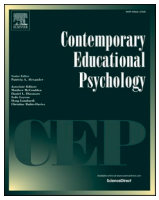
Contemporary Educational Psychology 81 (2025) 102373

Contents lists available at ScienceDirect

Contemporary Educational Psychology

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



Chatbots in education: Outperforming students but perceived as less trustworthy

Martin Laun [a,*], Leonard Puderbach [a,b], Katharina Hirt [c], Eva L. Wyss [c], Daniel Friemert [d], Ulrich Hartmann [d], Fabian Wolff [a]

[a] Department of Psychology, Bielefeld University, Universitätsstraße 25, Bielefeld 33615, Germany
[b] Department of Psychology, University of Koblenz, Universitätsstraße 1, Koblenz 56070, Germany
[c] Department of German Studies, University of Koblenz, Universitätsstraße 1, Koblenz 56070, Germany
[d] Department of Mathematics and Technology, Koblenz University of Applied Sciences, Joseph-Rovan-Allee 2, Remagen 53424, Germany

ARTICLE INFO

Keywords:
Chatbots
Education
Trustworthiness

ABSTRACT

In recent years, chatbots have emerged as potential individual assistance systems in educational settings. However, their successful implementation depends on their ability to deliver high-quality outcomes and gain students' trust. Against this background, this study investigated the perceived trustworthiness of chatbots in nursing education and whether this trust was justified in light of chatbots' actual performance in care plan creation. We conducted a study with 189 vocational nursing students who created medical care plans under two conditions: without ChatGPT (GPT-4) and with ChatGPT assistance. Additionally, ChatGPT was used to develop care plans without student involvement. Expert evaluations of these care plans allowed us to compare quality across conditions. We also examined students' trust by having them evaluate another care plan before and after experimentally manipulating information about its source (chatbot or peer-generated). Statistical analyses revealed that source disclosure significantly affected students' trust: Care plans believed to be chatbot-generated experienced a significant decrease in trust, while those attributed to peers showed no significant change. However, analyses of expert evaluations showed that ChatGPT-generated plans were of higher quality than those created by students, even when students used ChatGPT assistance. This discrepancy between perceived trustworthiness and actual performance of chatbots indicates that students' skepticism toward chatbot-generated content is, to some extent, exaggerated. While ChatGPT can enhance students' care plan quality, our findings emphasize the importance of addressing trust issues in educational settings where chatbots are implemented.

1. Introduction

The rapid advancement of artificial intelligence (AI) has profoundly transformed various sectors, including education (Hwang et al., 2020; Zhang & Aslan, 2021). One of the most notable developments in this domain is ChatGPT, an AI-driven language model capable of generating human-like responses in natural language interactions (Kasneci et al., 2023; OpenAI, 2023). While tools like ChatGPT offer promising capabilities for supporting complex educational tasks (Kabudi et al., 2021), their successful integration depends not only on technological capabilities but also on students' trust in these systems (Nazaretsky et al., 2025; Shahzad et al., 2024). This trust dimension is particularly critical in healthcare education, where errors can have far-reaching consequences, and concerns about AI reliability in high-stakes contexts remain prominent (Goh et al., 2024).

A key example of these responsibilities in nursing education is the creation of care plans, which serve as structured guidelines for medical treatment and play a vital role in ensuring continuity of care and patient safety (Kreikenbaum & Lay, 2022). Care plans typically include nursing diagnosis, patient resources, care objectives, and care measures (Lauber & Schmalstieg, 2017). While AI could potentially support students in creating these plans, questions remain about students' trust in AI-generated content and the quality of AI-assisted work.

Against this background, the present study examined how vocational nursing students perceived the trustworthiness of care plans when they assumed that either chatbots or human peers created the plans. Furthermore, this study investigated whether students' perceived trustworthiness in chatbots was justified by comparing the quality of

* Corresponding author.
E-mail address: martin.laun@uni-bielefeld.de (M. Laun).

https://doi.org/10.1016/j.cedpsych.2025.102373

Available online 22 April 2025

care plans produced under different conditions: by students alone, by students with ChatGPT assistance, and by ChatGPT without students. The insights expected from this study may have important implications for understanding chatbot integration in professional education and for developing strategies that promote students' reflective and critical approach to chatbot use.

2. Theoretical section

2.1. Trust perceptions of chatbot-generated content

Trust plays a pivotal role in shaping individuals' interactions with technological systems by influencing their engagement with new technologies (Cukurova et al., 2023; Freedy et al., 2007; Kaplan et al., 2023; Nadarzynski et al., 2019; Przegalinska et al., 2019). Empirical studies have identified trust as a significant predictor for accepting and using chatbots, also in learning contexts (Baek & Kim, 2023; Shahzad et al., 2024). In educational settings, where accuracy and reliable information are paramount, the trustworthiness of chatbots becomes particularly significant.

Several factors contribute to the debate over the trustworthiness of AI in education, including chatbots (Brummernhenrich et al., 2025; Kaplan et al., 2023; Nazaretsky et al., 2025). On the one hand, concerns exist about the potential for chatbots like ChatGPT to spread misleading information convincingly (Han et al., 2024; Spitale et al., 2023), which could have significant implications for professional practice and education. Analogous to internalizing other forms of misinformation (Bensley & Lilienfeld, 2017), students may unknowingly adopt incorrect solutions or inaccurate information from AI systems, potentially resulting in misconceptions and flawed understanding of critical professional concepts. Furthermore, algorithm aversion, a phenomenon particularly relevant in educational settings (Kaufmann, 2021), may undermine perceptions of trustworthiness, as people tend to lose confidence in algorithmic systems (Dietvorst et al., 2015). This tendency has been empirically demonstrated in educational contexts, as shown by Nazaretsky et al. (2024) in a study on student evaluations of feedback quality. Their research used a within-subject design, where students assessed two versions of feedback—one AI-generated and one human-created—on three dimensions: objectivity, usefulness, and genuineness. Initially, students rated the feedback without knowing the source, followed by a second evaluation after source disclosure. Results showed that ratings for AI-generated feedback significantly decreased after students learned its origin, whereas perceptions of human feedback slightly improved. This suggests that the observed changes in perception were driven by source disclosure rather than actual content quality.

Algorithm aversion can be due to several reasons, such as the belief that human experts possess unique knowledge unlikely to be replicated by algorithms, insufficient understanding of algorithmic performance, or asymmetrical forgiveness, where people judge algorithmic errors more harshly than human ones (Dietvorst et al., 2015). In educational contexts, students may reinforce this aversion upon encountering incorrect information from AI tools like ChatGPT, which can lead to a disproportionate loss of trust. This effect could be especially pronounced in hierarchical environments such as medical education, where authority figures are traditionally viewed with significant respect (Lempp & Seale, 2004). In these settings, the belief in human experts' unique knowledge and skepticism towards algorithms may be particularly entrenched. The "black box" nature of many AI algorithms (Rudin & Radin, 2019), including ChatGPT, may exacerbate this aversion. The lack of transparency in how these systems work and derive their responses could decrease trust due to unpredictability (Mayer et al., 1995). Additionally, personal data disclosure and usage concerns further complicate trust issues (Følstad et al., 2018; Kretzschmar et al., 2019; Nadarzynski et al., 2019).

On the other hand, however, there are also arguments in favor of a high perceived trustworthiness of chatbots. The perceived ease of use and user-friendly interfaces of chatbots such as ChatGPT (Sallam et al., 2024), coupled with their ability to provide quick access to information without the need for extensive research or website navigation (Dwivedi et al., 2023), align closely with the human desire to reduce complexity (Luhmann, 2014). According to the perfect automation theory (Dijkstra et al., 1998), individuals tend to trust computers more than humans because they perceive computers as more objective and rational, which is attributed to the expectation of human imperfection. Furthermore, research suggests that trust in chatbots is also influenced by perceived anthropomorphism and interaction quality (Hancock et al., 2011; Law et al., 2022; Seeger & Heinzl, 2017). For example, ChatGPT's capability for humanlike interaction (Orrù et al., 2023), by mimicking personal conversations, indistinguishable from human-written texts (Spitale et al., 2023), may further foster trust, enhancing users' comfort and confidence in the chatbots' responses. These opposing perspectives—ranging from concerns about misinformation and algorithm aversion to arguments supporting trust based on perceived anthropomorphism and interaction quality or reduced complexity—highlight the need for empirical investigation of chatbot trustworthiness in educational settings.

2.2. Student performance with and without chatbot assistance

As AI tools become increasingly integrated into educational and professional settings (Zhang & Aslan, 2021), including healthcare, effectively utilizing these technologies becomes crucial. Human-AI systems show considerable potential due to their complementary capabilities: While humans possess general intelligence for diverse problem-solving, AI systems can efficiently handle time-consuming, structured tasks (Vaccaro et al., 2024). In nursing education, AI-based chatbots like ChatGPT could augment students' task performance in care plan creation by providing rapid access to relevant information and structured guidance.

For chatbots to effectively support practical tasks in nursing education, their capability to handle medical content accurately is crucial. Recent research has demonstrated promising results regarding ChatGPT's medical knowledge and competencies. For instance, Gilson et al. (2023) found that ChatGPT (GPT-4) achieved accuracy levels near the passing thresholds for questions from the United States Medical Licensing Examination (USMLE). Similarly, Taira et al. (2023) reported that ChatGPT (GPT-3.5) met the passing criteria on the Japanese National Nurse Examinations. These findings suggest that ChatGPT possesses sufficient medical knowledge to support nursing students in practical tasks. However, both studies also identified limitations, including performance variations across different medical tasks and specialty areas.

While ChatGPT demonstrates promising capabilities in medical contexts, its impact on task performance may not be solely determined by its output quality but also by psychological mechanisms that influence how students interact with chatbots. In this context, cognitive load theory (Chandler & Sweller, 1991) may offer valuable insights into these interaction processes. When applied to students' performance in care plan creation with ChatGPT, this theoretical framework suggests both potential benefits and challenges of chatbot assistance. On the one hand, chatbots could reduce cognitive demands by handling information retrieval and organization in care plan creation. For example, the AI can quickly generate structured content and suggest relevant care measures, potentially reducing the mental effort required for basic task components. This freed cognitive capacity could allow students to focus more on critical evaluation and refinement of the care plans, potentially leading to higher-quality solutions. On the other hand, however, integrating chatbots may simultaneously introduce additional cognitive demands. Students must invest mental resources in formulating effective prompts and reconciling AI-generated suggestions with clinical guidelines, which can increase extraneous cognitive load. This is particularly relevant, considering that higher cognitive load is correlated with lower

intrinsic motivation (Evans et al., 2024). If students experience excessive cognitive demands when interacting with ChatGPT, their motivation to engage with the task may decrease, potentially offsetting the benefits of chatbot assistance. Thus, the interplay between reduced basic task demands and increased interaction demands may shape task performance outcomes, depending on how effectively students manage these different cognitive aspects during ChatGPT-assisted care plan creation.

Recent empirical evidence from medical practice provides valuable insights into the challenges of human-chatbot collaboration. For instance, in an experiment, Goh et al. (2024) examined the impact of ChatGPT on physicians' diagnostic reasoning. Their study revealed that while ChatGPT alone outperformed individual physicians, physicians using ChatGPT as a diagnostic aid did not significantly improve their performance compared to those without chatbot support. This finding underlines the notion that the mere availability of chatbot assistance does not automatically enhance professional task performance. Goh et al. (2024) concluded that users' ability to interact meaningfully with the system is a crucial determinant of effective chatbot utilization. Supporting this notion, research has demonstrated that the quality of chatbot-generated output is susceptible to prompt formulation (Nazari & Saadi, 2024), suggesting that users' ability to craft effective prompts may significantly influence task performance outcomes.

Beyond individual studies, broader perspectives on human-AI collaboration emerged from a *meta*-analysis by Vaccaro et al. (2024), which synthesized findings from 106 experimental studies across various domains, primarily in decision-making tasks where participants chose predefined options. Their analysis revealed that human-AI teams, on average, performed worse than the best-performing individual system, whether human or AI. However, Vaccaro et al. (2024) identified a critical research gap, particularly regarding generative AI tools like ChatGPT for creation tasks. They noted that much of the research on generative AI with human participants has primarily focused on user attitudes rather than task performance. This results in a lack of comprehensive comparisons between human, AI, and human-AI collaborative performance in creation tasks. This gap is particularly evident in educational settings, where AI tools are increasingly integrated into tasks such as care plan creation, yet empirical evidence on their actual impact remains limited.

In summary, the current research landscape underscores the need for empirical studies comparing performance outcomes across students working independently, students collaborating with chatbots, and chatbots alone in educational contexts, particularly in nursing education. Additionally, such analyses could help determine whether students' trust in content attributed to chatbots corresponds to the actual quality of work produced with chatbots.

3. The present study

Based on the considerations outlined in the theoretical section, this study addressed two main research questions to gain important insights into the potential and challenges of integrating chatbots in educational settings (particularly in nursing education):

1. How does the knowledge that a care plan is authored by a chatbot or a peer influence students' assessment of its trustworthiness?
2. Is the perceived trustworthiness of chatbots justified in light of chatbots' actual performance in care plan creation?

To address these questions, we conducted an exploratory empirical study involving students from vocational nursing schools. These students were asked to write medical care plans with and without a chatbot's assistance (ChatGPT4). Additionally, students evaluated the trustworthiness of an independent care plan following experimentally manipulated feedback indicating whether a peer or a chatbot created the plan. After completing their care plans with ChatGPT, students indicated their intentions to use the tool for future care plan creation.

4. Method

4.1. Sample

Our final sample consisted of N = 189 students from 13 classes in 10 different vocational nursing schools in Germany. All students were in their second and third years of nursing training (age: M = 25.66, SD = 9.30, 77.2 % female). Most students were native German speakers (62,96 %). Participation in our study was voluntary. Students were excluded from the study ($n = 32$) if they failed to pass our manipulation check of the experimental part (see below). Nevertheless, our findings were also replicated in additional analyses that included those students who had failed the manipulation check.

4.2. Procedure

The study was conducted from July to September 2023, during which 13 sessions were held. Each session was attended by 5–30 students and two study facilitators who provided instruction. The language used in the workshops was German. The local ethics commission declared the study ethically unobjectionable, and the school supervisory authority in Rhineland-Palatinate reviewed and approved it for implementation.

During the study, students in their classroom were provided with iPads (*9th generation*) and Smart Keyboards (*MX3L2D/A*) at the start. Each session lasted for approximately two hours. The study facilitators guided the students through the various phases of the study and ensured that the students worked on the different tasks simultaneously to avoid disturbances. The study started with a welcome and brief introduction, where students were informed about the study procedure. In particular, they were told that they would create two nursing care plans, one with and one without ChatGPT assistance. After the students were informed about the course of the study, they were asked to sign consent forms for voluntary participation and data use. Next, there was an orientation to ChatGPT (Model GPT4), during which the students learned about the usage and features of the chatbot. Instructions on toggling between ChatGPT and our survey tool (LimeSurvey, Version 3.28) and text copying and pasting techniques were provided. The students then completed two care tasks in a within-subjects design with double randomization. First, the order of the two care tasks was randomized for each student. Second, students were randomly assigned to use ChatGPT assistance either for their first task but not for their second task, or for their second but not for their first task. Each task had a 15-minute completion time frame, followed by a 15-minute break between tasks.

The task involved two standardized nursing scenarios from a textbook (Lauber & Schmalstieg, 2017), reflecting typical assignments that students would encounter in exams and have likely worked on previously. The first case was about Mrs. Firn, who underwent emergency gallbladder surgery. The second case focused on Mr. Holzer, an older man with a newly installed stoma (i.e., a surgically created opening on the abdomen) after rectal cancer surgery. For both scenarios, students had to develop a care plan by identifying (1) the nursing diagnosis, (2) resources, (3) care objectives, and (4) care measures from the text. To evaluate how the students performed against ChatGPT, we also had ChatGPT solve the tasks ten times independently. Furthermore, we analyzed how students performed when they used ChatGPT as a support tool in their tasks. This benchmarked student and chatbot performance and provided them with firsthand experience of ChatGPT's capabilities.

After completing their initial assignments, the students evaluated a third care plan derived from a textbook example (Lauber & Schmalstieg, 2017). This plan was about Pia, a premature infant who faced critical health issues upon admission to a hospital. In this case, the solutions to the categories were predefined. This scenario was chosen for its sensitivity and the serious consequences of incorrect care instructions. At this stage, the students were told that the plan could come from a chatbot or a peer and were asked to judge the trustworthiness of the plan.

Following their response, students received experimentally manipulated feedback that this plan was created either by a chatbot or a peer. Based on the new information, they reassessed the trustworthiness of the same plan. Then, a manipulation check was conducted, and the experimental part of our study concluded with a debriefing session, where participants were informed about the specifics of the experimental manipulation. Subsequently, the students answered questions about their socio-demographic data and intention to use ChatGPT for future care plan creation. Moreover, they responded to further questions, the results of which are beyond the scope of this particular study and are not discussed in this publication.

4.3. Measures

4.3.1. Trustworthiness

To assess how students judged the trustworthiness of the reviewed care plan, we used four items, which students answered on a Likert scale from 1 = *strongly disagree* to 7 = *strongly agree*: (1) "I find the care plan trustworthy," (2) "I would implement the care plan in practice," (3) "The care measures are sensible," and (4) "I believe that the care plan contains no errors." The reliability of these measures was good ($\alpha_{T1} = 0.89$, $\alpha_{T2} = 0.92$). To ensure robustness, we conducted additional analyses excluding the second item ("I would implement the care plan in practice"), which measured behavioral intention. These analyses replicated our findings.

4.3.2. Manipulation check

In our manipulation check, participants were asked: "Who was the author of the care plan?" with the options "chatbot" or "peer.".

4.3.3. Quality of the care plans

To determine the quality of the responses, two nursing experts (state-certified nurses with additional teaching qualifications) evaluated all three sets of the care plans (100 % double-coding): those created by students alone, those created by students with ChatGPT assistance, and all ten independently generated ChatGPT plans. The evaluation was conducted fully blinded, ensuring that raters were unaware of the authorship conditions of the care plans. The assessments were based on a standardized scoring schema. In this system, each category (nursing diagnosis, resources, care objectives, care measures) allowed for scoring from 0 to 3 points, thus enabling each plan to achieve a total score ranging from 0 to 12. For an incorrect answer, one point was subtracted. For each category, specific criteria needed to be addressed for students to earn points. This system involved identifying critical aspects within a broader set of possible answers. The Intraclass Correlation Coefficient (2,1) was employed to assess the agreement between the two independent raters. The calculated ICC value was 0.87, indicating a good agreement between the raters (Cicchetti, 1994).

4.3.4. Intention to use ChatGPT for future care plan creation

To assess students' intention to use ChatGPT for future care plan creation, we adapted three items from Venkatesh et al. (2003). Participants indicated their agreement on a 7-point Likert scale (1 = *strongly disagree*, 7 = *strongly agree*) with the following statements: (1) "I intend to continue using ChatGPT for creating care plans in the future," (2) "I expect to use ChatGPT for creating care plans in the future," and (3) "I plan to use ChatGPT for creating care plans in the future." The scale showed excellent internal consistency ($\alpha = 0.94$).

4.4. Analyses

We employed Mplus 8 software (Muthén & Muthén, 2017) for data analysis using the robust maximum likelihood estimator (Yuan & Bentler, 2000) for model estimation. We modeled trustworthiness as a latent variable to account for measurement error at the indicator level, while nursing care plan quality was specified as a manifest variable. Effect coding (Little et al., 2006) was applied for scaling and meaningful interpretation of latent variables. We evaluated the model fit using the common χ^2-statistic, CFI, and RMSEA.

To address our first research question, we specified a latent change score model (McArdle, 2009) to examine the effect of source disclosure (chatbot vs. peer) on change in nursing students' trust in care plans. The model is visualized in Fig. 1. We ensured scalar measurement invariance for comparability across time by constraining factor loadings and intercepts to be equal (Table 1). We allowed correlations between the same indicator variables over time to account for residual effects that could not be attributed to the latent constructs (Marsh & Hau, 1996). The change in trustworthiness was analyzed by decomposing the score at T2 into the score at T1 and the difference between the scores at T2 and T1. To this end, we fixed the path coefficient from the score at T1 and the change score to the score T2 to 1. The intercepts of the score at T1 and the change score were freely estimated, whereas the intercept of the score at T2 and the residual variance of the score at T2 were set to 0. The score at T1 and the change score were allowed to correlate. The change score was regressed on the Source Disclosure variable (coded as 1 = disclosed as peer-generated, 0 = disclosed as chatbot-generated), while the T1 score was correlated with the Source Disclosure variable to examine potential differences in the T1 scores between the two conditions. To further explore the change in trustworthiness scores based on the Source Disclosure condition, we conducted an additional multi-group latent change score analysis. Measurement invariance testing confirmed the comparability of our measures across both time points and groups (see Table S1 in the supplementary materials). We then examined whether the change scores differed significantly from zero for each disclosure condition.

To address our second research question, we conducted a regression analysis to assess the effect of authorship (students with ChatGPT vs. students alone vs. ChatGPT alone) on the quality of the nursing care plans. Following the standard structure of nursing care plans (Lauber & Schmalstieg, 2017), we analyzed four categories: nursing diagnosis, resources, care objectives, and care measures. We first calculated the score for each of these four components by averaging the ratings of both reviewers for each subtask. Then, we summed these four sub-scores to create a total score for each care plan. This total score served as our dependent variable in the analysis. Additionally, we used the "type = complex" option with student ID as a cluster variable to adjust for non-independence of our observations. For consistency in our data structure, ChatGPT in the "ChatGPT Alone" condition was also assigned a student ID. The total score was regressed on two dummy variables representing two of the three authorship conditions: "Students Alone" (1 = students alone, 0 = other) and "ChatGPT Alone" (1 = ChatGPT alone, 0 = other). Thus, the third condition, "Students with ChatGPT," served as the reference condition. To directly compare the effects of the Students Alone and ChatGPT Alone conditions, we conducted a Wald test by calculating the difference between their coefficients and testing whether this difference significantly differed from zero using the "model test" command in Mplus. We also controlled for the potential effects of the two case scenarios by regressing the total score on a vignette variable (1 = Holzer, 0 = Firn).

In additional analyses, we examined whether students who performed better without ChatGPT also achieved higher scores when using ChatGPT. For this purpose, we analyzed the correlation between the quality of care plans created by students in the "Students Alone" and "Students with ChatGPT" conditions. Furthermore, we analyzed the correlation between students' intention to use ChatGPT for future care plan creation and the difference in care plan quality with and without chatbot assistance. This could provide an indication of whether students recognized improvements in their work with ChatGPT support. Additionally, we examined students' intention to use ChatGPT for future care plan creation and tested it against the scale's midpoint ($M = 4$) using a one-sample *t*-test.

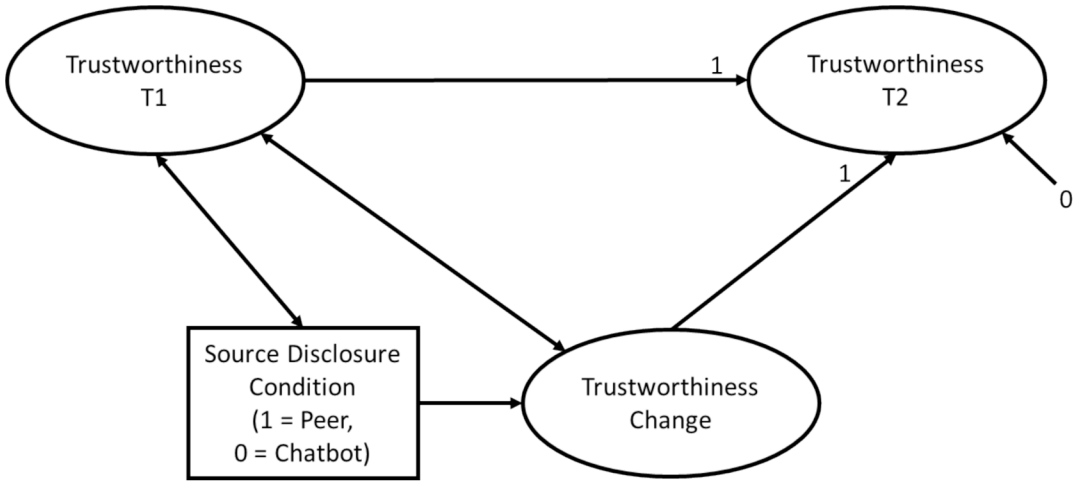


Fig. 1. Latent Change Score Model for Evaluating the Impact of Source Disclosure Condition on Perceived Care Plan Trustworthiness.

Table 1
Tests of Longitudinal Measurement Invariance for the Trustworthiness Scale.

	χ^2	*df*	*p*	CFI	RMSEA	$\triangle\chi^2$	$\triangle df$	*p*
Configural invariance	18.09	21	0.643	1.00	0.00			
Metric invariance	19.75	24	0.711	1.00	0.00	1.66	3	0.651
Scalar invariance	25.94	27	0.522	1.00	0.00	6.20	3	0.103

Note. N = 189.

5. Results

5.1. Research question 1: Changes in trust following source disclosure

The latent change score model demonstrated an excellent fit (χ^2 = 25.94, *df* = 27, *p* = 0.522, *RMSEA* = 0.00, *CFI* = 1.00). The trustworthiness rating at T1 in the total sample was *M* = 4.65 (*S.E.* = 0.09, p < 0.001), with a non-significant change score (Δ*M* = –0.09, *S.E.* = 0.05, *p* = 0.057) following source disclosure. Our primary analysis revealed that the Source Disclosure condition significantly affected changes in trustworthiness (*b* = 0.28, *S.E.* = 0.09, *p* = 0.003, β = *0.27*). Notably, the correlation between Source Disclosure and trustworthiness at T1 was not significant (*r* = 0.02, *p* = 0.834), indicating that the effect of source disclosure cannot be attributed to differences in initial trust levels. Additionally, the correlation between initial trust levels and change scores was not significant (*r* = –0.03, *p* = 0.606), suggesting that students' trust changed independently of their baseline trust levels. The additional multi-group latent change score analysis, conducted to explore these changes further, showed a good model fit (χ^2 = 53.90, *df* = 48, *p* = 0.259, *RMSEA* = 0.04, *CFI* = 0.99). This supplementary analysis revealed different patterns of change in trust for the students in the two Source Disclosure conditions: Students being told the text was peer-authored showed no significant change in trust (Δ*M* = 0.04, *S.E.* = 0.07, *p* = 0.552), while students being told the text was chatbot-

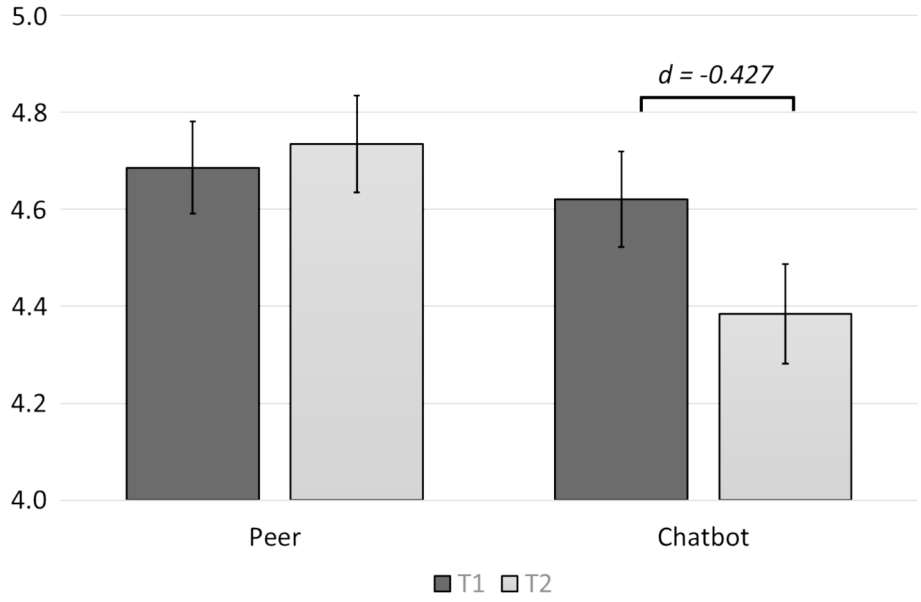


Fig. 2. Change in Perceived Trustworthiness.

authored experienced a significant decrease in trust ($\Delta M = -0.23$, *S.E.* = 0.06, $p < 0.001$, $d = -0.43$). Fig. 2 illustrates the estimated mean values for each condition at both time points.

5.2. Research question 2: Care plan quality across authorship conditions

Table 2 presents the results of the regression analyses comparing the quality of the care plans as a function of how the plans were created (by students alone, by students with ChatGPT assistance, or by ChatGPT alone). Compared to the reference condition (students with ChatGPT assistance), students working alone performed significantly worse, while ChatGPT alone produced significantly better results. The control variable vignette had no significant impact on the quality of care plans. The Wald test comparing the coefficients for the Students Alone and ChatGPT Alone conditions revealed a significant difference ($W = 844.85$, df = 1, $p < 0.001$).

5.3. Additional analysis

The additional analysis of the relationship between the quality of care plans created by students in the "Students Alone" and "Students with ChatGPT" conditions revealed a significant positive correlation ($r = 0.19$, $p = 0.009$). In contrast, the correlation between students' intention to use ChatGPT for future care plan creation and the difference in care plan quality between the conditions with and without ChatGPT assistance was not significant ($r = 0.07$, $p = 0.377$). The mean intention to use ChatGPT for future care plan creation was $M = 4.29$ ($SD = 1.62$), slightly but significantly above the scale's midpoint of 4, $t(188) = 2.43$, $p = 0.016$.

6. Discussion

This study investigated the perceived trustworthiness of chatbots and whether their actual performance justified this trustworthiness. Specifically, we examined how nursing students' trust in a care plan changed depending on whether it was attributed to being written by a chatbot or a peer while simultaneously assessing the actual quality of care plans created either by students without ChatGPT assistance, by students with ChatGPT assistance, and by ChatGPT alone.

Regarding our first research question—whether being informed that a chatbot or a peer authored a care plan influences students' assessment of its trustworthiness—we found that source disclosure significantly affected trustworthiness ratings: Students' trust decreased when the care plan was labeled as chatbot-generated, whereas it remained stable when attributed to a peer. This effect occurred independently of the initially perceived trustworthiness of the care plan (before source disclosure), as indicated by the non-significant correlation between initial trustworthiness and change scores.

The observed decrease in students' trustworthiness following chatbot authorship disclosure aligns with findings from Nazaretsky et al. (2024), who observed a similar pattern in educational settings, where students' evaluations of feedback objectivity, usefulness, and authenticity declined upon learning it was AI-generated. Our study extends these insights by not only assessing trustworthiness experimentally in a nursing education context but also incorporating objective performance evaluations to determine whether the lower trust in chatbot-generated content is justified by chatbot performance.

We examined differences in care plan quality between ChatGPT alone, students with ChatGPT, and students alone as part of our second research question. Interestingly, our analysis revealed a clear hierarchy in performance: ChatGPT alone produced the highest-quality plans, followed by students using ChatGPT, while students working alone achieved the lowest scores for the quality of care plans. This pattern remained consistent across both vignettes. Although students did not achieve the performance of ChatGPT working alone, they demonstrated a significant improvement when using ChatGPT compared to their performance without chatbot assistance.

This pattern of ChatGPT outperforming student-chatbot collaboration aligns with findings from Vaccaro et al.'s (2024) *meta*-analysis, which demonstrated that AI systems frequently outperform human-AI teams across various domains. In another medical context, Goh et al. (2024) found that ChatGPT outperformed individual physicians and physician-AI teams in diagnostic reasoning in a randomized clinical vignette study. However, our findings differ in a key aspect from those of Goh et al. (2024): While Goh et al. (2024) observed no improvement in physician performance with ChatGPT support, our students did benefit from ChatGPT assistance, even though they did not fully leverage the chatbot's potential.

A potential explanation for students' lower performance with ChatGPT compared to ChatGPT alone may lie in their underreliance on the tool. For instance, students who were overly skeptical of ChatGPT's output may have dismissed or underutilized its suggestions, leading to missed opportunities to enhance care plan quality. Instead of refining the chatbot-generated content, they may have relied solely on their own knowledge, which, as our findings suggest, resulted in lower-quality outcomes than fully leveraging ChatGPT's capabilities.

Another factor that may have contributed to the differences in care plan quality between students using ChatGPT and those working alone is the increase in extraneous cognitive load (Chandler & Sweller, 1991). While ChatGPT has the potential to simplify care plan creation, managing the tool itself may have introduced additional cognitive demands, which Evans et al. (2024) found to be negatively correlated with motivation. Students had to formulate effective prompts, critically evaluate ChatGPT-generated responses, and integrate them into their work simultaneously—a cognitively demanding process that may have increased mental workload, reduced motivation, and ultimately limited the benefits of chatbot assistance.

Moreover, prompt engineering skills are crucial for maximizing AI support (Nazari & Saadi, 2024). Students' limited ability to craft effective prompts may have resulted in suboptimal AI outputs and inefficient integration of ChatGPT-generated content into their workflow, further constraining the potential benefits of AI assistance.

Interestingly, we found a significant correlation between the quality of students' care plans with and without ChatGPT, indicating that those who produced higher-quality plans on their own also maintained higher quality when using chatbot assistance. This suggests that ChatGPT cannot fully compensate for lower domain expertise, as subject-specific knowledge may be crucial for formulating precise, discipline-specific prompts and effectively identifying and correcting potential misinformation.

ChatGPT's superior performance in care plan creation starkly contrasts our findings on trustworthiness. Despite generating higher-quality care plans, content attributed to ChatGPT was rated as less trustworthy than identical content attributed to peers. This discrepancy between objective quality and perceived trustworthiness suggests that factors beyond output quality influence trust in chatbot-generated content. Given ChatGPT's significant advantage over students in care plan creation—both with and without chatbot assistance— our results indicate that the quality of chatbot-generated plans did not justify the decline in trust following chatbot disclosure. Paradoxically, students placed the

Table 2
Prediction of the Care Plan Quality.

Predictors	*B*	*SE*	*p*	β
Students Alone (1 = Students alone, 0 = Other)	**–2.71**	**0.25**	**<0.001**	**–0.43**
ChatGPT Alone (1 = ChatGPT alone, 0 = Other)	**2.47**	**0.23**	**<0.001**	**0.12**
Vignette (1 = Holzer, 0 = Firn)	–0.13	0.24	0.587	–0.02

Note. N = 388. Significant values are highlighted in bold. The intercept reflects the average quality for care plans written by students with ChatGPT. Intercept = 8.35 (S.E. = 0.27, p < 0.001).

least trust in the source that objectively produced the highest-quality care plans.

Several psychological mechanisms warrant consideration to better understand this discrepancy between the quality of and trust in ChatGPT. One potential explanation is algorithm aversion (Dietvorst et al., 2015), which suggests that individuals tend to favor human judgment over algorithmic decisions, even when presented with clear evidence of AI's superior performance. The opacity of AI decision-making may further amplify this aversion. The "black box" nature of AI-generated text (Rudin & Radin, 2019) obscures how care plans are formulated, making it difficult for users to assess the reasoning behind AI-generated outputs. This concern may be exacerbated by ChatGPT's known tendency to hallucinate or generate misinformation (Han et al., 2024), reinforcing skepticism toward AI-generated content.

Another possible explanation for the trust-performance gap are students' perceptions of ChatGPT's benefits. Our findings indicate that students reported only moderate intentions to use ChatGPT for future care plan creation, and no significant correlation emerged between these intentions and the differences in care plan quality with and without ChatGPT assistance. This suggests that students may have underestimated ChatGPT's capabilities and how much they benefited from using the tool, potentially contributing to their reluctance to trust chatbot-generated content.

However, students' only moderate intention to use ChatGPT for future care plan creation may also reflect that real-world nursing care planning involves interpersonal and emotional factors that extend beyond our experimental setting. The importance of social processes in nursing may lead students to perceive ChatGPT as less applicable to professional practice, which could contribute to their lower intention to use it in this context. Since we did not explicitly assess whether students recognized improvements in their performance with ChatGPT assistance, further research is needed to examine whether awareness of ChatGPT's benefits influences trust in chatbot-assisted learning contexts. Future studies could address this by manipulating students' perceptions of ChatGPT's effectiveness, for instance, by framing their performance with ChatGPT as either highly effective or less impactful. Assessing trust levels after such an intervention could provide valuable insights into whether perceived chatbot effectiveness directly shapes students' trust in ChatGPT-generated content.

6.1. Implications

Our findings have several important implications. We identified an apparent discrepancy between ChatGPT's superior performance in care plan creation and students' decreased trust in chatbot-generated care plans following source disclosure. Given that trust is crucial for both the adoption of educational technology (Cukurova et al., 2023; Nazaretsky et al., 2025) and effective human-AI collaboration (Vaccaro et al., 2024), the persistence of mistrust despite ChatGPT's demonstrated performance highlights the need for targeted interventions in educational settings where chatbots are utilized. These interventions should address unwarranted skepticism while ensuring that students retain an appropriate level of critical evaluation, particularly in recognizing potential AI errors and limitations. Accordingly, it remains essential for students to continue developing domain-specific competencies, enabling them to recognize potential inaccuracies in chatbot-generated content and critically evaluate generative AI tools' trustworthiness.

In this context, a promising approach to addressing trust issues involves the development of targeted interventions for algorithm aversion (Dietvorst et al., 2015). Such interventions could enhance students' understanding of AI functionality through hands-on experience, allowing them to actively interact with ChatGPT, analyze its outputs, and critically assess its strengths and limitations. These approaches may help mitigate skepticism and improve trust by fostering transparency in AI-generated feedback.

However, while increasing AI literacy might seem like an obvious solution (Laupichler et al., 2022), recent research by Nazaretsky et al. (2025) highlights a more complex dynamic: A greater understanding of AI's internal workings can, in some cases, intensify ethical concerns rather than alleviate them. For instance, students who become more aware of issues such as bias in AI-generated decisions may develop greater skepticism rather than increased trust. This suggests that interventions should go beyond technical education and incorporate broader discussions on AI's ethical and societal implications in professional practice. A balanced approach that integrates technical aspects with ethical reflection may be important to fostering trust and critical engagement with generative AI tools.

Our findings regarding students' intention to use ChatGPT for future care plan creation offer another practical implication. The relatively reserved intentions to use ChatGPT in the future, combined with the lack of correlation between these intentions and the differences in the quality of care plans generated by students with and without chatbot assistance, suggest that students may underestimate both ChatGPT's capabilities and the potential benefits of using it for care planning tasks. This highlights the need for instructor feedback, providing students with clear insights into their AI-assisted performance and emphasizing how ChatGPT can enhance workflow efficiency in care plan creation. Moreover, training in prompt engineering could help nursing students engage more effectively with ChatGPT in care plan creation, as research has shown that the quality of prompts significantly impacts the output of large language models (Nazari & Saadi, 2024). Such training may also alleviate cognitive load, as greater familiarity with AI interaction could allow students to focus more on content than tool operation.

However, our findings suggest that students who performed well without ChatGPT also achieved better results with AI assistance, indicating that the chatbot did not bridge existing performance differences. This aligns with the broader perspective on AI in education—highlighted by Cukurova (2025)—that tools like ChatGPT should serve as complementary resources rather than substitutes for human expertise. Consequently, educational programs must continue prioritizing core nursing competencies while integrating AI as a supportive tool to prepare students for real-world practice.

6.2. Strengths and limitations

Our study exhibits several strengths. Firstly, students interacted with ChatGPT before evaluating the trustworthiness of an unfamiliar care plan written by either ChatGPT or a peer, providing them with firsthand experience they could draw upon in their assessment. Secondly, our experimental design enabled causal conclusions regarding students' perceptions of chatbot trustworthiness, which adds to the growing body of chatbot research that often relies on non-experimental approaches. Thirdly, our relatively large sample size provided high statistical power to examine changes in students' trust and differences in the quality of care plans created by ChatGPT alone, students with ChatGPT assistance, and students without ChatGPT assistance. Fourth, a strength of our study was the significant investment in resources to ensure a high-quality research environment. We purchased 30 subscriptions to a premium version of ChatGPT, which represents a key advantage over previous studies. Furthermore, each participant was provided with an iPad and a keyboard, facilitating a consistent and technologically advanced interface for interaction with ChatGPT. This level of resource commitment enhanced students' experience and ensured that the data collected were representative of the best possible use-case scenario for this technology.

Despite these strengths, several limitations of our study should be considered. First, the relatively high number of errors in the manipulation check suggests a need for careful consideration. The simple question "Who was the author of the care plan?" with the answer options "chatbot" or "peer" might have led some students to report their original beliefs about authorship rather than the experimentally assigned condition. However, the fact that our findings could be replicated in the entire sample suggests that these manipulation check failures did not

significantly affect our results. Second, it should be considered that using a premium version of ChatGPT might reflect something different than the typical resources available to the average student or educational institution. While our study demonstrated the performance of chatbot technology in an ideal setting, the findings might differ in less resource-intensive environments. However, as generative AI technology evolves, premium features might become more widely accessible, potentially broadening their educational impact. Third, the investigation of a sample of vocational nursing students in Germany, the focus on nursing care plans in this research, and the specific characteristics of ChatGPT may limit the generalizability of our findings to other student populations, healthcare contexts, and chatbot technologies. Furthermore, the importance of social processes in real-world nursing care raises questions about the extent to which our findings translate into clinical practice. Regarding the question of the broad applicability of our findings, future research should examine whether our findings are generalizable to diverse student populations, educational contexts, and real-world care settings.

Finally, our study did not cover all aspects that could be relevant with regard to the use of chatbots in education, which highlights several opportunities for future research. For example, while we examined student performance with and without chatbot support, the question of how chatbot assistance influences learning outcomes remains open. Given that elaborative strategies promote deeper cognitive processing (Hamilton, 1997; Martella et al., 2024) and that chatbots may facilitate such strategies through active discussions and question formulation, future research should investigate whether chatbot-assisted learning enhances conceptual understanding and long-term knowledge retention. Additionally, little is yet known about the role of individual differences in students' performance with chatbot support. Our findings showed that the quality of care plans created by students with ChatGPT assistance did not match the quality of care plans generated by ChatGPT alone, highlighting the need to investigate which factors influence students' ability to use chatbot assistance effectively. One potential factor could be students' AI-related self-concepts, given that domain-specific self-concepts of ability and achievements are strongly positively correlated (e.g., Möller et al., 2020; Schneider & Wolff, 2023; Wolff et al., 2021a) and have shown to influence each other (e.g., Marsh & Craven, 2006; Wigfield & Eccles, 2000; Wolff et al., 2021b). Specifically, students with a higher AI-related self-concept may engage more effectively with ChatGPT, which may allow them to interact more confidently with the tool, formulate better prompts, and ultimately perform better with ChatGPT assistance. Future studies should explore whether a stronger AI-related self-concept leads to improved engagement with ChatGPT and greater benefits from its use.

7. Conclusion

Our study revealed a significant gap between the perceived trustworthiness of chatbots and their actual performance in medical educational tasks, specifically in creating nursing care plans. Although ChatGPT generated higher-quality care plans than students working either alone or with ChatGPT assistance, students perceived plans as more trustworthy when they believed they were written by a peer rather than by a chatbot. This skepticism presents important challenges for educational practice, as it may influence how effectively students engage with and benefit from chatbot support. Furthermore, our findings suggest the importance of developing sophisticated approaches to AI integration that promote appropriate levels of trust in chatbots without promoting overreliance. In addition, it remains essential for students to continue developing domain-specific competencies, enabling them to recognize potential inaccuracies in chatbot-generated content and to evaluate the trustworthiness of generative AI tools critically. This underscores the importance of maintaining strong educational foundations across learning contexts, ensuring that chatbot integration enhances rather than replaces fundamental knowledge and critical thinking skills.

Author Statement

We thank all the nursing schools and students that took part in our study, our partner Aida Drews, as well as Susanne Čapková-Diouf, Christian Krause, Felicia Müller and Jan Oliver Schröder for their help with data collection and Miriam Fackler and Talina Wetterauer for their expert evaluation of the plans. Correspondence concerning this article should be addressed to Martin Laun, Department of Psychology, Bielefeld University, Universitätsstraße 25, 33,615 Bielefeld, Germany. E-mail address: martin.laun@uni-bielefeld.de.

Ethical Statement

This study received ethics approval from the institutional review board of the University of Koblenz. Informed consent was obtained from all participants.

CRediT authorship contribution statement

Martin Laun: Writing – original draft, Visualization, Software, Methodology, Investigation, Formal analysis, Conceptualization. **Leonard Puderbach:** Writing – review & editing. **Katharina Hirt:** Writing – review & editing. **Eva L. Wyss:** Writing – review & editing. **Daniel Friemert:** Writing – review & editing. **Ulrich Hartmann:** Writing – review & editing. **Fabian Wolff:** Writing – review & editing, Validation, Supervision, Resources, Project administration, Methodology, Conceptualization.

Declaration of Competing Interest

We have no conflicts of interest to disclose. Our research was funded by Grant 16DHBKI039 from the German Federal Ministry of Education and Research (BMBF). During the preparation of this work the authors used Grammarly and Claude Sonnet 3.5 in order to improve the readability of our publication. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Data availability

The data and supplemental material can be downloaded at: [*https://osf.io/wj49k/*].

References

Baek, T. H., & Kim, M. (2023). Is ChatGPT scary good? How user motivations affect creepiness and trust in generative artificial intelligence. *Telematics and Informatics, 83*. https://doi.org/10.1016/j.tele.2023.102030

Bensley, D. A., & Lilienfeld, S. O. (2017). Psychological misconceptions: Recent scientific advances and unresolved issues. *Current Directions in Psychological Science, 26*(4), 377–382. https://doi.org/10.1177/0963721417699026

Brummernhenrich, B., Paulus, C. L., & Jucks, R. (2025). Applying social cognition to feedback chatbots: Enhancing trustworthiness through politeness. *British Journal of Educational Technology*. https://doi.org/10.1111/bjet.13569

Chandler, P., & Sweller, J. (1991). Cognitive load theory and the format of instruction. *Cognition and Instruction, 8*(4), 293–332. https://doi.org/10.1207/s1532690xci0804_2

Cicchetti, D.v. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment, 6*(4), 284–290. https://doi.org/10.1037/1040-3590.6.4.284

Cukurova, M., Miao, X., & Brooker, R. (2023). Adoption of artificial intelligence in schools: Unveiling factors influencing teachers' engagement. In N. Wang, G. Rebolledo-Mendez, N. Matsuda, O. C. Santos, & V. Dimitrova (Eds.), Artificial intelligence in education. AIED 2023 (pp. 151–163). Springer. https://doi.org/10.1007/978-3-031-36272-9_13.

Cukurova, M. (2025). The interplay of learning, analytics and artificial intelligence in education: A vision for hybrid intelligence. *British Journal of Educational Technology, 56*(2), 469–488. https://doi.org/10.1111/bjet.13514

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General, 144*(1), 114–126. https://doi.org/10.1037/xge0000033

Dijkstra, J. J., Liebrand, W. B. G., & Timminga, E. (1998). Persuasiveness of expert systems. *Behaviour & Information Technology, 17*(3), 155–163. https://doi.org/10.1080/014492998119526

Dwivedi, Y. K., Kshetri, N., Hughes, L., Slade, E. L., Jeyaraj, A., Kar, A. K., Baabdullah, A. M., Koohang, A., Raghavan, V., Ahuja, M., Albanna, H., Albashrawi, M. A., Al-Busaidi, A. S., Balakrishnan, J., Barlette, Y., Basu, S., Bose, I., Brooks, L., Buhalis, D., & Wright, R. (2023). Opinion Paper: "So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. *International Journal of Information Management, 71*, Article 102642. https://doi.org/10.1016/j.ijinfomgt.2023.102642

Evans, P., Vansteenkiste, M., Parker, P., Kingsford-Smith, A., & Zhou, S. (2024). Cognitive load theory and its relationships with motivation: A self-determination theory perspective. *Educational Psychology Review, 36*(1), 7. https://doi.org/10.1007/s10648-023-09841-2

Følstad, A., Nordheim, C. B., & Bjørkli, C. A. (2018). What makes users trust a chatbot for customer service? An exploratory interview study. In S. Bodrunova (Ed.), Internet Science. INSCI 2018. Lecture Notes in Computer Science (Vol. 11193, pp. 194–208). Springer. https://doi.org/10.1007/978-3-030-01437-7_16.

Freedy, A., DeVisser, E., Weltman, G., & Coeyman, N. (2007). In *Measurement of trust in human-robot collaboration* (pp. 106–114). IEEE. https://doi.org/10.1109/CTS.2007.4621745.

Gilson, A., Safranek, C. W., Huang, T., Socrates, V., Chi, L., Taylor, R. A., & Chartash, D. (2023). How does ChatGPT perform on the United States Medical Licensing Examination (USMLE)? The implications of large language models for medical education and knowledge assessment. *JMIR Medical Education, 9*, Article e45312. https://doi.org/10.2196/45312

Goh, E., Gallo, R., Hom, J., Strong, E., Weng, Y., Kerman, H., Cool, J., Kanjee, Z., Parsons, A. S., Ahuja, N., Horvitz, E., Yang, D., Milstein, A., Olson, A. P., Rodman, A., & Chen, J. H. (2024). Influence of a large language model on diagnostic reasoning: A randomized clinical vignette study [Preprint]. medRxiv. https://doi.org/10.1101/2024.03.12.24303785.

Hamilton, R. J. (1997). Effects of three types of elaboration on learning concepts from text. *Contemporary Educational Psychology, 22*(3), 299–318. https://doi.org/10.1006/ceps.1997.0935

Han, Z., Battaglia, F., Udaiyar, A., Fooks, A., & Terlecky, S. R. (2024). An explorative assessment of ChatGPT as an aid in medical education: Use it with caution. *Medical Teacher, 46*(5), 657–664. https://doi.org/10.1080/0142159X.2023.2271159

Hancock, P. A., Billings, D. R., Schaefer, K. E., Chen, J. Y. C., de Visser, E. J., & Parasuraman, R. (2011). A meta-analysis of factors affecting trust in human-robot interaction. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 53*(5), 517–527. https://doi.org/10.1177/0018720811417254

Hwang, G.-J., Xie, H., Wah, B. W., & Gašević, D. (2020). Vision, challenges, roles and research issues of Artificial Intelligence in Education. *Computers and Education: Artificial Intelligence, 1*, Article 100001. https://doi.org/10.1016/j.caeai.2020.100001

Kabudi, T., Pappas, I., & Olsen, D. H. (2021). AI-enabled adaptive learning systems: A systematic mapping of the literature. *Computers and Education: Artificial Intelligence, 2*, Article 100017. https://doi.org/10.1016/j.caeai.2021.100017

Kaplan, A. D., Kessler, T. T., Brill, J. C., & Hancock, P. A. (2023). Trust in artificial intelligence: Meta-analytic findings. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 65*(2), 337–359. https://doi.org/10.1177/00187208211013988

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

Kaufmann, E. (2021). Algorithm appreciation or aversion? Comparing in-service and pre-service teachers' acceptance of computerized expert models. *Computers and Education: Artificial Intelligence, 2*, Article 100028. https://doi.org/10.1016/j.caeai.2021.100028

Kreikenbaum, J., & Lay, R. (2022). *Pflegeplanung leicht gemacht: Pflegeprozesssteuerung im Pflegealltag [Care planning made easy: Managing the nursing process in everyday care]* ((9. Aufl.).). Elsevier.

Kretzschmar, K., Tyroll, H., Pavarini, G., Manzini, A., & Singh, I. (2019). Can your phone be your therapist? Young people's ethical perspectives on the use of fully Automated conversational Agents (Chatbots) in mental health support. *Biomedical Informatics Insights, 11*. https://doi.org/10.1177/1178222619829083

Lauber, A., & Schmalstieg, P. (2017). Band 3: Pflegerische Interventionen [Vol. 3: Nursing interventions] (3rd ed.). Georg Thieme Verlag.

Laupichler, M. C., Aster, A., Schirch, J., & Raupach, T. (2022). Artificial intelligence literacy in higher and adult education: A scoping literature review. *Computers and Education: Artificial Intelligence, 3*, Article 100101. https://doi.org/10.1016/j.caeai.2022.100101

Law, E. L., Følstad, A., & Van As, N. (2022). Effects of humanlikeness and conversational breakdown on trust in chatbots for customer service. In Proceedings of the Nordic Human-Computer Interaction Conference (pp. 1–13). ACM. https://doi.org/10.1145/3546155.3546665.

Lempp, H., & Seale, C. (2004). The hidden curriculum in undergraduate medical education: Qualitative study of medical students' perceptions of teaching. *BMJ, 329*(7469), 770–773. https://doi.org/10.1136/bmj.329.7469.770

Little, T. D., Slegers, D. W., & Card, N. A. (2006). A non-arbitrary method of identifying and scaling latent variables in SEM and MACS models. *Structural Equation Modeling: A Multidisciplinary Journal, 13*(1), 59–72. https://doi.org/10.1207/s15328007sem1301_3

Luhmann, N. (2014). *Vertrauen: Ein Mechanismus der Reduktion sozialer Komplexität [Trust: A mechanism for reducing social complexity]* (p. (5.). Aufl.). UVK Verlagsgesellschaft.

Marsh, H. W., & Craven, R. G. (2006). Reciprocal effects of self-concept and performance from a multidimensional perspective: Beyond seductive pleasure and unidimensional perspectives. *Perspectives on Psychological Science, 1*(2), 133–163. https://doi.org/10.1111/j.1745-6916.2006.00010.x

Marsh, H. W., & Hau, K.-T. (1996). Assessing goodness of fit. *The Journal of Experimental Education, 64*(4), 364–390. https://doi.org/10.1080/00220973.1996.10806604

Martella, A. M., Swisher, M., & Mayer, R. E. (2024). How much active teaching should be incorporated into college course lectures to promote active learning? *Contemporary Educational Psychology, 102316*. https://doi.org/10.1016/j.cedpsych.2024.102316

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. *Academy of Management Review, 20*(3), 709–734. https://doi.org/10.5465/amr.1995.9508080335

McArdle, J. J. (2009). Latent variable modeling of differences and changes with longitudinal data. *Annual Review of Psychology, 60*(1), 577–605. https://doi.org/10.1146/annurev.psych.60.110707.163612

Möller, J., Zitzmann, S., Helm, F., Machts, N., & Wolff, F. (2020). A meta-analysis of relations between achievement and self-concept. *Review of Educational Research, 90*(3), 376–419. https://doi.org/10.3102/0034654320919354

Muthén, L. K., & Muthén, B. O. (2017). *Mplus user's guide* (8th ed.). Los Angeles, CA: Muthén & Muthén.

Nadarzynski, T., Miles, O., Cowie, A., & Ridge, D. (2019). Acceptability of artificial intelligence (AI)-led chatbot services in healthcare: A mixed-methods study. *DIGITAL HEALTH, 5*. https://doi.org/10.1177/2055207619871808

Nazaretsky, T., Mejia-Domenzain, P., Swamy, V., Frej, J., & Käser, T. (2024). AI or human? Evaluating student feedback perceptions in higher education. In R. Ferreira Mello, N. Rummel, I. Jivet, G. Pishtari, & J. A. Ruipérez Valiente (Eds.), *Technology enhanced learning for inclusive and equitable quality education* (Vol. 15159, pp. 243–254). Springer. https://doi.org/10.1007/978-3-031-72315-5_20.

Nazaretsky, T., Mejia-Domenzain, P., Swamy, V., Frej, J., & Käser, T. (2025). The critical role of trust in adopting AI-powered educational technology for learning: An instrument for measuring student perceptions. *Computers and Education: Artificial Intelligence, 8*, Article 100368. https://doi.org/10.1016/j.caeai.2025.100368

Nazari, M., & Saadi, G. (2024). Developing effective prompts to improve communication with ChatGPT: A formula for higher education stakeholders. *Discover Education, 3*(1), 45. https://doi.org/10.1007/s44217-024-00122-w

OpenAI. (2023). GPT-4 technical report. arXiv. https://doi.org/10.48550/arxiv.2303.08774.

Orrù, G., Piarulli, A., Conversano, C., & Gemignani, A. (2023). Human-like problem-solving abilities in large language models using ChatGPT. *Frontiers in Artificial Intelligence, 6*. https://doi.org/10.3389/frai.2023.1199350

Przegalinska, A., Ciechanowski, L., Stroz, A., Gloor, P., & Mazurek, G. (2019). In bot we trust: A new methodology of chatbot performance measures. *Business Horizons, 62*(6), 785–797. https://doi.org/10.1016/j.bushor.2019.08.005

Rudin, C., & Radin, J. (2019). Why are we using black box models in AI when we don't need to? A lesson from an explainable AI competition. *Harvard Data Science Review, 1*(2). https://doi.org/10.1162/99608f92.5a8a3a3d

Sallam, M., Elsayed, W., Al-Shorbagy, M., Barakat, M., el Khatib, S., Ghach, W., Alwan, N., Hallit, S., & Malaeb, D. (2024). ChatGPT usage and attitudes are driven by perceptions of usefulness, ease of use, risks, and psycho-social impact: A study among university students in the UAE. *Frontiers in Education, 9*. https://doi.org/10.3389/feduc.2024.1414758

Schneider, R., & Wolff, F. (2023). The formation of subject-specific values as a two-step process: Self-concepts mediate the relation between achievement and values. *Contemporary Educational Psychology, 75*, Article 102223. https://doi.org/10.1016/j.cedpsych.2023.102223

Seeger, A., & Heinzl, A. (2017). Human versus machine: Contingency factors of anthropomorphism as a trust-inducing design strategy for conversational agents. In *In Lecture notes in information systems and organisation* (pp. 129–139). https://doi.org/10.1007/978-3-319-67431-5_15

Shahzad, M. F., Xu, S., & Javed, I. (2024). ChatGPT awareness, acceptance, and adoption in higher education: The role of trust as a cornerstone. *International Journal of Educational Technology in Higher Education, 21*(1), 46. https://doi.org/10.1186/s41239-024-00478-x

Spitale, G., Biller-Andorno, N., & Germani, F. (2023). AI model GPT-3 (dis)informs us better than humans. *Science Advances, 9*(26). https://doi.org/10.1126/sciadv.adh1850

Taira, K., Itaya, T., & Hanada, A. (2023). Performance of the large language model ChatGPT on the National Nurse Examinations in Japan: Evaluation study. *JMIR Nursing, 6*, Article e47305. https://doi.org/10.2196/47305

Vaccaro, M., Almaatouq, A., & Malone, T. (2024). When combinations of humans and AI are useful: A systematic review and meta-analysis. *Nature Human Behaviour, 8*(12), 2293–2303. https://doi.org/10.1038/s41562-024-02024-1

Venkatesh, V., Morris, M. G., Davis, G. B., & Davis, F. D. (2003). User acceptance of information technology: Toward a unified view. MIS Quarterly. *Management Information Systems, 27*(3). https://doi.org/10.2307/30036540

Wigfield, A., & Eccles, J. S. (2000). Expectancy–value theory of achievement motivation. *Contemporary Educational Psychology, 25*(1), 68–81. https://doi.org/10.1006/ceps.1999.1015

Wolff, F., Lüdtke, O., Helm, F., & Möller, J. (2021). Integrating the big-fish-little-pond effect, the basking-in-reflected-glory effect, and the internal/external frame of reference model predicting students' individual and collective academic self-concepts. *Contemporary Educational Psychology, 65*, Article 101952. https://doi.org/10.1016/j.cedpsych.2021.101952

Wolff, F., Sticca, F., Niepel, C., Götz, T., van Damme, J., & Möller, J. (2021). The reciprocal 2I/E model: An investigation of mutual relations between achievement and self-concept levels and changes in the math and verbal domain across three countries. *Journal of Educational Psychology, 113*(8), 1529–1549. https://doi.org/10.1037/edu0000632

Yuan, K., & Bentler, P. M. (2000). 5. Three likelihood-based methods for mean and covariance structure analysis with nonnormal missing data. Sociological Methodology, 30(1), 165–200. https://doi.org/10.1111/0081-1750.00078.

Zhang, K., & Aslan, A. B. (2021). AI technologies for education: Recent research & future directions. *Computers and Education: Artificial Intelligence, 2*, Article 100025. https://doi.org/10.1016/j.caeai.2021.100025