our university to take part in our experiment. The experiment was conducted as a web experiment facilitated by the behavioral lab of our university, and thus, designed and reviewed according to the ethical guidelines of the lab and the university. After randomization, we counted 58 valid results in the treatment and 61 in the control group. Participants of the treatment group had an average age of 23.89 (SD= 3.07), 30 were male, 28 were female. In the control group, participants' average age was 23.80 (SD= 3.11), 35 were male, 26 were female. All participants were compensated with an equivalent of about 12 USD for a 25 to 30 minutes experiment.

4.2 Experimental Design

The treatment group used ELEA to do the writing exercise[2], while participants in the control group used an alternative collaborative learning application based on [77] (Figure 4). To control for the differences and similarities in the design between the alternative tool and ELEA, we implemented our own alternative collaborative learning approach. For the design, we followed the approach of *NeuroMessenger* by [77], since it is a recent theory-based approach and empirically proven to foster the empathy skills of users through text recommendations. The learning tool supports the writing process of users with dictionary-based text recommendations (see Figure 4, *F1* and *F2*). Users can use the recommendations and improvement suggestions to correct their texts. In our approach, we derived a German dictionary list for cognitive and emotional empathetic writing in business model peer reviews based on the 14-page annotation guideline of the corpus of [98] with inspiration by [77]. To keep ELEA and the alternative learning approach consistent with each other, there are many functions that are shared between them, e.g., the introduction text and the scripted text input fields (i.e., for strengths, weaknesses, suggestions) are the same across both apps. The explanation buttons and the user interaction correspond respectively to the same ones in ELEA.

[2]ELEA was designed in German to provide German students with feedback on German texts. However, for ease of understanding in this paper, we translated parts of our user interface into English (e.g., see Figure 1).

4.3 Experimental Procedure

The experiment consisted of three main parts: 1) pre-test, 2) peer writing exercise, and 3) post-test. The pre-test and post-test phases were consistent for all participants. In the writing phase, the treatment group used ELEA to write a business model peer review, whereas participants of the control group used the alternative tool.

1) Pre-test: The experiment started with a pre-survey of eight questions. Here, we tested two different constructs to assess whether the randomization was successful. First, we asked four items to test the personal innovativeness of the participants in the domain of information technology following [3]. The items were *"I like to experiment with new information technologies"*, *"If I heard about a new information technology, I would look for ways to experiment with it,"*, *"In general, I am hesitant to try out new information technologies"*, and *"Among my peers, I am usually the first to try out new information technologies".* Second, we tested the construct of feedback-seeking of individuals following [5]. Items are: "*It is important for me to receive feedback on my performance.*", "*I find feedback on my performance useful.*", "*I would like to get more feedback on what behaviors would help me to conduct a task better.*", and "*It is important to me to receive feedback on my progress potential for skill learning.*" Both constructs were measured with a 1- to 7-point Likert scale (1: totally disagree to 7: totally agree, with 4 being a neutral statement).

2) Peer review exercise: In the peer writing phase of the experiment, participants were asked to read an essay about a business model of a peer student. Afterwards, they were asked to write a business model review for the peer by elaborating on the strengths, weaknesses, and improvement suggestions of the particular business model. The participants were told to spend at least 15 minutes writing this review. A countdown indicated the remaining time. They were only able to continue the experiment after the countdown was finished. The treatment group used ELEA to write the review, the control group used the reference tool. We did not provide any introduction to any of the tools. The students using ELEA retrieved individual and adaptive feedback based on our feedback algorithms. Participants using the reference tool retrieved help by dictionary-based recommendations during the writing process.

3) Post-test: In the post-survey, we measured perceived usefulness, intention to use, and ease of use, following the technology acceptance model of [90, 91]. Example items for the three constructs are: "*Imagine the tool was available in your next course, would you use it?*", "*The use of the empathy tool enables me to write more empathetic texts.*", or "*I would find the tool to be flexible to interact with*". Moreover, we measured the perceived level of enjoyment of the students by asking the following items: *"The interaction with the learning tool was exciting"* and *"It is fun to interact with the learning tool"* [39]. Also, we asked the participants to judge their perceived empathy skill learning (PESL) by asking two items that cover cognitive and emotional empathy skills based on [24, 86]: *"I assume that the tool would help me improve my ability to give appropriate emotional feedback."* and *"I assume that the tool would help me improve my ability to empathize with others when writing reviews."*

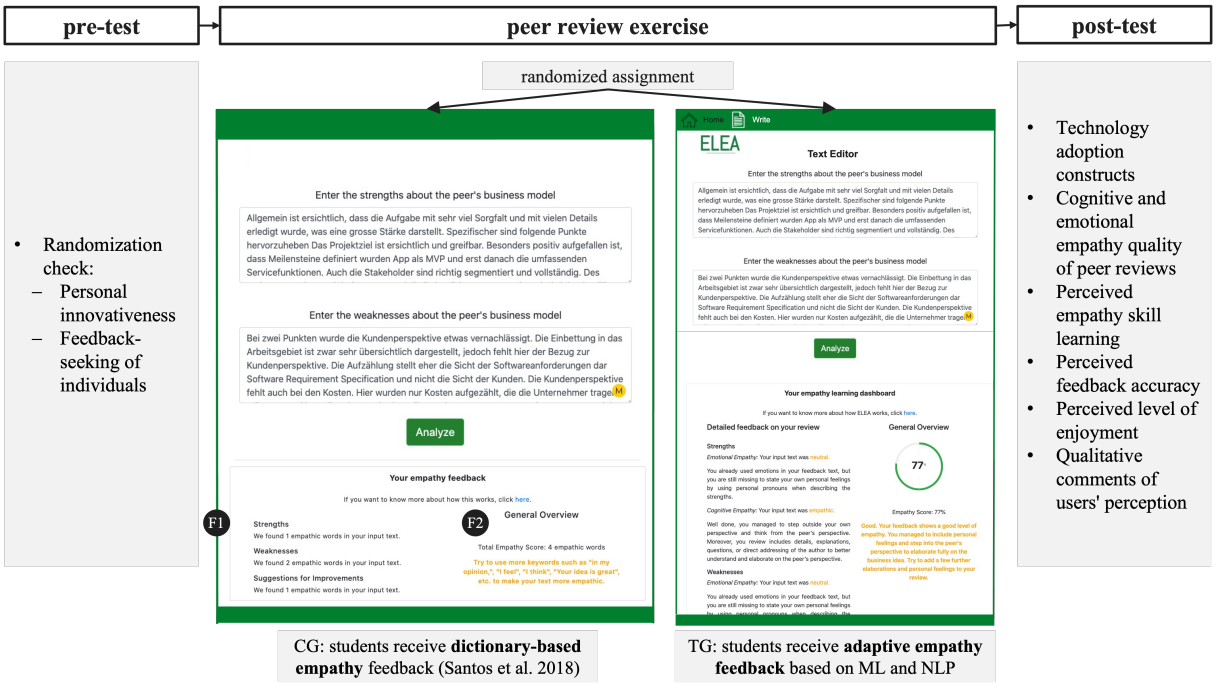

Figure 4: Overview of the experimental setup of our study: participants of the control group (CG) receive dictionary-based empathy feedback based on the approach of [77]. Students in the treatment group (TG) receive adaptive empathy feedback during their peer review writing process with our tool ELEA.

Finally, we surveyed the perceived feedback accuracy (PFA) [65] of both learning tools by asking three items: *"The feedback I received reflected my true performance.", "The tool accurately evaluated my performance."* and *"The feedback I received from the tool was an accurate evaluation of my performance".* All constructs were measured with a 1- to 7-point Likert scale (1: totally disagree to 7: totally agree, with 4 being a neutral statement). Further, we were asking three qualitative questions: "*What did you particularly like about the use of the tool?*", "*What else could be improved?*", and "*Do you have any other ideas?* and captured the demographics. In total, we asked 24 questions.

4.4 Behavioral Measurements

There is a vivid discussion in prior research about the value of using perception-based measures to grasp perceived learning outcomes [12, 83]. Therefore, we decided to enrich the perception-based measurement of empathy skill learning with a more objective approach. This is in line with the argumentation by [84] that it is most desirable when perceived as well as objective learning outcomes can be observed. Thus, besides measuring the perceived constructs in our post-test, we also had two judges measuring the empathy levels of the written texts from both groups to evaluate our main hypothesis. Therefore, we measured two variables: 1) the emotional empathy writing level and 2) the cognitive empathy writing level [44, 86]. To do so, we applied the annotation scheme of [98] based on their 14-pages guideline to judge the emotional empathy writing level as well as the cognitive empathy writing level of the received texts from the peer writing exercise on a document-level. We used these guidelines since an annotator agreement was already conducted during the corpus collection process for the same kind of texts in the same domain (student peer reviews). We relied on two annotators, who independently judge the empathic writing levels of the texts with document-level scores. The objective was to judge how empathic the given text is according to the defined emotional and cognitive empathy dimensions on a Likert scale from 1 to 5 points (1: low, 5: high). Finally, we took the mean of both annotators as a final variable for the cognitive and emotional empathy level of the texts.

5 EVALUATION AND RESULTS

To evaluate our hypothesis that adaptive feedback on students' empathy will help them to foster their empathy skills, our objective was to answer two research questions (RQ):

RQ1: *How effective is ELEA at helping users to foster their empathy skills in peer writing exercises compared to the traditional approach?*

RQ2: *Do students perceive ELEA to be useful, enjoyable, and easy to use, and would they continue to use it in the future?*

To evaluate our first research question, we compared the perceived empathy skill learning between the treatment and the control group. Moreover, we compared the cognitive and emotional empathy writing levels between the written text of the treatment and the control group. Therefore, we applied analysis of variance (ANOVA) to evaluate whether the differences between the groups are statistically significant. We checked their assumptions visually with a

test for normality and a test for homoscedasticity: all assumptions were met. Moreover, we calculated the *Cohen's d* to measure the effect size between the means of the perceived constructs [20]. Furthermore, we controlled for differences in the perceived feedback accuracy between both tools to investigate if adaptive empathy feedback has a significant effect on the user perception of feedback quality.

The second research question will be answered by comparing the constructs of perceived usefulness, intention to use, ease of use, and level of enjoyment for participants using ELEA compared to participants using the alternative tool. Again, we performed an ANOVA to assess whether differences between both groups are statistically significant and calculated the *Cohen's d* to measure the effect sizes. Moreover, we will compare the results of ELEA to the midpoints scale to validate a general positive technology acceptance as done in [96]. Finally, to ensure that the randomization resulted in randomized groups and to control for potential effects of interfering variables with our small sample size, we compared the differences in the means of the two constructs included in the pre-test. For both constructs, we received p values larger than 0.05 between the treatment and the control group (for personal innovativeness p = 0.8676, TG mean= 4.30, SD= 0.57, CG mean= 4.29, SD= 0.59; for feedback-seeking of individuals p = 0.6702, TG mean= 6.13, SD= 0.57, CG mean= 6.10, SD= 0.63).

Group	**Emotional Empathy** (on a 1-to-5 Likert Scale)	**Cognitive Empathy** (on a 1-to-5 Likert Scale)	**Perceived Empathy Skill Learning** (on a 1-to-7 Likert Scale)
Mean ELEA	2.75	3.24	5.03
Mean reference tool	2.20	3.25	3.93
SD ELEA	1.12	1.01	1.05
SD reference tool	0.73	1.07	1.50
p value	0.0027	0.9299	<0.001
effect size (cohen's d)	0.5699 (medium effect)	0.0162 (negligible effect)	0.8411 (large effect)
degrees of freedom	116	116	117

Table 3: Overview of results of the cognitive and emotional empathy level of students using ELEA and students using the reference tool.

5.1 Cognitive and Emotional Empathy Level of Written Texts

We found that students who used ELEA wrote their text with a significantly higher level of emotional empathy compared to participants who used the reference tool (mean value ELEA = 2.75, mean value reference tool = 2.20, p = 0.0027, p<0.01) The calculated *Cohen's d* of 0.5699 indicates a medium effect size. Cohen suggested that d = 0.2 be considered a 'small' effect size, 0.5 represents a 'medium' effect size and 0.8 a 'large' effect size. This means that if two groups' means don't differ by 0.2 standard deviations or more,

Group	**Intention to use** (on a 1-to-7 Likert Scale)	**Perceived usefulness** (on a 1-to-7 Likert Scale)	**Perceived ease of use** (on a 1-to-7 Likert Scale)	**Level of enjoyment** (on a 1-to-7 Likert Scale)
Mean ELEA	5.14	5.05	5.72	5.31
Mean reference tool	3.77	3.95	5.61	4.35
SD ELEA	1.14	0.58	0.92	1.10
SD reference tool	1.41	1.44	1.23	1.61
p value	<0.001	<0.001	0.5588	<0.001
effect size (cohen's d)	1.0654 (large effect)	0.8108 (large effect)	0.1075 (negligible effect)	0.6953 (medium effect)
degrees of freedom	117	117	117	117

Table 4: Overview of results of the perceived constructs of students using ELEA and students using the reference tool.

the difference is trivial, even if it is statistically significant [20]. However, we did not find any difference in the cognitive empathy level between the texts of both groups (mean value ELEA = 3.24, mean value reference tool = 3.25, p = 0.9299). The results indicate that adaptive feedback on students' empathy level helps them to write emotionally more empathic texts. The results show that students' using ELEA wrote texts with a higher level of emotional empathy compared to the ones using the reference tool. However, adaptive cognitive empathy feedback seems to have no significant influence on students' cognitive empathy writing levels compared to non-adaptive cognitive empathy feedback.

5.2 Perceived Empathy Skill Learning of Students

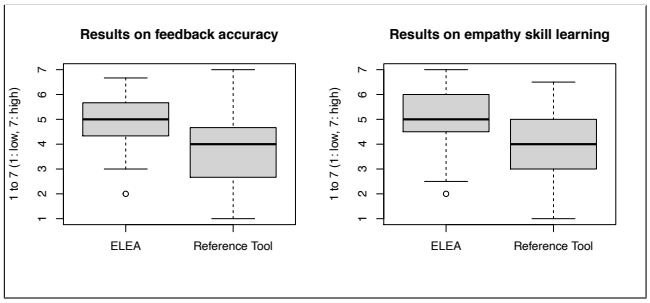

Figure 5: Results of empathy feedback accuracy (left) and empathy skill learning (right) between both tools.

Participants using ELEA judged their empathy skill learning with a mean of 5.03 (SD= 1.05). Participants using the alternative tool judged their empathy skill learning with a mean of 3.93 (SD= 1.50) (see Figure 5 and Table 3). An ANOVA confirmed that the treatment group perceived their empathy skill learning to be significantly higher compared to the control group ($p<0.001$). Moreover, we calculated a *Cohen's d* of 0.8410 indicating a large effect size [20]. This proves our hypothesis that individual feedback on students' empathy levels helps them foster their empathy skills. The results show that students using ELEA judged their empathy skill learning to be significantly higher compared to the ones using the traditional approach. Moreover, we compared the results for perceived feedback accuracy (PFA) between both learning tools. Participants using ELEA rated the PFA with a mean of 4.93 (SD= 0.94), whereas participants from the control group rated the PFA with a mean of 3.69 (SD= 1.36). The difference is statistically significant ($p<0.001$, cohen's d = 1.0470), indicating that the adaptive feedback approach has a significant impact on students' perception of the feedback accuracy compared to the dictionary-based feedback approach.

5.3 Technology Acceptance

For the technology acceptance, we calculated the average of every construct. The answers were provided on a 1- to 7-point Likert scale (1: totally disagree, 7: totally agree). First, we compared the results of ELEA with the results of the alternative tool. The perceived usefulness of ELEA was rated with a mean value of 5.05 (SD= 0.58) and the average perceived level of enjoyment of ELEA was 5.31 (SD= 1.10). The mean value of intention to use of participants using ELEA as a writing support tool was 5.14 (SD= 1.14) (see Table 4). These values are significantly better than the results of the alternative approach. For perceived usefulness, we observed a mean value of 3.77 (SD= 1.41, $p<0.001$), and for a perceived level of enjoyment the value was 4.35 (SD= 1.61, $p<0.001$) for participants from the control group. The mean value for the intention to use was 3.77 (SD= 1.41, $p<0.001$). The results clearly show that the participants of our experiment rated the technology acceptance of ELEA as an adaptive feedback tool positively compared to the usage of the alternative application. Moreover, the mean values of ELEA are also very promising when comparing the results to the midpoints. All results are better than the neutral value of 4. Especially the perceived usefulness for writing better empathetic texts and the intention to use ELEA as a writing support tool in learning scenarios show promising results. Also, the high level of enjoyment for ELEA as a learning tool provides promising results since enjoyment during a learning process has a major influence on the adoption of IT tools [45] and on the learning success of students [59]. A positive technology acceptance is especially important for learning tools to ensure students are perceiving the usage of the tool as helpful, useful, and easy to interact with. This will foster motivation and engagement to use the learning application. The perceived usefulness and intention to use provides promising results to use this tool as a feedback application in different learning settings.

Moreover, we calculate a mean value of 5.72 (SD= 0.92) for the perceived ease of use of ELEA and a mean of 5.61 (SD= 1.23) for participants from the control group. Both values are very high and therefore promising for future usage. In fact, we did not expect a difference in the perceived ease of use between both tools, since the look and feel of the user interaction were purposely designed the same.

5.4 Analysis of Gender Differences

As explained before, past literature has investigated differences in the context of empathy between certain populations. Even though our objective was to design an empathy learning tool for the average user of the particular scenario of peer reviews at our university, we want to investigate if significant differences in the perception and learning through empathy feedback exists. One of the prominent literature streams discussed in past studies is gender differences cornering empathy (e.g., [19, 51, 56]). Since we also captured the gender of students (65 male, 54 female, 0 non-disclosure), we investigated the differences of our measured constructs between males and females for both empathy learning tools.

We found two significant differences between male and female users. Females rated the intention to use of an empathy learning tool significantly higher than males (mean females = 4.53, SD = 0.048; mean males 4.34, SD = 4.34, p = 0.048). Moreover, females rated the perceived usefulness of an empathy learning tool significantly better than male participants across both treatments (mean female = 4.65, SD = 1.03; mean male = 4.29 SD = 1.10, p = 0.01248). For all other perceived constructs, we received p-values > 0.05 between male and female users (perceived ease of use p = 0.2743, level of enjoyment p = 0.8978, perceived empathy skill learning p = 0.116 - female mean = 4.52; SD = 1.13, male mean 4.399, SD = 1.15). For the behavioral variables, we also found no significant effects between female and male participants (emotional empathy p = 0.8296. cognitive empathy p = 0.8253).

5.5 Qualitative User Feedback

We also asked open questions in our survey to receive the participants' opinions about the tool they used. The general attitude for ELEA was very positive. Participants positively mentioned the simple and easy interaction with ELEA, the distinction between cognitive and emotional empathy feedback, and the overall empathy score together with the adaptive feedback message several times. However, participants also said that ELEA should provide even more detailed feedback based on more categories and provide concrete text examples on how to improve their empathy score. We translated the responses from German and clustered the most representative responses in Table 5. To further control for potential problems, harm, or unintended consequences of our empathy feedback and scripts, we separately conducted an analysis of all qualitative user comments. While no particular cluster of answers could be derived, two male users mentioned potential limitations and risks of empathy feedback, such as *"I'm a bit worried that an algorithm will tell me if I'm empathic or not, I don't think real empathy can be put into 0's and 1's ;-)"* and *"Danger with the "machine" that you can seem empathetic just by inserting personal pronouns and some adjectives/adverbs."*

Cluster	Feature
On empathy feedback reaction	*"In my case, I was empathetic on a cognitive level but not on an emotional one. This is also consistent with experiences from my everyday life. I am empathetic but basically more interested in objective-rational solutions. I think that this tool could help me not only to put myself in the position of a person in terms of content and make suggestions but also to communicate to them better"*
On the feedback for skill learning	*"The empathy feedback was clear and could be easily implemented. I had the feeling I learned something."*
On the user interaction	*"The tool was very easy to use and the feedback was helpful! Simple handling."*
On the speed of the tool	*"Clear evaluation and fast feedback. Would use it again!"*
On cognitive and emotional empathy	*"It was helpful that a distinction was made between the two categories of empathy. This again clearly showed me that I do not show emotional empathy enough. It was also useful that the tool said how to show emotional empathy (feelings when reading the business idea etc.)."*
On the percentage score and progress indication	*"I also liked how the tool immediately showed me how my text became more empathetic based on the percentage score."*
Improvements on feedback granularity	*"It would be better if the feedback was more selective or with detailed categories about empathy."*
Improvements on feedback recommendations	*"Even more detailed information on how I can improve my empathy writing would be helpful, e.g., with review examples."*

Table 5: Representative examples of qualitative user responses for ELEA.

6 DISCUSSION

6.1 Theoretical and Practical Contributions

Individual support and feedback for students to learn effectively is still an ongoing challenge (e.g., [104]). Advances in NLP and ML-based algorithms might be able to provide individualization at scale in many distance learning scenarios (such as Massive Open Online Courses) or at larg-scale university lectures, in which individual interaction between students and instructors might be naturally limited. Our research provides insight into the potential of NLP and ML to foster students' learning outcomes in specific pedagogical scenarios independent of an educator, time, and location. By designing a particular adaptive learning system for empathy skills in German peer reviews on business models, we found that students receiving adaptive empathy writing support based on ML and NLP wrote their peer reviews with significantly higher emotional empathy levels compared to the ones receiving dictionary-based feedback. Moreover, the students judged their empathy skill learning to be significantly higher. We believe that self-regulated learning theory is a suitable basis to explain this effect. The tailored feedback on a student's skills, such as empathetic writing skills, seems to foster the *"ability to simply understand the other person's perspective and to react to the observed experiences of another"* ([24], p.1). This is also reflected in the feedback accuracy of our tool, which is perceived as very high. Users judged the adaptive feedback of ELEA as very accurate, which is a necessary pre-condition for them to foster learning according to self-monitoring and self-evaluation [8, 110]. The high degree of enjoyment and perceived usefulness of ELEA as an adaptive empathy learning tool reflects the promise of our adaptive empathy feedback technology. This, combined with the high perceived usefulness, indicates that our learning tool has the potential to be successful in a real-world scenario. Positive technology acceptance is crucial for students to perceive the tool's use as beneficial, practical, and simple to interact with. This will increase motivation, engagement, and long-term use of the learning software.

We hypothesize that positive perceptions of skill development and technology acceptance in a possible continuous use case will lead to user self-efficacy and inspire them to learn and improve their skills [8]. Self-regulated learning theory supports our underlying hypothesis that individual and personal feedback on a student's empathy level motivates the student to improve their skill level [14, 34]. Furthermore, the results of our analysis of gender differences partly provide support for what has been observed in prior literature. Females participants showed increased perceptions of usefulness as well as intentions to use such a tool in the future. This mirrors the findings that female users value empathy or empathy skills more than male users [19]. However, we could not observe performance differences comparing female and male participants. The reasons for these observations could be manifold, making it an interesting area of future research.

Therefore, our work makes several contributions to current research in human-computer interaction and computer-supported collaborative learning. To the best of our knowledge, this study is one of the first to present evaluated design knowledge on how to build a learning tool to train cognitive and emotional empathy skills based on adaptive and intelligent feedback [74]. Past research has mostly built empathy learning tools based on syntactical analysis (e.g., [77]) or leveraged emotional modeling in other contexts than education (e.g., [60, 87]). Our research provides a basis for researchers who also aim to develop learning tools to train metacognition skills to compare their solutions with ours. Educators can now use our design evaluation to create their own learning tools for providing adaptive and intelligent support of empathy skills in their large-scale or distance-learning scenarios. We believe, our findings could not only be applied in peer review scenarios, but also in collaborative teamwork, online discussions, or other collaborative learning settings where empathy might be important.

6.2 Limitations and Future Work

Nevertheless, our work faces several limitations. First, the design findings and the empirical evaluation of the impact of our adaptive empathy learning tool are limited to the very specific pedagogical scenario of student peer reviews on business models in German language. As stated, the entire design, such as the formulation of scripts as well as the predictive model is biased from a Western European perspective. The predictions might be confusing or even harmful for other contexts such as other user groups, pedagogical concepts, or cultures. More research is needed to investigate the design, the empathy nuances and replicate the empirical findings on the effect of these educational systems' in other pedagogical domains, and in different languages (for example, English, French,

or Chinese). Moreover, future work should especially investigate different user groups and how different perceptions of empathy affect their learning experience, such as individuals from the autism spectrum. Although we tried to pay special attention to potential harm or unintended consequence in the application of our tool (e.g., by analyzing the qualitative user comments), further research is needed, to investigate how "average" learning tools influence general empathy nuances of people, for example, at a certain workplace, gender or minorities. Moreover, the effect on more nuanced constructs and dimensions of students' perception of an AI-based learning tool during the learning process, such as trust and reliance, as well as the implications of erroneous learning advice [79], should be regarded in future research. Concerning the transferability of our model, corpora and annotation schemes are needed in other languages and domains to transfer and test our findings in different languages. The study of [98] might help scholars as a start to investigate how to model cognitive and emotional empathy skills in different domains and languages.

Second, our research is limited when it comes to the long-term impact on the empathy abilities of students. In our experiment, we investigated the short-term influence of ELEA on the perceived skill learning and the emotional empathy skills of students. In future work, longitudinal studies will be critical in determining the long-term influence of empathy learning on learning outcomes.

Third, another limitation of our results evolves from the measurements of the empathy learning outcomes. Since metacognition skills are complex pedagogical constructs, we relied on a proven method to measure metacognition skills in textual data (e.g., done for argumentation skills in [2, 96, 101] or for problem-solving skills in [103]). The expert assessment of metacognition skills has been widely used, however, is limited when it comes to capturing more sophisticated dimensions of metacognition skill learning in different domains. More refined empathy skill assessments are required to more precisely examine the influence of individual empathy feedback and self-evaluation on students' skill levels. These assessments should be able to capture students' empathic writing competencies in various pedagogical domains (e.g., business, ethics, and education) before and after a specific treatment on various empathy skill dimensions to represent students' empathy skills with a higher level of granularity. Future research is needed to build, validate, and assess such test measurements in order to more precisely measure the impact of various pedagogical learning method designs on metacognition skill learning outcomes.

Fourth, we did not investigate any significant effect of adaptive empathy feedback on students' cognitive empathy writing level in peer reviews. A possible reason for this might lie in the complex structures of perspective talking. While it could be comparably easy to support students in writing emotionally more empathic feedback, becoming better in cognitive empathy might require more training with stricter guidance. A possible human-computer interaction could be to increase social presence, e.g, by embedding a conversational agent as an empathy learning tool. Past research (e.g., [100, 109] has shown that conversational agents in pedagogical scenarios are able to increase social prescience. Future research could investigate this potential by embedding adaptive empathy writing support in a conversational learning tool and comparing it against our non-conversational approach.

Finally, although all interviews of the design requirement collection process were recorded, transcribed and abstract categories were formatted in the form of user stores, one limitation arises by the fact that this process has only been conducted by one researcher. Nevertheless, the derived requirements in the form of user stories based on the transcriptions have been discussed with two senior researchers in three workshops to ensure the validity of the design findings. The main improvement suggestion from users in the qualitative feedback was that the feedback of ELEA could encompass even more empathy dimensions and concrete examples on how to improve empathetic writing in a certain domain. We call for future work to enrich current corpora on student-written texts for more precise empathy feedback. Moreover, we aim to embed more scripted feedback recommendations with more accurate and transparent action steps on how to achieve a higher rating and examples on how to improve empathy writing in a certain domain. Besides, we want to ensure that the empathy feedback stays transparent and understandable for the users. Therefore, we will guarantee that empathy theory is even better explained with multimedia elements such as illustrations and videos.

7 CONCLUSION

In this research project, we designed, built, and evaluated ELEA, an adaptive IT tool that provides students with feedback on the cognitive and emotional empathy level of a text by leveraging the recent advances of NLP and ML algorithms. We compared ELEA to a proven empathy writing support approach in a rigorous use study with 119 participants. We found students using ELEA wrote more empathic texts with a higher level of emotional empathy compared to the ones using the alternative approach. The high perceived empathy skill learning, the technology acceptance, and the level of enjoyment for ELEA provide promising results to use this tool as a feedback application in traditional learning settings. Our results also offer design suggestions to further improve educational feedback applications based on intelligent algorithms. With NLP and ML becoming more powerful, we hope our work will attract other researchers to design and build more intelligent tutoring systems for other learning scenarios or metacognition skills.

ACKNOWLEDGMENTS

We thank the Swiss National Science Foundation for supporting this research collaboration (grant 200207). We further thank the German Federal Ministry of Education and Research for supporting this research through the project Komp-HI (grant 16DHBKI073). Last but not least, we thank Corinne Ruckstuhl for supporting the tool development with her thesis.

REFERENCES

[1] Ahmed Abbasi, Hsinchun Chen, and Arab Salem. 2008. Sentiment analysis in multiple languages: Feature selection for opinion classification in Web forums. *ACM Transactions on Information Systems* 26, 3 (6 2008), 1–34. https://doi.org/10.1145/1361684.1361685

[2] Tazin Afrin, Omid Kashefi, Christopher Olshefski, Diane Litman, Rebecca Hwa, and Amanda Godley. 2021. Effective Interfaces for Student-Driven Revision Sessions for Argumentative Writing. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–13. https://doi.org/10.1145/3411764.3445683

[3] Ritu Agarwal and Elena Karahanna. 2000. Time Flies When You're Having Fun: Cognitive Absorption and Beliefs about Information Technology Usage. *MIS*

Quarterly 24, 4 (12 2000), 665. https://doi.org/10.2307/3250951

[4] Firoj Alam, Morena Danieli, and Giuseppe Riccardi. 2018. Annotating and modeling empathy in spoken conversations. *Computer Speech and Language* 50 (2018), 40–61. https://doi.org/10.1016/j.csl.2017.12.003

[5] S. J. Ashford. 1986. Feedback-Seeking in Individual Adaptation : A Resource Perspective. *Academy of Management Journal* 29, 3 (9 1986), 465–487. https://doi.org/10.2307/256219

[6] Jeremy N. Bailenson, Nick Yee, Jim Blascovich, Andrew C. Beall, Nicole Lundblad, and Michael Jin. 2008. The use of immersive virtual reality in the learning sciences: Digital transformations of teachers, students, and social context. , 102–141 pages. https://doi.org/10.1080/10508400701793141

[7] Albert Bandura. 1977. Self-efficacy: Toward a unifying theory of behavioral change. *Psychological Review* 84, 2 (3 1977), 191–215. https://doi.org/10.1037/0033-295X.84.2.191

[8] Albert Bandura. 1991. Social cognitive theory of self-regulation. *Organizational Behavior and Human Decision Processes* 50, 2 (12 1991), 248–287. https://doi.org/10.1016/0749-5978(91)90022-L

[9] Simon Baron-Cohen and Sally Wheelwright. 2004. *The Empathy Quotient: An Investigation of Adults with Asperger Syndrome or High Functioning Autism, and Normal Sex Differences.* Technical Report 2.

[10] C. Daniel Batson, Jim Fultz, and Patricia A. Schoenrade. 1987. Distress and Empathy: Two Qualitatively Distinct Vicarious Emotions with Different Motivational Consequences. *Journal of Personality* 55, 1 (3 1987), 19–39. https://doi.org/10.1111/j.1467-6494.1987.tb00426.x

[11] Hope Bell. 2018. Creative Interventions for Teaching Empathy in the Counseling Classroom. *Journal of Creativity in Mental Health* 13, 1 (1 2018), 106–120. https://doi.org/10.1080/15401383.2017.1328295

[12] Raquel Benbunan-Fich. 2017. Is Self-Reported Learning a Proxy Metric for Learning? Perspectives From the Information Systems Literature. *https://doi.org/10.5465/amle.9.2.zqr321* 9, 2 (11 2017), 321–328. https://doi.org/10.5465/AMLE.9.2.ZQR321

[13] Eric Bergemann. 2009. *Exploring psychotherapist empathic attunement from a psychoneurobiological perspective: Is empathy enhanced by yoga and meditation?* . Ph.D. Dissertation.

[14] Robert A. Bjork, John Dunlosky, and Nate Kornell. 2013. Self-regulated learning: Beliefs, techniques, and illusions. *Annual Review of Psychology* 64, November 2012 (2013), 417–444. https://doi.org/10.1146/annurev-psych-113011-143823

[15] Paul Black and Dylan Wiliam. 2009. Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability* 21, 1 (2009), 5–31. https://doi.org/10.1007/s11092-008-9068-5

[16] Marcela Borge, Yann Shiou Ong, and Carolyn Penstein Rosé. 2018. Learning to monitor and regulate collective thinking processes. *International Journal of Computer-Supported Collaborative Learning* 13, 1 (3 2018), 61–92. https://doi.org/10.1007/s11412-018-9270-5

[17] Sven Buechel, Anneke Buffone, Barry Slaff, Lyle Ungar, and João Sedoc. 2018. Modeling empathy and distress in reaction to news stories. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018* (2018), 4758–4765. https://doi.org/10.18653/v1/d18-1507

[18] Sven Buechel and Udo Hahn. 2018. Emotion Representation Mapping for Automatic Lexicon Construction (Mostly) Performs on Human Level. (2018), 2892–2904. http://www.julielab.dehttp://arxiv.org/abs/1806.08890

[19] Leonardo Christov-Moore, Elizabeth A. Simpson, Gino Coudé, Kristina Grigaityte, Marco Iacoboni, and Pier Francesco Ferrari. 2014. Empathy: Gender effects in brain and behavior. *Neuroscience and biobehavioral reviews* 46, Pt 4 (10 2014), 604. https://doi.org/10.1016/J.NEUBIOREV.2014.09.001

[20] Jacob Cohen. 1988. *Statistical Power Analysis for the Behavioral Sciences.* https://doi.org/10.4324/9780203771587

[21] Mike Cohn. 2004. *User Stories Applied For Agile Software Development.* Technical Report.

[22] Harris M. Cooper. 1988. Organizing knowledge syntheses: A taxonomy of literature reviews. *Knowledge in Society* 1, 1 (1988), 104–126. https://doi.org/10.1007/BF03177550

[23] Benjamin M.P. Cuff, Sarah J. Brown, Laura Taylor, and Douglas J. Howat. 2016. Empathy: A review of the concept. , 144–153 pages. https://doi.org/10.1177/1754073914558466

[24] Mark H. Davis. 1983. Measuring individual differences in empathy: Evidence for a multidimensional approach. *Journal of Personality and Social Psychology* 44, 1 (1983), 113–126. https://doi.org/10.1037//0022-3514.44.1.113

[25] Frederique de Vignemont and Tania Singer. 2006. The empathic brain: how, when and why? *Trends in Cognitive Sciences* 10, 10 (10 2006), 435–441. https://doi.org/10.1016/j.tics.2006.08.008

[26] Jean Decety and Philip L. Jackson. 2004. The functional architecture of human empathy. , 71–100 pages. https://doi.org/10.1177/1534582304267187

[27] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. (2018). http://arxiv.org/abs/1810.04805

[28] Pierre Dillenbourg, Sanna Järvelä, and Frank Fischer. 2009. The Evolution of Research on Computer-Supported Collaborative Learning. In *Technology-Enhanced Learning: Principles and Products*, Nicolas Balacheff, Sten Ludvigsen, Ton de Jong, Ard Lazonder, and Sally Barnes (Eds.). Springer Netherlands, Dordrecht, 3–19. https://doi.org/10.1007/978-1-4020-9827-7_1

[29] Nancy Eisenberg. 2000. Emotion, regulation, and moral development. *Annual review of psychology* 51, 1 (2000), 665–697.

[30] Karen E. Gerdes, Elizabeth A. Segal, Kelly F. Jackson, and Jennifer L. Mullins. 2011. Teaching empathy: A framework rooted in social cognitive neuroscience and social justice. *Journal of Social Work Education* 47, 1 (12 2011), 109–131. https://doi.org/10.5175/JSWE.2011.200900085

[31] Jochen. Glaeser and Grit. Laudel. 2010. *Experteninterviews und qualitative Inhaltsanalyse : als Instrumente rekonstruierender Untersuchungen.* VS Verlag fuer Sozialwiss. http://www.springer.com/de/book/9783531172385

[32] Alvin I Goldman. 1993. Ethics and Cognitive Science. *Ethics* 103, 2 (1993), 337–360. http://www.jstor.org/stable/2381527

[33] Saurabh Gupta and Robert P. Bostrom. 2009. Technology-Mediated learning: A comprehensive theoretical model. *Journal of the Association for Information Systems* 10, 9 (2009), 686–714. https://doi.org/10.17705/1jais.00207

[34] John Hattie and Helen Timperley. 2007. The power of feedback. , 81–112 pages. https://doi.org/10.3102/003465430298487

[35] Meirav Hen and Marina Goroshit. 2011. Emotional competencies in the education of mental health professionals. *Social Work Education* 30, 7 (10 2011), 811–829. https://doi.org/10.1080/02615479.2010.515680

[36] Robert Hogan. 1969. Development of an empathy scale. *Journal of Consulting and Clinical Psychology* 33, 3 (6 1969), 307–316. https://doi.org/10.1037/H0027580

[37] Ann Jones and Kim Issroff. [n.d.]. Learning technologies: Affective and social issues in computer-supported collaborative learning. ([n. d.]). https://doi.org/10.1016/j.compedu.2004.04.004

[38] Hamed Khanpour, Cornelia Caragea, and Prakhar Biyani. 2017. *Identifying Empathetic Messages in Online Health Communities.* Technical Report. 246–251 pages. https://csn.cancer.org

[39] Soomin Kim, Joonhwan Lee, and Gahgene Gweon. 2019. Comparing data from chatbot and web surveys effects of platform and conversational style on survey response quality. *Conference on Human Factors in Computing Systems - Proceedings* (2019), 1–12. https://doi.org/10.1145/3290605.3300316

[40] Sara H. Konrath, Edward H. O'Brien, and Courtney Hsing. 2011. Changes in dispositional empathy in American college students over time: A meta-analysis. *Personality and Social Psychology Review* 15, 2 (2011), 180–198. https://doi.org/10.1177/1088868310377395

[41] Timothy Koschmann. 1996. *Paradigm Shifts and Instructional Technology.* Technical Report. 1–23 pages. http://opensiuc.lib.siu.edu/meded_books/4

[42] Dana Kralicek, Lisa Von Rabenau, Swati Shelar, and Paulo Blikstein. 2018. Inside out:Teaching empathy and social-emotional skills. In *IDC 2018 - Proceedings of the 2018 ACM Conference on Interaction Design and Children.* Association for Computing Machinery, Inc, New York, NY, USA, 525–528. https://doi.org/10.1145/3202185.3213525

[43] Severin Landolt, Thiemo Wambsganß, and Matthias Söllner. 2021. A taxonomy for deep learning in natural language processing. In *Proceedings of the Annual Hawaii International Conference on System Sciences.* https://doi.org/10.24251/hicss.2021.129

[44] E. J. Lawrence, P. Shaw, D. Baker, S. Baron-Cohen, and Anthony S. David. 2004. Measuring empathy: Reliability and validity of the Empathy Quotient. *Psychological Medicine* 34, 5 (7 2004), 911–919. https://doi.org/10.1017/S0033291703001624

[45] Matthew K O Lee, Christy M K Cheung, and Zhaohui Chen. 2005. Acceptance of Internet-based learning medium: The role of extrinsic and intrinsic motivation. *Information and Management* 42, 8 (2005), 1095–1104. https://doi.org/10.1016/j.im.2003.10.007

[46] Bing Liu. 2015. *Sentiment analysis: Mining opinions, sentiments, and emotions.* Cambridge University Press. 1–367 pages. https://doi.org/10.1017/CBO9781139084789

[47] Benjamin Lok and Adriana E. Foster. 2019. Can Virtual Humans Teach Empathy? In *Teaching Empathy in Healthcare.* Springer International Publishing, 143–163. https://doi.org/10.1007/978-3-030-29876-0_9

[48] Joseph Luca and Pina Tarricone. 2001. Does Emotional Intelligence Affect Successful Teamwork? *Proceedings of the 18th Annual Conference of the Australasian Society for Computers in Learning in Tertiary Education* December 2001 (2001), 367–376.

[49] Scott W. McQuiggan and James C. Lester. 2007. Modeling and evaluating empathy in embodied companion agents. *International Journal of Human Computer Studies* 65, 4 (4 2007), 348–360. https://doi.org/10.1016/j.ijhcs.2006.11.015

[50] Jack Mezirow. 1991. *Transformative dimensions of adult learning.* Jossey-Bass, San Francisco, CA 94104-1310.

[51] Kalina J. Michalska, Katherine D. Kinzler, and Jean Decety. 2013. Age-related sex differences in explicit measures of empathy do not predict brain responses across childhood and adolescence. *Developmental Cognitive Neuroscience* 3, 1 (1 2013), 22–32. https://doi.org/10.1016/J.DCN.2012.08.001

[52] Mario Mikulincer and Phillip R. Shaver. 2001. Attachment theory and intergroup bias: Evidence that priming the secure base schema attenuates negative reactions to out-groups. *Journal of Personality and Social Psychology* 81, 1 (2001), 97–115.

https://doi.org/10.1037/0022-3514.81.1.97

[53] Mario Mikulincer and Phillip R. Shaver. 2007. Boosting attachment security to promote mental health, prosocial values, and inter-group tolerance. *Psychological Inquiry* 18, 3 (2007), 139–156. https://doi.org/10.1080/10478400701512646

[54] Aurelien Miralles, Marine Grandgeorge, and Michel Raymond. 2021. Exploring neurodiversity through biodiversity: Empathy of people with autism towards living beings mostly differs for a single species, ours. *PsyArXiv* (2021). https://doi.org/10.31234/OSF.IO/49QXJ

[55] David L. Neumann, Raymond C.K. Chan, Gregory J. Boyle, Yi Wang, and H. Rae Westbury. 2015. Measures of Empathy: Self-Report, Behavioral, and Neuroscientific Approaches. *Measures of Personality and Social Psychological Constructs* (1 2015), 257–289. https://doi.org/10.1016/B978-0-12-386915-9.00010-3

[56] Ed O'Brien, Sara H. Konrath, Daniel Grühn, and Anna Linda Hagen. 2013. Empathic Concern and Perspective Taking: Linear and Quadratic Effects of Age Across the Adult Life Span. *The Journals of Gerontology: Series B* 68, 2 (3 2013), 168–175. https://doi.org/10.1093/GERONB/GBS055

[57] OECD. 2018. The Future of Education and Skills - Education 2030. https://doi.org/2018-06-15

[58] Alissa D. Parr, Samuel T. Hunter, and Gina Scott Ligon. 2013. Questioning universal applicability of transformational leadership: Examining employees with autism spectrum disorder. *The Leadership Quarterly* 24, 4 (8 2013), 608–622. https://doi.org/10.1016/J.LEAQUA.2013.04.003

[59] Reinhard Pekrun and Elizabeth J. Stephens. 2012. Academic emotions. *APA educational psychology handbook, Vol 2: Individual differences and cultural and contextual factors.* (10 2012), 3–31. https://doi.org/10.1037/13274-001

[60] Zhenhui Peng, Qingyu Guo, Ka Wing Tsang, and Xiaojuan Ma. 2020. Exploring the Effects of Technological Writing Assistance for Support Providers in Online Mental Health Community. *Conference on Human Factors in Computing Systems - Proceedings* (4 2020). https://doi.org/10.1145/3313831.3376695

[61] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global vectors for word representation. In *EMNLP 2014 - 2014 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference.* Association for Computational Linguistics (ACL), 1532–1543. https://doi.org/10.3115/v1/d14-1162

[62] Verónica Pérez-Rosas, Rada Mihalcea, Kenneth Resnicow, Satinder Singh, and Lawrence An. 2017. Understanding and predicting empathic behavior in counseling therapy. *ACL 2017 - 55th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)* 1 (2017), 1426–1435. https://doi.org/10.18653/v1/P17-1131

[63] Robin T. Peterson and Yam Limbu. 2009. The convergence of mirroring and empathy: Communications training in business-to-business personal selling persuasion efforts. *Journal of Business-to-Business Marketing* 16, 3 (7 2009), 193–219. https://doi.org/10.1080/10517120802484551

[64] Jean Piaget, Terrance Brown, and Kishore Julian Thampy. 1986. The Equilibration of Cognitive Structures: The Central Problem of Intellectual Development. Jean Piaget , Terrance Brown , Kishore Julian Thampy. *American Journal of Education* 94, 4 (8 1986), 574–577. https://doi.org/10.1086/443876

[65] Philip M. Podsakoff and Jiing Lih Farh. 1989. Effects of feedback sign and credibility on goal setting and task performance. *Organizational Behavior and Human Decision Processes* 44, 1 (8 1989), 45–67. https://doi.org/10.1016/0749-5978(89)90034-4

[66] Claire Polo, Kristine Lund, Christian Plantin, and Gerald P. Niccolai. 2016. Group emotions: the social and cognitive functions of emotions in argumentation. *International Journal of Computer-Supported Collaborative Learning* 11, 2 (2016), 123–156. https://doi.org/10.1007/s11412-016-9232-8

[67] Mathis Poser and Eva A. C. Bittner. 2020. Hybrid Teamwork: Consideration of Teamwork Concepts to Reach Naturalistic Interaction between Humans and Conversational Agents. In *WI2020*. GITO Verlag. https://doi.org/10.30844/wi_2020_a6-poser

[68] Corinne Reid, Helen Davis, Chiara Horlin, Mike Anderson, Natalie Baughman, and Catherine Campbell. 2013. The Kids' Empathic Development Scale (KEDS): a multi-dimensional measure of empathy in primary school-aged children. *The British journal of developmental psychology* 31, Pt 2 (6 2013), 231–256. https://doi.org/10.1111/BJDP.12002

[69] Rachel Carlos Duque Reis, Seiji Isotani, Carla Lopes Rodriguez, Kamila Takayama Lyra, Patrícia Augustin Jaques, and Ig Ibert Bittencourt. 2018. Affective states in computer-supported collaborative learning: Studying the past to drive the future. *Computers and Education* 120 (2018), 29–50. https://doi.org/10.1016/j.compedu.2018.01.015

[70] Eric Ries. 2011. The Lean Startup: How Today's Entrepreneurs Use Continuous Innovation to Create Radically Successful Businesses.

[71] Roman Rietsche and Matthias Söllner. 2019. Insights into Using IT-Based Peer Feedback to Practice the Students Providing Feedback Skill. *Proceedings of the 52nd Hawaii International Conference on System Sciences* (2019). https://doi.org/10.24251/hicss.2019.009

[72] Saman Rizvi, Bart Rienties, and Jekaterina Rogaten. 2018. Temporal dynamics of MOOC learning trajectories. In *ACM International Conference Proceeding Series*. Association for Computing Machinery, New York, NY, USA, 1–6. https://doi.org/10.1145/3279996.3280035

[73] Henry L Roediger and Jeffrey D Karpicke. 2006. *Test-Enhanced Learning Taking Memory Tests Improves Long-Term Retention.* Technical Report. http://learninglab.psych.purdue.edu/downloads/2006_Roediger_Karpicke_PsychSci.pdf

[74] Carolyn Rosé, Yi Chia Wang, Yue Cui, Jaime Arguello, Karsten Stegmann, Armin Weinberger, and Frank Fischer. 2008. Analyzing collaborative learning processes automatically: Exploiting the advances of computational linguistics in computer-supported collaborative learning. *International Journal of Computer-Supported Collaborative Learning* 3, 3 (2008), 237–271. https://doi.org/10.1007/s11412-007-9034-0

[75] Sara Rosenthal, Noura Farra, and Preslav Nakov. 2018. SemEval-2017 Task 4: Sentiment Analysis in Twitter. (2018), 502–518. https://doi.org/10.18653/v1/s17-2088

[76] D Royce Sadler. 1989. Formative assessment and the design of instructional systems. *Instructional Science* 18, 2 (1989), 119–144. https://doi.org/10.1007/BF00117714

[77] Breno Santana Santos, Methanias Colaqo Junior, and Janisson Gois De Souza. 2018. An Experimental Evaluation of the NeuroMessenger: A Collaborative Tool to Improve the Empathy of Text Interactions. *Proceedings - IEEE Symposium on Computers and Communications* 2018-June (2018), 573–579. https://doi.org/10.1109/ISCC.2018.8538442

[78] Katharina Scheiter and Peter Gerjets. 2007. Learner control in hypermedia environments. *Educational Psychology Review* 19, 3 (2007), 285–307. https://doi.org/10.1007/s10648-007-9046-3

[79] Anuschka Schmitt, Thiemo Wambsganss, Andreas Janson, and Matthias Söllner. 2021. Towards a Trust Reliance Paradox ? Exploring the Gap Between Perceived Trust in and Reliance on Algorithmic Advice. In *Forty-Second International Conference on Information Systems*. Austin, Texas, 1–17.

[80] Julia E. Seaman, I. E. Allen, and Jeff Seaman. 2018. *Higher Education Reports - Babson Survey Research Group.* Technical Report. http://www.onlinelearningsurvey.com/highered.htmlhttps://www.onlinelearningsurvey.com/highered.html

[81] Ashish Sharma, Adam S Miner, David C Atkins, and Tim Althoff. 2020. A Computational Approach to Understanding Empathy Expressed in Text-Based Mental Health Support. (2020), 5263–5276. http://arxiv.org/abs/2009.08441

[82] John Sinclair. 2005. Developing Linguistic Corpora: a Guide to Good Practice. (2005). http://www.ahds.ac.uk/creating/guides/linguistic-corpora/index.htm

[83] Traci Sitzmann, Katherine Ely, Kenneth G. Brown, and Kristina N. Bauer. 2017. Self-Assessment of Knowledge: A Cognitive Learning or Affective Measure? *https://doi.org/10.5465/amle.9.2.zqr169* 9, 2 (11 2017), 169–191. https://doi.org/10.5465/AMLE.9.2.ZQR169

[84] Matthias Söllner, Philipp Bitzer, Andreas Janson, and Jan Marco Leimeister. 2018. Process is King: Evaluating the Performance of Technology-mediated Learning in Vocational Software Training. *Journal of Information Technology* 33, 3 (9 2018), 233–253. https://doi.org/10.1057/s41265-017-0046-6

[85] Elliot Soloway, Mark Guzdial, and Kenneth E Hay. 1994. Learner-Centered Design The Challenge For WC1 In The Xst Century. *Interactions* (1994), 36–48.

[86] R. Nathan Spreng, Margaret C. McKinnon, Raymond A. Mar, and Brian Levine. 2009. The Toronto empathy questionnaire: Scale development and initial validation of a factor-analytic solution to multiple empathy measures. *Journal of Personality Assessment* 91, 1 (2009), 62–71. https://doi.org/10.1080/00223890802484381

[87] Michael J. Tanana, Christina S. Soma, Vivek Srikumar, David C. Atkins, and Zac E. Imel. 2019. Development and Evaluation of ClientBot: Patient-Like Conversational Agent to Train Basic Counseling Skills. *Journal of medical Internet research* 21, 7 (7 2019). https://doi.org/10.2196/12529

[88] Edward Titchener. 1909. *Lectures on the experimental psychology of the thought-processes.* Macmillan, New York.

[89] Miranda A.L. Van Tilburg, Marielle L. Unterberg, and Ad J.J.M. Vingerhoets. 2002. Crying during adolescence: The role of gender, menarche, and empathy. *British Journal of Developmental Psychology* 20, 1 (3 2002), 77–87. https://doi.org/10.1348/026151002166334

[90] Viswanath Venkatesh and Hillol Bala. 2008. Technology acceptance model 3 and a research agenda on interventions. *Decision Sciences* 39, 2 (5 2008), 273–315. https://doi.org/10.1111/j.1540-5915.2008.00192.x

[91] Viswanath Venkatesh, Michael G Morris, Gordon B Davis, and Fred D Davis. 2003. User Acceptance of Information Technology: Toward a Unified View. *MIS Quarterly* 27, 3 (2003), 425–478.

[92] Jan vom Brocke, Wolfgang Maaß, Peter Buxmann, Alexander Maedche, Jan Marco Leimeister, and Günter Pecht. 2018. Future Work and Enterprise Systems. *Business and Information Systems Engineering* 60, 4 (2018), 357–366. https://doi.org/10.1007/s12599-018-0544-2

[93] Jan vom Brocke, Alexander Simons, Kai Riemer, Bjoern Niehaves, Ralf Plattfaut, and Anne Cleven. 2015. Standing on the shoulders of giants: Challenges and recommendations of literature search in information systems research. *Communications of the Association for Information Systems* 37, 1 (8 2015), 205–224. https://doi.org/10.17705/1cais.03709

[94] Lev Semenovich Vygotsky. 1980. *Mind in society: The development of higher psychological processes*. Harvard university press.

[95] Thiemo Wambsganss, Tobias Kueng, Matthias Soellner, and Jan Marco Leimeister. 2021. ArgueTutor: An Adaptive Dialog-Based Learning System for Argumentation Skills. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3411764.3445781

[96] Thiemo Wambsganss, Christina Niklaus, Matthias Cetto, Matthias Söllner, Jan Marco Leimeister, and Siegfried Handschuh. 2020. AL : An Adaptive Learning Support System for Argumentation Skills. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–14.

[97] Thiemo Wambsganss, Christina Niklaus, Matthias Söllner, Siegfried Handschuh, and Jan Marco Leimeister. 2020. A Corpus for Argumentative Writing Support in German. In *28th International Conference on Computational Linguistics (Coling)*. Barcelona, Spain. https://doi.org/10.18653/v1/2020.coling-main.74

[98] Thiemo Wambsganss, Christina Niklaus, Matthias Söllner, Siegfried Handschuh, and Jan Marco Leimeister. 2021. Supporting Cognitive and Emotional Empathic Writing of Students. In *59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*. 4063–4077. https://doi.org/10.18653/v1/2021.acl-long.314

[99] Thiemo Wambsganss, Florian Weber, and Matthias Söllner. 2021. Design and Evaluation of an Adaptive Empathy Learning Tool. In *Hawaii International Conference on System Sciences*.

[100] Florian Weber, Thiemo Wambsganss, Dominic Rüttimann, and Matthias Söllner. 2021. Pedagogical Agents for Interactive Lernaing : A Taxonomy of Conversational Agents in Education. In *Forty-Second International Conference on Information Systems*. Austin, Texas, 1–17.

[101] Armin Weinberger and Frank Fischer. 2006. A framework to analyze argumentative knowledge construction in computer-supported collaborative learning. *Computers and Education* 46, 1 (2006), 71–95. https://doi.org/10.1016/j.compedu.2005.04.003

[102] Susanne Weis and Heinz-Martin Süß. 2005. Social Intelligence—A Review and CriticalDiscussion of Measurement Concepts. In *Emotional intelligence.An international handbook*, Ralf Schulze and Richard D Roberts (Eds.). Hogrefe, 203–230.

[103] Rainer Winkler, Sebastian Hobert, Antti Salovaara, Matthias Söllner, and Jan Marco Leimeister. 2020. Sara, the Lecturer: Improving Learning in Online Education with a Scaffolding-Based Conversational Agent. In *Conference on Human Factors in Computing Systems - Proceedings*. https://doi.org/10.1145/3313831.3376781

[104] Rainer Winkler, Matthias Söllner, and Jan Marco Leimeister. 2021. Enhancing problem-solving skills with smart personal assistant technology. *Computers and Education* 165 (2021). https://doi.org/10.1016/j.compedu.2021.104148

[105] Lauren Wispé. 1987. History of the concept of empathy. In *Empathy and its development*. Cambridge University Press, New York, NY, US, 17–37.

[106] Bo Xiao, Dogan Can, Panayiotis G Georgiou, David Atkins, and Shrikanth S Narayanan. 2012. Analyzing the Language of Therapist Empathy in Motivational Interview based Psychotherapy. *Signal and Information Processing Association Annual Summit and Conference (APSIPA), ... Asia-Pacific. Asia-Pacific Signal and Information Processing Association Annual Summit and Conference* 2012 (2012). http://www.ncbi.nlm.nih.gov/pubmed/27602411http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5010859

[107] Parvaneh Yaghoubi Jami, Behzad Mansouri, Stephen J. Thoma, and Hyemin Han. 2018. An investigation of the divergences and convergences of trait empathy across two cultures. *Journal of Moral Education* 48, 2 (4 2018), 214–229. https://doi.org/10.1080/03057240.2018.1482531

[108] Qing Zhao, David L. Neumann, Yuan Cao, Simon Baron-Cohen, Chao Yan, Raymond C.K. Chan, and David H.K. Shum. 2019. Culture-sex interaction and the self-report empathy in Australians and mainland Chinese. *Frontiers in Psychology* 10, MAR (2019), 396. https://doi.org/10.3389/FPSYG.2019.00396/BIBTEX

[109] N. Zierau, T Wambsganss, Andreas Janson, Sofia Schöbel, and Jan Marco Leimeister. 2020. The Anatomy of User Experience with Conversational Agents : A Taxonomy and Propositions of Service Clues. In *International Conference on Information Systems (ICIS)*. 1–17.

[110] Barry J Zimmerman and Dale H Schunk. 2001. *Self-regulated learning and academic achievement: Theoretical perspectives*. Routledge.