This is the accepted version of the following article: Pompedda, F., Segal, A., Haginoya, S., Bakaitytė-Bagdonė, A., Ustinavičiūtė-Klenauskė, L., Kaniušonytė, G., Žukauskienė R., & Santtila, P. (2025). Experience and Long-Term Training Effects in Simulated Child Sexual Abuse Interviews. *Applied Cognitive Psychology*. 39:e70051., which has been published in final form at [https://doi.org/10.1002/acp.70051].

Abstract

Previous research has shown that simulated interviews with avatars can improve question quality, but evidence regarding long-term effects and the role of experience remains limited. We investigated both short and long-term impact of avatar training with feedback in Child Protection Services (CPS) workers and student groups. Thirty-one CPS workers and 35 novice students interviewed four child avatars, with half receiving feedback after each interview. After four months, the training was repeated with all participants receiving feedback. Training with feedback improved investigative interview quality in the short term, with no substantial decline after four months. Experience had no effects on interview quality, nor did it moderate training impact. These findings suggest that avatar-based interview training with feedback could effectively improve investigative interviewing skills across different experience levels, maintaining its effects over a 4-month period.

Keywords: child sexual abuse investigation, interviewer training, simulation-based training, long-term effects, experience, feedback

Experience and Long-Term Training Effects in Simulated Child Sexual Abuse Interviews

Introduction

Background on Child Interviewing and Training

Interviewing children is a complex task that requires consistent training accompanied by feedback (see Korkman et al., 2024). In recent years, studies have focused on ways to provide cost-effective but impactful types of training including virtual reality (VR) elements (Krause et al., 2024) or avatars that can respond automatically using artificial intelligence (AI) (Haginoya et al., 2023; Røed et al., 2023). These technological advances add to the body of previous research that has shown immediate post-training effects after simulated interviews with students (e.g., (Haginoya et al., 2020; Krause et al., 2024; Pompedda et al., 2015, 2017), police officers (e.g., Kask et al., 2022), psychologists (e.g., Haginoya et al., 2021; Pompedda et al., 2021), teachers (e.g., Brubacher et al., 2015) and mixed groups (Benson & Powell, 2015; Pompedda et al., 2022) of participants and transfer of the training effects achieved in simulations into interviews with real children (Benson & Powell, 2015; Kask et al., 2022).

Long-Term Effects of Training

As mentioned above, even if such positive immediate changes in interviewer behaviour have been demonstrated by several research groups around the world, a critical issue that has been given insufficient attention, is the long-term effects of these training programs. Past research on the implementation of semi-structured interview protocols has highlighted how field interviewers are likely to discontinue practices they have learned as soon as the training and feedback are discontinued (Lamb et al., 2002). This highlights the importance of investigating how long training effects last.

Kask and colleagues (2022) showed how monthly booster sessions might maintain previously acquired skills in subsequent real interviews. However, design issues in the study do not allow firm causal conclusions regarding the maintenance of training effects to be drawn. In one of the few studies showing clear maintenance of training effects, Brubacher and colleagues' (2022) results showed that training effects lasted for nine months after a hybrid mix of lectures

and practice. However, due to dropouts in the control group, the role of booster sessions in the maintenance of the training effect could not be unambiguously identified. More recently, Haginoya and Santtila (2023) reported preliminary findings demonstrating the maintenance of training effects over several months using simulated interviews with AI-driven avatars. The relative lack of research on long term effects of training has been highlighted also in a recent systematic review (Akca et al., 2021) summarising the efficacy of investigative interview training programs.

Theoretical Framework: Cognitive Load Theory

While there is a relative lack of formal tests of long term-effects, training programs in this field have been developed based on theoretical principles aimed at maximising retention of the acquired skills. The training programme in the present study was based on previous work (Pompedda, 2018; Pompedda et al., 2015) that used the cognitive-load theory (CLT) to guide its development (e.g., Paas et al., 2003).

CLT focuses on how the capacity limitations of working memory can impact learning of complex skills. Cognitive load is categorized into three types: intrinsic load, which relates to the complexity of the task; extraneous load, which relates to how the information is presented; and germane load which relates to the mental effort needed for the construction of schemas. In other words, the goal of applying the CLT when planning an investigative interview training is to reduce unnecessary extraneous cognitive load and manage intrinsic load in a way that facilitates cognitive schema construction and maintenance over time (Paas & van Merriënboer, 2020). Several instructional design principles of avatar training should do exactly this, for example, the provision of feedback helping in schemata constructions (Young et al., 2014), repeated practice favouring schemata automation (Young et al., 2014) and the use of avatars with varying response patterns increasing transfer (e.g., see Kirschner & Van Merrienboer, 2008).

The Role of Feedback in Investigative Interviews

Practice, feedback, and supervision are the cornerstone of behavioural change (Hattie & Timperley, 2007) and this is not different in investigative interviews of children (Cederborg et al., 2013). A recent paper of recommendations based on evidence from the literature (Korkman et al., 2024), highlights how regular feedback is pivotal for an effective training program. Research has highlighted how the way in which feedback is provided might influence its efficacy. To maximise

its effects, feedback must be continuous, immediate, and detailed (Lamb et al., 2002; Smith, 2008). Studies using an avatar setup showed how the combination of outcome (feedback on the ground truth about the case) and process feedback (feedback on the questions used) was more efficient than the two types of feedback provided separately or no feedback at all (Pompedda et al., 2017). In addition, there is consensus and evidence from the literature on the importance of providing feedback beyond the training period (Benson & Powell, 2015; Krause et al., 2017; Pompedda et al., 2015; Powell et al., 2016) in the form of refresher sessions (Price & Roberts, 2011). What is still not known, is when refresher sessions should be utilised, and how much training and supervision is necessary within a training that involves avatars, while there is some evidence from more traditional training that the effect of training with feedback disappears almost immediately after training and supervision is discontinued (Lamb et al., 2002), while the above mentioned study by Benson and Powell (2015) showed effects present at 6.

The Role of Experience in Investigative Interviews

Another important aspect of training in investigative interviews is the role of practical experience particularly in the context of avatar-based training. Most avatar training studies conducted in recent years have included relatively inexperienced groups of participants, for example, teachers (Brubacher et al., 2015) or psychologists without specific experience in the forensic field (Haginoya et al., 2021; Pompedda et al., 2021) and non-professional groups, for example, students (e.g., Krause et al., 2017). In contrast, traditional investigative interview studies have been conducted with graduate trainees within psychology and social work, police officers and child protection workers (Røed et al., 2023), forensic interviewers, medical professionals, social workers, and others (Baugerud et al., 2024). When examining avatar-based interview training specifically, the relationship between interviewer experience and interview quality shows varied patterns. Some studies have shown no differences as a function of experience (Benson & Powell, 2015; Haginoya et al., 2021, Ko et al., 2025), while some have shown positive associations (Lafontaine & Cyr, 2016) while still others have suggested negative relationships (Lafontaine & Cyr, 2017; Powell et al., 2014). An important differentiation needs to be made between laboratory studies, for example, simulated avatar interviews, and field studies where quality of interviews in actual investigations has been researched. Experience seems to have a more beneficial effect on at least certain

measures of quality, for example, use of open-ended questions, in laboratory studies compared to field studies where the association is less clear (Lamb et al., 2018).

The Intersection among Feedback, Learning, and Experience

The relationship between investigator experience and training outcomes remains an important area of inquiry, separate from the broader question of how experience relates to interviewing quality. There is less conclusive evidence on the interaction between feedback and experience with children in general as potential moderators as experience of interacting with children both in an informal role (e.g., babysitter) or formal (interviewing children) has not been extensively studied (Pompedda et al., 2022).

Previous studies have proposed that, due to the proactive interference theory (Keppel & Underwood, 1962), more experienced interviewers might have more difficulties in changing behaviours, compared to lay people, as the previously acquired information might interfere with the new learning (Powell et al., 2014). This could also extend to training programs that utilise avatars. However, evidence in this area is limited. Røed and colleagues (2023) showed how professionals had good engagement in interaction with a simple avatar within a training program, however, as lay interviewers were not included in the sample, and behavioural changes were not the aim of the study, it is not possible to determine if the same effect would be present in inexperienced interviewers and how this would interact with learning. While the software used in this experiment aimed to create an interview that closely resembles a real interview, as this increases the possibility of transfer of skills from the simulation to the real interviews (Blume et al., 2010), some differences are present (e.g., complex body language is limited within the avatars). Participants that have interviewed children before, have done it in a different context (e.g., face to face) compared to an interview with avatars, the limited presence of certain features might hinder their learning processes compared to someone that has not interviewed children before, and will do this for the first time with an avatar. On the other hand, the multi-interview structure of the training may explain why experienced interviewers might struggle initially (compared to student group) but demonstrate improved performance after they adapt to the differences between avatar and real-life interviews.

However, this question can be approached through the lens of CLT, which suggests that participants with prior interviewing experience should learn new skills more rapidly than novices (Sweller, 1998). This advantage stems from experienced interviewers' ability to more efficiently allocate cognitive resources and leverage existing mental schemas. While novice interviewers must divide their cognitive capacity between learning new interviewing techniques and managing the basic interview dynamics and emotional demands, experienced interviewers can focus more of their cognitive resources on implementing new techniques, having already automated the basic management of interview situations.

However, experience might not uniformly facilitate learning. The expertise reversal effect suggests that when feedback reinforces known but unimplemented principles (such as the importance of open-ended questions), the intervention might paradoxically prove more beneficial for inexperienced interviewers (Chen et al., 2017). Pompedda and colleagues (2022), in a similar setup that used avatars similar to the present study, showed how individuals with parenting experience asked more not recommended questions, however, they improved more over time compared to those without parenting experience. In the same study, the authors showed how professional experience, coded as having interviewed vs not having interviewed children, was a positive predictor of number of recommended questions and improvements over rounds of practice compared to participants with no professional experience, which provide support for the CLT.

While previous studies have demonstrated that training interventions can improve the use of open questions by experienced professionals, and certain studies showed how experience in interviewing children might be beneficial, more evidence is needed to conclusively determine whether these improvements occur more quickly or slowly compared to inexperienced interviewers.

This creates an interesting tension in predicting how experience might moderate training effects in the present study. On one hand, CPS workers might show less improvement simply because they start at a higher baseline level of performance if there is an overall positive effect of experience. Alternatively, experienced interviewers may show less improvement in their techniques (such as open-ended questioning) because interviewing experience usually correlates with years on the job. The longer someone has been working, the more established their work patterns become, making

these habits harder to change (Powell et al., 2014). However, we do not have information on years of experience on the job in this experiment. On the other hand, their experience in managing interviews might reduce their cognitive load during training, allowing them to focus more attention on implementing the feedback they receive. These competing possibilities made it difficult to form a directional hypothesis about the moderating role of experience on training effects.

In the present study, Lithuanian Child Protection Service (CPS) workers were included as an experienced group given that all of them had either received training in investigative interviews of children and/or had as their primary role working with potentially endangered children in situations where their task was to gather sufficient information to be able to decide on further procedural steps (Intė et al., 2023). In Lithuania, CPS workers are charged with communicating with children to conduct a preliminary assessment of whether they have or are in danger of being abused even if they are not tasked with the formal investigation of alleged abuse cases as this is reserved to specialized personnel: psychologists or judges with the support of a psychologist (Banevičienė et al., 2023). However, for all practical intents and purposes they engage in gathering information from the child which from a psychological perspective amount to an investigative interview.

Aims and Hypotheses

In the present study, participants completed two training sessions approximately four months apart, with four simulated interviews in each session. The design included two key independent variables: experience (CPS workers vs. students) and feedback (feedback vs. no feedback). However, experience could be operationalised in different ways. As we did not have information on how many interviews with children our participants have done, which could be referred as expertise, we operationalised experience in three different ways: a) CPS vs students b) regardless of the professional role, we coded as experienced all those participants that had conducted interviews with children in general or in a CSA context c) regardless of the professional role, we coded as experienced all those participants that had conducted interviews with children in general and in a CSA context. In the main text we report results of the first operationalisation, and in supplementary materials the results of the second operationalisation. The third operationalisation created unbalanced

groups and is not reported. Regardless of the operationalisation used, results were similar, and the small discrepancies are reported in the supplementary materials (see Appendix A).

During the first session, half of the participants received feedback (experimental group) after each interview while the control group received no feedback. Importantly, in the second session (starting from the fifth interview), all participants, including those initially in the control group, received feedback after each interview.

Our first aim was to replicate the effects of feedback on questioning quality. Previous research with similar interview simulations (e.g., Pompedda et al., 2022) has demonstrated that training coupled with feedback improved the quality of investigative interviews across various groups of professionals and students. We expected to replicate these findings in our sample.

Our second aim was to examine the impact of experience, both independently and in combination with feedback, on questioning style. Based on previous literature showing positive effects of experience in laboratory studies, we expected CPS workers to demonstrate overall higher quality interviews compared to students. However, we had competing predictions about how experience might moderate the effects of feedback on skill development. While CPS workers might show less improvement due to higher baseline performance, on the other hand, their experience in managing interviews could reduce cognitive load during training, potentially allowing them to more effectively implement the feedback received.

Our final aim was to investigate the long-term effects of training. Given that our intervention was designed to maximize skill acquisition and transfer; considering Kask and colleagues (2022) evidence of transfer effects after similar training, we expected the training effects to persist across the four-month interval between sessions.

For all hypotheses, interview quality was operationalized as participants posing more recommended questions, fewer not recommended questions, and a higher proportion of recommended questions, which in turn would elicit better quality information (i.e., more relevant details and fewer wrong details) from the avatars. While recommended questions and relevant details are positively correlated, these correlations vary among interviews and avatars, making it still an important measure of interview quality. In addition, in previous studies, results between questions asked and details

obtained varied in their effect sizes, potentially showing that there might be two different processes happening. Specifically, we hypothesized:

Hypothesis 1: A main effect of experience, with CPS workers (vs. students) asking better quality questions and eliciting better quality information from the avatars.

Hypothesis 2: A feedback x number of interviews interaction, with participants receiving feedback showing greater improvement in question quality and information elicitation over successive interviews. Pairwise comparisons at each interview would identify where differences between feedback and no-feedback groups emerged.

Given the competing predictions about experience effects, we did not formulate a directional hypothesis regarding experience as a moderator of the feedback x number of interviews interaction. Instead, we conducted exploratory analyses to investigate such moderation effects.

Hypothesis 3: Maintenance of training effects at four months, with participants in the feedback group showing no significant decline in question quality or information elicitation between their final interview in the first session and their first interview in the second session.

Methods

Participants

Sixty-six participants aged 19 to 62 years (57 women, M_{age} = 33.1, *SD* = 12.8) were recruited, thirty-one professionals from the Lithuanian Child Protection Services (CPS) (M_{age} = 40.6, *SD* = 11.4) and thirty-five university students (M_{age} = 26.4, *SD* = 10.0). CPS workers replied voluntarily to a recruitment ad shared through their place of work. Students were recruited by sending an ad through internal university channels. All participants received a certificate of participation in the training and a €10 voucher.

Design

The study had two between-subjects variables: experience with two levels (CPS workers vs students) and feedback with two levels (no feedback vs feedback). The combination of these variables produced four separate groups: CPS control [*n* = 15], CPS feedback [*n* = 16], student control [*n* = 17], and student feedback [*n* = 18]. All participants

participated in two training sessions with four avatar interviews in each. In the first session, half of the participants (experimental group) received feedback after each interview whereas the other half (control group) did not. In the second session, about four months later (116 days [range 83-163]), all participants received feedback after each interview. Due to attrition and technical difficulties the number of participants in the second round went down to 48, CPS control [n = 13], CPS feedback [n = 12], student control [n = 11], and student feedback [n = 12]. The design can be seen in Figure 1. Procedures were equal for all participants apart from the feedback provided in the first round of interviews. The study was approved by the Ethics Committee for Psychological Research at Mykolas Romeris University (Decision Nr. 7/2021).

PLEASE, INSERT FIGURE 1 HERE

Materials and Procedure

Avatar Selection. For both sessions, four out of sixteen possible avatars were randomly selected using a Latin Square method, ensuring no participant interviewed the same avatar twice. Both sets of four avatars contained an equal split of abuse and no-abuse scenarios, ages (4 vs 6), and gender (male vs female).

First Interview Session. Upon arrival, participants signed an informed consent form. Before each interview, participants received a background scenario about the avatar and reported their preliminary assumption about the presence of sexual abuse, along with their confidence (range 50% - 100%). Interviews lasted up to ten minutes but could be terminated earlier at the interviewer's discretion. After each interview, participants again rated their judgment about abuse presence and confidence. Participants in the experimental group received feedback on two recommended (e.g., What happened next? Who touched you?) and two not recommended questions (e.g., Did he touch you? Was it your stepdad?) they had used, plus information about what actually happened to the avatar. Control group participants proceeded directly to the next interview without feedback.

Second Interview Session. Conducted approximately four months later, following the same procedures as the first session, but with all participants receiving feedback after each interview and were debriefed at the end of the 8th interview.

Avatar Training System. Interviews were conducted using the Lithuanian version of CSA interview simulation software (Avatar Training; Segal et al., 2023a). The system features 16 avatars varying in age (4 vs 6 years), sex (male vs

female), abuse status (yes vs no), and emotionality (crying vs no crying). The Lithuanian version uniquely employs actors speaking in children's voices rather than computerized text-to-speech (Segal et al., 2023b).

The training focused on developing open-ended questioning skills through practice and individualized feedback, without theoretical instruction. Participants received case information and were asked to "find out what happened" within a 10-minute limit. They were instructed to skip rapport-building (described as already established) and focus on investigating the situation. Within well-established semi structured protocols, this would roughly be the “allegation phase” of the ten-step interview (Lyon, 2014) or the “substantive phase” of the NICHD protocol (Lamb et al., 2007, Hershkowitz et al., 2014)

Feedback consisted of reviewing two recommended and two not recommended questions from the interview, providing verbatim examples and explanations for each. For subsequent interviews, feedback prioritized new question types not previously addressed, while maintaining the two-category limit for not recommended questions. For example, if during the interview 1 the interviewer only asks option posing questions and unclear questions, while during the interview two they would ask, together with the first two categories used during the interview 1, suggestive and multiple-choice questions, feedback would be provided only on suggestive and multiple-choice questions.

For what concerns the feedback provided, if the interviewer asked this question: "Do you play football with dad?", this would be identified as an option-posing question. The interviewer, in this case, will receive the verbatim example of the question, the classification (option posing) with an explanation that such closed questions increase the probability of wrong details while focusing on previously unmentioned details typically eliciting yes/no responses (see also Table 1). The feedback was delivered live by trained raters who listened to each interview in real-time. When participants did not use any recommended techniques, raters demonstrated how to reformulate their actual questions into recommended ones (e.g., changing "Do you play football with dad?" to "What games do you play with dad?" when playing was previously mentioned), thus making the feedback individualized. The feedback duration usually lasted around 5-10 minutes.

Question Coding. Questions were coded in real-time by one of four operators using a scheme adapted from previous experiments (Pompedda, 2018; Segal et al., 2023a; see Table 1). All operators achieved at least 95% coding

reliability (64 out of 67 questions) after training. In addition, pseudo-open questions (e.g., can you tell me what happened?), were coded as not-recommended questions (in this case option-posing).

Response Algorithms. Separate algorithms governed responses for 4- and 6-year-old avatars. While recommended questions were more likely to elicit narrative details (mimicking real children's behaviour), the probabilistic nature of the algorithms meant this wasn't guaranteed. Responses included narrative details, yes/no answers, "I don't know," and resistance behaviours (e.g., "Can I go to play?"). Operators could also select responses providing basic avatar information.

Avatar

Details. Avatars provided two types of predefined narrative details: (1) Relevant details: Substantive information about the suspected abuse situation, including either abuse descriptions or alternative explanations for suspicious circumstances; (2) Neutral details: Non-substantive information about the child, family, or school. They do not refer to the emotional content of the reply.

Details were revealed in a fixed logical order, with the final four relevant details providing conclusive information about abuse status. To enhance realism, operators could manually select additional details at the beginning of the interview in relation to name age and regardless of the experimental condition, if an open question requesting these pieces of information was asked, participants would have received the correct answer (e.g., I am 4 years old). These questions are excluded from calculations. While avatars did not spontaneously provide false information, inappropriate questioning could elicit false details (e.g., asking "Did John touch you?" in a non-abuse case might elicit a "Yes" response).

[PLEASE INSERT FIGURE 2 HERE]

[PLEASE INSERT TABLE 1 HERE]

Statistical analyses

Hypotheses 1, 2 and 3 were tested conducting a series of experience (2) by feedback (2) by number of interviews (5) mixed measures ANOVAs on proportion of recommended questions, total number of recommended questions, total number of not recommended questions, which are direct measures of interviewers' behaviours, and on total number of

relevant details, and total number of wrong details which are based on the type of questions asked by the interviewer and mediated by the algorithm. As mentioned before, participants in the control group started receiving feedback after each interview in the second session. This means that they had not received any feedback before the first five interviews: four in the first round and one in the second session. For this reason, only the first five interviews out of eight were included in the analyses testing the hypotheses. However, a series of experience (2) by number of interviews (8) exploratory ANOVAs limited to the original feedback group were conducted to assess whether experience moderated the efficacy of the feedback intervention.

We employed both frequentist and Bayesian approaches to analyse training maintenance for several reasons: (a) The null hypothesis that skills are perfectly maintained is likely unrealistic (b) We needed to quantify evidence both for and against meaningful decline (c) Bayes factors allow us to evaluate the relative evidence for maintenance versus decline without requiring a specific threshold for what constitutes meaningful decline. Given that it is unrealistic to expect that the interviewers would be at the exact same level after four months, Bayes factors allow to determine how much more likely, based on the data at hand, one model is over the other (Lakens et al., 2020).

We decided to use a noninformative prior (Lakens et al., 2020) as we have no previous evidence of long-term effects within this setup. In addition, we decided to keep the area of potential effect size large as previous research (Pompedda et al., 2022) has shown large effects of feedback compared to control groups (d = 1.1). As we did not know if the effect of training would remain stable, increase, or if the participants would go back to baseline performance, this seemed the most objective approach. For this reason, we also used a non-directional approach. The uninformative prior was described by a Cauchy distribution and centered around zero with a width parameter of 1 (Schmalz et al., 2023) with a 70% probability that the effect size would be between -2 and 2.

In some of the analyses Mauchly's Test of Sphericity was violated, values of Epsilon (ε) varied too, hence based on a recommendation from the literature (Girden, 1992), when $\varepsilon < .75$, a Greenhouse-Geisser correction for degrees of freedom was used while when $\varepsilon > .75$, a Huynh-Feldt correction was applied. In cases where sphericity was not violated,

no correction was applied. No corrections for the main analyses testing the hypotheses have been applied, but Bonferroni correction was applied for follow-up tests.

In all analyses, according to Shapiro-Wilk tests, normality was violated in certain cells of the combination of variables, there were outliers, and Levene's test was also violated in certain instances. Given that normality and Levene's test were violated only in certain cells of the analysis we decided to run the analyses without any transformation nor to move towards non-parametric testing. In addition, as the outliers can be considered to represent true data in this experiment, we decided not to remove them.

Results

Descriptive Statistics

Descriptive statistics of experience with investigative interviews of children preceding the training with avatars are presented in Table 2. A series of Fisher's exact tests showed that there were statistically significant associations in the expected direction between previous experience and belonging to the CPS worker and student groups: 20/31 professionals were caregivers vs 8/35 students $\chi^2(1, N = 66) = 11.7$, $p < .001$, $\varphi = .42$, $p < .001$; 20/31 professionals had received previous training vs 8/35 students χ^2 $(1, N = 66) = 11.7$, $p < .001$, $\varphi = .42$, $p < .001$; 28/31 professionals had experience in interviewing children compared to vs 2/35 students χ^2 $(1, N = 66) = 47.5$, $p < .001$, $\varphi = .85$, $p < .001$; and 19/31 professionals had experience in interviewing children in CSA cases vs 0/31 students χ^2 $(1, N = 66) = 30.1$, $p < .001$, $\varphi = .68$, $p < .001$.

Table 3 provides means and standard deviations for split by our independent variables, namely, experimental conditions (Feedback VS No Feedback), Experience (CPS VS Students), and their interactions, over time (eight interviews).

PLEASE INSERT TABLE 2 AND 3 HERE

Testing Hypotheses 1 and 2 Using the First Five Interviews

Proportion of recommended questions. Analysis regarding the proportion of recommended questions adjusted with Huynh-Feldt correction revealed significant main effects for feedback ($F[1, 44] = 29.60$, $p < .001$, $\eta_{p^2} = .40$) and for

number of interviews (F[3.51,154.32] = 6.21., p < .001, η_{p^2} = .12), as well as a significant number of interviews*feedback interaction (F[3.51,154.32] = 6.63., p < .001, η_{p^2} = .13. All other effects were not significant (e.g., ps ≥ .428)

To follow up the statistically significant interaction, pairwise comparisons with Bonferroni corrections were computed. As shown in Figure 3 and in Table 4 feedback increased the proportion of recommended questions significantly from the second interview onwards.

Number of recommended questions. Analysis on the number of recommended questions adjusted with Huynh-Feldt correction revealed significant main effects for feedback (F[1,44] = 11.96., p = .001, η_{p^2} = .21) and number of interviews (F[3.84,168.83] = 8.90., p < .001, η_{p^2} = .17), as well as a significant number of interviews*feedback interaction (F[3.84,168.83] = 4.12., p = .004, η_{p^2} = .09). All other effects were not significant (e.g., ps ≥ .423)

Feedback increased the number of recommended questions significantly from the second interview onwards.

Number of not recommended questions. Analysis on the number of not recommended questions during the first five interviews adjusted with Huynh-Feldt correction revealed significant main effect in the expected direction for feedback (F[1,44] = 8.26., p = .006, η_{p^2} = .16), and number of interviews*feedback (F[4,176] = 2.82., p = .027, η_{p^2} = .06) interaction.

All other effects were not significant (e.g., ps ≥ .092)

The effect of feedback was significant from the second interview onwards but absent during the fifth interview.

Number of relevant details. Analysis on the number of relevant details with Huynh-Feldt correction revealed significant main effects for feedback (F[1,44] = 14.30, p < .001, η_{p^2} = .24) and for number of interviews (F[4,176] = 4.99., p < .001, η_{p^2} = .10) as well as a significant number of interviews*feedback interaction (F[4,176] = 2.84., p = .026, η_{p^2} = .06). All other effects were not significant (e.g., ps ≥ .119)

The effect of feedback was significant from the third interview onwards but absent during the fifth interview.

Number of wrong details. Analysis on the number of wrong details with Huynh-Feldt correction revealed a significant main effect in the expected direction for feedback (F[1,44] = 6.14, p = .017, η_{p^2} = .12). Participants in the

feedback group (M = 5.06, SE = .64) retrieved statistically fewer wrong details compared to the control group (M = 7.30, SE = .64). All other effects were not significant (e.g., ps ≥ .095)

These results fail to support Hypothesis 1 of a main effect of experience while they provide robust support for Hypothesis 2 of an improvement in interview quality and elicited information over the interviews in the feedback but not in the control group. There was no evidence of experience moderating the effect of feedback over the five interviews.

PLEASE INSERT TABLE 4 HERE

PLEASE INSERT FIGURE 3 HERE

Exploratory Analyses: Experience as a Moderator of the Effect of Feedback Using All Eight Interviews

In all these analyses, only participants in the original feedback group were included as these participants received feedback consistently throughout the whole experiment.

Proportion of recommended questions. Analysis on the proportion of recommended questions adjusted with Huynh-Feldt correction revealed a significant effect for the number of interviews (F[6.16, 129.27] = 18.69, p < .001, η_{p^2} = .47). All other effects were not significant (e.g., ps ≥ .737).

Number of recommended questions. Analysis on the number of recommended questions adjusted with Greenhouse-Geisser correction revealed a significant effect for number of interviews (F[3.68, 77.39] = 11.19, p < .001, η_{p^2} = .35). All other effects were not significant (e.g., ps ≥ .173).

Number of not recommended questions. Analysis on the number of not recommended questions adjusted with Huynh-Feldt correction revealed a significant effect for number of interviews (F[4.37, 124.54] = 6.07, p < .001, η_{p^2} = .22). All other effects were not significant (e.g., ps ≥ .583).

Number of relevant details. Analysis on the number of relevant details adjusted with Huynh-Feldt correction revealed a significant effect for number of interviews (F[6.94, 145.76] = 8.63, p < .001, η_{p^2} = .29). All other effects were not significant (e.g., ps ≥ .175).

Number of wrong details. Analysis on the number of wrong details adjusted with Huynh-Feldt correction revealed no significant effects (e.g., ps ≥ .100).

These results further confirm the previous analyses showing that experience did not moderate the effects of the feedback intervention.

PLEASE, INSERT TABLE 5 HERE

Testing of Hypothesis 3 Using the Last Interview of the First Session and First Interview of the Second Session

While the mixed measures ANOVA provided initial evidence regarding the training effect being maintained over the four-month break, we conducted additional Bayesian analyses to more rigorously evaluate the degree of evidence for maintenance versus decline. This complementary approach was particularly important given that traditional null hypothesis testing cannot provide positive evidence for the null hypothesis of maintained effects. To confirm the nonsignificant results of the mixed measures ANOVA, a series of Bayesian paired *t*-tests and Bayes factor analyses were run for all the variables of interest comparing the last interview of session one and the first interview of session two within the feedback group. We preferred this method over using the average of all four interviews in the first session as this would result in a more conservative measure of the learning during that session making it less likely to observe a decline in interviewing skills during the four-month break. The results in Table 4 provided some evidence in favour of the null hypothesis of no decline. However, Bayes factors ranged from 2.54 to 6.31 in the feedback group which according to common cut offs based on Jeffery (1998) are between anecdotal and moderate evidence (see also Beard et al., 2016). The posterior distribution means were all around zero apart from not-recommended questions and wrong details and all 90% credible intervals contained zeros. Robust checks suggested that the results would not qualitatively change based on the width of the distribution (see Figure 5 for proportion of recommended questions and total number of recommended questions).

PLEASE INSERT TABLE 6 HERE

PLEASE INSERT FIGURE 5 HERE

Discussion

The aim of the present study was three-fold: To replicate previous results on the effect of feedback in simulated interviews with avatars, test the effects of experience (both overall and as a moderator of learning during the avatar training), and finally test the long-term effects of the training.

Contrary to Hypothesis 1, experience had no impact on the quality of investigative interviews on any of the measures we evaluated. Also, experience did not interact with the feedback manipulation in either the analyses of the first five interviews or in the analyses of all eight interviews (the latter restricted to the experimental group only). This means that the CPS workers did not improve differently from the students. Given that the CPS workers were not better than the students overall or in the first interview, they had equally as much room to improve as the students. This means that we had a relatively powerful test of the idea that the CPS workers would be able to take better advantage of the training by not having to allocate attentional resources to manage the interview process itself.

An explanation for not finding any advantage of experience may be that, while CPS-workers have more contact than students with abused and neglected children, and they receive more training and instructions compared to students in communicating with children, they do not receive regular feedback on their performance. One of the solid results from the literature in this area is that the quality and characteristics of training has a strong impact on retention of information and transfer of the skills acquired during training (e.g., Lamb et al., 2018). Similar to how simulation-based learning is also used in training clinical reasoning (Decormeille et al., 2025), as clinical reasoning is a complex skill like interviewing. One important characteristic of training is that participants need to receive immediate and detailed feedback (Lamb et al., 2018; Smith, 2008), without this previous theoretical training is unlikely to have lasting effects. Most importantly, in real cases, it is rare that the ground truth of what happened is known (Lyon et al., 2020), which means that it is difficult to know if someone's decisions are correct or wrong, even in cases where there is a judicial outcome. This in turn means that it is rare that in real cases someone can receive detailed feedback about the correctness of specific details reported by a child witness.

Given that Lithuanian CPS workers do not conduct formal investigative interviews but rather understand their task as to be able to listen to the child, another reason for absence of an effect could be lack of practice. While they may have deepened theoretical knowledge of children behaviour, rapport-building, question types, and children development, which are all important theoretical information that might help to conduct interviews of good quality, this knowledge does not necessarily transfer into appropriate interview skills. The literature also warns us about the difference between theory and practice (e.g., Sternberg et al., 2001). Finally, experience might have a small effect size that could only be detected in studies with larger samples. Potential supporting evidence for this argument is that Pompedda et al. (2022) mega-analytic study suggested some positive effect of experience. It could also be that proactive interference theory and/or expertise reversal effects are at place here. However, as we do not have information on how many interviews have been conducted by our participants, we cannot say for sure if more expertise would hinder even more learning. Finally, it is possible that the CPS workers would have had an advantage in actual interviews with children and that the avatar training did not simulate important aspects of such interviews (Akca et al., 2021).

While the CLT would predict faster learning in experienced participants, our results suggest a more nuanced pattern. The similar pace in performance across both groups might be explained by the social-evaluative nature of the task. Both experienced and non-experienced participants likely faced stress during the simulated interviews, potentially interfering with their learning and performance. In fact, experienced participants may have encountered additional pressure due to heightened expectations of their expertise, potentially amplifying their stress response. This social-evaluative stress could have partially counteracted the cognitive advantages that their prior experience would theoretically provide. These findings align with research showing that even highly skilled professionals can experience performance decrements under social-evaluative pressure (Yu, 2015).

In line with previous literature, training with avatars coupled with feedback improved the quality of the interviews regarding the proportion of recommended questions, number of not recommended questions, and number of relevant details. In the analyses regarding the number of wrong details the interaction with number of interviews, while in the predicted direction, failed in yielding a statistically significant effect. However, the overall effect of feedback was

significant also for this variable. The pairwise comparisons showed that variables measuring question types started to show statistically significant differences between feedback and no-feedback groups earlier than variables measuring the quality of information elicited from the avatars. This is understandable given that the quality of information elicited from the avatars is only probabilistically related to the quality of the questions asked meaning that robust differences are required in the quality of questions for these to be also reflected in the quality of the information received from the avatars.

This is also important in relation to wrong details, that is, while they are correlated with utilising a not recommended questions style, their frequency is not related to the number of not recommended questions asked directly, but only probabilistically due to the algorithms being probabilistic. Giving a concrete example, one participant could ask 20 not recommended questions and receive one wrong detail, while another participant could ask 10 not recommended questions and receive three wrong details.

Within group differences showing change over the interviews were present in the feedback group and not in the no feedback group over the first four interviews. Surprisingly, there was some evidence of positive changes in the no feedback group from the fourth to the fifth interview. This result goes against both theoretical expectations and the bulk of previous empirical evidence. A possible explanation is that participants from feedback and no feedback groups, even if instructed not to discuss the nature of the training, talked with each other during the four-month break and shared effective questioning strategies.

As previously discussed, there is clear evidence that receiving feedback during training sessions plays a critical role in improving questioning skills (see also Haginoya et al., 2025). However, it is equally important to emphasize the foundational role of simulated interviews with avatars in enabling effective feedback delivery. The Avatar Training software is designed around simulated interviews with avatars developed to replicate real children, based on empirically validated response patterns. Before or after these simulations, interviewers are offered various types of interventions that provide opportunities and cues to refine their questioning techniques, thereby enhancing their skills. Therefore, while previous research, including the present study, has consistently shown that repeated avatar interviews alone do not

significantly improve questioning skills, this does not imply that avatar interviews as such would be an unnecessary feature of the training. On the contrary, simulated interviews with avatars that realistically mimic children in a cost-effective way are essential for ensuring that interventions such as feedback function properly. A combination of realistic simulations and well-structured interventions is crucial for achieving optimal training outcomes, provided they remain feasible and accessible.

Moreover, simulated interviews with avatars offer learners a psychologically safe environment. This risk-free setting allows them to practice and refine their interviewing techniques without the fear of making mistakes (see also Brubacher et al., 2025). Summarising, avatar training within this setup combines principles from the CLT aimed at favouring schemata constructions, automation and transfer within a safe environment where mistakes are part of the learning and help in providing individualised feedback. In addition, it allows plenty of possibilities for training with limited monetary and time burden. In addition, the software is a medium for completing interviews whenever needed, and with response algorithms not reproducible by humans with the same consistency, making the technology a fundamental part of the learning short term but also long term. Other studies have also looked at emotional responses utilising the same software, showing that the avatars are able to elicit emotions that have also behavioural impact (e.g., Segal et al, 2024a). Results from previous studies that utilised a similar software and training have shown that avatar training can elicit emotional reactions (Segal et al., 2024a) with behavioural consequences (Segal et al., 2023a; Segal et al., 2024b), short term effects in simulated interviews (e.g., Pompedda et al., 2022), and short- and long-term effects in field interviews (e.g., Kask et al., 2022).

However, it is worth noting that Avatar Training, including the current study, does not formally incorporate debriefing sessions in which learners reflect on their series of avatar interviews. An exception is the study by Krause and colleagues (2017), which examined the effects of reflection, asking participants to give specific examples of the questions they used after receiving feedback on question types. Although no significant improvements in questioning skills were observed in their study, the potential benefits of self-reflection on past behaviours remain underexplored in the

investigative interviewer training. Future research should therefore investigate the role of debriefing sessions following simulated interviews as part of the training process.

The present study adds more evidence and provides a replication of previous studies in a new group of participants and a new set of professionals that has not been tested before. More importantly, besides the preliminary findings reported in Haginoya and Santtila (2023), this is the first study to test long-term effect of a short training intervention with feedback and avatars without a significant theoretical component (as was the case in Brubacher et al., 2022). Evaluating the long-term effects of training manipulations is relevant to other skills-based fields like technical apprenticeship (Schriek et al., 2025). The combination of frequentist and Bayesian analyses provided a more complete picture of training maintenance than either approach alone. While the frequentist analysis showed no significant decline, the Bayesian analysis allowed us to quantify the strength of evidence for maintenance, revealing moderate support for sustained effects in most measures. In sum, we found a statistically significant difference between participants that conducted simulated interviews with avatars while receiving feedback and participants that did not receive feedback after four-months. Contrasts at the fifth interviews between feedback and no feedback groups, showed statistically significant differences in favour of the feedback group for the proportion of recommended questions, and the total number of recommended questions. While effect sizes were moderate, the number of not recommended questions, relevant details, and wrong details were not significant but in the expected direction.

We believe we are the first to test long-term effects of training utilising Bayes factors, hence it is difficult to compare the magnitude of these results with other studies. Based on the literature, participants tend to go back to baseline after feedback is discontinued (Lamb et al., 2002), with some exceptions (Brubacher et al., 2022). However, Brubacher and colleagues utilised a more complex and long intervention compared to the one utilised in this study.

Conclusion

To conclude, our results suggest that the training had a sustained impact over the four-month period without utilising intensive and spread practice. However, given that the performance of the interviewer was slightly worse, we recommend, as other colleagues in the field (e.g., Korkman et al., 2024), that refresher sessions should be conducted to

maintain the training effects potentially even after four months and to improve the quality of investigative interviews. These findings have important implications for training design. The effectiveness of our relatively brief intervention, combined with its sustained impact over a four-month period, suggests that well-structured simulation training with immediate feedback can benefit both novice and interviewers with field experience. However, the observed slight decline in performance over time indicates that periodic booster sessions may be valuable for maintaining and optimising interviewing skills.

Limitations and Future Directions

Due to organizational constraints, particularly the number of CPS-workers enrolled in the training, we did not conduct formal power calculations for main analysis, so the results should be treated with caution. Given that the sample size is small this can impair our ability to detect smaller effect sizes. Operators, while they received training and reached 95% agreement before the study, were not blind to the conditions and this could have impacted the results. It is unknown if the skills achieved will remain after four months. However, the structure of the training allows for booster sessions that do not require much time strain for the participants, which can counteract potential decline in the quality of investigative interviews. Future studies should aim to replicate these results within a bigger sample, and with a longer intervention interval. As we do not expect these training effects to remain stable indefinitely, a longer retention interval, or several, can help in mapping at which stage in time booster sessions might be useful. Future research should examine how additional interventions combined with feedback affect the retention of interviewing skills, particularly considering how interviewers' prior knowledge and experience levels influence their long-term skill maintenance.

Acknowledgments

We used Claude 3.5 Sonnet (Anthropic, 2024) to assist in restructuring sections of this manuscript. The LLM was used solely to reorganize existing content to improve logical flow and remove redundancies, without generating new content or modifying the original text beyond minor transitional adjustments. The reorganization was reviewed and approved by the authors to ensure it maintained the intended meaning and academic rigor of the original manuscript. All original content, analyses, and conclusions remain the work of the authors. The data collection was funded by the European Regional

Development Fund (project UNCOVABUSE No 01.2.2-LMTK-718-03-0067) under grant agreement with the Research Council of Lithuania. The work for Aleksandr Segal, Aistė Bakaitytė and Rita Žukauskienė was funded by the State Budget titled “Establishment of Centers of Excellence at Mykolas Romeris University,” which is implemented under the initiative “Centers of Excellence Initiative” initiated by the Ministry of Education, Science and Sports of the Republic of Lithuania.

References

Akca, D., Larivière, C. D., & Eastwood, J. (2021). Assessing the efficacy of investigative interviewing training courses: A systematic review. *International Journal of Police Science & Management*, *23*(1), 73–84. https://doi.org/10.1177/14613557211008470

Assembly, UN General. (1989). Convention on the rights of the child. *United Nations, Treaty Series, 1577*(3), 1-23.

Banevičienė, A., Vasiliauskienė, V., & Zuzevičiūtė, V. (2023). *Mapping child protection systems in the EU (27)*. Retrieved June 10, 2024, from https://fra.europa.eu/sites/default/files/fra_uploads/lt_-_report_-_mapping_child_protection_systems_-_2023.pdf

Baugerud, G. A., Johnson, M. S., Dianiska, R., Røed, R. K., Powell, M. B., Lamb, M. E., ... & Quas, J. (2024). Using an AI-based avatar for interviewer training at Children's Advocacy Centers: Proof of concept. *Child Maltreatment*, 10775595241263017. https://doi.org/10.1177/10775595241263017

Beard, E., Dienes, Z., Muirhead, C., & West, R. (2016). Using Bayes factors for testing hypotheses about intervention effectiveness in addictions research. *Addiction*, *111*(12), 2230–2247. https://doi.org/10.1111/add.13501

Benson, M. S., & Powell, M. B. (2015). Evaluation of a comprehensive interactive training system for investigative interviewers of children. *Psychology, Public Policy, and Law*, *21*(3), 309–322. https://doi.org/10.1037/law0000052

Blume, B. D., Ford, J. K., Baldwin, T. T., & Huang, J. L. (2010). Transfer of training: a meta-analytic review. *Journal of Management*, *36*(4), 1065–1105. https://doi.org/10.1177/0149206309352880

Brubacher, S. P., Powell, M., Skouteris, H., & Guadagno, B. (2015). The effects of e-simulation interview training on teachers' use of open-ended questions. *Child Abuse & Neglect*, *43*, 95–103. https://doi.org/10.1016/j.chiabu.2015.02.004

Brubacher, S. P., Shulman, E. P., Bearman, M. J., & Powell, M. B. (2022). Teaching child investigative interviewing skills: Long-term retention requires cumulative training: Psychology, Public Policy, and Law. *Psychology, Public Policy, and Law*, *28*(1), 123–136. https://doi.org/10.1037/law0000332

Brubacher, S. P., Powell, M. B., Johnson, M. S., Lopez Cano, M.-C., Hassan, S. Z., Riegler, M. A., Halvorsen, P., & Baugerud, G. A. (2025). Experts' views on artificial intelligence-based child chatbots to train investigative interviewing skills. Applied Cognitive Psychology.

Cederborg, A. C., Alm, C., Lima da Silva Nises, D., & Lamb, M. E. (2013). Investigative interviewing of alleged child abuse victims: An evaluation of a new training programme for investigative interviewers. Police Practice and Research, 14(3), 242–254. https://doi.org/10.1080/15614263.2012.712292

Chen, O., Kalyuga, S., & Sweller, J. (2017). The expertise reversal effect is a variant of the more general element interactivity effect. *Educational Psychology Review*, *29*, 393-405. https://doi.org/10.1007/s10648-016-9359-1

Decormeille, G., Geeraerts, T., Descoins, M., & Huet, N. (2025). Screen-based simulation in nursing school: help use and self-regulated learning. Applied Cognitive Psychology.

Girden, E. R. (1992). *ANOVA: Repeated Measures*. SAGE.

Haginoya, S., Ibe, T., Yamamoto, S., Yoshimoto, N., Mizushi, H., & Santtila, P. (2023). AI avatar tells you what happened: The first test of using AI-operated children in simulated interviews to train investigative interviewers. *Frontiers in Psychology*, *14*. https://doi.org/10.3389/fpsyg.2023.1133621

Haginoya, S., Sun, Y., Yamamoto, S., Mizushi, H., Yoshimoto, N., & Santtila, P. (2025). Improving questioning skills and use of supportive statements in simulated child sexual abuse interviews. Applied Cognitive Psychology, 39(1), e70031. https://doi.org/10.1002/acp.70031

Haginoya, S., & Santtila, P. (2023, July 4-7). *Retention or Decay?: The effect of a one-month interval on improved interviewing skills in child sexual abuse interviews with AI-driven avatars* [Oral presentation]. Annual Conference of the European Association of Psychology and Law, Cluj-Napoca, Transylvania, Romania.

Haginoya, S., Yamamoto, S., Pompedda, F., Naka, M., Antfolk, J., & Santtila, P. (2020). Online simulation training of child sexual abuse interviews with feedback improves interview quality in Japanese university students. *Frontiers in Psychology*, *11*. psyh. https://doi.org/10.3389/fpsyg.2020.00998

Haginoya, S., Yamamoto, S., & Santtila, P. (2021). The combination of feedback and modeling in online simulation training of child sexual abuse interviews improves interview quality in clinical psychologists. *Child Abuse & Neglect*, *115*. psyh. https://doi.org/10.1016/j.chiabu.2021.105013

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

Jeffreys, H. (1998). *The Theory of Probability*. OUP Oxford.

Kask, K., Pompedda, F., Palu, A., Schiff, K., Mägi, M.-L., & Santtila, P. (2022). Transfer of Avatar training effects to investigative field interviews of children conducted by police pfficers. *Frontiers in Psychology*, *13*. https://doi.org/10.3389/fpsyg.2022.753111

Hershkowitz, I., Lamb, M. E., & Katz, C. (2014). Allegation rates in forensic child abuse investigations: Comparing the revised and standard NICHD protocols. Psychology, Public Policy, and Law, 20(3), 336. https://psycnet.apa.org/doi/10.1037/a0037391

Keppel, G., & Underwood, B. J. (1962). Proactive inhibition in short-term retention of single items. Journal of Verbal Learning and Verbal Behavior, 1(3), 153–161. https://doi.org/10.1016/S0022-5371(62)80023-1

Kirschner, P., & Van Merrienboer, J. J. G. (2008). Ten steps to complex learning: a new approach to instruction and instructional design. In T. Good, *21st Century Education: A Reference Handbook* (p. I-244-I–253). SAGE Publications, Inc. https://doi.org/10.4135/9781412964012.n26

Ko, H., Hovden, E. A. S., Köpp, U. M. S., Johnson, M. S., & Baugerud, G. A. (2025). Using an AI-driven child chatbot avatar as a training tool for information gathering skills of dental and medical professionals: a pilot study. Applied Cognitive Psychology, 39(1), e70022. https://doi.org/10.1002/acp.70022

Korkman, J., Otgaar, H., Geven, L. M., Bull, R., Cyr, M., Hershkowitz, I., Mäkelä, J.-M., Mattison, M., Milne, R., Santtila, P., Van Koppen, P., Memon, A., Danby, M., Filipovic, L., Garcia, F. J., Gewehr, E., Gomes Bell, O., Järvilehto, L., Kask, K., … Volbert, R. (2024). White paper on forensic child interviewing: Research-based recommendations by the European Association of Psychology and Law. *Psychology, Crime & Law*, 1–44. https://doi.org/10.1080/1068316X.2024.2324098

Krause, N., Gewehr, E., Barbe, H., Merschhemke, M., Mensing, F., Siegel, B., Müller, J. L., Volbert, R., Fromberger, P., Tamm, A., & Pülschen, S. (2024). How to prepare for conversations with children about suspicions of sexual abuse? Evaluation of an interactive virtual reality training for student teachers. *Child Abuse & Neglect*, *149*, 106677. https://doi.org/10.1016/j.chiabu.2024.106677

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. *Applied Cognitive Psychology*, *31*(2), 187–198. https://doi.org/10.1002/acp.3316

Laajasalo, T., Korkman, J., Pakkanen, T., Oksanen, M., Tuulikki, L., Peltomaa, E., & Aronen, E. T. (2018). Applying a research-based assessment model to child sexual abuse investigations: model and case descriptions of an expert center. *Journal of Forensic Psychology Research and Practice*, *18*(2), 177–197. https://doi.org/10.1080/24732850.2018.1449496

Lafontaine, J., & Cyr, M. (2016). A Study of the relationship between investigators' personal characteristics and adherence to interview best practices in training. *Psychiatry, Psychology and Law*, *23*(5), 782–797. https://doi.org/10.1080/13218719.2016.1152925

Lafontaine, J., & Cyr, M. (2017). The relation between interviewers' personal characteristics and investigative interview performance in a child sexual abuse context. *Police Practice & Research, 18*(2), 106–118. https://doi.org/10.1080/15614263.2016.1242423

Lakens, D. (2017). Equivalence tests. *Social Psychological and Personality Science*, *8*(4), 355–362. https://doi.org/10.1177/1948550617697177

Lakens, D., McLatchie, N., Isager, P. M., Scheel, A. M., & Dienes, Z. (2020). Improving inferences about null effects with bayes factors and equivalence tests. *The Journals of Gerontology: Series B*, *75*(1), 45–57. https://doi.org/10.1093/geronb/gby065

Lamb, M. E., Brown, D. A., Hershkowitz, I., Orbach, Y., & Esplin, P. W. (2018). *Tell me what happened: Questioning children about abuse*. John Wiley & Sons.

Lamb, M. E., Orbach, Y., Hershkowitz, I., Esplin, P. W., & Horowitz, D. (2007). A structured forensic interview protocol improves the quality and informativeness of investigative interviews with children: A review of research using the NICHD Investigative Interview Protocol. Child abuse & neglect, 31(11-12), 1201-1231. https://doi.org/10.1016/j.chiabu.2007.03.021

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? *Applied Developmental Science*, *6*(1), 35–41. https://doi.org/10.1207/S1532480XADS0601_04

Lamb, M. E., Sternberg, K. J., Orbach, Y., Hershkowitz, I., Horowitz, D., & Esplin, P. W. (2002). The effects of intensive training and ongoing supervision on the quality of investigative interviews with alleged sex abuse victims. *Applied Developmental Science*, *6*(3), 114–125. https://doi.org/10.1207/S1532480XADS0603_2

Lyon, T. D. (2014). Interviewing children. Annual review of law and social science, 10(1), 73-89. https://doi.org/10.1146/annurev-lawsocsci-110413-030913

Lyon, T. D., Williams, S., & Stolzenberg, S. N. (2020). Understanding expert testimony on child sexual abuse denial after New Jersey v. J.L.G.: Ground truth, disclosure suspicion bias, and disclosure substantiation bias. *Behavioral Sciences & the Law*, *38*(6), 630–647. https://doi.org/10.1002/bsl.2490

Paas, F., Renkl, A., & Sweller, J. (2003). Cognitive load theory and instructional design: Recent developments. *Educational Psychologist*. https://doi.org/10.1207/S15326985EP3801_1

Paas, F., & van Merriënboer, J. J. G. (2020). Cognitive-Load Theory: methods to manage working memory load in the learning of complex tasks. *Current Directions in Psychological Science*, *29*(4), 394–398. https://doi.org/10.1177/0963721420922183

Pompedda, F. (2018). *Training in investigative interviews of children: Serious gaming paired with feedback improves interview quality*. [Doctoral Dissertation, Abo Akademi] https://www.doria.fi/handle/10024/152565

Pompedda, F., Antfolk, J., Zappalà, A., & Santtila, P. (2017). A combination of outcome and process feedback enhances performance in simulations of child sexual abuse interviews using avatars. *Frontiers in Psychology*, *8*(SEP). https://doi.org/10.3389/fpsyg.2017.01474

Pompedda, F., Palu, A., Kask, K., Schiff, K., Soveri, A., Antfolk, J., & Santtila, P. (2021). Transfer of simulated interview training effects into interviews with children exposed to a mock event. *Nordic Psychology*, *73*(1), 43–67. psyh. https://doi.org/10.1080/19012276.2020.1788417

Pompedda, F., Zappalà, A., & Santtila, P. (2015). Simulations of child sexual abuse interviews using avatars paired with feedback improves interview quality. *Psychology, Crime & Law*, *21*(1), 28–52. https://doi.org/10.1080/1068316X.2014.915323

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. *Journal of Police and Criminal Psychology*, *37*(3), 485–498. https://doi.org/10.1007/s11896-022-09509-7

Powell, M. B., Guadagno, B., & Benson, M. (2016). Improving child investigative interviewer performance through computer-based learning activities. Policing and Society, 26(4), 365–374. https://doi.org/10.1080/10439463.2014.942850

Powell, M. B., Hughes-Scholes, C. H., Smith, R., & Sharman, S. J. (2014). The relationship between investigative interviewing experience and open-ended question usage. *Police Practice and Research*, *15*(4), 283–292. https://doi.org/10.1080/15614263.2012.704170

Price, H. L., & Roberts, K. P. (2011). The effects of an intensive training and feedback program on police and social workers' investigative interviews of children. Canadian Journal of Behavioural Science, 43(3), 235–244. https://doi.org/10.1037/a0022541

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. *Frontiers in Psychology*, *14*. https://doi.org/10.3389/fpsyg.2023.1198235

Røed, R. K., Powell, M. B., Riegler, M. A., & Baugerud, G. A. (2023). A field assessment of child abuse investigators' engagement with a child-avatar to develop interviewing skills. Child Abuse & Neglect, 143, 106324. https://doi.org/10.1016/j.chiabu.2023.106324

Schmalz, X., Biurrun Manresa, J., & Zhang, L. (2023). What is a Bayes factor? *Psychological Methods*, *28*(3), 705–718. https://doi.org/10.1037/met0000421

Schriek, S., Berthold, K., & Hefter, M. (2025). Retrospective focus prompts facilitate learning from video tutorials for technical apprenticeship. Applied Cognitive Psychology.

Segal, A., Bakaitytė, A., Kaniušonytė, G., Ustinavičiūtė-Klenauskė, L., Haginoya, S., Zhang, Y., Pompedda, F., Žukauskienė, R., & Santtila, P. (2023a). Associations between emotions and psychophysiological states and confirmation bias in question formulation in ongoing simulated investigative interviews of child sexual abuse. *Frontiers in Psychology*, *14*. https://doi.org/10.3389/fpsyg.2023.1085567

Segal, A., Kaniušonytė, G., Bakaitytė, A., Žukauskienė, R., & Santtila, P. (2023b). The effects of emotions on the assessment of child sexual abuse interviews. *Journal of Police and Criminal Psychology*, *38*(4), 826-837. https://doi.org/10.1007/s11896-022-09571-1

Segal, A., Pompedda, F., Haginoya, S., Kaniušonytė, G., & Santtila, P. (2024a). Avatars with child sexual abuse (vs. no abuse) scenarios elicit different emotional reactions. *Psychology, Crime & Law*, *30*(3), 250-270. https://doi.org/10.1080/1068316X.2022.2082422

Segal, A., Bakaitytė, A., Kaniušonytė, G., Ustinavičiūtė-Klenauskė, L., Haginoya, S., Žukauskienė, R., & Santtila, P. (2024b). Are emotions and psychophysiological states experienced when observing a child sexual abuse interview associated with confirmation bias in subsequent question formulation? *Journal of Investigative Psychology and Offender Profiling*, e1643. https://doi.org/10.1002/jip.1643

Smith, M. C. (2008). Pre-professional mandated reporters' understanding of young children's eyewitness testimony: Implications for training. *Children and Youth Services Review*, *30*(12), 1355–1365. https://doi.org/10.1016/j.childyouth.2008.04.004

Sternberg, K. J., Lamb, M. E., Davies, G. M., & Westcott, H. L. (2001). The memorandum of good practice: theory versus application. *Child Abuse & Neglect*, *25*(5), 669–681. https://doi.org/10.1016/S0145-2134(01)00232-0

Sweller, J. (1988). Cognitive load during problem solving: effects on learning. *Cognitive Science*, 12(2), 257-285. https://doi.org/10.1207/s15516709cog1202_4

Young, J. Q., Van Merrienboer, J., Durning, S., & Ten Cate, O. (2014). Cognitive load theory: implications for medical education: AMEE Guide No. 86. *Medical teacher*, *36*(5), 371-384. https://doi.org/10.3109/0142159X.2014.889290

Yu, R. (2015). Choking under pressure: the neuropsychological mechanisms of incentive-induced performance decrements. *Frontiers in Behavioral Neuroscience*. 9:19. https://doi:10.3389/fnbeh.2015.00019

Figures

Figure 1. Experimental design.

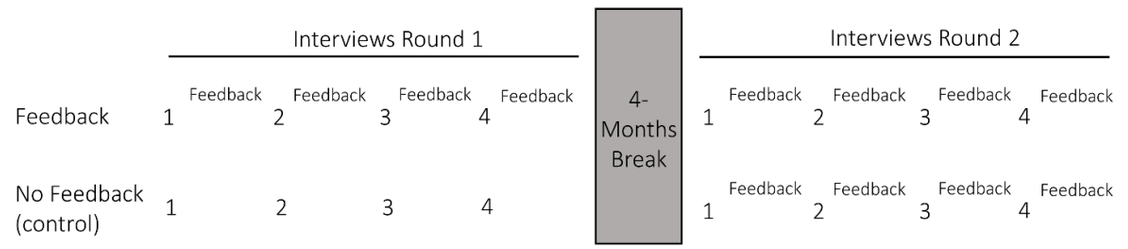


Figure 1 shows the experimental design utilised in the experiment. As clarified in the image, both groups received feedback in the second round of interviews. Essentially, the control group completed 5 interviews without feedback, four in round one and the first in round two, and three interviews with feedback in round two.

Figure 2

Computer-generated image of Miglė. One of the avatars used in the simulation

Figure 3

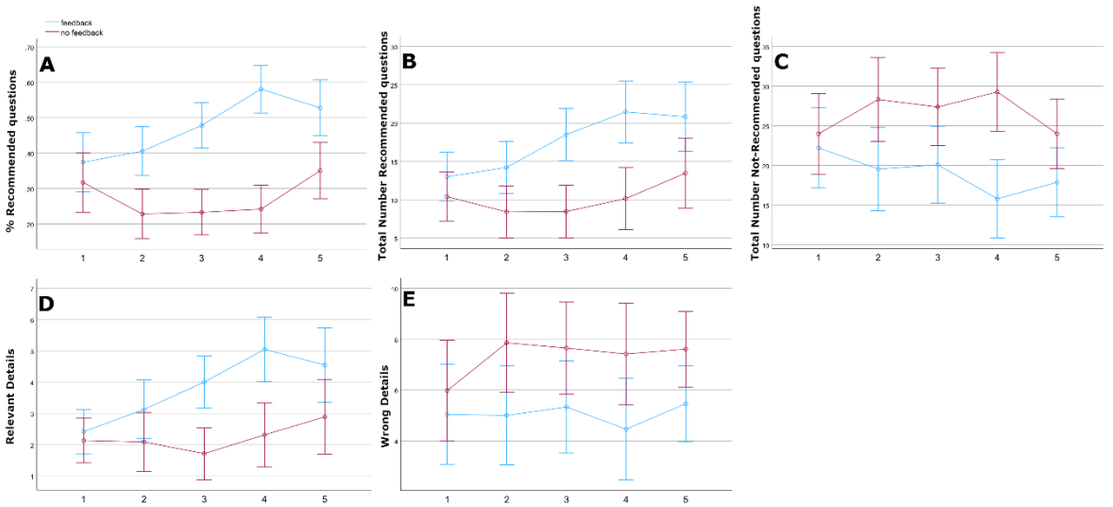


Figure 3 shows the effect of Feedback and Number of Interviews on all dependent variables. On the x axis the Number of Interviews (1-5) and on the y axis represent the estimated marginal means for: Panel A shows the proportion of recommended questions posed, Panel B shows the total number of recommended questions, Panel C shows the total number of not recommended questions, Panel D shows the total number of relevant or substantive details, and Panel E shows the total number of wrong detailsWhiskers show 95 CI, when CIs do not overlap there is a statistically significant difference.

Figure 4

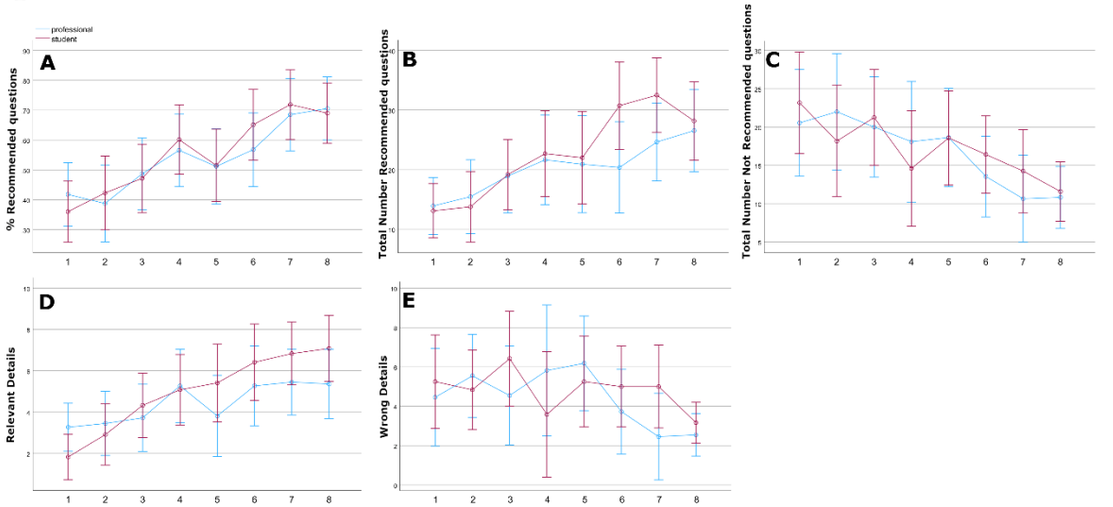
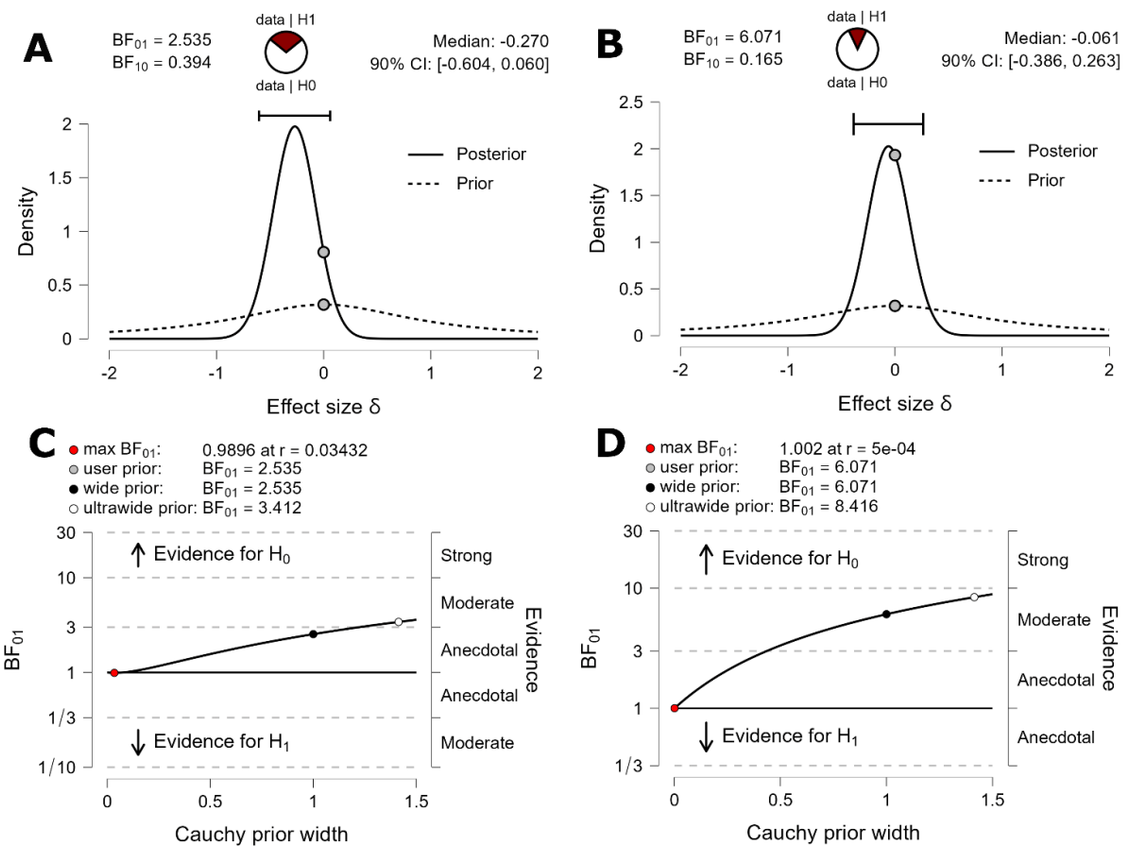


Figure 4 shows the effect of Experience within the Feedback group and Number of Interviews on all dependent variables. On the x axis the Number of Interviews (1-8) and on the y axis represent the estimated marginal means for: Panel A shows the proportion of recommended questions posed, Panel B shows the total number of recommended questions, Panel C shows the total number of not recommended questions, Panel D shows the total number of relevant or substantive details, and Panel E shows the total number of wrong details. Whiskers show 95 CI, when CIs do not overlap there is statistically a significant difference.

Figure 5. Prior and Posterior distributions and Bayes Factor Robustness Check for proportion of recommended questions and total number of recommended questions



Panels A and B show the prior and post distribution. On the left for the proportion of recommended questions and on the right for the total number of recommended questions. On the x axis the effect size while on the y axis the density of prior and posterior distributions. BF01 is the bayes factor in favour of the null hypothesis (means will be similar in the population) while BF10 is the bayes factor in favour of the alternative hypothesis (means will be different in the population). The pizza slot showed evidence in favour of the two hypotheses. Being the part in red (data in favour of H1) smaller than the white part (data in favour of H0) we can conclude that even if participants had a decline, this effect would not be significant in the population. Dots show effect size zero on both distributions. CI = Credible Interval
Panels C and D show the strength of the model. On the left of for the proportion of recommended questions and on the right for the total number of recommended questions. Width, Cauchy distribution, was set at 1 and it is not visible as it is under the black dot. The graph shows how the model can be affected by the prior utilised however, the qualitative interpretation remains similar

Tables

Table 1
Question-types coding utilised in the simulation and based on previous experiments

Category	Description	Examples
	Recommended Questions	
Facilitators	Open-ended and non-suggestive utterances that encourage the child to continue with the previous answer. Includes also requests for clarification, and echoing of the child's words.	"Ok" "Continue"
Invitations	Open-ended utterances (questions, statements, or imperatives) used to elicit free recall responses from the child. Invitations could be broad or narrow	"Tell me everything that happened from the very beginning to the end", "You mentioned that you like playing, tell me more about it"
Directive	Open-ended and non-suggestive questions that focus the child's attention on a previously mentioned detail asking for a focalized explanation.	"Where did you play football?"
	Not Recommended Questions	
Option-posing	Closed-ended questions that focus on unmentioned details, without implying a particular type of response or on a mentioned detail asking the child to provide a yes/no answer.	"Do you play with dad?"
Specific suggestive	Open/Closed-ended questions that are based on an unmentioned detail and express the expected response from the child.	"David did something bad, didn't he?", "Someone abused you. Tell me who!"
Unspecific suggestive	Open/Closed-ended questions that are **not** based on an unmentioned detail but express the expected response from the child (e.g., social and peer pressure).	"This is not what your mother has told me!"
Repetitions	Repetitions of a previously recommended or not recommended question more than once in a row	
Too-long/Unclear	Questions that use a language too complicated for the cognitive level of the child, and/or formulated haphazardly, and/or contains more than one concept at the time.	"Do you play with dad? What type of games do you play?"
Multiple choices	Questions that provide a predetermined list from which the child is requested to pick from.	"Did you play this game with Luke or Matthew?"
Time	Open/Closed-ended questions that require the child to provide or recollect precise time-related information.	"How many months ago this happened?"
Fantasy	Open/Closed-ended questions that can activate the child fantasy or move the discussion from the reality to the fantasy level.	"Imagine you are the princess, and dad is the king. What would the king have done to you?
Feelings	Open/Closed-ended questions that require the child to provide accounts regarding own or others feelings	"How did your mum feel when you told her about the games you played with dad?"

Table 2

Descriptive Statistics of the preceding experience divided by experimental groups

	N				Frequency (%)
Professionals	31	CSA Interviews	Feedback	yes	7 (23%)
				no	9 (29%)
			No feedback	yes	12 (39%)
				no	3 (10%)
		Raising Children	Feedback	yes	9 (29%)
				no	7 (23%)
			No feedback	yes	11 (35%)
				no	4 (13%)
		Training Received	Feedback	yes	9 (29%)
				no	7 (23%)
			No feedback	yes	11 (35%)
				no	4 (13%)
		Interviewing children	Feedback	yes	15 (48%)
				no	1 (3%)
			No feedback	yes	13 (42%)
				no	2 (6%)
Students	35	CSA Interviews	Feedback	yes	0
				no	18 (51%)
			No feedback	yes	0
				no	17 (49%)
		Raising Children	Feedback	yes	4 (11%)
				no	14 (40%)
			No feedback	yes	4 (11%)
				no	13 (37%)
		Training Received	Feedback	yes	4 (11%)
				no	14 (40%)
			No feedback	yes	4 (11%)
				no	13 (37%)
		Interviewing children	Feedback	yes	2 (6%)
				no	16 (46%)
			No feedback	yes	0
				no	17 (49%)

Note. In parenthesis the percentage out of the total number of participants for each group rounded at the nearest whole number

Table 3 Descriptive statistics for all dependent variables over time divided by Feedback VS No Feedback and Professionals VS Students

	Interview Number	*Total Number Recommended Questions*		*Total Number Not Recommended Questions*		*Total number of Questions*		*Percentage Recommended Questions*		*Relevant Details*		*Wrong details*	
		M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Feedback	1	14.27	11.49	21.94	10.65	36.21	14.56	0.38	0.18	2.55	1.94	5.03	3.57
	2	14.09	8.77	20.09	11.73	34.18	15.02	0.40	0.19	3.15	2.18	5.29	3.68
	3	18.15	9.87	19.50	9.54	37.65	13.37	0.46	0.19	3.74	2.51	5.24	3.70
	4	20.59	10.57	16.06	12.73	36.65	16.22	0.58	0.18	4.85	2.69	4.12	4.70
	5	20.83	12.80	17.88	10.43	38.71	16.30	0.53	0.21	4.54	3.15	5.46	3.88
	6	25.78	13.07	15.04	8.34	40.83	15.54	0.61	0.20	5.87	3.06	4.39	3.41
	7	28.74	10.96	12.52	9.03	41.26	14.05	0.70	0.19	6.17	2.57	3.78	3.67
	8	27.39	10.76	11.22	6.31	38.60	9.63	0.70	0.16	6.26	2.75	2.87	1.71
No feedback	1	9.53	6.83	22.69	15.30	32.22	19.36	0.34	0.23	1.91	1.44	5.44	5.48
	2	8.13	5.80	26.75	15.27	34.87	17.70	0.25	0.16	1.88	2.11	7.53	5.74
	3	8.25	5.89	25.16	13.36	33.41	16.56	0.26	0.15	1.78	1.36	6.69	4.59
	4	10.06	6.85	28.03	13.57	38.09	16.09	0.27	0.21	2.06	1.88	7.75	4.93
	5	13.58	8.44	24.54	12.27	38.12	15.85	0.35	0.17	2.92	2.64	7.79	3.89
	6	19.67	11.20	20.46	14.67	40.12	15.88	0.50	0.22	4.13	2.36	5.75	4.76
	7	24.52	8.31	14.43	8.89	38.96	11.52	0.64	0.17	5.39	2.33	4.70	2.91
	8	25.30	11.78	12.78	7.94	38.09	11.70	0.65	0.19	5.74	2.05	3.52	3.38
Professionals	1	11.81	6.76	24.84	12.43	36.64	14.82	0.34	0.17	2.77	1.77	5.16	4.40
	2	12.68	9.40	25.39	14.70	38.06	16.79	0.34	0.20	2.90	2.44	6.58	4.92
	3	13.61	9.76	23.71	14.03	37.32	15.64	0.38	0.22	2.55	2.34	5.74	4.28
	4	15.71	10.83	24.42	16.08	40.13	16.82	0.43	0.26	3.52	2.49	6.35	5.56
	5	17.20	10.19	24.12	12.93	41.32	17.23	0.43	0.19	3.44	2.48	7.84	4.58
	6	19.96	10.55	19.35	13.72	39.30	15.63	0.51	0.19	4.74	2.85	5.39	4.87
	7	24.95	11.21	13.23	9.47	38.18	13.04	0.66	0.20	5.45	2.40	3.73	3.16
	8	25.77	9.85	13.27	7.92	39.04	9.68	0.66	0.18	5.36	2.44	3.73	3.27
Students	1	12.06	11.89	20.00	13.34	32.06	18.85	0.38	0.24	1.74	1.56	5.29	4.79
	2	9.89	6.39	21.49	13.02	31.37	15.31	0.32	0.18	2.20	1.98	6.20	4.92
	3	13.11	9.47	20.94	9.44	34.06	14.53	0.35	0.19	3.00	2.18	6.11	4.16
	4	15.29	10.07	19.60	12.46	34.88	15.15	0.42	0.23	3.49	2.92	5.46	4.73
	5	17.22	12.69	18.04	9.65	35.26	14.03	0.45	0.23	4.04	3.48	5.30	2.85
	6	25.25	13.68	16.33	10.57	41.58	15.72	0.60	0.23	5.21	2.87	4.79	3.44

	7	28.17	8.36	13.71	8.57	41.87	12.51	0.68	0.16	6.08	2.52	4.71	3.43
	8	26.88	12.51	10.83	6.28	37.71	11.55	0.69	0.17	6.58	2.28	2.71	1.92
	Interview Number	*Total Number Recommended Questions*		*Total Number Not Recommended Questions*		*Total number of Questions*		*Percentage Recommended Questions*		*Relevant Details*		*Wrong details*	
Feedback* Professionals	1	13.91	7.71	20.55	10.37	33.56	13.48	0.38	0.17	3.27	2.05	4.45	5.37
	2	15.45	12.36	22.00	15.68	37.19	18.48	0.41	0.23	3.45	2.62	5.55	3.62
	3	18.91	11.94	20.00	13.42	37.06	14.34	0.48	0.21	3.73	2.72	4.55	3.78
	4	21.64	14.28	18.09	17.02	38.25	19.83	0.55	0.24	5.27	2.69	5.82	6.70
	5	20.91	13.22	18.64	10.63	36.83	19.35	0.54	0.20	3.82	3.12	6.18	4.73
	6	20.36	10.75	13.55	7.46	33.91	12.55	0.57	0.19	5.27	3.38	3.73	2.83
	7	24.64	13.42	10.64	8.51	35.27	13.34	0.69	0.23	5.45	3.01	2.45	2.84
	8	26.55	9.48	10.82	5.86	37.36	9.92	0.71	0.16	5.36	2.84	2.55	1.13
No Feedback * Professionals	1	12.09	7.06	29.73	14.28	39.93	15.93	0.29	0.16	2.45	1.37	5.55	4.99
	2	10.27	6.90	32.09	14.20	39.00	15.38	0.26	0.14	2.00	1.90	8.73	6.39
	3	9.64	6.67	29.55	14.31	37.60	17.43	0.27	0.17	1.09	0.70	7.36	5.41
	4	11.73	6.50	30.82	9.38	42.13	13.29	0.31	0.24	2.45	1.86	8.09	4.83
	5	15.27	4.38	31.45	11.13	45.46	14.56	0.33	0.10	3.45	2.02	10.09	3.45
	6	19.45	11.35	23.64	16.54	44.25	17.02	0.45	0.18	4.00	2.24	6.55	6.04
	7	25.27	9.14	15.82	10.06	41.09	12.67	0.63	0.18	5.45	1.75	5.00	3.07
	8	25.00	10.60	15.73	9.17	40.73	9.60	0.61	0.20	5.36	2.11	4.91	4.25
Feedback * Students	1	13.08	7.50	23.17	11.73	38.71	15.50	0.38	0.20	1.83	1.64	5.25	1.91
	2	13.75	6.92	18.17	7.54	31.50	10.98	0.40	0.16	2.92	2.31	4.83	3.13
	3	19.17	7.36	21.25	6.61	38.17	12.85	0.45	0.18	4.33	2.46	6.42	4.21
	4	22.67	9.40	14.58	6.16	35.22	12.62	0.60	0.10	5.08	2.97	3.58	3.58
	5	22.00	12.73	18.58	9.88	40.58	13.19	0.52	0.22	5.42	3.15	5.25	2.80
	6	30.75	13.42	16.42	9.17	47.17	15.76	0.65	0.20	6.42	2.78	5.00	3.88
	7	32.50	6.65	14.25	9.52	46.75	12.83	0.72	0.16	6.83	1.99	5.00	4.02
	8	28.17	12.19	11.58	6.93	39.75	9.65	0.69	0.17	7.08	2.50	3.17	2.12
No Feedback * Students	1	9.36	8.70	17.73	13.85	25.41	19.97	0.37	0.28	1.73	1.68	6.18	6.42
	2	7.45	5.50	25.45	13.52	31.24	19.24	0.25	0.18	1.64	1.29	7.64	5.39
	3	7.55	6.33	23.73	11.44	29.71	15.31	0.25	0.14	2.27	1.74	7.91	4.57
	4	8.55	5.66	26.36	13.51	34.53	17.84	0.24	0.18	2.09	1.97	6.91	4.25
	5	12.00	10.90	17.45	9.82	29.45	13.06	0.37	0.23	2.55	3.33	5.36	3.04
	6	20.09	12.57	16.00	12.80	36.00	14.16	0.55	0.26	4.09	2.62	4.55	3.24
	7	24.64	7.65	13.55	8.16	37.00	10.52	0.65	0.17	5.82	2.40	4.64	2.91

8	25.64	13.88	10.00	6.03	35.67	13.30	0.69	0.17	6.09	2.12	2.27	1.74

Note. Highlighted by a shade of grey the interviews conducted with feedback by the original control group. Between interview 4 and 5 happened the 4-months-break.

*Table 4 Pairwise comparison with Bonferroni Correction for significant interactions Feedback*Number of Interviews during the first five interviews*

Interview Number	*Total Number of Recommended Questions*					*Total Number of Not Recommended Questions*					*Percentage of Recommended Questions*				
	Feedback		*No Feedback*			*Feedback*		*No Feedback*			*Feedback*		*No Feedback*		
	M	*SE*	*M*	*SE*	η_p^2	*M*	*SE*	*M*	*SE*	η_p^2	*M*	*SE*	*M*	*SE*	η_p^2
1	$13.00_{\phi, a}$	1.58	10.41_{ϕ}	1.58	.03	$22.21_{\phi,}$	2.51	$23.98_{\phi,}$	2.52	.01	$0.37_{\phi, a}$	0.04	$0.32_{\phi,}$	0.04	.02
2	$14.21_{\phi, a}$	1.68	8.42_{ω}	1.69	.12	$19.54_{\phi,}$	2.61	$28.30_{\omega,}$	2.62	.11	$0.41_{\phi, ab}$	0.03	$0.23_{\omega,}$	0.03	.23
3	$18.50_{\phi, b}$	1.69	$8.46_{\omega, a}$	1.70	.28	$20.08_{\phi,}$	2.41	$27.36_{\omega,}$	2.42	.09	$0.48_{\phi, ac}$	0.03	$0.23_{\omega, a}$	0.03	.40
4	$21.46_{\phi, b}$	2.01	10.16_{ω}	2.01	.26	$15.79_{\phi,}$	2.45	$29.26_{\omega,}$	2.46	.25	$0.58_{\phi, d}$	0.03	$0.24_{\omega,}$	0.03	.54
5	$20.83_{\phi, b}$	2.24	$13.46_{\omega, b}$	2.25	.11	$17.87_{\phi,}$	2.16	$24.00_{\phi,}$	2.17	.08	$0.53_{\phi, bcd}$	0.04	$0.35_{\omega, b}$	0.04	.19

Interview Number	*Total Number of Relevant Details*					*Total Number of Wrong Details*[1]				
	Feedback		*No Feedback*			*Feedback*		*No Feedback*		
	M	*SE*	*M*	*SE*	η_p^2	*M*	*SE*	*M*	*SE*	η_p^2
1	$2.42_{\phi, a}$	0.35	$2.13_{\phi,}$	0.35	.01	5.04	0.98	5.98	0.98	#
2	$3.12_{\phi, abc}$	0.46	$2.09_{\phi,}$	0.47	.05	5.00	0.97	7.86	0.97	#
3	$4.00_{\phi, bd}$	0.41	$1.71_{\omega,}$	0.41	.26	5.33	0.90	7.65	0.90	#
4	$5.04_{\phi, d}$	0.51	$2.31_{\omega,}$	0.51	.25	4.46	0.99	7.42	0.99	#
5	$4.54_{\phi, cd}$	0.59	$2.89_{\phi,}$	0.59	.08	5.46	0.74	7.60	0.74	#

Note. *N*=48. Feedback *n*= 24, No feedback *n*=24, Professional *n*=25, Students *n*=23. Estimated marginal means are reported. For raw means refer to table 3. Means sharing different subscripts differ significantly at .05 as indicated by Bonferroni Correction. Roman letters are used for column comparisons (within) while Greek letters are used for row comparisons (between). η_p^2 represents the effect size of the between contrasts. 95%. Confidence Intervals are visually reported in the Figures. [1]For wrong details only Estimated Marginal means have been computed because the interaction effect was not significant.

Table 5 Pairwise Comparisons with Bonferroni Correction for the significant effect of Interview number in the feedback group.

Interview Number	*Total Number Recommended Questions*		*Total Number Not Recommended Questions*		*Percentage Recommended Questions*		*Relevant Details*		*Wrong details*	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
1	13.48 $_{a}$	7.44	21.91 $_{a}$	10.93	0.39 $_{a}$	0.17	2.52 $_{a}$	1.95	4.87	3.88
2	14.57 $_{ab}$	9.70	20.00 $_{ab}$	12.00	0.41 $_{ab}$	0.20	3.17 $_{ab}$	2.42	5.17	3.31
3	19.04 $_{abc}$	9.58	20.65 $_{ab}$	10.20	0.48 $_{bc}$	0.19	4.04 $_{ac}$	2.55	5.52	4.03
4	22.17 $_{cd}$	11.71	16.26 $_{abc}$	12.40	0.58 $_{ce}$	0.19	5.17 $_{c}$	2.77	4.65	5.30
5	21.48 $_{abce}$	12.68	18.61 $_{ab}$	10.01	0.51 $_{abcd}$	0.20	4.65 $_{bc}$	3.17	5.70	3.78
6	25.78 $_{cf}$	13.07	15.04 $_{abc}$	8.34	0.61 $_{cde}$	0.20	5.87 $_{c}$	3.06	4.39	3.41
7	28.74 $_{def}$	10.96	12.52 $_{bd}$	9.03	0.70 $_{e}$	0.19	6.17 $_{c}$	2.57	3.78	3.67
8	27.39 $_{cf}$	10.76	11.22 $_{cd}$	6.31	0.70 $_{e}$	0.16	6.26 $_{c}$	2.75	2.87	1.71

Note. N=23, Professionals n=11, students n=12. Means sharing different subscripts differ significantly at .05 as indicated by Bonferroni Correction. Roman letters are used for column comparisons (within).

For wrong details only descriptive statistics are reported are effects were not significant.

Table 6

Bayesian Paired sample t-test with Bayes factors for all dependent variables comparing the first interview after the break with the last interview before the break the Feedback group

									Posterior Distribution	
	Variable	*N*	M^{diff}	*SD*	BF_{01}	*t*	*df*	*p*	*[90 Credible L]*	*[90 Credible U]*
Feedback	% Recommended	24	-0.05	0.18	2.53	-1.41	23	.172	-0.12	0.01
	Recommended	24	-0.63	9.6	6.07	-0.32	23	.752	-4.15	2.90
	Not Recommended	24	2.08	12.0	4.51	0.85	23	.403	-2.32	6.48
	Neutral details	24	-0.08	2.8	6.31	-0.15	23	.885	-1.11	0.94
	Wrong details	24	1.00	4.4	3.54	1.12	23	.275	-0.61	2.61

% Recommended = Proportion Recommended Question. BF_{01} *shows evidence in favour of the null hypothesis that the two means will not be different in the population. For example, concerning recommended questions, with a* BF_{01} *= 6.07, this result suggests that the data are 6 times more likely in favour of the null hypothesis. T test presented is a frequentist t-test.*