TU Delft

Delft University of Technology

Document Version
Final published version

Licence
CC BY

Citation (APA)
Al Owayyed, M., Tielman, M. L., & Brinkman, W. P. (2026). Simulation-based training for child helpline counsellors: Impact of simulation, feedback, and reflection in a randomised controlled trial. *International Journal of Human-Computer Studies*, Article 103923. https://doi.org/10.1016/j.ijhcs.2026.103923

Important note
To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Copyright
In case the licence states “Dutch Copyright Act (Article 25fa)”, this publication was made available Green Open Access via the TU Delft Institutional Repository pursuant to Dutch Copyright Act (Article 25fa, the Taverne amendment). This provision does not affect copyright ownership.
Unless copyright is transferred by contract or statute, it remains with the copyright holder.

Sharing and reuse
Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

Takedown policy
Please contact us and provide details if you believe this document breaches copyrights.
We will remove access to the work immediately and investigate your claim.

This work is downloaded from Delft University of Technology.

Int. J. Human–Computer Studies 217 (2027) 103923



Contents lists available at ScienceDirect

International Journal of Human - Computer Studies

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



Simulation-based training for child helpline counsellors: Impact of simulation, feedback, and reflection in a randomised controlled trial

Mohammed Al Owayyed [a, b,*], Myrthe L. Tielman [a], Willem-Paul Brinkman [a]

[a] Interactive Intelligence, Delft University of Technology, Delft, 2628XE, the Netherlands
[b] College of Computer Science, King Saud University, Riyadh, 11362, Saudi Arabia

HIGHLIGHTS

- Simulation-only training improves trainees' knowledge.
- Adding guidance feedback to a simulation improves trainees' knowledge and performance.
- Adding a reflection task to simulation with feedback improves reflective ability but slows conversational performance gains.
- Learning outcomes increase across multiple training sessions.
- Across trainees, we observed that they use both task and process level feedback.

ARTICLE INFO

Keywords:
BDI
Child helpline
Simulation-based training
Conversational agents
Feedback
Reflection
RCT

ABSTRACT

Child helplines require well-trained counsellors to support children in need. Such training typically involves role-playing, which is effective but costly and difficult to organise at scale. A promising addition, therefore, is simulation-based training, where trainees interact with a conversational agent that mimics a child contacting the helpline. However, interacting with a simulation alone may not be sufficient, as augmenting it with feedback and reflection can provide more guided learning. The learning effects of sequentially adding these elements—simulation, feedback, and reflection—to training systems are not yet fully understood. In this paper, we extend Lilobot, a conversational BDI-based virtual child, to investigate the effects of these elements on learning outcomes. In a randomised controlled online trial (N = 346), participants were randomly assigned to one of four conditions: no intervention, simulation only, simulation with feedback, or simulation with feedback and reflection. Participants interacted with the interventions in five sessions over fourteen days. Compared with the no-intervention condition, simulation training improved participants' task knowledge and reflective-writing capability. Adding feedback led to further gains in knowledge and conversational outcomes (performance). Additionally, adding reflection on top of this further increased reflective-writing capability, though it dampened improvements in conversational outcomes. These findings clarify how simulation, feedback, and reflection each contribute to learning in early-stage counsellor training.

1. Introduction

Child helplines provide support and advice to adolescents and children on various issues, including mental health, sexuality, bullying, and family problems. Typically, children can access these services by interacting with a counsellor via phone or chat. The international child helpline network reported that its more than 150 member helplines around the world received over 6 million contacts in 2023, more than 2 million of which were counselling contacts (Child Helpline International, 2024). The British Childline receives a contact through its various channels every 45 seconds on average (NSPCC, 2025). Similarly, the Dutch child helpline, De Kindertelefoon, received an average of 1000 contacts per day in 2023 (Stichting De Kindertelefoon, 2023) and 900 per day in 2024 (Stichting De Kindertelefoon, 2024). To cope with this demand, De Kindertelefoon trained more than 640 new volunteers over these two years.

For such interactions to be effective, counsellors are trained to apply counselling protocols and theories. However, training counsellors

* Corresponding author at: Interactive Intelligence, Delft University of Technology, Delft, 2628XE, the Netherlands.
Email address: M.AlOwayyed@tudelft.nl (M. Al Owayyed).

https://doi.org/10.1016/j.ijhcs.2026.103923
Received 11 December 2025; Received in revised form 28 June 2026; Accepted 14 August 2026
Available online 17 August 2026
1071-5819/© 2026 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

typically requires multiple sessions rather than a single interaction. For example, a one-day training workshop on Motivational Interviewing is not sufficient for developing sustained skills or improving client outcomes (Miller and Mount, 2001; Miller et al., 2004). Training counsellors' skills often includes a role-playing setting, where a counsellor takes on the role of the person contacting the helpline (Froehle et al., 1983; Larson and Daniels, 1998). However, reinforcing these skills through role-playing requires repeated practice (Maier, 2002), which is costly, difficult to coordinate, and time-intensive to scale. For instance, both trainers and trainees must be present simultaneously to conduct role-play sessions. An online simulation training tool has the potential to overcome these limitations when incorporated with traditional training. These tools could provide a controlled, readily available, and affordable training environment (Gratch et al., 2016; Mascarenhas et al., 2010; Reger et al., 2020).

Several conversational agent (chatbot) simulations have been developed with similar aims, such as training on interrogation techniques (Bruijnes, 2016), diagnostic skills (Hirumi et al., 2016), or sales-talk skills (Muller et al., 2012). Nevertheless, limited research has focused on training helpline counsellors using conversational agents, especially for training counsellors at children's helplines. Demasi et al. (2020) proposed a chatbot to train suicide prevention hotline counsellors, evaluating how participants perceive the chatbot across several dimensions (e.g., coherence and usefulness). Another example is ClientBot, which is designed to train basic counselling skills, including open questions and reflective statements (Tanana et al., 2019). Simulated child-avatar interview training, particularly in the context of suspected child sexual abuse investigations, has also been shown to improve interview quality (Pompedda et al., 2022). In the specific context of children's helplines, Veldhuizen-Ochodničanová et al. (2023) presented a serious game to train counsellors in skills such as empathy and communication. This system follows a choice-based script in which trainees select from predefined replies, rather than generating their own. As a result, it provides less practice in formulating responses from scratch compared to open-ended interaction.

Before helplines consider integrating such a simulation tool into their curriculum, they need to understand what it should include and how effective it will be. Specifically, this involves identifying which components make a simulation training system effective and how different elements can enhance trainees' skills and performance. The ARTES architecture (Al Owayyed et al., 2025a) identifies agent simulation and educational components as key elements of social skills training systems. Building on this distinction, we examine the incremental addition of three elements: simulation, feedback, and reflection.

In this study, *simulation* refers to open-ended practice with a virtual child that mimics a real-life counselling interaction. *Feedback* then provides system-generated guidance to help trainees interpret the child's state, evaluate their own responses, and identify ways to improve their knowledge and skills. This is complemented by *reflection*, a deliberate written activity in which trainees review what happened during the interaction, what they learned, and how they could improve in future conversations. To examine the incremental contribution of these three elements, we conducted a randomised controlled online trial (N = 346), in which potential new trainees were assigned to a control condition or one of three versions of the tool across five sessions over fourteen days: simulation only, simulation with feedback, or simulation with feedback and reflection. We evaluated the incremental addition by assessing trainees' task knowledge, that is, their understanding of how to apply a communication protocol when counselling the virtual child; reflection capability, that is, their ability to critically review their training experience, explain what they learned, and identify how they could improve; conversational outcomes, that is, their progress in applying the communication protocol during the simulation; trainees' perceptions of the tool; and behavioural patterns associated with different feedback levels.

2. Hypotheses and research questions

2.1. Simulating humans for social skills training

Training with simulations can help people develop important skills. Social skills training systems have been applied in various domains, including aggression de-escalation (Bosse and Provoost, 2014), medical education (Washburn et al., 2020), and job interviews (Anderson et al., 2013). A recent literature review (Ding et al., 2024) found that most of these systems report positive outcomes on measures such as improved conversational skills. However, these systems are often evaluated as complete packages that combine simulation with additional elements such as feedback or guidance, rather than isolating the role of the simulation itself. This distinction is important because it clarifies whether improvements stem from the simulation itself or from elements that enrich and build upon it. Still, studies on simulation-only training systems have found positive outcomes, particularly in how participants perceive and engage with the simulation (Maicher et al., 2017; Dupuy et al., 2020).

Interacting with a simulation can align with the behaviourist perspective (Watson, 2017), which focuses on reinforcing desired learner behaviours through repetition and feedback. At the same time, simulations engage trainees at higher levels of Bloom's taxonomy (Bloom et al., 1956) (applying and analysing) because they require active decision-making, testing of different possibilities, and adaptation of actions based on simulated outcomes rather than simple recall of information. This active engagement encourages trainees to think critically about their actions, evaluate strategies, and reflect on their performance. In this way, simulation provides trainees with a concrete experience of applying the communication protocol in context. Although simulation-only training does not include an explicit reflection activity, it still provides trainees with richer, application-based material for reflection than a theoretical explanation alone. Such deeper application can foster metacognitive abilities (Adams, 2015; Parwata et al., 2023), including an understanding of the task and the quality of reflection (Kitchener, 1983; Knight et al., 2022). This suggests that simulation practice improves not only task knowledge but also individuals' ability to reflect on their experience. Therefore, our first hypothesis is as follows.

Hypothesis 1 (H_1)**.** Compared to those who only received the theoretical explanation, people who interact with a simulation demonstrate (1-a) a higher increase in task knowledge and (1-b) a higher reflection capability.

2.2. Feedback in simulation-based training systems

Learning solely through simulation aligns with a behaviourist perspective, which focuses on enforcing desired learner behaviour (e.g., through repetition) (Watson, 2017). Adding guidance to the simulation introduces a cognitivist perspective, emphasising the development of knowledge through mental processes (Ertmer and Newby, 1993). Blending these perspectives can enhance the learning process (Ertmer and Newby, 1993). Also, incorporating tutoring techniques into the simulation could lead to a more guided learning experience (Kirschner et al., 2006).

Delivering feedback only through the simulation may not be sufficient. For example, feedback embedded in the agent's behaviour (e.g., leaving the conversation) helped participants recognise what had happened and whether mistakes were made, but they still reported a need for more clarification about what went wrong (Nouraei et al., 2025). Several simulation-based training systems have incorporated different forms of feedback that have been shown to enhance trainees' knowledge and performance. This includes delayed feedback provided after the interaction (e.g., Anderson et al., 2013) and immediate feedback delivered during the interaction (e.g., Murali et al., 2021). Beyond timing, prior work has suggested that effective feedback design should also consider what the feedback contains, such as clarifying errors, explaining

consequences, and providing guidance on how to improve (Hattie and Timperley, 2007; Shute, 2008).

Feedback has been shown to improve both cognitive and motivational outcomes (Wisniewski et al., 2020), and maintain long-term training effects (Lamb et al., 2002). One study that examined this addition is the MACH system, a job interview training simulation (Hoque et al., 2013). Participants were assigned to one of three conditions: watching an educational video (control), using MACH and reviewing their own interview video, or using MACH, reviewing their interview video, and receiving automated visual feedback. Human experts rated participants' performance, and results showed higher improvement in the feedback condition compared to both the simulation-only and the control conditions, while no difference was reported between the simulation-only and control conditions. MACH is relevant to this study because it isolates the added value of feedback beyond simulation alone, though in a single-session job-interview setting rather than longitudinal, repeated helpline counselling training. Similar evidence for the value of feedback was found with ClientBot (Tanana et al., 2019), where adding real-time feedback to a simulated patient improved participants' performance compared to using the simulation alone. However, the feedback focused on the trainee's demonstrated skills and on suggesting the next skill to use, rather than on the overall counselling process or session structure (see the task-level and process-level discussion in Section 2.4). Similarly, Røed et al. (2023) found that, in a simulated interview about child physical abuse, training with a child avatar chatbot led to greater improvements in open-ended questioning when paired with direct feedback than when used without feedback.

Although feedback is widely used in agent-based social skills training systems, its design and impact on learning in helpline training simulations are not yet well understood. This is partly because helpline training requires simulating a conversational agent using text only. Altogether, we propose the following second hypothesis:

Hypothesis 2 (H_2)**.** Compared to a simulation only, people who interact with a feedback-integrated simulation demonstrate (2-a) a higher increase in task knowledge, (2-b) a higher reflection capability and (2-c) higher conversational outcomes (performance) over time.

2.3. Reflection moments in simulation-based training systems

Adding reflection moments to training, in the form of self-reflection, could enhance the learning process. This is supported by Kolb's Learning Cycle, which shows the importance of reflection in experiential learning (Kolb, 1984). Several researchers have established links between effective learning and reflection (e.g., Lew and Schmidt, 2011; Watts, 2019). Incorporating self-reflection moments into a training system adds another learning perspective, constructivism, where learners construct their own knowledge (Ertmer and Newby, 1993).

Hoang et al. (2025) showed that incorporating self-reflection prompts into an online clinical skills training improved learners' knowledge acquisition, interpersonal skills, and performance compared to training without self-reflection prompts. Additionally, in a medical education context, Maqsood et al. (2025) found that integrating self-reflection with feedback led to higher learning outcomes than reflection alone, highlighting the importance of combining both. Despite this, self-reflection is not commonly integrated into training simulations.

As mentioned earlier, training systems can improve reflection. Reflection can be further enhanced through repeated opportunities, helping learners develop stronger reflective abilities over time. Therefore, we propose the third hypothesis:

Hypothesis 3 (H_3)**.** Compared to a feedback-integrated simulation, people who interact with a feedback-simulation extended with reflection demonstrate (3-a) a higher increase in task knowledge, (3-b) a higher reflection capability and (3-c) higher conversational outcomes (performance) over time.

Building on H_1–H_3, we expect a cumulative effect, such that the sequential incorporation of simulation, feedback, and reflection leads to progressively higher learning outcomes. Therefore, we propose the following overall hypothesis:

Hypothesis 4 (H_4)**.** The subsequent incorporation of simulation, feedback, and reflection shows an overall (4-a) higher increase in task knowledge, (4-b) higher reflection capability, and (4-c) higher conversational outcomes (performance) over time.

2.4. How trainees use task- and process-level feedback

Although feedback and reflection are useful for training, they can also overwhelm trainees if not carefully designed. According to Hattie and Timperley (2007), feedback can occur at four levels: *task* (correctness of work), *process* (strategies and methods), *self-regulation* (self-monitoring and control), and *self-level* (personal evaluation). In training contexts, *task*-level feedback supports improvement of immediate actions, whereas *process*-level feedback supports understanding of the underlying strategy and structure. These two levels are particularly important because they help students refine tasks and better understand the processes involved. However, cognitive load theory suggests that too much information can overwhelm cognitive resources (Sweller, 2011), which could happen with feedback (Hattie and Timperley, 2007). Learners may also differ in how they value different feedback levels. For example, Brooks et al. (2019) found that students perceived self-regulatory feedback as less helpful, and that students' perceptions of task- and process-level feedback were closely interrelated. Furthermore, Van Boekel et al. (2023) found that students most frequently endorsed task-level feedback when describing the content of their ideal feedback, even though they most often described the purpose of feedback in process-level terms, such as supporting improvement and strategy development. This supports examining feedback-level preferences in training simulations.

Multiple mechanisms influence the levels of feedback learners require during training. Gagné's Hierarchy of Learning presents a bottom-up progression in which learners first acquire prerequisite capabilities, such as stimulus-response, before building on more complex skills, such as problem solving (Gagné, 1985). From this framework, one might infer that feedback should initially focus on lower-level task performance and then shift toward process-level feedback as learners advance. At the same time, some instructional frameworks, such as advanced organisers and meaningful learning theory (Ausubel, 1963), suggest beginning with higher-level understanding (similar to process-level feedback) before focusing on more detailed information (task-level). Existing recommendations and studies on effective feedback (e.g., Hattie and Timperley, 2007; Shute, 2008) often recommend combining the different levels of feedback.

Given these perspectives, people may engage with different levels of feedback in different ways, and specific usage patterns may be associated with better performance. For example, if trainees prefer a certain type of feedback, introducing that type early in the training may be more effective. Previous studies in social skills training have examined the effects of feedback timing (immediate vs. delayed) (Chollet et al., 2022), but not its levels. Here, we therefore propose the following exploratory research questions:

RQ1: How do people use different levels of feedback over time?
RQ2: How does it influence their task knowledge, reflection capability, and conversational outcomes?

3. Method

We conducted a mixed-method experiment with 346 participants: a between-subjects design to test the hypotheses and a within-subjects design to analyse the research questions. All participants watched a video explaining the communication protocol. They were then randomly assigned to one of four conditions: no additional intervention (*C*, i.e.,

Table 1
Summary of participant demographics by condition (control, simulation only, simulation and feedback, simulation, feedback and reflection). Full details are in Table A.6 in the appendices.

Demographic	Control (n = 66)	Simulation (n = 66)	Simulation and feedback (n = 64)	Simulation, feedback, and reflection (n = 150)
Age Average	$M = 31.41$, $SD = 9.58$	$M = 31.50$, $SD = 9.13$	$M = 31.75$, $SD = 8.59$	$M = 31.90$, $SD = 10.24$
Gender	Men: 30 (45.5%) Women: 36 (54.5%)	Men: 28 (42.4%) Women: 38 (57.6%)	Men: 32 (50%) Women: 32 (50%)	Men: 62 (41.3%) Women: 81 (54%) Other: 4 (2.7%) Prefer not to say: 3 (2%)
Education	Community/Technical: 3 (4.5%) Bachelor's: 44 (66.7%) Master's: 15 (22.7%) PhD: 4 (6.1%)	Community/Technical: 3 (4.5%) Bachelor's: 49 (74.2%) Master's: 12 (18.2%) PhD: 2 (3%)	Community/Technical: 6 (9.4%) Bachelor's: 41 (64.1%) Master's: 16 (25%) PhD: 1 (1.6%)	Community/Technical: 13 (8.7%) Bachelor's: 99 (66%) Master's: 31 (20.7%) PhD: 7 (4.7%)
Has Children	Yes: 30 (45.5%) No: 34 (51.5%) Prefer not to say: 2 (3%)	Yes: 32 (48.5%) No: 30 (45.5%) Prefer not to say: 4 (6.1%)	Yes: 36 (56.2%) No: 25 (39.1%) Prefer not to say: 3 (4.7%)	Yes: 82 (54.7%) No: 62 (41.3%) Prefer not to say: 6 (4%)
Experienced Bullying	Yes: 60 (90.9%) No: 4 (6.1%) Prefer not to say: 2 (3%)	Yes: 60 (90.9%) No: 6 (9.1%)	Yes: 57 (89.1%) No: 6 (9.4%) Prefer not to say: 1 (1.6%)	Yes: 132 (88%) No: 15 (10%) Prefer not to say: 3 (2%)
Chatbot Experience	NA	Once or twice: 4 (6.1%) Occasionally: 25 (37.9%) Frequently: 37 (56.1%)	Once or twice: 12 (18.8%) Occasionally: 18 (28.1%) Frequently: 34 (53.1%)	Never: 1 (0.7%) Once or twice: 13 (8.7%) Occasionally: 54 (36%) Frequently: 82 (54.7%)

Control); interaction with the simulation only (*S*); simulation with feedback (*SF*); or simulation with feedback and reflection (*SFR*). Participants in conditions *S*, *SF*, and *SFR* interacted with the training tool in five sessions, each conducted every third day (i.e., with two days between sessions), to assess the effects of interaction over time. Before the experiment, we conducted a pilot study with 16 participants to refine the tool's instructions and reflection questions, as well as to improve how the system recognised participants' inputs. These participants completed all five sessions, and their data were not included in the final analysis. The study was approved by the university's Human Research Ethics Committee (ID: 3946) and was pre-registered prior to data collection.[1]

3.1. Participants

To estimate the required sample size, we conducted a priori power analysis (using Frequentist methods). This resulted in at least 64 participants per condition to achieve statistical significance at the 95% confidence level with 0.8 power, assuming a medium effect size for an independent *t*-test (Cohen, 1992). For the exploratory research questions (i.e., those concerning condition *SFR*), we conducted 1000 simulations of the model using synthetic data to estimate the necessary sample size. These simulations, based on the models we planned to test, suggested that 150 participants would be sufficient to achieve a desired power of 0.8 with 95% confidence. Therefore, we assigned participants using a weighted randomisation (43% *SFR*; 19% each for the other conditions) to ensure target sample sizes of at least 150 in *SFR* and 64 in each of the remaining conditions.

In total, we recruited 738 participants through the online platform Prolific, of whom 346 completed the experiment and comprised the final sample. The inclusion criteria were fluency in English, a Prolific approval rate of at least 95% with five or more completed studies, and holding at least a Bachelor's degree in any field—a prerequisite for counsellors starting training at the Dutch children's helpline. A summary of the demographics per condition is shown in Table 1. Fig. 1 illustrates the participant flow and exclusion reasons per condition. Demographics of excluded participants are reported in Appendix Table A.7. Overall, there were no clear differences in demographic characteristics between excluded participants and the final sample, except that excluded participants were slightly more likely to have children; however, this difference had a small effect size (see the Appendix for details). All participants were compensated at Prolific's minimum payment guideline (£6 per hour) for the first four sessions, and £10 per hour for the final session.

[1] https://osf.io/j9k4u.

3.2. Training system design

To examine the impact of sequentially adding simulation, feedback, and reflection on learning outcomes, we updated Lilobot's simulation to include these components. Therefore, we ended up with three versions of the system: Lilobot alone (condition *S*), Lilobot with feedback (*SF*), and Lilobot with feedback and reflection (*SFR*). In this section, we address the following questions: How did each component work? And what did the overall structure of these components within a learning session look like?

3.2.1. Simulation through lilobot

Lilobot is a virtual child for training counsellors-in-training at helplines (Grundmann et al., 2025). With this tool, a trainee interacts with a conversational agent mimicking a virtual child, whom we call Lilobot. Lilobot simulates children reaching out to a helpline to discuss issues, such as being bullied at school, through an open-ended chat-based channel. Lilobot's simulation was evaluated with experienced counsellors at the helpline; the findings suggested its potential as a supplementary training tool (Grundmann et al., 2025).

The trainee aims to counsel Lilobot by applying communication protocols, such as the five-phase model (Sindahl, 2013), which was developed for children's helplines. The phases of this model are building rapport, clarifying the child's story, defining the conversation goal, working toward the goal, and wrapping up the conversation (Sindahl, 2013). This protocol guides the conversation while ensuring that it remains child-focused. We used this model because it was developed for the child-helpline context and provides a concrete, phase-based protocol that can be taught, practised, and assessed consistently in the simulation and the experiment.

Lilobot's decision-making process is based on a rule-based beliefs-desires-intentions (BDI) model, which has been used to simulate various interactive agents for training purposes (e.g., Bruijnes, 2016; Anderson et al., 2013). In brief, the agent holds several beliefs about itself and the world (e.g., "Lilobot thinks the trainee can be trusted"). These beliefs are influenced by a predefined mapping of the participant's inputs. For example, if the participant tells Lilobot, "If I understand you correctly, you are being bullied, and you want it to stop, is that correct?", Lilobot's trust belief in the trainee increases. Lilobot's desires represent the child's goals during the conversation (e.g., "Lilobot wants to talk about their problem"), while intentions reflect the currently active desire. Lilobot's current beliefs trigger these active desires. The agent then selects an appropriate response based on the participant's most recent input and the active desire from the knowledge base. The conversation can be concluded in two ways: either when the participant successfully applies the five-phase model, or Lilobot perceives the interaction as unhelpful or intimidating. Fig. 2 illustrates an example interaction in

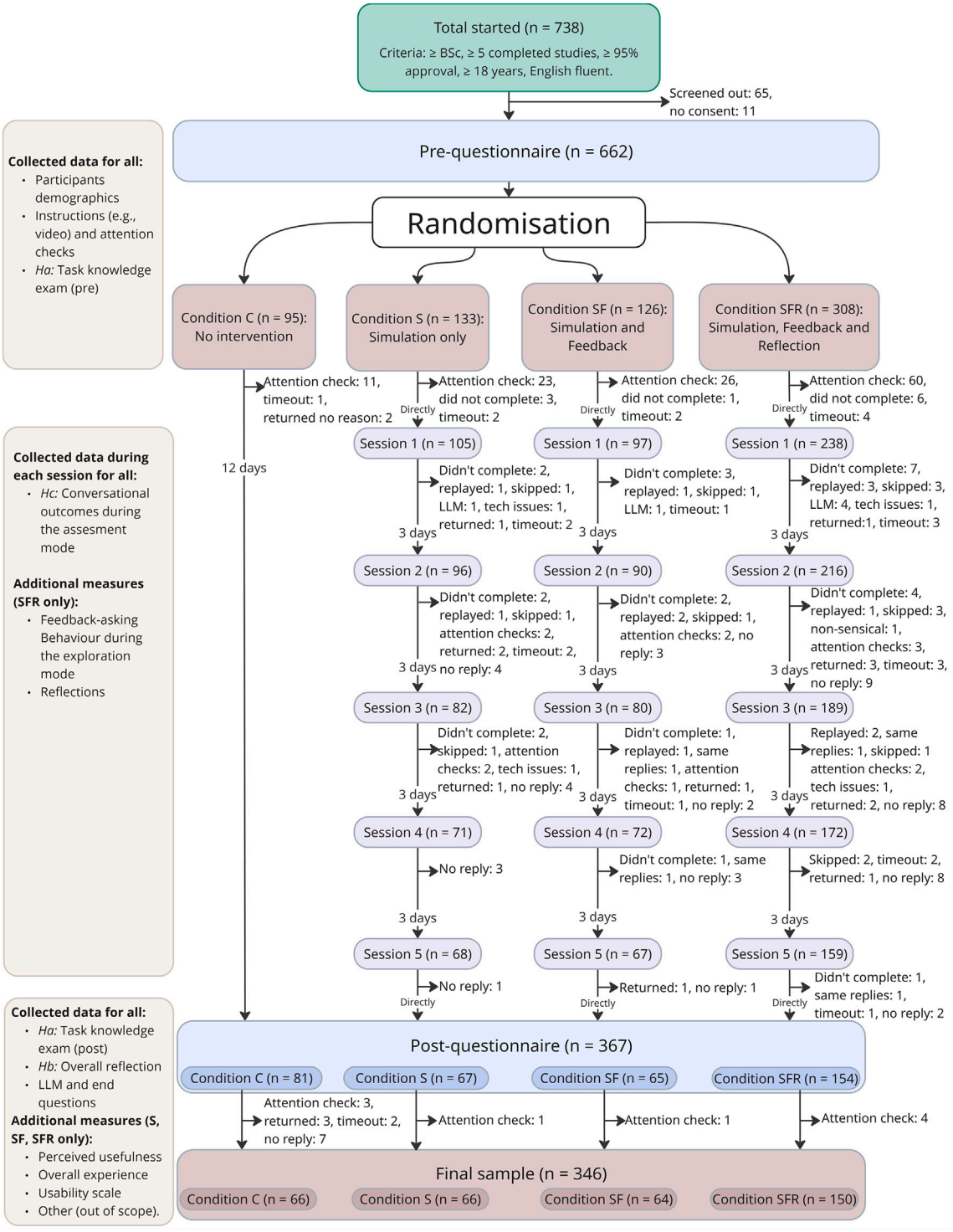


Fig. 1. A Flowchart of participants' progression through the experiment, including recruitment, condition assignment, exclusions, session activities, and measures. Each session followed the structure in Fig. 3. Invitations for the next session were sent three days after each session.

which a participant moves from the end of phase two (clarifying the child's story) to phase three (setting up the conversation goal). A detailed description of the simulation is provided by Grundmann et al. (2025).

Grundmann et al. (2025) introduced only a single scenario in Lilobot: Lilo being bullied at school and asking the trainee (i.e., participant) to call their school to stop the bullying. In the current study, we extended the training to twelve different scenarios, all centred on bullying but varying along two dimensions: the setting (e.g., school, home, neighbourhood, online) and the child's initial goal in the conversation (e.g., asking the participant to call their school, not knowing what to do, or seeking revenge). By introducing multiple scenarios, we hoped to improve the generalisability of both the experimental design and the training outcomes. Relying on a single scenario risked overfitting the design and participants' learning to a specific situation. Moreover, exposure to varied scenarios limited the potential learning effects of repetition, while diversity enhanced training quality. We also introduced scenarios with challenging goals (e.g., wanting to hit back),

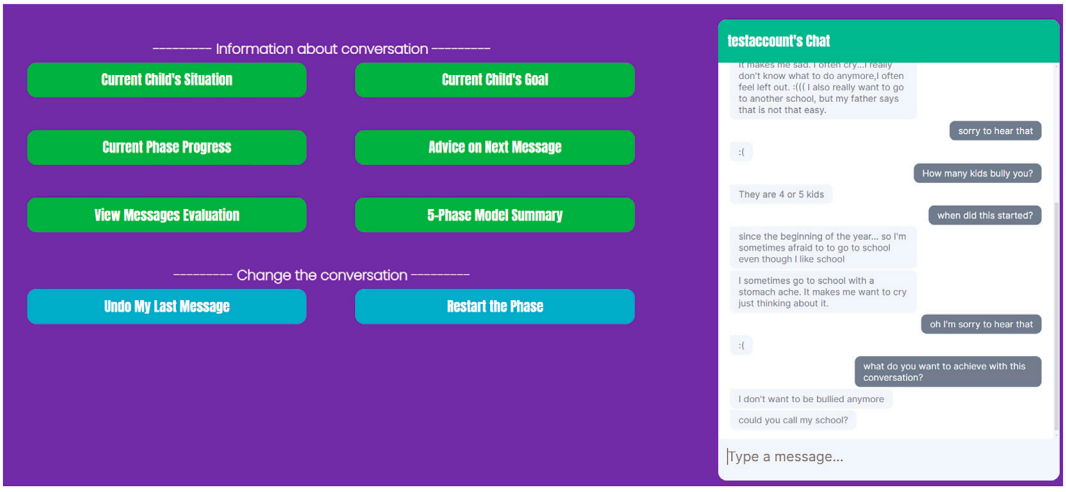


Fig. 2. Screenshot of the training system. The right panel displays a chat simulation with Lilobot during the transition from phase 2 to phase 3 of the five-phase model, while the left panel provides immediate feedback options for the participant to use.

Table 2
Scenario settings of twelve virtual children, including the bullying setting, the wrong goal attributed to the child, and the corresponding child names.

Setting	Challenging Goal			No Goal	Non-challenging Goal
	Call School/ Parents	Stop Going/ Playing	Hit Back/ Revenge		
School break	Lilo				Miles
School music					Maya
School class		Daisy			
Sports club			Tim	Jim	
At home				Lyla	
Neighborhood	Alex				Ori
Online gaming		Kai			
Social media			Ava		
Playdate				Dania	

non-challenging goals (e.g., the child is ready to talk), and no goal (e.g., not knowing what to do), so that participants would not develop biases towards a particular scenario during training. The scenarios focused primarily on bullying because it is multifaceted (Kaufman et al., 2018), can take many forms (Cameron et al., 2023), and is a frequent issue in helpline contexts (Mathieu et al., 2021), while still allowing for grounded and comparable solutions across cases. Table 2 presents the twelve virtual child scenario settings and their names.

To generate the virtual children's responses, we first provided GPT-4o with utterances from the original scenario in (Grundmann et al., 2025), along with a new scenario setup and child persona, and asked it to produce corresponding utterances for the new scenario. These dialogues were then refined into rule-based responses. We applied the same process across all scenarios, and experts reviewed a subset of the responses and conversations. We also set the BDI values to reflect the specific settings and goals of the virtual children. Although the BDI values varied slightly across scenarios, we expected the underlying knowledge gained from interaction to be transferable. Participants, therefore, encountered different interactions within the same session to avoid repeating the same inputs.

3.2.2. *Delivering feedback*

The system provided feedback to guide participants in improving their performance and gaining a better understanding of the situation. In the system, we implemented feedback at only the two aforementioned levels: task-level, which focused on the current state and actions within the interaction, and process-level, which related to the broader structure and strategy of the interaction. The feedback was grounded in the five-phase model and was designed to be interpretable, explicitly linking the participant's actions to the child's behaviour.

The system delivered feedback both immediately during the interaction and after the conversation had ended, as combining both approaches has been recommended for effective learning (Chollet et al., 2022; Hattie and Timperley, 2007). It addressed the three feedback questions defined by Hattie and Timperley (2007) by making expectations explicit (feed up), assessing where the interaction fell short (feedback), and offering suggestions on how to move forward (feed forward). This shifted participants' focus from guessing what to do next toward understanding which strategies were effective and why. We designed the feedback following guidelines on formative feedback (Shute, 2008; Wiggins, 2012). This feedback operated independently of specific scenarios, focusing instead on the applicability of the five-phase model and on how the child's behaviour and thinking changed. Appendix B.1 includes examples of immediate and delayed feedback taken from the system.

Immediate feedback. During the interaction with the virtual child, participants could request immediate feedback through a set of feedback options (Fig. 2). These options provided guidance on the child's state, corrective input on the participant's performance, and control over the progress of the conversation. Eight options were available: three at the task level, four at the process level, and one providing general information. Table 3 lists these feedback options, their descriptions, levels, and focus areas. Immediate feedback was learner-controlled rather than assigned by the system: throughout the chat interaction, participants could access any of the eight feedback options at any time and as often as needed, with no restrictions on which options they selected or in what order, thereby determining the feedback they received. Through this design, we aimed to make information easier to process and more manageable by giving participants agency over when and how to receive support. Prior work suggests that learner-controlled feedback, particularly in terms of timing and performance, is effective in supporting learning outcomes (Chiviacowsky and Wulf, 2002, 2005). Furthermore, not all trainees are alike, and they might use feedback differently (Carless, 2020). This design also allowed us to observe when participants chose task- or process-level feedback.

Delayed Feedback. Participants automatically received delayed feedback after completing an interaction in the guided session. This feedback covered both task and process levels, highlighting potential shortcomings during the conversation (e.g., not showing enough empathy) and their consequences (e.g., the child leaving due to a lack

Table 3
Description of the immediate feedback options in the training system and how they guide trainees (i.e., participants) during training. Each feedback option provides feedback at either the task or process level. Table B.8 shows how the feedback logic was implemented.

Feedback Option	Description	Level	Focus
Current Child's Goal	Displays what the child wants to achieve in the conversation (i.e., their desire from the BDI).	Process	Child's thinking
Current Child's Situation	Explains what the child currently thinks or feels based on their beliefs in the BDI.	Task	Child's thinking
Current Phase Progress	Shows the trainee's progress in the current phase (percentage completed) and indicates what is needed to proceed to the next phase.	Process	Trainee's interaction
Advice on Next Message	Provides a hint about what the trainee should type next in the interaction.	Task	Trainee's interaction
View Messages Evaluation	Displays which trainee messages might work better in a different phase. Shows the message, the phase it was used in, and the phase where it would be more appropriate.	Process	Trainee's interaction
5-Phase Model Summary	Provides a summary of the five-phase model, which remains the same throughout the training session.	Overall summary	Static information
Restart the Phase	Resets the entire current phase of the five-phase model. This deletes all messages exchanged during the current phase and restores the child's BDI to its state at the beginning of that phase.	Process	Conversation
Undo My Last Message	Deletes only the most recent trainee message and the corresponding virtual child reply, then restores the child's BDI to the state just before that exchange.	Task	Conversation

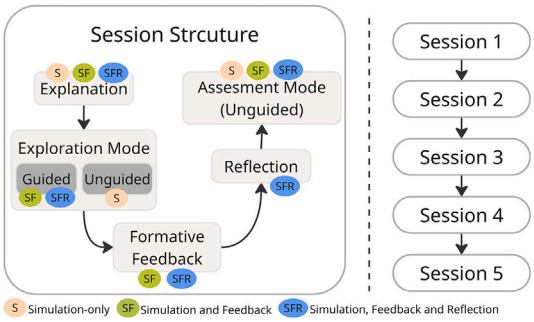


Fig. 3. Structure and flow of a learning session within the training system, highlighting moments for exploration (exploring the simulation with or without feedback), reflection, and unguided assessment (assessing performance with the simulation only). Each session of the five sessions had the same structure. The design follows Kolb's experiential learning cycle (Kolb, 1984) and Brookhart (2017). Each session maintained the same overall structure, while the specific content varied across the three conditions (*S*, *SF*, and *SFR*).

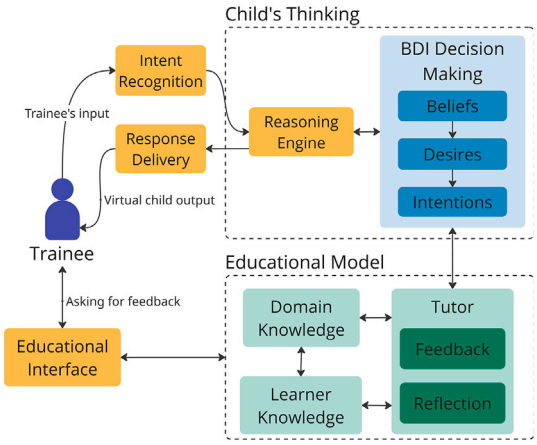


Fig. 4. System architecture based on the ARTES framework (Al Owayyed et al., 2025a). It integrates the virtual child's thinking (BDI-based reasoning and response generation) with the educational model, where the tutor component provides feedback and reflection moments informed by learner knowledge, the BDI model, and domain knowledge.

of trust). It also provided suggestions for improvement in subsequent interactions, such as asking more clarifying questions.

3.2.3. Trainees' reflections

After completing the interaction and receiving feedback, participants were asked to reflect on the interaction, the feedback, and how these elements shaped their overall learning process. The focus was mainly on their reflections about the exercise itself and the learning outcomes of the conversation. We followed Ash and Clayton's (2004) recommendation to write reflective questions that guided learners to consider what they learned, how they learned it, why it mattered, and how they could improve.

3.2.4. A training session flow

A model illustrating the structure of a learning session with the training system is presented on the left side of Fig. 3. This model is based on Kolb's learning cycle (Kolb, 1984) and the Formative Learning Cycle, which integrates feedback, by Brookhart (2017) (adapted from (Moss and Brookhart, 2012)). The session began with an explanation of the task. This was followed by an exploration mode, in which participants applied the knowledge gained from the explanation step. The purpose of the exploration mode was to provide a task that resembled real interactions while allowing participants to explore without performance assessment, thereby helping them understand and apply the concepts. In the *S* condition, participants received only the simulation and were asked to explore. In the *SF* and *SFR* conditions, they interacted with a simulation that incorporated immediate guiding feedback and received delayed feedback based on their performance during practice. Participants in the three conditions completed the exploration mode twice using the same scenario. In the next step, participants in the *SFR* condition reflected on their learning during the exploration mode. This reflection was meant to help them review and consolidate what they had learned, deepening their understanding of the task. Afterwards, the participants in the three conditions completed the assessment mode, which evaluated their performance as they interacted with the simulation without guidance to apply their knowledge independently.

To bring all components together, Fig. 4 presents an architecture based on the ARTES architecture (Al Owayyed et al., 2025a). The virtual child's thinking included the simulation parts, which controlled the BDI model and the reasoning process (explained in detail in (Grundmann et al., 2025)). The educational model included the tutor, which provided guiding feedback and determined when to introduce reflective moments. To deliver feedback and guidance, the tutor drew on three sources: the learner's knowledge component, to assess the participant's current knowledge and performance; the BDI decision-making, to obtain the virtual child's state and goals; and the domain knowledge component, to extract relevant feedback from the learning material (e.g., the five-phase model).

3.3. Measures

We classified the measures as primary and secondary. We used the primary measures to test the hypotheses and address the research

questions, and the secondary measures to explore participants' perceptions of the conditions. All questionnaire items and rubrics are referenced here or provided in Appendix C.

3.3.1. Primary measures

Task Knowledge (a). This measure assessed participants' knowledge of applying the five-phase model. We developed two versions of a knowledge assessment exam, based on training materials from the Dutch children's helpline and a handbook on the five-phase model (Sindahl, 2013). Each exam version contained different multiple-choice questions, with a maximum score of twelve points. The questions were designed to assess one of four levels in Bloom's taxonomy (Bloom et al., 1956) (remembering, understanding, applying, and analysing), and this structure was the same across both versions. For each participant, the order of the two exam versions was randomly assigned to serve as pre- and post-tests. After each exam, participants rated their confidence in their answers on a scale from 0 (not confident at all) to 10 (extremely confident), to provide additional insights.

Reflection Capability (b). We used this measure to compare participants' reflections between conditions at the end of the experiment. Following a similar approach to Moussa-Inaty (2015), who categorised participants' reflections into levels, we assigned reflection scores to participants. To derive these scores, we developed a rubric adapted from the levels of reflection by Fleck and Fitzpatrick (2010); Hatton and Smith (1995). Instead of treating the levels as a hierarchy, we used them as five dimensions of reflection: revisiting, explanation, exploring relationships, fundamental changes, and wider implications. Each dimension was scored from 0 to 3, depending on whether no statement was present (0), or one, two, or three or more statements were present in the participant's written reflection. The maximum total score was 15 points.

Conversational Outcomes (c). The performance measure assessed how successful participants were at applying the five-phase model during their conversations. It represents a score for the participant's interaction with a virtual child in an unguided assessment task, resulting in a separate score for each of the five sessions. To measure the score, we calculated the deviation of the virtual child's belief values from the target belief values at the end of the conversation, that is, what the child's BDI values should be when a trainee applies the five-phase model correctly. In the BDI model, all belief values range from 0 to 1. For binary beliefs (restricted to 0 or 1), a fixed penalty of 0.2 was applied instead of a variation score when the condition to change the belief was not met. All deviations were summed to yield a total score. The optimal deviation score is zero, indicating no divergence from the targeted conversational path. Higher scores indicate greater divergence from the targeted path. Participants were not penalised for achieving better belief values (e.g., showing more empathy than required), as this still demonstrated correct application of the model. This measure is therefore not intended as a direct estimate of real-world counselling quality, but as a proxy for whether trainees can apply the five-phase model to move virtual children toward desirable conversational states. Because the system was implemented across twelve different scenarios, the score reflects application of the protocol rather than performance on a single fixed scenario.

Feedback-Asking Behaviour (RQ1-RQ2). We tracked participants' use of the eight feedback options (Fig. 2) throughout the experiment. In each guided interaction, we recorded how often each feedback option was selected. This resulted in a count of how many times an option was selected per session. To ensure consistency in the within-subject analysis, we collected this measure only in the condition with all features (simulation, feedback, and reflection).

3.3.2. Secondary measures

We examined secondary measures to describe participants' perceived experience. In addition, we collected data outside the scope of this study (e.g., participants' learning strategies, self-efficacy, and perceived values from the virtual child), which we leave for future research (see preregistration for details).

Perceived Usefulness. This measures participants' perception of the training's usefulness. We developed a six-item questionnaire focusing on how useful the information provided by the tool was and how it supported the application of the five-phase model. The questionnaire was adapted from the performance expectancy construct in the Unified Theory of Acceptance and Use of Technology model (Venkatesh et al., 2003) and from the work of Fitrianie et al. (2021). Participants rated each item on a 7-point Likert scale ranging from 1 (strongly disagree) to 7 (strongly agree), with intermediate points at 2 (disagree), 3 (somewhat disagree), 4 (neither agree nor disagree), 5 (somewhat agree), and 6 (agree). An example item is: "The exploration mode provides me with information about the five-phase model".

Overall Experience with the Agent. This qualitative measure assessed how participants perceived their interaction with the agent. We used the 24-item short version of the Artificial Social Agent Questionnaire (ASAQ) (Fitrianie et al., 2022, 2025). Following the authors' recommendations, responses were rated on a scale from -3 (Disagree), through 0 (Neither agree nor disagree), to $+3$ (Agree). In all items, we used the term "virtual child" to refer to the agent.

Usability of the Training System. We measured the usability of the training system using the System Usability Scale (SUS) questionnaire (Brooke et al., 1996). Ratings were made on a 5-point Likert scale: 0 (strongly disagree), 1 (somewhat disagree), 2 (neither agree nor disagree), 3 (somewhat agree), and 4 (strongly agree). In the questionnaire, we referred to the system as "the training system."

3.4. Procedure

Participants were screened to ensure they met the recruitment criteria. Upon confirmation, they were redirected to Qualtrics, an online survey platform, to answer demographic questions. They were then presented with the learning objective of the experiment: "This experiment will train you to apply the five-phase model, a conversational strategy that new counsellors in child helplines are trained on." They also received a short description of what they would do during the experiment. Participants then watched an educational video on the five-phase model, which explained the model's merits and applications. Afterwards, participants completed an attention check test to verify whether they had watched and understood the video. If a participant failed this test, they had to rewatch the video and retake the attention test. We excluded participants who failed on their second attempt. Following the attention checks, participants completed one version of a knowledge assessment exam, which was randomly assigned, to measure their baseline knowledge of the five-phase model before the experiment.

Participants were then assigned to one of four conditions: no intervention (*C*), simulation only (*S*), simulation with feedback (*SF*), or simulation with feedback and reflection (*SFR*). Participants in the no-intervention condition (condition *C*) did not receive the simulation-based training. For the intervention conditions (*S*, *SF*, and *SFR*), participants received instructions on how to interact with the tool, depending on their assigned condition. For example, participants in the simulation-only (*S*) condition did not receive instructions on the feedback options and what each option was intended to do. We also explicitly instructed participants in *S*, *SF*, and *SFR* not to use large language models (LLMs) to interact with the virtual child. After completing comprehension checks, participants began the first training session.

Each session began with an exploration mode and ended with an assessment mode, as explained in Section 3.2.4. The modes were clearly labelled to participants as the "exploration" and "assessment" modes. In the exploration mode, participants interacted with a virtual child either without feedback (condition *S*) or with feedback options in a guided interaction setting (conditions *SF* and *SFR*). They completed two interactions with the same virtual child scenario within this mode. Participants in conditions *SF* and *SFR* received immediate on-demand feedback

during the interactions and delayed feedback after each interaction; both conditions used the same feedback mechanisms, as explained in Section 3.2.2. We recorded participants' use of feedback options during the exploration mode for condition *SFR*. After the exploration, participants in condition *SFR* were asked to reflect on their learning experience. Next, all participants in conditions *S*, *SF*, and *SFR* moved to the assessment mode, where they engaged in an unguided interaction (without feedback options) with a new virtual child scenario, applying the five-phase model. After completing this mode, the interaction session ended. At the end of the experiment session, participants in all four conditions (including *C*) filled out a self-efficacy questionnaire (outside the scope of this study) and had the option to provide further comments on the session.

After completing a session, participants were invited to the next session, reassigned to the same condition as before, and followed the same procedure. We invited participants in condition *C* to complete post-questionnaires again after twelve days, with 48 hours to complete the session (i.e., fourteen days in total). In the other three conditions, each participant completed five sessions over fourteen days. For each session, participants had up to 48 hours to begin after receiving the invitation and were required to complete it in a single sitting without a timeout. After the 48-hour window, there was a 24-hour break before participants were invited to the next session.

Within the five sessions, we randomised two factors to reduce potential bias and counterbalance order effects. First, participants received 10 out of 12 virtual child scenarios in a random order, such that each scenario was shown only once (two scenarios per session: one in the exploration mode and one in the assessment mode). Second, for participants in the feedback conditions (*SF* and *SFR*), we randomised the order position of the feedback option buttons (Fig. 2) between participants, but kept the order identical for each participant across all five sessions. To ensure coherence, relevant options were paired together according to their focus (Table 3; e.g., "Undo My Last Message" and "Restart the Phase").

Upon finishing the fifth session, participants completed the post-questionnaires. Here, they first took the alternate version of the knowledge assessment exam as a post-exam. They were then asked to reflect on their overall learning experience and complete the secondary measures questionnaires. Before concluding the experiment, participants were asked two quality control measures: whether they had used an LLM during the experiment (and if so, why), and whether they recommended that their data be used for scientific purposes. Fig. 1 provides an overview of the experiment's procedure and the measures used at each stage.

3.5. Data analysis

For quality control measures, all participants recommended using their data for scientific purposes. Regarding the question of whether they had used LLMs, twelve participants reported using LLMs to help with brainstorming during the chat (n = 2), writing reflections (n = 6), getting information about the five-phase model (n = 2), answering the knowledge test (n = 1), and translation (n = 1). We included their data for analysis, as they did not use LLMs to generate responses to the virtual child.

To obtain a final score for the questionnaire measures, we calculated the mean scores per scale, reverse-coded items where appropriate, and followed the guidelines provided for the questionnaires. The conversational outcomes score reflects deviation from the optimal path, where higher scores indicate worse performance. For easier interpretation and analysis, we reversed the score and transformed the data into a performance index using a linear transformation based on the possible score range, i.e., $(\text{max score} - \text{observed score})/\text{score range} \times 100$. This index is presented on a percentage scale: 0 represents no performance improvement, i.e., the child's starting position, and 100 represents the best performance, correctly applying all learning goals. For the SUS questionnaire, one item was not collected due to an administrative error ("*I thought the system was easy to use*"). We corrected this by multiplying the resulting score of the other nine items by 2.78 (instead of 2.5) to obtain a final score out of 100. For task knowledge and reflection, item scores were summed to form total scores, ranging from 0 to 12 and from 0 to 15, respectively.

We used a local LLM (LLaMA 3.1 8B Instruct model) to evaluate participants' reflections, given the large number of reflections (346 reflections). Our approach was inspired by Chiang and Lee (2023), who tested LLMs to score open-ended story texts using dimensions such as likeability, relevance, and cohesiveness. They found the results to be consistent between human experts and LLM evaluations. We also followed Bavaresco et al.'s (2024) recommendation to check the reliability of LLM evaluations against human evaluators. To do this, we prompted the LLM with the scoring rubric, explanations of the rubric dimensions, and five example reflections that included human-assigned scores and justifications. After receiving this prompt, both the LLM and the first author independently scored a different set of 100 reflections to assess consistency between human and model ratings. The resulting LLM scores showed a correlation of $r = 0.80$ with the human-coded final reflection scores, which is classified as a high correlation by Hinkle et al. (2003). To further check for individual biases, a second coder (a PhD candidate in computer science) independently scored the same 100 reflections after training on the five examples that were given to the LLM. Agreement between the first and second human coders was 0.77 (high correlation), while agreement between the LLM and the second coder was 0.69 (moderate correlation), following the same classification (Hinkle et al., 2003). Given this reliability, we used the LLM to score all the reflections in our analysis.

First, we checked whether people improved over time. We analysed the within-condition changes in task knowledge (pre- vs. post-test) and conversational outcomes (first session vs. last session) using Bayesian paired-samples *t*-tests. As for hypothesis testing between conditions, we also analysed the data using Bayesian methods. To predict the task knowledge post-test, we fitted a multilevel model with fixed effects for condition and pre-test scores, and random intercepts for participants. For conversational outcomes collected across multiple sessions, we tested the added value of each condition per session. Thus, we fitted a multilevel model with fixed effects for session, condition, and their interaction, and random intercepts for participants to account for the repeated measures. To compare the overall increase in conversational outcomes between conditions, we computed the delta per participant (i.e., the difference between the indices of the first and last sessions). Then we fitted a model to assess how the conditions predict the index delta for the conditions. For reflection and other post-only measures (e.g., perceived usefulness), we fitted linear models with condition as a fixed effect. To test for an overall increase with the subsequent addition of elements (H_4), we fitted the same models using an ordered categorical predictor with a monotonic effect for conditions. Regarding the two exploratory research questions, we examined trends in feedback options and performance over time, as well as the associations between feedback option usage and learning outcomes.

For all Bayesian analyses, we used uninformative priors and interpreted posterior probabilities to evaluate hypotheses, following the guidelines of Chechile (2020); Andraszewicz et al. (2015). We report 95% Highest Density Intervals (HDIs) to indicate the range of credible parameter values. We inferred effect size probabilities based on the Sequential Effect eXistence and sIgnificance Testing (SEXIT) framework (Makowski et al., 2019a). Following the framework, we report the proportion of the posterior distribution lying beyond predefined thresholds for "small" and "large" effects (set at |0.10| and |0.30| of the outcome's standard deviation, respectively) (Funder and Ozer, 2019; Makowski et al., 2019a). We conducted all analyses in R (version 4.3.1) using the `brms` package (Bürkner, 2017) for multilevel modelling, the `bayestestR` package (Makowski et al., 2019b) to calculate HDIs and effect sizes, and the `BayesFactor` package (Morey and Rouder, 2018) for Bayesian

Table 4
Descriptive statistics and Bayesian within-condition test results for task knowledge (pre- vs. post-test) and conversational outcomes (session 1 vs. session 5) across experimental conditions.

Condition	Measure	% Posterior Probability > 0	Interpretation	Estimate [95% HDI] (SD)	% Effect Size >	
					Small	Large
C	Task Knowledge M_{pre} = 6.41 (1.66) M_{post} = 5.92 (1.84)	0.03	Good bet against - too good to disregard	−0.48 [−0.98, 0.01] (0.25)	0.86	0.30
	Reflection Capability M_{post} = 3.67 (3.01)	–	–	–	–	–
S	Task Knowledge M_{pre} = 6.70 (1.55) M_{post} = 7.06 (1.95)	0.91	A promising but risky bet	0.36 [−0.18, 0.88] (0.27)	0.70	0.13
	Conversations M_{s1} = 41.09 (20.04) M_{s5} = 45.50 (22.03)	0.92	A promising but risky bet	4.34 [−1.79, 10.34] (3.09)	0.73	0.15
	Reflection Capability M_{post} = 4.97 (3.61)	–	–	–	–	–
SF	Task Knowledge M_{pre} = 6.38 (1.73) M_{post} = 7.38 (1.90)	0.9998	Nearing certainty	0.99 [0.46, 1.52] (0.27)	>0.99	0.89
	Conversations M_{s1} = 46.64 (20.57) M_{s5} = 64.13 (23.87)	>0.99995	Virtually certain	17.26 [10.56, 24.10] (3.45)	>0.99	>0.99
	Reflection Capability M_{post} = 4.97 (2.80)	–	–	–	–	–
SFR	Task Knowledge M_{pre} = 6.71 (1.73) M_{post} = 7.39 (2.00)	0.9996	Nearing certainty	0.68 [0.28, 1.07] (0.2)	0.98	0.37
	Conversations M_{s1} = 46.64 (22.95) M_{s5} = 58.56 (26.70)	>0.99995	Virtually certain	11.85 [6.58, 17.07] (2.68)	>0.99	0.77
	Reflection Capability M_{post} = 5.43 (3.50)	–	–	–	–	–

Note. C = Control; S = Simulation only; SF = Simulation and Feedback; SFR = Simulation, Feedback and Reflection. Reflection capability was measured at post-test only; therefore, no within-condition change statistics are reported.

t-tests. The data, R code for the analysis, generated markdown reports, Docker image for reproducibility, and prompts are publicly available via the 4TU Research Data Repository.[2] The system code is also publicly available on GitHub.[3]

4. Results

4.1. Hypotheses results

Before testing the hypotheses, we examined whether the four conditions showed improvement over time. Table 4 presents the results for the two primary repeated measures (task knowledge and conversational outcomes) within each condition. The table also shows the means and standard deviations for all three measures in each condition. A summary of the four hypothesis results is presented in Table 5, including the probability that each hypothesis is true, its interpretation, and the likelihood that the observed effect sizes are small or large.[4]

Within-condition results: Posterior probabilities in Table 4 show credible support that participants in the base control condition (*C*) decreased in task knowledge over time, with a 30% probability of a large effect size. The condition *S*, which involved only the simulation, showed some support for improvement: there was a 91% posterior probability that task knowledge increased over time, with a 13% probability of a large effect size, and a 92% posterior probability for an increase in conversational outcomes over time, with a 15% probability of a large effect size. For participants in conditions *SF* and *SFR*, the results show credible support for improvements over time in both task knowledge and conversational outcomes. For *SF*, the probabilities of a large effect size are 89% and 99.45%, respectively, for task knowledge and conversational outcomes. For *SFR*, the probabilities of a large effect size are 37% for task knowledge and 77% for conversational outcomes.

***H_4*: Overall: subsequent incorporation leads to higher measures:** We begin with this hypothesis to test whether there is an overall effect of sequentially adding components. Across all measures, the results showed credible support for the subsequent incorporation of components (Table 5). The corresponding posterior probabilities (task knowledge: > 99.99%, reflection capability: 99.93%, and conversational outcomes: 96%) provide evidence for overall positive effects. Hence, H_4 is supported. Regarding the likelihood of a large effect size, it is almost certain for task knowledge (99%), somewhat lower for reflection capability (87%), and lower for conversational outcomes (35%), for which a small effect appears more likely (81%).

***H_1*: S is higher than C:** Participants in the simulation-only condition showed higher improvement than those in the control condition, both in task knowledge and reflection capability (Table 4). As shown in Table 5, the posterior probabilities (99.9% and 98.9%) provide credible support that the *S* condition improved both measures more than the *C* condition, supporting H_1. The likelihood of large effect sizes also supports this outcome, with 92% for task knowledge and 70% for reflection capability.

***H_2*: SF is higher than S:** Table 5 shows 89% posterior probability that participants in the simulation with feedback condition (*SF*) had higher task knowledge than the simulation only (*S*) condition. Reflection capability was similar between the two conditions, with a posterior probability of 50%. The conversational outcomes showed credible support that the *SF* condition improved more across sessions than

[2] https://data.4tu.nl/datasets/635d08b5-814e-45d3-99c1-c330f3c9813d.
[3] https://github.com/alowayyedm/LilobotTraining/tree/BDIFeedbackFinal.
[4] All model comparison results and additional analyses are provided in Table D.11 in the appendices.

Table 5
Summary of Bayesian test results for the four hypotheses, comparing conditions across the three outcome measures: task knowledge, reflection capability, and conversational outcomes.

Hypothesis	Measure	% Posterior Probability > 0	Interpretation	Estimate [95% HDI] (SD)	% Effect Size >	
					Small	Large
***H_4*: Overall Increase**	a: Task Knowledge	>0.99995	Virtually certain	1.38 [0.82, 1.92] (0.28)	>0.99	>0.99
	b: Reflection Capability	0.99935	Very strong bet	1.56 [0.62, 2.52] (0.49)	0.99	0.87
	c: Conversations Overall	0.93	A promising but risky bet	7.12 [−2.34, 16.13] (4.73)	0.81	0.35
	c: Conversations per Session	0.96	Good bet for - too good to disregard	1.22 [−0.14, 2.24] (0.60)	0.01	<0.01
***H_1*: S > C**	a: Task Knowledge	0.9992	Very strong bet	1.05 [0.41, 1.70] (0.33)	>0.99	0.92
	b: Reflection Capability	0.989	Good bet for - too good to disregard	1.31 [0.18, 2.44] (0.58)	0.95	0.70
***H_2*: SF > S**	a: Task Knowledge	0.89	Only a casual bet	0.41 [−0.26, 1.05] (0.33)	0.72	0.26
	b: Reflection Capability	0.5	Not worth betting	−0.006 [−1.14, 1.10] (0.58)	0.28	0.04
	c: Conversations Overall	0.9937	Strong bet	13.11 [2.85, 23.30] (5.22)	0.98	0.78
	c: Conversations per Session	0.9954	Very strong bet	2.99 [0.72, 5.25] (1.16)	0.68	<0.01
***H_3*: SFR > SF**	a: Task Knowledge	0.40	Not worth betting	−0.07 [−0.63, 0.47] (0.28)	0.32	0.03
	b: Reflection Capability	0.83	Only a casual bet	0.47 [−0.49, 1.44] (0.49)	0.60	0.14
	c: Conversations Overall	0.11	Only a casual bet against	−5.54 [−14.00, 3.68] (4.50)	0.71	0.22
	c: Conversations per Session	0.14	Only a casual bet against	−1.06 [−2.92, 0.92] (0.98)	0.08	<0.01

Note. C = Control; S = Simulation only; SF = Simulation and Feedback; SFR = Simulation, Feedback and Reflection.

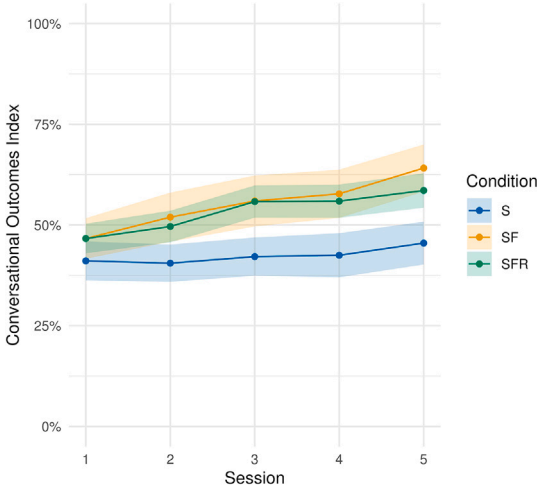


Fig. 5. Average conversational outcome deviation indices per condition across sessions. The measure was collected during each assessment mode interaction, yielding one index per session per participant (see Fig. 3). Shaded areas represent 95% credible intervals.

the *S* condition, with a 99.5% posterior probability and an estimated average difference of 2.99 per session (95% HDI [0.72, 5.25]). This was further supported by a 78% likelihood that the overall increase in conversational outcomes reflects a large effect size. The pattern is also illustrated in Fig. 5, where the average indexes per session for the *SF* condition remain consistently better than those for *S*. Overall, these results provide credible support for an effect of adding feedback to simulation on conversational outcomes, but only some support for additional improvements in task knowledge, supporting H_2.

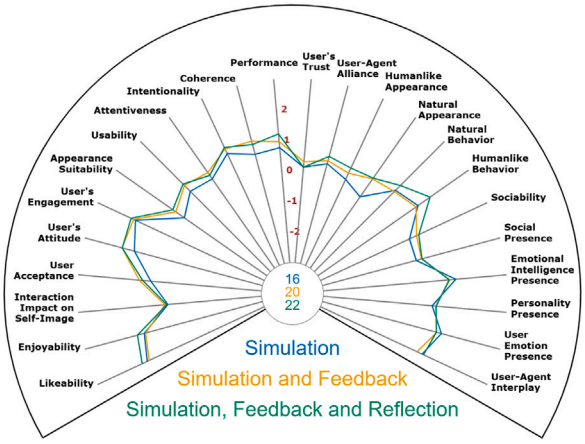
***H_3*: SFR is higher than SF:** As shown in Table 5, participants in the simulation with feedback and reflection condition (*SFR*) performed differently from those in the simulation with feedback condition (*SF*) in some measures. For task knowledge, the posterior probability is 40%, interpreted as not worth betting on. Reflection capability showed some support for improvement in the *SFR* condition, with an 83% posterior probability and a 14% likelihood of the effect size being large. In contrast, conversational outcomes provided some support against the *SFR* condition, with a 14% posterior probability (only a casual bet against; 86% support for the opposite) that *SFR* increases conversational outcomes compared to *SF*. This pattern is also visible in Fig. 5, where participants in *SF* scored on average higher than those in *SFR* in the final session. Taken together, these findings suggest that adding reflection to feedback improved reflection capability (some support for H3b), but led to a reduction in conversational outcomes improvements (some support against H3c), and no additional task knowledge improvement (no support for H3a). Therefore, H_3 as a whole does not seem to hold.

4.2. Secondary results

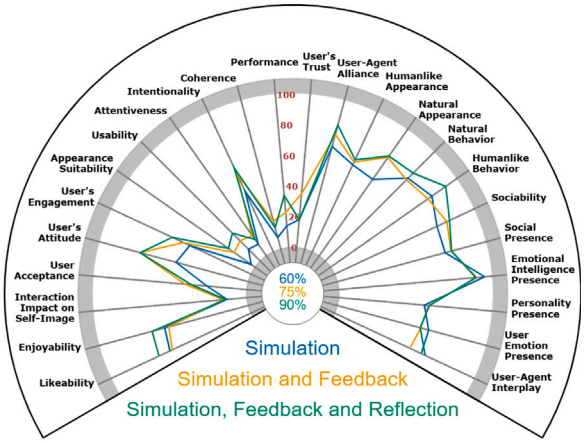
For the secondary analysis, we asked participants in the conditions *S*, *SF*, and *SFR* to rate their perceived usefulness of the training tool, their overall perceived experience with the agents (through the short version of ASAQ), and the usability (SUS) of the system. Participants perceived the usefulness of the three conditions differently (*SFR:* $M = 6.16$, $SD = 0.78$; *SF:* $M = 6.12$, $SD = 0.83$; *S:* $M = 5.36$, $SD = 1.01$). The analysis shows credible support that adding feedback increases perceived usefulness (posterior probability: > 99.99%, 95% HDI [0.46, 1.04]), but no support for adding reflection to feedback (posterior probability: 66%, 95% HDI [−0.2, 0.29]). Next, the overall experience ratings (ASAQ) also differed (*SFR:* $M = 0.91$, $SD = 1.2$; *SF:* $M = 0.83$, $SD = 1.21$; *S:* $M = 0.68$, $SD = 1.03$), with some support for adding feedback (posterior probability: 78%, 95% HDI [−0.24, 0.56]), and no support for adding reflection (posterior probability: 68%, 95% HDI [−0.25, 0.44]). Fig. 6(a) shows how the three conditions compare across each construct of ASAQ, and Fig. 6(b) compares the three conditions to the 29 artificial social agents in the ASAQ Representative Set 2024 (Fitrianie et al., 2025). The *S* condition scored with overall ASAQ higher than 60% of the dataset, while SF scored 75%, and SFR 90%. Regarding SUS, participants' scores also differ (*SFR:* $M = 78.33$, $SD = 14.33$; *SF:* $M = 77.08$, $SD = 14.54$; *S:* $M = 73.19$, $SD = 15.44$). All three scores fall within the "good" usability range (Bangor et al., 2009). The statistical analysis on SUS scores shows that adding feedback increased usability (posterior probability: 93%, 95% HDI [−1.06, 8.97]), and there is some weak support for adding reflection compared to simulation and feedback (posterior probability: 72%, 95% HDI [−3.00, 5.54]).

4.3. Exploratory results

RQ1: Using feedback levels over time. To examine whether participants used task- or process-level feedback options differently over time, we analysed their usage patterns across sessions in condition *SFR*. The results did not indicate a clear tendency toward either type (posterior probability = 0.72, $M = 0.92$, 95% HDI = [−2.21, 4.06]). Therefore, we visualised the data to explore potential trends across specific feedback options and participants (Fig. 7). As shown in the

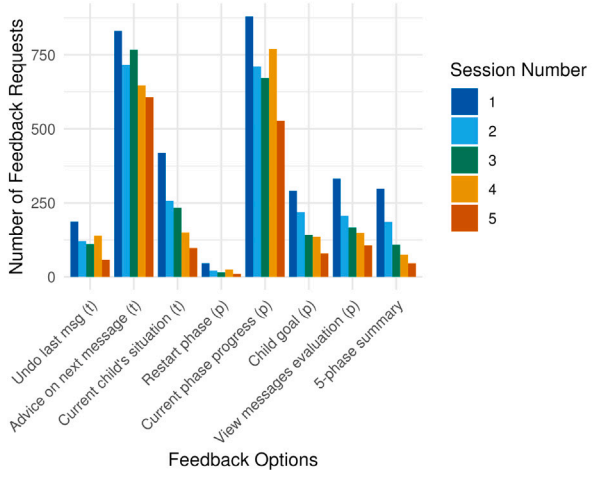


(a) Comparison of the raw ratings for the short ASAQ between the three conditions.

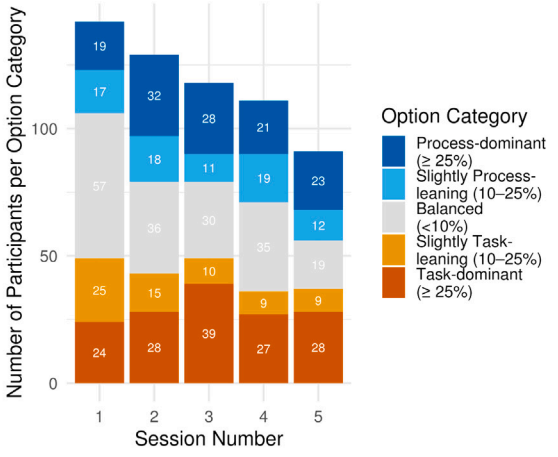


(b) Percentile chart of each condition on the short ASAQ constructs, normalised against the 29 agents in the ASAQ Representative Set 2024 (Fitrianie et al., 2025). Grey bands indicate values lower than or higher than those observed for all agents in the reference set.

Fig. 6. Two comparative views based on the short ASAQ constructs. The charts were generated using the ASAQ visualisation tool.



(a) Number of times each feedback option was clicked per session.



(b) Each interaction in the exploration mode was categorised into one of five categories based on how many times a participant clicked task- or process-level feedback. This figure illustrates the distribution of these categories across the five sessions.

Fig. 7. Overall description of feedback option usage across the five sessions. The left figure shows how many times each feedback button was clicked per session, while the right figure illustrates the relative dominance of task- or process-level feedback clicks and how this distribution varied across the sessions.

figures, participants generally used both feedback types less frequently over time, which is consistent with the analysis of feedback option requests over time (posterior probability = 0.99, $M = -2.20$, 95% HDI = $[-4.02, -0.36]$). Fig. 7(a) shows that the most frequently used feedback options were *Advice on Next Message* (task level) and *Current Phase Progress* (process level), whereas options that directly affected the conversation—*Undo last message* (task level) and *Restart phase* (process level)—were used the least. In Fig. 7(b), we categorise each interaction in the exploration mode (one per participant per session) into one of five feedback option categories based on the dominance of task- or process-level feedback: process-dominant (≥ 25% process-level use), slightly process-leaning (10–25% process level), balanced (within ±10%), slightly task-leaning (10–25% task level), and task-dominant (≥ 25% task level). In the first session, some participants used more task- or process-level feedback, but the largest proportion (40%) fell into the balanced category. Over time, the proportion of participants with balanced feedback requests decreased to 21% in session five, as participants increasingly shifted their feedback requests toward either task- or process-level categories.

RQ2: Influence of feedback level on knowledge, reflection capability and conversational outcomes.

We tested whether using more task-level feedback than process-level feedback (or vice versa) was associated with better outcome measures. No apparent association effects were found for any of the three measures: task knowledge, reflection capability, and conversational outcomes. Specifically, when comparing the relative use of each level, the posterior distributions showed weak tendencies toward more task-level feedback for knowledge (74%, $M = 0.64$, 95% HDI = $[-1.27, 2.55]$), but a weak tendency toward more process-level feedback for reflection (61%, $M = -0.38$, 95% HDI = $[-3.16, 2.32]$), and conversational outcomes

over the five sessions as an interaction effect (68%, $M = -1.21$, 95% HDI $= [-6.08, 3.72]$). All three distributions are interpreted as not worth betting on. However, when examining only the main effect on conversational outcome indexes (without the interaction with session), there is some support that using task-level feedback led to better conversational outcomes compared to process-level (93%: a promising but risky bet, $M = 12.52$, 95% HDI $= [-3.96, 29.17]$). This effect appeared most prominently in the first session. The overall patterns for these three measures are shown in Fig. D.11 in the Appendix, which illustrates how the feedback option categories scored across the three outcome measures.

5. Discussion and conclusions

In this paper, we examined how the subsequent addition of simulation (H_1), feedback (H_2), and reflection (H_3) shapes trainees' learning in simulation-based training. Across five sessions, we assessed how these elements contributed to improvements in task knowledge, reflection capability, and conversational outcomes (performance). Overall, the results support the cumulative effect proposed in H_4: adding elements sequentially leads to overall increases across the three learning measures. Still, each element contributed differently, and these differences highlight important considerations for the design and impact of simulation-based training.

Participants interacting with the simulation only (*S*) showed increases over time in both task knowledge and conversational outcomes, whereas the control group (*C*) showed a decrease in task knowledge. H_1 results indicate that adding simulation increases knowledge and reflection more than the control condition. This improvement may be attributed to the effect of repetition in several sessions, which supports learning (Walsh et al., 2023). Participants in condition *S* needed to repeatedly decide what to ask, how to phrase questions, and how to move the conversation forward across five sessions, whereas the control condition received only theoretical explanations. This lack of repeated engagement may also explain the observed decrease in task knowledge in condition *C*. The results of condition *S* align with Reger et al. (2020), who found that training with a virtual patient for motivational interviewing—compared to academic study alone—led to increased skill outcomes.

Adding feedback to the simulation (condition *SF*) improved task knowledge and conversational outcomes, while maintaining similar reflection capability scores compared to the simulation-only condition (H_2). The need for feedback was further highlighted by participant comments, with several participants in condition *S* explicitly commenting at the end of the sessions that they wished for more guidance, e.g., "I wish there was a redo button…" and "It would be good to have feedback as well". On the other hand, the results for reflection capability were similar for the two conditions. One possible explanation is that participants across both *S* and *SF* received similar amounts of interaction experience with the tool across five sessions, which may have influenced their reflections, even without targeted reflection training. For conversational outcomes, participants across conditions did not start from similar baselines after the first session. Participants in the *S* condition performed worse in the first session compared to those in *SF* (Fig. 5). This is likely because the two conditions with feedback benefited from guidance that helped them navigate their conversations more effectively from the beginning. In turn, this inflated the standard deviation used to compute the thresholds for small and large effect sizes, which resulted in a lower probability of observing effects when comparing improvements in conversational outcomes between *S* and *SF* across the five sessions. These baseline differences should be considered when interpreting the conversational outcome results across conditions.

Although adding reflection moments (condition *SFR*) improved reflection capability as expected, it dampened conversational outcomes compared to simulation with feedback (H_3). The latter may be attributed to two factors. First, the reflection task introduced a break between exploration and assessment modes. This temporal shift in focus, from applying the task in the exploration to reflecting on it, may have hindered the immediate carry-over of learning into the assessment interaction — the SF group moved directly from exploration to assessment, while the SFR group completed a reflection task in between. Second, reflection requires effort, motivation, and practice (Hartmann et al., 2023; Anseel et al., 2009; Moon, 2013). Yet despite the effort required, many participants in the *S* and *SF* conditions were intrinsically motivated to write reflective comments in the optional comments field after each session, even without being prompted. Thus, reflection may have occurred spontaneously for a subset of participants in *S* and *SF*. The findings align with Krause et al. (2017) child-avatar interview training study, in which feedback improved interviewing quality, but adding a post-feedback reflection task did not further enhance performance beyond feedback alone.

For the secondary measures, participants reported better perceived usefulness and overall experience as more elements were added. This suggests that trainees valued not only the learning content but also the added structure, guidance, and support within the system. The exploratory research questions on how participants used task- and process-level feedback did not reveal clear patterns; thus, a plausible explanation is that both types should be used depending on the situation and what information trainees need to obtain. This interpretation is also supported by recommendations to combine the levels of feedback (Hattie and Timperley, 2007), and that the level should focus on the current problem being solved (Smit et al., 2024). Task-level feedback, however, was associated with better conversational outcomes when examining individual sessions. This may be due to the "Advice on next message" option, as it directly suggests what to consider next, which can help trainees progress further through the conversation, thereby making the assessment interaction easier to navigate. Future work could examine whether such differences persist days or weeks after the training, as directive guidance might not have long-term benefits for knowledge transfer (Roll et al., 2018).

5.1. Limitations and future work

A first limitation concerns the extent to which Prolific participants resemble real helpline volunteer trainees. Helpline trainees are typically intrinsically motivated (Millette and Gagné, 2008), for example, by commitment to the helpline and its mission. Although Prolific crowd workers have been found to provide higher-quality data than participants from other crowdsourcing platforms and undergraduate samples (Douglas et al., 2023), their motivation may still differ from that of helpline trainees. To reduce self-interest and loss-aversion biases, we informed participants that their payment would not depend on their performance or interactions (Draws et al., 2021). However, crowdsourcing workers' motivation extends beyond payment alone and is also shaped by factors relevant to this experiment, such as enjoyment, skill development (Ye and Kankanhalli, 2017), and contributing to a social cause (Cappa et al., 2019). The results of this study suggest that participants improved their skills over time. Nevertheless, paid participants may have felt more obligated to complete all learning tasks, whereas helpline trainees might skip tasks they consider unnecessary. Therefore, future research should include real helpline volunteers to rule out potential motivational effects.

Relatedly, participant retention may have been influenced by motivation. Motivated participants (either extrinsically through compensation or intrinsically due to interest) may have been naturally retained across sessions. This may limit generalisability to less motivated trainees. However, in the case of children's helplines, trainees are typically volunteers who are themselves motivated to learn, which makes this characteristic of the sample less problematic.

Another limitation to generalisability is that excluded participants were slightly more likely to be parents than included participants, although the effect size of this difference was small. This may reflect the five-session design, which required commitment and availability and

may therefore have been harder to sustain for participants with children. At the same time, most child helpline counsellors are volunteers who also need to commit time, which may make the retained sample more similar to the target group. Still, this does not threaten internal validity, as parental status did not differ across the four included conditions.

While repeated practice improves learning and retention (Karpicke and Roediger III, 2007), this study does not address the transfer of skills to real helpline communications. In particular, the conversational outcomes measure should be interpreted as a proximal, protocol-aligned performance measure rather than as direct evidence of transfer to real help-seekers. Thus, results might differ if assessed weeks or months later or in settings involving real help-seekers, especially since reflection supports knowledge retention (Chang, 2019; Larsen et al., 2016). Future research could compare simulation-based training with traditional training methods and evaluate knowledge transfer in actual helpline contexts. Another avenue for future work is to examine the extent to which trainees ask open versus closed questions in conversations across conditions and sessions, thereby adding a counselling-relevant behavioural dimension to the conversational outcomes measure.

Because immediate feedback was learner-controlled rather than experimentally controlled, participants in the *SF* and *SFR* conditions may have differed in the type and amount of feedback they viewed. This variability limits the extent to which the effects of specific feedback levels can be isolated. However, this design reflects a realistic self-guided training context, in which trainees decide when they need support and which guidance to consult. Therefore, despite the limitation, this choice increased ecological validity and allowed us to study natural feedback-use patterns, which would not have been possible if all participants had received the same fixed sequence or amount of feedback. Importantly, the fact that we observed effects even though participants were not required to view all available feedback suggests that the effects may have been stronger if feedback exposure had been experimentally controlled. As an additional check, we examined whether the number of feedback options used across the five sessions differed between the two conditions with feedback (*SF*: $M = 69.05$, $SD = 76.51$; *SFR*: $M = 76.06$, $SD = 74.14$). A Bayesian t-test did not provide credible support for a difference (posterior probability = 73%, 95% HDI = $[-28.03, 14.25]$).

Several participants reported misunderstandings by the virtual child. We believe these stem from the rule-based intent recognition component, which did not perform well with verbose input. This limitation might be addressed by integrating Large Language Models (LLMs), as Al Owayyed et al. (2025b) showed that an LLM-integrated virtual child was perceived more positively than its rule-based counterpart.

Participants in the *SFR* condition showed higher reflection capability. However, it remains unclear whether this improvement reflects growth in counselling-specific reflections or simply a general increase in reflective writing skills. It is possible that participants were practising reflection across the five sessions without necessarily developing deeper counselling insight, as practice improves reflection ability (Radović et al., 2023). Future work could test this by adding a condition in which participants reflect on an unrelated topic, such as cycling or daily routines, to differentiate domain-specific reflective growth from general reflective writing skills. Furthermore, since the experimental design was not fully factorial, the independent contribution of reflection (without simulation or feedback) remains unknown and could be examined in future work.

Beyond the five-phase model used in this study, the incremental benefits of simulation, feedback, and reflection are not tied to this protocol alone. They may also extend to training other communication frameworks, such as motivational interviewing (Hettema et al., 2005). Likewise, the approach may generalise to other helpline and training contexts beyond children's helplines, such as suicide prevention or mental-health crisis lines, where volunteers must learn to navigate emotionally complex and sensitive conversations. Although the scenarios in this study focused on supporting a child to find a solution, the underlying mechanisms of simulation, guided practice, and supported reflection seem relevant across domains involving interpersonal interactions, e.g., sales, negotiations, or customer service.

5.2. Ethical considerations

Children exhibit diverse behaviours. Modelling only limited behaviours and scenarios in simulations (e.g., only bullying) could risk reinforcing stereotypes about how children act when contacting helplines (Al Owayyed et al., 2025a). Therefore, training should include a wider range of perspectives, behavioural patterns, and helpline topics (e.g., sexuality, family issues, relationships (Fukkink et al., 2016)) to avoid narrowing trainees' expectations. For example, it is uncommon for children to ask helplines to call their school on their behalf. Yet, repeated exposure to such a scenario might give trainees the false impression that this is typical.

5.3. Conclusion

This study shows that simulation-based training can support the skill development of trainee helpline counsellors. Adding feedback enhances the learning process further by helping trainees better understand the child's perspective and their own decision-making within the conversation. Furthermore, adding reflective practice to simulation and feedback strengthens reflection capability, although it appears to slow the rate of performance improvement compared to simulation and feedback alone. When combined, simulation and feedback create a training environment that enhances participants' learning, performance, and experience with the training tool, while incorporating reflection requires more careful consideration. Based on these findings, we hope that simulation-based training, particularly when combined with feedback, can serve as a scalable addition to helpline education, accelerating early-stage counsellor development.

CRediT authorship contribution statement

Mohammed Al Owayyed: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Myrthe L. Tielman:** Writing – review & editing, Visualization, Supervision, Resources, Project administration, Methodology, Formal analysis, Conceptualization. **Willem-Paul Brinkman:** Writing – review & editing, Visualization, Supervision, Resources, Project administration, Methodology, Formal analysis, Conceptualization.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work, generative AI tools (ChatGPT and Grammarly) were used to enhance language and grammar. All content was subsequently reviewed and edited by the authors.

Funding

This work and the first author's research are supported by King Saud University and the Saudi Arabian Cultural Mission (SACM).

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We want to thank Morita Tarvirdians and Nele Albers for their valuable contributions. We also thank the PROTECtME Consortium and De Kindertelefoon for their helpful input and support.

Appendix A. Participant demographics

There were no clear demographic differences between included and excluded participants (Table A.7), except for parental status. Excluded participants were more likely to have children than included participants (posterior probability = 99.9%), and this difference was more likely to be small (79.9%) than large (0%).

Table A.6
Full participant demographics by condition (control, simulation only, simulation and feedback, simulation, feedback and reflection).

Demographic	Control (n = 66)	Simulation (n = 66)	Simulation and feedback (n = 64)	Simulation, feedback, and reflection (n = 150)
Age	M = 31.41, SD = 9.58 18–23: 7 (10.6%) 24–29: 28 (42.4%) 30–39: 24 (36.4%) 40–49: 5 (7.6%) 70+: 2 (3%)	M = 31.50, SD = 9.13 18–23: 14 (21.2%) 24–29: 17 (25.8%) 30–39: 26 (39.4%) 40–49: 4 (6.1%) 50–59: 5 (7.6%)	M = 31.75, SD = 8.59 18–23: 13 (20.3%) 24–29: 17 (26.6%) 30–39: 24 (37.5%) 40–49: 6 (9.4%) 50–59: 4 (6.2%)	M = 31.90, SD = 10.24 18–23: 30 (20%) 24–29: 48 (32%) 30–39: 48 (32%) 40–49: 12 (8%) 50–59: 6 (4%) 60–69: 5 (3.3%) 70+: 1 (0.7%)
Gender	Men: 30 (45.5%) Women: 36 (54.5%)	Men: 28 (42.4%) Women: 38 (57.6%)	Men: 32 (50%) Women: 32 (50%)	Men: 62 (41.3%) Women: 81 (54%) Other: 4 (2.7%) Prefer not to say: 3 (2%)
Education	Community/Technical: 3 (4.5%) Bachelor's: 44 (66.7%) Master's: 15 (22.7%) PhD: 4 (6.1%)	Community/Technical: 3 (4.5%) Bachelor's: 49 (74.2%) Master's: 12 (18.2%) PhD: 2 (3%)	Community/Technical: 6 (9.4%) Bachelor's: 41 (64.1%) Master's: 16 (25%) PhD: 1 (1.6%)	Community/Technical: 13 (8.7%) Bachelor's: 99 (66%) Master's: 31 (20.7%) PhD: 7 (4.7%)
Has Children	Yes: 30 (45.5%) No: 34 (51.5%) Prefer not to say: 2 (3%)	Yes: 32 (48.5%) No: 30 (45.5%) Prefer not to say: 4 (6.1%)	Yes: 36 (56.2%) No: 25 (39.1%) Prefer not to say: 3 (4.7%)	Yes: 82 (54.7%) No: 62 (41.3%) Prefer not to say: 6 (4%)
English Fluency	Native: 27 (40.9%) Fluent: 39 (59.1%)	Native: 27 (40.9%) Fluent: 39 (59.1%)	Native: 25 (39.1%) Fluent: 39 (60.9%)	Native: 53 (35.3%) Fluent: 97 (64.7%)
Field of Study	Business: 6 (9.1%) Law: 6 (9.1%) ICT: 10 (15.2%) Engineering: 13 (19.7%) Natural Sciences: 2 (3%) Social Sciences: 2 (3%) Medical Sciences: 3 (4.5%) Humanities: 7 (10.6%) Education: 6 (9.1%) Other: 11 (16.7%)	Business: 10 (15.2%) Law: 2 (3%) ICT: 13 (19.7%) Engineering: 5 (7.6%) Natural Sciences: 6 (9.1%) Social Sciences: 2 (3%) Medical Sciences: 6 (9.1%) Humanities: 1 (1.5%) Education: 8 (12.1%) Other: 10 (15.2%) Prefer not to say: 3 (4.5%)	Business: 12 (18.8%) Law: 4 (6.2%) ICT: 8 (12.5%) Engineering: 5 (7.8%) Natural Sciences: 1 (1.6%) Social Sciences: 2 (3.1%) Medical Sciences: 11 (17.2%) Humanities: 3 (4.7%) Education: 9 (14.1%) Other: 8 (12.5%) Prefer not to say: 1 (1.6%)	Business: 20 (13.3%) Law: 8 (5.3%) ICT: 36 (24%) Engineering: 17 (11.3%) Natural Sciences: 7 (4.7%) Social Sciences: 8 (5.3%) Medical Sciences: 11 (7.3%) Humanities: 7 (4.7%) Education: 12 (8%) Other: 23 (15.3%) Prefer not to say: 1 (0.7%)
Experienced Bullying	Yes: 60 (90.9%) No: 4 (6.1%) Prefer not to say: 2 (3%)	Yes: 60 (90.9%) No: 6 (9.1%)	Yes: 57 (89.1%) No: 6 (9.4%) Prefer not to say: 1 (1.6%)	Yes: 132 (88%) No: 15 (10%) Prefer not to say: 3 (2%)
Chatbot Experience	NA	Once or twice: 4 (6.1%) Occasionally: 25 (37.9%) Frequently: 37 (56.1%)	Once or twice: 12 (18.8%) Occasionally: 18 (28.1%) Frequently: 34 (53.1%)	Never: 1 (0.7%) Once or twice: 13 (8.7%) Occasionally: 54 (36%) Frequently: 82 (54.7%)

Table A.7
Summary of demographics for excluded and included participants who started the pre-questionnaire and provided consent.

Demographic	Excluded Participants (n = 316)	Included Participants (n = 346)
Age	M = 31.63, SD = 9.73 18–23: 67 (21.2%) 24–29: 95 (30.1%) 30–39: 107 (33.9%) 40–49: 27 (8.5%) 50–59: 12 (3.8%) 60–69: 5 (1.6%) 70+: 3 (0.9%)	M = 31.70, SD = 9.58 18–23: 64 (18.5%) 24–29: 110 (31.8%) 30–39: 122 (35.3%) 40–49: 27 (7.8%) 50–59: 15 (4.3%) 60–69: 5 (1.4%) 70+: 3 (0.9%)
Gender	Men: 140 (44.3%) Women: 170 (53.8%) Other: 3 (0.9%) Prefer not to say: 3 (0.9%)	Men: 152 (43.9%) Women: 187 (54.0%) Other: 4 (1.2%) Prefer not to say: 3 (0.9%)
Education	Community/Technical: 19 (6.0%) Bachelor's: 213 (67.4%) Master's: 67 (21.2%) PhD: 17 (5.4%)	Community/Technical: 25 (7.2%) Bachelor's: 233 (67.3%) Master's: 74 (21.4%) PhD: 14 (4.0%)
Has Children	Yes: 206 (65.2%) No: 103 (32.6%) Prefer not to say: 7 (2.2%)	Yes: 180 (52.0%) No: 151 (43.6%) Prefer not to say: 15 (4.3%)
English Fluency	Native: 126 (39.9%) Fluent: 190 (60.1%)	Native: 132 (38.2%) Fluent: 214 (61.8%)
Field of Study	Business: 24 (7.6%) Law: 13 (4.1%) ICT: 76 (24.1%) Engineering: 30 (9.5%) Natural Sciences: 14 (4.4%) Social Sciences: 19 (6.0%) Medical Sciences: 26 (8.2%) Humanities: 17 (5.4%) Education: 40 (12.7%) Other: 53 (16.8%) Prefer not to say: 4 (1.3%)	Business: 48 (13.9%) Law: 20 (5.8%) ICT: 67 (19.4%) Engineering: 40 (11.6%) Natural Sciences: 16 (4.6%) Social Sciences: 14 (4.0%) Medical Sciences: 31 (9.0%) Humanities: 18 (5.2%) Education: 35 (10.1%) Other: 52 (15.0%) Prefer not to say: 5 (1.4%)
Experienced Bullying	Yes: 286 (90.5%) No: 26 (8.2%) Prefer not to say: 4 (1.3%)	Yes: 309 (89.3%) No: 31 (9.0%) Prefer not to say: 6 (1.7%)

Appendix B. Lilobot training system feedback

B.1. Lilobot screenshots

See Figs. B.8–B.10.

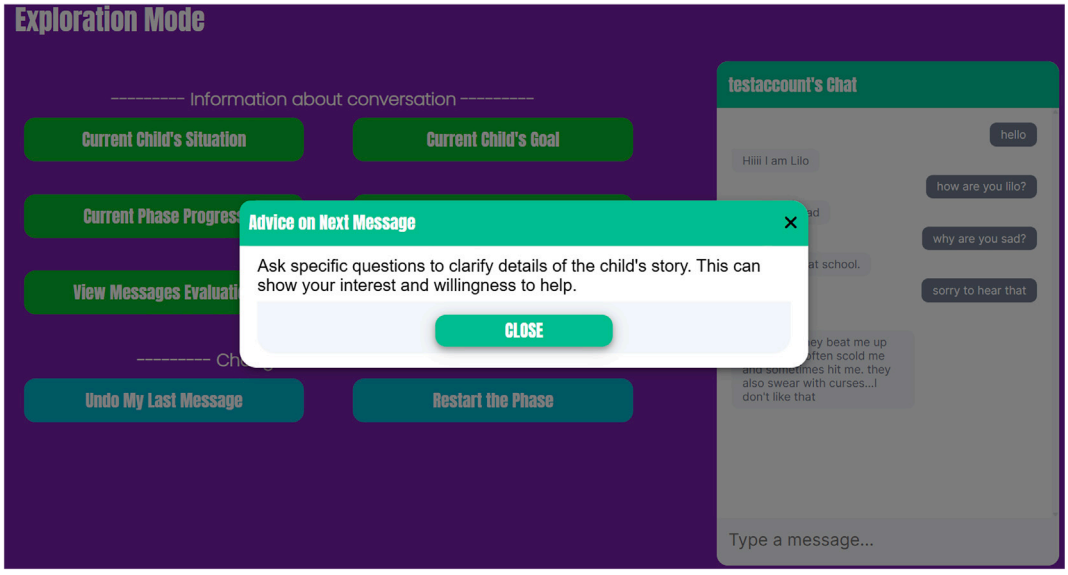


Fig. B.8. An example of immediate feedback when a trainee chooses ”Advice on Next Message”, which provides task-level feedback.

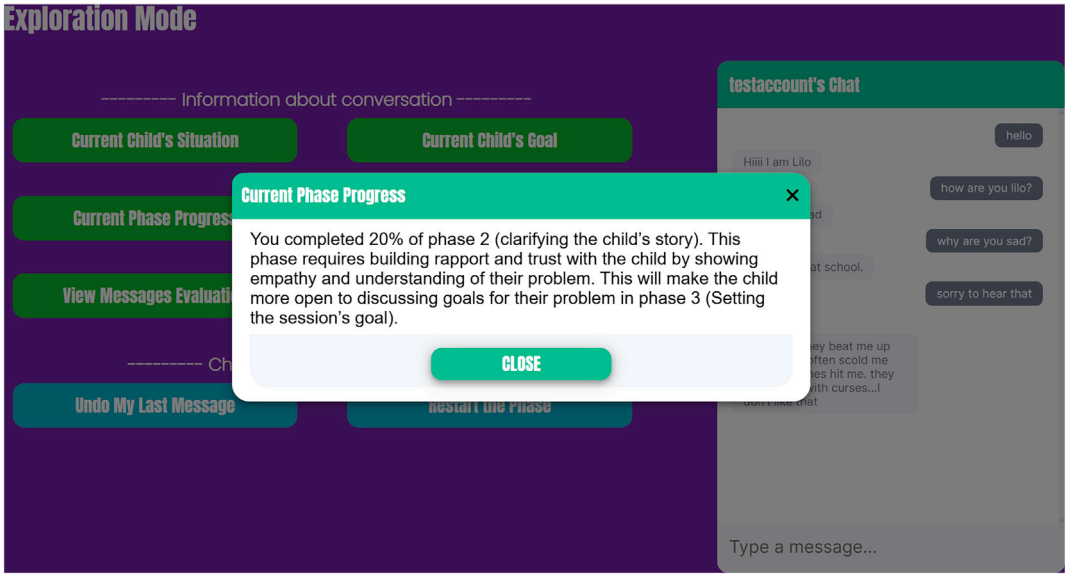


Fig. B.9. An example of immediate feedback when a trainee chooses ”Current Phase Progress”, which provides process-level feedback.

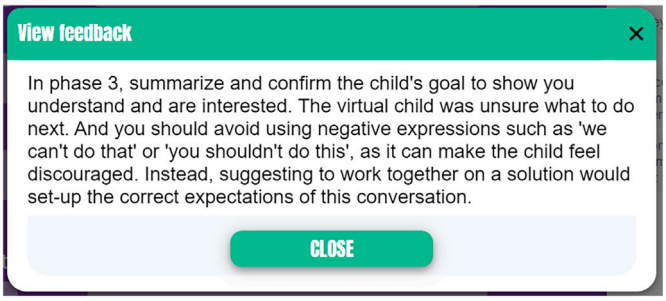


Fig. B.10. An example of the delayed feedback at the end of the interaction.

B.2. Feedback generation mechanisms

Table B.8
Generation mechanism of the feedback options in the training system.

Feedback option	Generation mechanism
Current Child's Goal	Retrieved from the currently active desire in the child's BDI model.
Current Child's Situation	Generated from the child's current belief values, compared with the expected belief values along an optimal scenario path.
Current Phase Progress	Calculated by comparing the current belief state with the target belief state for the ongoing phase, and extracting the remaining steps to reach the next phase (according to the optimal path) as a progress percentage.
Advice on Next Message	Generated from the trainee's current position in the optimal path within the phase, suggesting a next utterance that would move the conversation forward.
View Messages Evaluation	Generated by classifying trainee utterances according to the five-phase model and checking whether each utterance fits the current phase. Utterances that better fit another phase are stored in a list and displayed to the trainee when this option is used.
5-Phase Model Summary	Provided as a static textual summary of the five-phase model.
Restart the Phase	Based on the stored history of BDI states and phase changes. The system removes all messages and belief changes from the current phase and restores the BDI state to its state at the start of the current phase.
Undo My Last Message	Based on the stored history of previous BDI states. The system removes the most recent trainee–agent message pair and restores the BDI state prior to that exchange.
Delayed Feedback	Generated after the interaction by comparing the final BDI state and phase progress with the optimal path. The system also checks why the child left to explain what went wrong, and uses the optimal path and BDI values to suggest how the trainee can improve in the next interaction.

Appendix C. Questionnaires

C.1. Task knowledge

Participants were given two task knowledge exams in a random order. The exams had the same structure (e.g., number of questions), but the question order was displayed randomly.

C.1.1. Exam version A

Description *(if pre)*. As the first part of the experiment, you will be given a test to measure your current knowledge level of the five-phase model. We want to have an honest understanding of your current knowledge of the five-phase model (e.g., low, medium, high). Please don't Google the answers.

Q1. What is the purpose of phase 1?

- Obtaining background information from the child
- Asking the child for their name
- Informing the child about how the conversation is going
- Giving the child a warm welcome

Q2. What is the purpose of phase 3?

- To determine the child's desired state
- To provide the child with possible solutions
- To explain to the child what they can do
- To contact the child's school about bullying

Q3. You are now wrapping up the conversation. What happened just before this?

- You worked with the child on their desired state and goal
- You listened to the child's story
- You asked if the child trusts someone to tell
- You asked about the child's wishes

Q4. What should a counsellor NOT do in phase 2?

- Make the situation clearer by asking further questions
- Ask about one specific situation in which the problem occurs
- Acknowledge the child's problem
- Ask about important people in the child's life

Q5. What should a counsellor NOT do in phase 4?

- Ask about important people in the child's life
- Ask what the child can try to solve the problem
- Ask how the child can implement the agreed solution
- Ask what the child's desired state is

Q6. You have just asked the child about their wish. What happened before this?

- You welcomed the child and calmed them
- You asked about important people in the child's life
- You listened to the child's background story
- You explained to the child what steps they could take

Q7. The child told you that they are being bullied. What question can you ask now in phase 2?

- "Who can you talk to about this problem?"
- "Did you talk to your parents about the bullying?"
- "Can you tell me more about your situation?"
- "What do you want to achieve from this conversation?"

Q8. Which of the following can you say to help you navigate the conversation from phase 3 to phase 4?

- "What would you like to achieve with this conversation?"
- "Is there someone whom you can talk to about the bullying?"
- "What do you think about finding a solution together?"
- "To make it clear, you are being bullied at school and you want this to stop, is that correct?"

Q9. What can you ask to get a clearer view of the child's story during phase 2?

- "Please tell me more about yourself."
- "Since when have you been struggling with this?"
- "Did you start at your school this year?"
- "Have you told anyone that you are being bullied?"

Table C.9
Statements for question 11 (exam version A).

Statement	In which phase? (options)
The counsellor did not properly stimulate the child's problem-solving skills to find a solution	phase1, phase2, phase3, phase4, phase5
The counsellor did not ask enough questions about the details of the child's problem	phase1, phase2, phase3, phase4, phase5
The counsellor did not show enough empathy and understanding of the child's concerns	phase1, phase2, phase3, phase4, phase5
The counsellor didn't suggest a solution that the child could perform	phase1, phase2, phase3, phase4, phase5

Table C.10
Statements for question 11 (exam version B).

Statement	In which phase? (options)
The counsellor did not ask enough questions about the details of the child's problem	phase1, phase2, phase3, phase4, phase5
The counsellor did not show enough empathy and understanding of the child's concerns	phase1, phase2, phase3, phase4, phase5
The counsellor did not ask enough questions about the details of applying the goal	phase1, phase2, phase3, phase4, phase5
The counsellor suggested a goal that the child can't perform	phase1, phase2, phase3, phase4, phase5

Q10. The child asks you to call their parents and tell them about bullying. What should you say next?

- "I can call them later today. Can you send me their number?"
- "You should talk to your parents and tell them about the bullying"
- "I suggest we look for another solution together. What do you think?"
- "Sorry, the children's helpline cannot call your parents."

Q11. The child kept saying 'I don't know' and then left when the counsellor continuously asked about the child's goal with the question: 'What can you do to solve this problem?'. What is the most likely mistake the counsellor made, and in which phase did they make it? (Only choose one statement from the left and one phase number (Table C.9).

C.1.2. Exam version B

Description *(if pre)*. As the first part of the experiment, you will be given a test to measure your current knowledge level of the five-phase model. We want to have an honest understanding of your current knowledge of the five-phase model (e.g., low, medium, high). Please don't Google the answers.

Q1. What is the purpose of phase 2?

- To determine the child's wish and work with the child on that wish
- To reassure the child and tell them it will be okay
- To listen to the child and gather information about their story.
- To ask about important people in the child's life to contact them

Q2. What is the purpose of phase 4?

- To ask about the child's strengths
- To build towards the child's desired state and goal
- To provide possible solutions that the child can take
- To make the problem clearer by asking further questions

Q3. You have just given the child a warm welcome. What should happen next?

- Listen to the child's story
- Ask about the desired state of the child
- Explain to the child what steps they should take
- Find a solution together with the child

Q4. What should a counsellor NOT do in phase 1?

- Acknowledge the child's problem and show empathy
- Give a warm welcome
- Make contact at the beginning of the conversation
- Ask about the child's desired state

Q5. What should a counsellor NOT do in phase 3?

- Investigate how the desired state would be for the child
- Agree with the child on the goal details
- Work towards applying a goal
- Summarise the child's problem and desired goal

Q6. What can you say to the child in phase 5?

- "What score would you give the conversation?"
- "If I understand correctly, you will talk to your teacher about the problem, right?"
- "How did you find the conversation?"
- "I'll arrange it for you! You'll hear back from me later today!"

Q7. Which of these questions is NOT part of phase 4?

- "What do you hope will be different after this conversation?"
- "Shall we agree that you will talk to your teacher about your situation?"
- "What are you going to tell your teacher?"
- "Who can you talk to and who listens to you?"

Q8. You have just listened to the child's story details. What can you say next to the child according to the five-phase model?

- "Why were you bullied at school?"
- "Who can you talk to about this problem?"
- "What have you already tried that made you feel happier at school?"
- "Should we look for a solution together?"

Q9. You agreed with the child that they would talk to their teacher about the problem. What would be the most suitable thing to say next?

- "Imagine a ruler from 0 to 10. 0 is your problem at its worst, 10 is solved. Where do you stand now?"
- "I suggest ending the conversation now so you can get started."
- "Do you promise to talk to your teacher?"
- "Is there someone else you can talk to if your teacher doesn't listen to you?"

Q10. Which of the following can you say to help you navigate the conversation from phase 2 to phase 3?

- "Is there someone you can talk to about the bullying?"
- "If I understand you correctly, you are being bullied at school, is that correct?"
- "What would you like to achieve with this conversation?"
- "Is there anything I can do to help you?"

Q11. The child agreed to a goal but was not confident in performing it by the end of the conversation. What is the most likely mistake the counsellor made, and in which phase did they make it? (only choose one statement and one phase number (Table C.10).

C.2. Reflection question

Reflection on learning experience: Please take some time to reflect on your overall learning experience since starting the experiment. We encourage you to think deeply, as we value your reflective capabilities. So please, take your time.

You may want to reflect on aspects such as: what you have learned, how you learned it, how it changed your thinking, what the outcomes were, and how they align with your initial goals.

C.2.1. Assessment rubric

We assessed each reflection on five dimensions. The dimensions follow the reflection levels in (Fleck and Fitzpatrick, 2010).

D1: Description (Revisiting)—Max 3 points

- 0: Not mentioned
- + 1: Statement about an event
- + 1: Another statement
- + 1: Another statement

D2: Reflective Description (Revisiting with Explanation)—Max 3 points

- 0: Not mentioned
- + 1: Includes reasoning
- + 1: Additional reasoning
- + 1: Further reasoning

D3: Dialogic Reflection (Exploring Relationships)—Max 3 points

- 0: Not mentioned
- + 1: Explored one perspective or dialogue within the self
- + 1: Explored more than one perspective and reasoning
- + 1: Another exploration

D4: Transformative Reflection (Fundamental Change)—Max 3 points

- 0: Not mentioned
- + 1: One fundamental change mentioned
- + 1: Reasoning about it
- + 1: Another fundamental change

D5: Critical Reflection (Wider Implications)—Max 3 points

- 0: Not mentioned
- + 1: One wider implication mentioned
- + 1: Reasoning about it
- + 1: Another wider implication

Total reflection score: D1 + D2 + D3 + D4 + D5 = X/15

C.3. Perceived usefulness questionnaire

Please rate the following statements about the exploration mode on a scale from strongly disagree to strongly agree.

- The exploration mode provides me with information about the five phase model.
- The exploration mode helps me to apply the five phase model better.
- The exploration mode provides me with clear insight into the five phase model.
- The exploration mode provides me enough control over my own learning.
- Using the exploration mode, I can apply the five phase model independently.
- The exploration mode fits my learning strategies.

Appendix D. Extra analysis

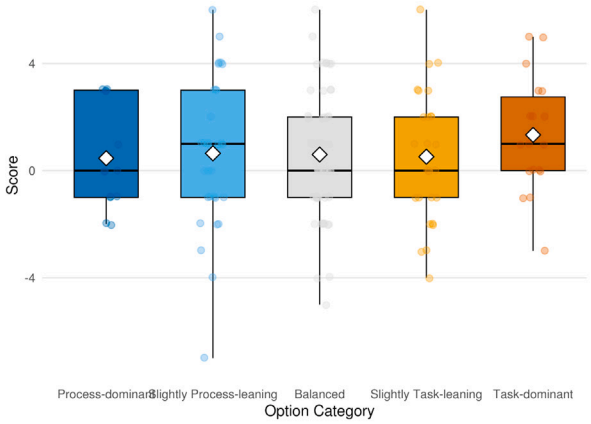
D.1. Cross-condition analysis of primary measures

Table D.11
Summary of Bayesian test results for three primary measures (Task Knowledge, Reflection Capability, Conversational Outcomes), comparing each pair of conditions.

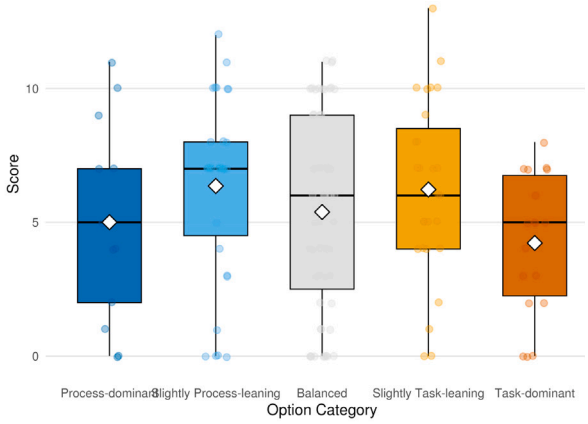
Measure	Comparison	Posterior Probability	Interpretation	Mean [95% HDI] (SD)
Ha: Task Knowledge	S > C	0.999	Very strong bet	1.05 [0.41, 1.70] (0.33)
	SF > C	>0.999	Virtually certain	1.46 [0.81, 2.14] (0.34)
	SF > S	0.891	Only a casual bet	0.41 [−0.25, 1.05] (0.33)
	SFR > C	>0.999	Virtually certain	1.39 [0.80, 1.94] (0.29)
	SFR > S	0.882	Only a casual bet	0.33 [−0.23, 0.88] (0.28)
	SFR > SF	0.395	Not worth betting	−0.07 [−0.63, 0.47] (0.28)
Hb: Reflection Capability	S > C	0.989	Good bet - too good to disregard	1.31 [0.18, 2.44] (0.58)
	SF > C	0.987	Good bet - too good to disregard	1.30 [0.08, 2.41] (0.59)
	SF > S	0.496	Not worth betting	−0.00 [−1.14, 1.10] (0.58)
	SFR > C	>0.999	Nearing certainty	1.77 [0.79, 2.72] (0.50)
	SFR > S	0.827	Only a casual bet	0.46 [−0.46, 1.45] (0.49)
	SFR > SF	0.830	Only a casual bet	0.47 [−0.49, 1.44] (0.49)
Hc: Conversations1	SF > S	0.994	Strong bet - irresponsible to avoid	13.1 [2.85, 23.3] (5.22)
	SFR > S	0.958	Good bet - too good to disregard	7.57 [−0.73, 16.5] (4.40)
	SFR > SF	0.109	Only a casual bet for the opposite	−5.54 [−14.0, 3.68] (4.50)
Hc: Conversations2	SF > S	0.995	Very strong bet	2.99 [0.72, 5.25] (1.16)
	SFR > S	0.976	Good bet - too good to disregard	1.93 [−0.02, 3.78] (0.97)
	SFR > SF	0.140	Only a casual bet	−1.06 [−2.92, 0.92] (0.98)

Note. C = Control; S = Simulation only; SF = Simulation and Feedback; SFR = Simulation, Feedback and Reflection.

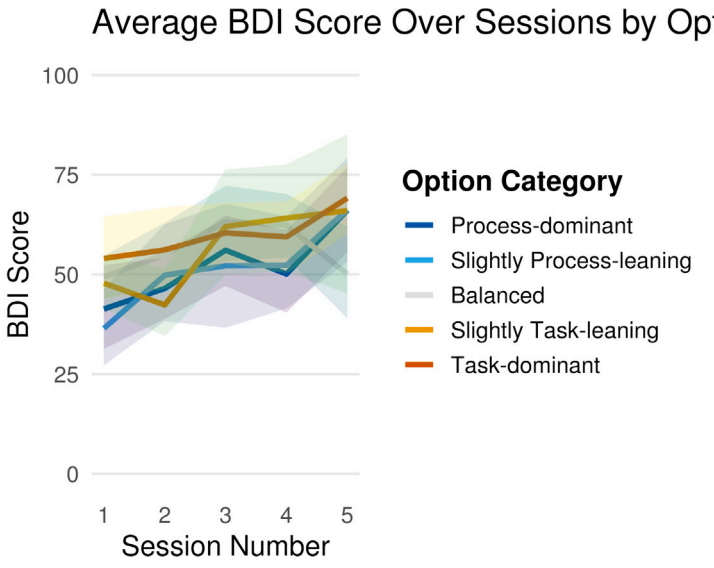
D.2. Exploring feedback option usage in relation to task knowledge, reflection, and conversational outcomes



(a) How the different option categories (Figure 7b) improved in the knowledge test exam.



(b) How the different option categories (Figure 7b) scored in the reflection measure.



(c) How the different option categories (Figure 7b) scored in the conversational outcomes measure per session.

Fig. D.11. How the different option categories in Fig. 7(b) scored in the three outcome measures.

Data availability

Data, R code, and prompts are available in the 4TU Research Data Repository. Source code for the training system is available on GitHub.

References

Adams, N.E., 2015. Bloom's taxonomy of cognitive learning objectives. J. Med. Libr. Assoc. JMLA 103, 152.

Al Owayyed, M., Tielman, M., Hartholt, A., Specht, M., Brinkman, W.-P., 2025a. Agent-based social skills training systems: the ARTES architecture, interaction characteristics, learning theories and future outlooks. Behav. Inf. Technol. 44, 1787–1814.

Al Owayyed, M., Denga, A., Brinkman, W.-P., 2025b. Controlled yet natural: a hybrid BDI-LLM conversational agent for child helpline training. In: Proceedings of the 25th ACM International Conference on Intelligent Virtual Agents, pp. 1–10.

Anderson, K., André, E., Baur, T., Bernardini, S., Chollet, M., Chryssafidou, E., Damian, I., Ennis, C., Egges, A., Gebhard, P., et al., 2013. The TARDIS framework: intelligent virtual agents for social coaching in job interviews. In: International Conference on Advances in Computer Entertainment Technology. Springer, pp. 476–491.

Andraszewicz, S., Scheibehenne, B., Rieskamp, J., Grasman, R., Verhagen, J., Wagenmakers, E.-J., 2015. An introduction to Bayesian hypothesis testing for management research. J. Manag. 41, 521–543.

Anseel, F., Lievens, F., Schollaert, E., 2009. Reflection as a strategy to enhance task performance after feedback. Organ. Behav. Hum. Decis. Process. 110, 23–35.

Ash, S.L., Clayton, P.H., 2004. The articulated learning: an approach to guided reflection and assessment. Innov. High. Educ. 29, 137–154.

Ausubel, D.P., 1963. The Psychology of Meaningful Verbal Learning. Grune & Stratton.

Bangor, A., Kortum, P., Miller, J., 2009. Determining what individual SUS scores mean: adding an adjective rating scale. J. Usability Stud. 4, 114–123.

Bavaresco, A., Bernardi, R., Bertolazzi, L., Elliott, D., Fernández, R., Gatt, A., Ghaleb, E., Giulianelli, M., Hanna, M., Koller, A., et al., 2024. LLMs instead of human judges? A large scale empirical study across 20 NLP evaluation tasks. arXiv preprint arXiv:2406.18403.

Bloom, B.S., Engelhart, M.D., Furst, E.J., Hill, W.H., Krathwohl, D.R., et al., 1956. Taxonomy of Educational Objectives: The Classification of Educational Goals. Handbook 1: Cognitive Domain. Longman New York.

Bosse, T., Provoost, S., 2014. Towards aggression de-escalation training with virtual agents: a computational model. In: International Conference on Learning and Collaboration Technologies. Springer, pp. 375–387.

Brooke, J., et al., 1996. SUS-A quick and dirty usability scale. Usability eval. Ind. 189, 4–7.
Brookhart, S.M., 2017. How to Give Effective Feedback to Your Students. Ascd.
Brooks, C., Huang, Y., Hattie, J., Carroll, A., Burton, R., 2019. What is my next step? School students' perceptions of feedback. Front. Educ. 4, https://doi.org/10.3389/feduc.2019.00096
Bruijnes, M., 2016. Believable Suspect Agents: Response and Interpersonal Style Selection for an Artificial Suspect (Ph.D. thesis). University of Twente.
Bürkner, P.-C., 2017. brms: an R package for Bayesian multilevel models using stan. J. Stat. Softw. 80, 1–28.
Cameron, D.L., Kovac, V.B., Horverak, M.O., 2023. Online chat-support for bullying prevention and follow-up: an analysis of the characteristics of chat-users and their experiences. Int. J. Adolesc. Youth 28, 2292060.
Cappa, F., Rosso, F., Hayes, D., 2019. Monetary and social rewards for crowdsourcing. Sustainability 11, 2834.
Carless, D., 2020. Longitudinal perspectives on students' experiences of feedback: a need for teacher–student partnerships. High. Educ. Res. Dev. 39, 425–438.
Chang, B., 2019. Reflection in learning. Online Learn. 23, 95–110.
Chechile, R.A., 2020. Bayesian Statistics for Experimental Scientists: A General Introduction Using Distribution-Free Methods. MIT Press.
Chiang, C.-H., Lee, H.-Y., 2023. Can large language models be an alternative to human evaluations? arXiv preprint arXiv:2305.01937.
Child Helpline International, 2024. Voices of children & young people around the world: global child helpline data from 2023. https://childhelplineinternational.org/voices-2023-data/ (Accessed: 1 April 2025).
Chiviacowsky, S., Wulf, G., 2002. Self-controlled feedback: does it enhance learning because performers get feedback when they need it? Res. Q. Exerc. Sport 73, 408–415.
Chiviacowsky, S., Wulf, G., 2005. Self-controlled feedback is effective if it is based on the learner's performance. Res. Q. Exerc. Sport 76, 42–48.
Chollet, M., Marsella, S., Scherer, S., 2022. Training public speaking with virtual social interactions: effectiveness of real-time feedback and delayed feedback. J. Multimodal User Interfaces 16, 17–29.
Cohen, J., 1992. A power primer. Psychol. Bull. 112 (1), 155–159.
Demasi, O., Li, Y., Yu, Z., 2020. A multi-persona chatbot for hotline counselor training. In: Findings of the Association for Computational Linguistics: EMNLP 2020, pp. 3623–3636.
Ding, D., Remeijsen, P., Song, Z., Neerincx, M.A., Brinkman, W.-P., 2024. Technology-supported social skills training systems: a systematic literature review. In: 2024 27th International Conference on Computer Supported Cooperative Work in Design (CSCWD). IEEE, pp. 3249–3256.
Douglas, B.D., Ewell, P.J., Brauer, M., 2023. Data quality in online human-subjects research: comparisons between MTurk, prolific, CloudResearch, qualtrics, and SONA. PLoS One 18, e0279720.
Draws, T., Rieger, A., Inel, O., Gadiraju, U., Tintarev, N., 2021. A checklist to combat cognitive biases in crowdsourcing. In: Proceedings of the AAAI Conference on Human Computation and Crowdsourcing, vol. 9. pp. 48–59.
Dupuy, L., Micoulaud-Franchi, J.-A., Cassoudesalle, H., Ballot, O., Dehail, P., Aouizerate, B., Cuny, E., de Sevin, E., Philip, P., 2020. Evaluation of a virtual agent to train medical students conducting psychiatric interviews for diagnosing major depressive disorders. J. Affect. Disord. 263, 1–8.
Ertmer, P.A., Newby, T.J., 1993. Behaviorism, cognitivism, constructivism: comparing critical features from an instructional design perspective. Perform. Improv. Q. 6, 50–72.
Fitrianie, S., Horsch, C., Beun, R.J., Griffioen-Both, F., Brinkman, W.-P., 2021. Factors affecting user's behavioral intention and use of a mobile-phone-delivered cognitive behavioral therapy for insomnia: a small-scale UTAUT analysis. J. Med. Syst. 45, 110.
Fitrianie, S., Bruijnes, M., Li, F., Abdulrahman, A., Brinkman, W.-P., 2022. The artificial-social-agent questionnaire: establishing the long and short questionnaire versions. In: Proceedings of the 22nd ACM International Conference on Intelligent Virtual Agents, pp. 1–8.
Fitrianie, S., Bruijnes, M., Abdulrahman, A., Brinkman, W.-P., 2025. The artificial social agent questionnaire (ASAQ)—development and evaluation of a validated instrument for capturing human interaction experiences with artificial social agents. Int. J. Hum.-Comput. Stud. 199, 103482.
Fleck, R., Fitzpatrick, G., 2010. Reflecting on reflection: framing a design landscape. In: Proceedings of the 22nd Conference of the Computer-Human Interaction Special Interest Group of Australia on Computer-Human Interaction, pp. 216–223.
Froehle, T.C., Robinson, S.E., De Kurpius, W.J., 1983. Enhancing the effects of modeling through role-play practice. Couns. Educ. Superv. 22, 197–206.
Fukkink, R.G., Bruns, S., Ligtvoet, R., 2016. Voices of children from around the globe; an international analysis of children's issues at child helplines. Child. Soc. 30, 510–519.
Funder, D.C., Ozer, D.J., 2019. Evaluating effect size in psychological research: sense and nonsense. Adv. Methods Pract. Psychol. Sci. 2, 156–168.
Gagné, R.M., 1985. The Conditions of Learning and Theory of Instruction. Holt.
Gratch, J., DeVault, D., Lucas, G., 2016. The benefits of virtual humans for teaching negotiation. In: International Conference on Intelligent Virtual Agents. Springer, pp. 283–294.
Grundmann, S., Al Owayyed, M., Bruijnes, M., Vroonhof, E., Brinkman, W.-P., 2025. Lilobot: a cognitive conversational agent to train counsellors at children's helplines: design and initial evaluation. J. Med. Syst. 49, 5.
Hartmann, A., Vinke-de Kruijf, J., Van Weesep, R., 2023. Asking the right questions: the role of reflection for learning in and between projects. Int. J. Proj. Manag. 41, 102494.
Hattie, J., Timperley, H., 2007. The power of feedback. Rev. Educ. Res. 77, 81–112.
Hatton, N., Smith, D., 1995. Reflection in teacher education: towards definition and implementation. Teach. Teach. Educ. 11, 33–49.
Hettema, J., Steele, J., Miller, W.R., 2005. Motivational interviewing. Annu. Rev. Clin. Psychol. 1, 91–111.
Hinkle, D.E., Wiersma, W., Jurs, S.G., et al., 2003. Applied Statistics for the Behavioral Sciences, vol. 663. Houghton Mifflin Boston.
Hirumi, A., Johnson, T., Reyes, R.J., Lok, B., Johnsen, K., Rivera-Gutierrez, D.J., Bogert, K., Kubovec, S., Eakins, M., Kleinsmith, A., et al., 2016. Advancing virtual patient simulations through design research and inter PLAY: Part II—integration and field test. Educ. Technol. Res. Dev. 64, 1301–1335.
Hoang, A., Hepburn, S.-J., Morawska, A., Sanders, M.R., 2025. The effect of self-reflection on the outcomes of online clinical skills training: a comparative study. Adv. Health Sci. Educ. 1–19.
Hoque, M., Courgeon, M., Martin, J.-C., Mutlu, B., Picard, R.W., 2013. Mach: my automated conversation coach. In: Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing, pp. 697–706.
Karpicke, J.D., Roediger III, H.L., 2007. Repeated retrieval during learning is the key to long-term retention. J. Mem. Lang. 57, 151–162.
Kaufman, T.M., Kretschmer, T., Huitsing, G., Veenstra, R., 2018. Why does a universal anti-bullying program not help all children? Explaining persistent victimization during an intervention. Prev. Sci. 19, 822–832.
Kirschner, P., Sweller, J., Clark, R.E., 2006. Why unguided learning does not work: an analysis of the failure of discovery learning, problem-based learning, experiential learning and inquiry-based learning. Educ. Psychol. 41, 75–86.
Kitchener, K.S., 1983. Cognition, metacognition, and epistemic cognition: a three-level model of cognitive processing. Hum. Dev. 26, 222–232.
Knight, J.K., Weaver, D.C., Peffer, M.E., Hazlett, Z.S., Andrews, T.C., 2022. Relationships between prediction accuracy, metacognitive reflection, and performance in introductory genetics students. CBE—Life Sci. Educ. 21, ar45. https://doi.org/10.1187/cbe.21-12-0341
Kolb, D., 1984. Experiential Learning: Experience as the Source of Learning and Development. Prentice-Hall Englewood Cliffs, New Jersey.
Krause, N., Pompedda, F., Antfolk, J., Zappalá, A., Santtila, P., 2017. The effects of feedback and reflection on the questioning style of untrained interviewers in simulated child sexual abuse interviews. Appl. Cogn. Psychol. 31, 187–198.
Lamb, M.E., Sternberg, K.J., Orbach, Y., Esplin, P.W., Mitchell, S., 2002. Is ongoing feedback necessary to maintain the quality of investigative interviews with allegedly abused children? Appl. Dev. Sci. 6, 35–41.
Larsen, D.P., London, D.A., Emke, A.R., 2016. Using reflection to influence practice: student perceptions of daily reflection in clinical education. Perspect. Med. Educ. 5, 285–291.
Larson, L.M., Daniels, J.A., 1998. Review of the counseling self-efficacy literature. Couns. Psychol. 26, 179–218.
Lew, M.D., Schmidt, H.G., 2011. Self-reflection and academic performance: is there a relationship? Adv. Health Sci. Educ. 16, 529–545.
Maicher, K., Danforth, D., Price, A., Zimmerman, L., Wilcox, B., Liston, B., Cronau, H., Belknap, L., Ledford, C., Way, D., et al., 2017. Developing a conversational virtual standardized patient to enable students to practice history-taking skills. Simul. Healthc. 12, 124–131.
Maier, H.W., 2002. Role playing: structures and educational objectives. Int. Child Youth Care Netw. 36.
Makowski, D., Ben-Shachar, M.S., Chen, S.A., Lüdecke, D., 2019a. Indices of effect existence and significance in the Bayesian framework. Front. Psychol. 10, 2767.
Makowski, D., Ben-Shachar, M.S., Lüdecke, D., 2019b. bayestestR: describing effects and their uncertainty, existence and significance within the Bayesian framework. J. Open Source Softw. 4, 1541.
Maqsood, Z., Sajjad, M., Yasmin, R., 2025. Effect of feedback-integrated reflection, on deep learning of undergraduate medical students in a clinical setting. BMC Med. Educ. 25, 66.
Mascarenhas, S., Dias, J., Prada, R., Paiva, A., 2010. A dimensional model for cultural behavior in virtual agents. Appl. Artif. Intell. 24, 552–574.
Mathieu, S.L., Uddin, R., Brady, M., Batchelor, S., Ross, V., Spence, S.H., Watling, D., Kölves, K., 2021. Systematic review: the state of research into youth helplines. J. Am. Acad. Child Adolesc. Psychiatry 60, 1190–1233.
Miller, W.R., Mount, K.A., 2001. A small study of training in motivational interviewing: does one workshop change clinician and client behavior? Behav. cogn. psychother. 29, 457–471.
Miller, W.R., Yahne, C.E., Moyers, T.B., Martinez, J., Pirritano, M., 2004. A randomized trial of methods to help clinicians learn motivational interviewing. J. Consult. Clin. Psychol. 72, 1050.
Millette, V., Gagné, M., 2008. Designing volunteers' tasks to maximize motivation, satisfaction and performance: the impact of job characteristics on volunteer engagement. Motiv. Emot. 32, 11–22.
Moon, J.A., 2013. Reflection in Learning and Professional Development: Theory and Practice. Routledge.
Morey, R.D., Rouder, J.N., 2018. BayesFactor: Computation of Bayes Factors for Common Designs. https://CRAN.R-project.org/package=BayesFactor.
Moss, C.M., Brookhart, S.M., 2012. Learning Targets: Helping Students Aim for Understanding in Today's Lesson. ASCD.
Moussa-Inaty, J., 2015. Reflective writing through the use of guiding questions. Int. J. Teach. Learn. High. Educ. 27, 104–113.
Muller, T., Heuvelink, A., van den Bosch, K., Swartjes, I., 2012. Glengarry glen ross: using bdi for sales game dialogues. In: Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment, vol. 8. pp. 167–172.
Murali, P., Trinh, H., Ring, L., Bickmore, T., 2021. A friendly face in the crowd: reducing public speaking anxiety with an emotional support agent in the audience. In: Proceedings of the 21st ACM International Conference on Intelligent Virtual Agents, pp. 156–163.

Nouraei, F., Rebello, K., Fallah, M., Murali, P., Matuszak, H., Jap, V., Parker, A., Paasche-Orlow, M., Bickmore, T., 2025. Virtual agent-based communication skills training to facilitate health persuasion among peers. Proc. ACM Hum.-Comput. Interact. 9, 1–24.

NSPCC, 2025. Childline and NSPCC helpline statistics. https://learning.nspcc.org.uk/research-resources/statistics-briefings/childline-nspcc-helpline-statistics (Accessed: 6 November 2025).

Parwata, I., Jayanta, I.N.L., Widiana, I.W., 2023. Improving metacognitive ability and learning outcomes with problem-based revised bloom's taxonomy oriented learning activities. Emerg. Sci. J. 7, 569–577.

Pompedda, F., Zhang, Y., Haginoya, S., Santtila, P., 2022. A mega-analysis of the effects of feedback on the quality of simulated child sexual abuse interviews with avatars. J. Police crim. Psychol. 37, 485–498.

Radović, S., Firssova, O., Hummel, H.G., Vermeulen, M., 2023. Improving academic performance: strengthening the relation between theory and practice through prompted reflection. Act. Learn. High. Educ. 24, 139–154.

Reger, G.M., Norr, A.M., Rizzo, A.S., Sylvers, P., Peltan, J., Fischer, D., Trimmer, M., Porter, S., Gant, P., Baer, J.S., 2020. Virtual standardized patients vs academic training for learning motivational interviewing skills in the US department of veterans affairs and the US military: a randomized trial. JAMA Netw. Open 3, e2017348.

Røed, R.K., Baugerud, G.A., Hassan, S.Z., Sabet, S.S., Salehi, P., Powell, M.B., Riegler, M.A., Halvorsen, P., Johnson, M.S., 2023. Enhancing questioning skills through child avatar chatbot training with feedback. Front. Psychol. 14, 1198235.

Roll, I., Butler, D., Yee, N., Welsh, A., Perez, S., Briseno, A., Perkins, K., Bonn, D., 2018. Understanding the impact of guiding inquiry: the relationship between directive support, student attributes, and transfer of knowledge, attitudes, and behaviours in inquiry learning. Instr. Sci. 46, 77–104.

Shute, V.J., 2008. Focus on formative feedback. Rev. Educ. Res. 78, 153–189.

Sindahl, T.N., 2013. Chat Counselling for Children and Youth: A Handbook.

Smit, R., Bachmann, P., Dober, H., Hess, K., 2024. Feedback levels and their interaction with the mathematical reasoning process. Curric. J. 35, 184–202.

Stichting De Kindertelefoon, 2023. De kindertelefoon jaarverslag 2023. https://jaarverslag.kindertelefoon.nl/2023 (accessed 13 October 2025).

Stichting De Kindertelefoon, 2024. De kindertelefoon jaarverslag 2024. https://jaarverslag.kindertelefoon.nl/2024 (accessed 13 October 2025).

Sweller, J., 2011. Cognitive load theory. In: Psychology of Learning and Motivation, vol. 55. Elsevier, pp. 37–76.

Tanana, M.J., Soma, C.S., Srikumar, V., Atkins, D.C., Imel, Z.E., 2019. Development and evaluation of ClientBot: patient-like conversational agent to train basic counseling skills. J. Med. Internet Res. 21, e12529.

Van Boekel, M., Hufnagle, A.S., Weisen, S., Troy, A., 2023. The feedback I want versus the feedback I need: investigating students' perceptions of feedback. Psychol. Sch. 60, 3389–3402.

Veldhuizen-Ochodničanová, E., Cotton, R., Putney, C., Brennan, M., Haahr, M., 2023. The use of serious games in knowledge-and skills-based digital counselling: applications to trafficking in persons in the war in Ukraine. In: Joint International Conference on Serious Games. Springer, pp. 290–298.

Venkatesh, V., Morris, M.G., Davis, G.B., Davis, F.D., 2003. User acceptance of information technology: toward a unified view. MIS Q. 425–478.

Walsh, M.M., Krusmark, M.A., Jastrembski, T., Hansen, D.A., Honn, K.A., Gunzelmann, G., 2023. Enhancing learning and retention through the distribution of practice repetitions across multiple sessions. Mem. Cogn. 51, 455–472.

Washburn, M., Parrish, D.E., Bordnick, P.S., 2020. Virtual patient simulations for brief assessment of mental health disorders in integrated care settings. Soc. Work Ment. Health 18, 121–148.

Watson, J.B., 2017. Behaviorism. Routledge.

Watts, L., 2019. Reflective practice, reflexivity, and critical reflection in social work education in Australia. Aust. Soc. Work 72, 8–20.

Wiggins, G., 2012. Seven keys to effective feedback. Feedback 70, 10–16.

Wisniewski, B., Zierer, K., Hattie, J., 2020. The power of feedback revisited: a meta-analysis of educational feedback research. Front. Psychol. 10, 3087.

Ye, H.J., Kankanhalli, A., 2017. Solvers' participation in crowdsourcing platforms: examining the impacts of trust, and benefit and cost factors. J. Strateg. Inf. Syst. 26, 101–117.