Transfer of avatar training effects to mock investigative interviews with adult witnesses

Mari-Liis Tohvelmann, Kristjan Kask, Shumpei Haginoya, Pekka Santtila

https://doi.org/10.3013/vpag3r71

Abstract

Interviewing witnesses is a delicate and complex process that must follow certain principles. Previous studies have noted that investigative interviews with adult victims and witnesses are often of low quality. Research also suggests that interviewing skills improve when the interviewers are given feedback on their performance. We have created a web-based adult witness avatars (AWA) software to simulate investigative interviews with adult victims and witnesses that would help interviewers train question types that will allow the witnesses to recall more accurate accounts. We examined whether avatar interviews coupled with feedback (versus no feedback) would result in improvements in interviews with human mock witnesses. Half of the 58 participants received process feedback after each of four simulated interviews. The avatars revealed predefined memories and made errors as a function of algorithms formulated based on previous empirical research on the response behaviour of adult witnesses in experimental studies. We found that feedback was linked to a high proportion of recommended questions within the AWA training itself. Receiving feedback after the simulated interviews increased the proportion of recommended questions (i.e. free recall and open questions) in mock interviews with human interviewees compared with not receiving feedback (67 % versus 49 %, respectively). However, there were no significant differences between feedback and control groups in terms of correct or incorrect details elicited from the mock witnesses. Future steps and implications for this type of software in training investigative interviewing skills will be discussed.

Keywords: interviewer training, adult witnesses, avatars, serious gaming, feedback, transfer effect.

1. Introduction

Interviews with victims and witnesses are an important source of evidence, especially in tackling organised crime such as high-risk criminal networks. Unfortunately, these statements can be influenced by inappropriate interviewing methods, potentially resulting in hampered investigations or even miscarriages of justice (Milne and Bull, 1999). Therefore, it is important that witnesses are interviewed appropriately.

Asking questions that elicit accurate answers is an important element of recent best practice guidelines (Principles on effective interviewing for investigations and information gathering, 2021; United Nations, 2024; Halley et al., 2023; UK Ministry of Justice, 2022; Bjerknes and Fahsing, 2018; Griffiths and Rachlew, 2018; Fisher and Geiselman, 1992). These guidelines all stress asking questions that facilitate good communication with the witness and will help them recall as much information as possible. Unfortunately, interviewers have difficulties in applying these principles in practice (Walsh et al., 2025).

Questions can be categorised in different ways (Nunan et al., 2020) as productive and unproductive (Griffiths and Milne, 2006), appropriate and inappropriate (Phillips et al., 2011), or recommended and not recommended (Pompedda et al., 2017). It is preferable to enhance witnesses' recall by letting them freely recollect what they have experienced or direct their recall using open questions (Oxburgh et al., 2010). On the contrary, option-posing, repeated, suggestive or misleading questions are not recommended as they increase the proportion of inaccurate information (Otgaar et al., 2023). Therefore, the wording of questions posed to witnesses is central to ensuring the detail and accuracy of witness statements.

Theory-based training is the most common way of training interviewers (Pompedda, 2018). It increases the interviewers' knowledge of the topic but may not actually improve their interviewing skills (Lamb, 2016). Immediate personalised continuous feedback is needed to change interviewer behaviour (Casey and Powell, 2021). As technology adds new possibilities, serious gaming solutions in virtual environments have been found to be effective in improving learners' performance (Graafland et al., 2012; van der Zee et al., 2012). Such solutions are particularly attractive as they make the provision of immediate feedback quick and cost-effective (see Kask et al., 2024). However, research has focused on child interviewing; therefore, there is a lack of methods for training in adult witness interviewing (Tohvelmann and Kask, 2022).

Tohvelmann and Kask (2022) therefore proposed a serious gaming solution targeted at improving the quality of interviews with adult witnesses. Tohvelmann et al. (2025) created a web-based solution called adult witness avatars (AWA) to simulate investigative interviews with adult victims and witnesses. They examined whether simulated interviews with avatars, coupled with feedback (versus no feedback), would improve interview quality. Interviews with repeated feedback included a higher proportion of recommended questions (i.e. free recall and open questions) than those without feedback (90.3 % versus 72.6 %, respectively). However, despite these promising results, it is not known whether the effect of this training in avatar interviews would transfer to subsequent interviews with real witnesses.

1.1. Current study

Given that Tohvelmann et al. (2025) showed that repeated feedback after a series of simulated investigative interviews increased the proportion of recommended questions in avatar interviews, we examined whether this training effect would transfer to interviews with human mock witnesses. More precisely, we examined whether AWA training sessions coupled with feedback (versus no feedback) would be effective in changing (novice) interviewer behaviour both during training and in subsequent interviews with a human mock witness. First, we hypothesised that receiving feedback (versus not receiving feedback) regarding questions asked during the avatar interviews would result in (a) an increased number of recommended questions, (b) a decreased number of non-recommended questions, and (c) a higher proportion of recommended questions overall. Second, we hypothesised that having received feedback (versus not having received feedback) during the avatar interviews would result in (a) an increased number of recommended questions, (b) a decreased number of non-recommended questions, and (c) a higher proportion of recommended questions in subsequent interviews with human mock witnesses. Third, we hypothesised that the participants who had received feedback (versus those who had not received feedback) during the avatar training session would elicit (a) more correct details, (b) fewer incorrect details, and (c) a higher proportion of correct details overall. Fourth, we also hypothesised that participants asking a higher proportion of recommended questions during avatar interviews would (a) ask a higher proportion of recommended questions in interviews with human mock witnesses; and (b) elicit a higher proportion of correct details in these interviews.

2. Method

We preregistered the desired sample size, including variables, hypotheses and planned analyses on AsPredicted ([13]). The dataset containing all primary study variables can be obtained from the authors.

2.1. Participants

We recruited 60 interviewer participants but two of them were excluded as their interviewees could not be reached for the online interview. Therefore, the final number of interviewer participants was 58 (see 'Characteristics' in Table 6.1). The experiment was conducted in Estonian. The participants were neither law students nor third-year psychology students (since legal psychology is an optional course for the third-year students at the university) and had conducted no prior investigative interviews. The feedback and control groups did not differ significantly in terms of gender, $\chi2(2) = 1.47$, $p = 0.48$, age, $t(56) = -1.355$, $p = 0.18$, or education level, $\chi2(3) = 2.83$, $p = 0.24$. We also recruited 58 interviewee participants (see 'Characteristics' in Table 6.1). The feedback and control groups did not differ significantly in terms of gender, $\chi2(1) = 0.17$, $p = 0.681$, age, $t(53) = 0.066$, $p = 0.947$, or education level, $\chi2(3) = 2.96$, $p = 0.40$. The participants did not receive any form of recompense for their participation.

([13]) https://aspredicted.org/blind.php?x=JQN_DX3.

Table 6.1. Demographics of interviewers and interviewees

Characteristics	Interviewers	Interviewees
Mean age	23 years (SD ([14]) = 2.8, range 19–29)	32.8 years (SD = 10.4, range 18–56)
Gender	30 males, 27 females, 1 non-binary	26 males, 31 females, 1 did not disclose gender
Native language	55 Estonian, 3 Russian	57 Estonian, 1 did not disclose native language
Highest education level	45 secondary education, 11 vocational education, 2 higher education	2 basic education, 14 secondary education, 6 vocational education, 35 higher education

2.2. Study design

We used a between-subjects design with two groups: (a) a feedback group (n = 29), where participants received process feedback after interviewing each avatar; and (b) a control group (n = 29), where participants did not receive any feedback. Each participant performed four interviews with avatars and one with the human mock witness. The four avatars were the same for each participant, but their order was randomised. The mock witness was interviewed through Google Meet.

2.3. Materials

Four AWAs (two males and two females, comprising three victims and one bystander) were created based on real-life crime scenarios (for more details, see Tohvelmann et al., 2025). Response algorithms were created based on a systematic review of experimental studies examining how actual adult witnesses behave during interviews (see Tohvelmann et al., 2025), thereby allowing us to simulate how a real adult would respond to different types of questions. Responses consisted of answers containing correct details (the details mentioned by the victims and witnesses based on the information from actual court verdicts) and incorrect details (created by the manuscript authors and known to be false).

2.4. Procedure

Participants were tested individually in Tallinn University's experimental psychology laboratory. The Board of Research Ethics at Tallinn University approved the study, with all the participants signing a fully informed consent form stipulating the possibility to withdraw from the study at any time during the experiment.

Participants were seated in front of a laptop computer where the videos of the AWAs were displayed. First, each participant had to read a description of best practices in investigative interviewing of adults and answer two questions about what they read. If they answered incorrectly, the same questions were asked again until both questions were answered correctly. Before the interviewing process started, each participant was also presented with 10 pairs of questions (randomised between the versions) and had to choose the correctly formed one (four different versions, see Appendix 6.1) and mark their selection. No feedback was given on these decisions.

([14]) SD refers to standard deviation.

Each participant interviewed four avatars (for more detail, see Tohvelmann et al., 2025). Before each interview, participants read a short description of the criminal case of the AWA to be interviewed (e.g. 'a security guard reported that two men left a store without paying for a teddy bear, then one of them punched the security guard in the face'). Participants then had 10 minutes to find out what happened, a time frame previously shown to be sufficient for producing a training effect in increasing the proportion of recommended questions (see Pompedda et al., 2022). If they were satisfied that they had found out what had happened, they could finish the interview earlier. The participants' questions were audio-recorded. To ask a question, the participants had to press a red button that initiated the recording. After asking a question, they had to press a black square that ended the recording. The operator then manually coded the question type (free or cued recall, open, closed, forced-choice, suggestive or misleading questions) and the detail type that the question addressed. The latter included general information (yes, no, don't know, etc.), avatar personal information (name, etc.), location, time, action (who did what), objects (e.g. knife), offender characteristics, victim/witness characteristics and the manner in which the crime was committed (e.g. 'he attacked with a knife'). The software then launched a video clip with the avatar's response, which included either correct or incorrect details, depending probabilistically on the type of question asked. If there were two possible incorrect answers available, the operator chose manually which answer to play depending on the storyline (but not accuracy) of previous AWA answers. At the end of the interview, the participants reported what they believed the AWA had witnessed/experienced. After interviewing four avatars, participants repeated the task of choosing the correctly formed question from among the different question pairs. Again, no feedback was given on their decisions.

After each interview, the control group (n = 29) proceeded to learn the story of the next avatar and started the next interview. The feedback group (n = 29) received feedback on four of the questions (two recommended and two non-recommended questions) they had used during the preceding interview. Over the four interviews, as many different question types as possible were covered. Specifically, if the interviewer used several different new question types during the interview, these were prioritised during the feedback session. Regardless, feedback was always provided for at least two recommended and two non-recommended questions. However, if the interviewer did not ask any questions of a particular type at all (either recommended or non-recommended), feedback was given on only two questions. As a result, the number of questions the participant received feedback on could vary. Example feedback to recommended questions would be: 'You asked the avatar, "Where were you?", which is an open question; carry on with these types of questions'. For non-recommended questions, feedback might be 'You asked "Was there one or two persons?", which is an option-posing question; try to phrase this question in a more open manner next time'.

After the avatar training session, the interviewers conducted an interview with a human interviewee using Google Meet. The interviewees first signed an informed consent form and were shown one of the two half-minute-long video clips, selected at random (one video clip was of a mock theft where two women interacted with a man, and one of the women stole the man's mobile phone; the other was of two women interacting with a man, and by the end of video clip the man thought that something was missing from his backpack). Roughly a week later, they were interviewed with a mean (M) delay of 6.7 days (SD = 0.8, range from 6 to 10 days). There were no differences between the feedback and control groups, $t(56) = -0.82$, $p = 0.419$ (feedback M = 6.6, SD = 0.6 versus control M = 6.8, SD = 1.0). Finally, the interviewers gave the AWA a realism rating on a 10-point Likert scale with anchors 1 = not at all realistic and 10 = very realistic. The experiment lasted about 120 minutes. Finally, participants were briefed about the aim of the study.

2.4.1. Coding of interviews

During the interviews with the AWAs, the operator (Mari-Liis Tohvelmann) coded all the questions asked by the participants into one of seven categories (see Table 6.2; see also Tohvelmann et al., 2025). Free and cued recall and open questions were later grouped as recommended questions, and closed, forced-choice, suggestive and misleading questions as non-recommended questions. The coded question types were saved by the programme together with the correct and incorrect details that the AWAs provided in response to the questions.

Table 6.2. Question categories with definitions and examples

Free recall	These utterances help the interviewee to recall a response with minimal influence from the interviewer.	'Tell me everything.'
Cued recall	These utterances are similar to free recall, but related to a previous statement elicited from the interviewee.	'You mentioned a bag. Tell me everything about the bag.'
Open questions	These questions focus the interviewee's attention on a particular detail and ask for an explanation (usually using words such as 'who', 'where', 'what' and 'how').	'Where did you go with him?'
Closed questions	These questions focus on the details the interviewee may have mentioned and do not require a particular type of response, although they can be answered 'yes' or 'no'.	'Did it happen in the living room?'
Forced-choice questions	These questions provide response options from which the interviewee must choose.	'Did it happen once, twice or three times?'
Suggestive questions	These questions imply what kind of response is expected, using details that the interviewee may have mentioned before.	'He touched you, didn't he?'
Misleading questions	These questions strongly indicate the kind of response expected from the interviewee but mention a detail that was not included in the interviewee's previous responses.	'So I understand that the hat was red' (if the colour of the hat was not mentioned before and is not known).

After the avatar interviews, an independent research assistant re-coded the question type classifications (recommended or not recommended) for a random sample of 48 interview transcripts. The interrater reliability was Cohen's $\kappa = 0.84$ (95 % confidence intervals (CI) 0.82 to 0.90), $p < 0.001$. Similarly, after the mock interviews, an independent research assistant re-coded the question type classifications (recommended or not recommended) for a random sample of 12 interview transcripts with an interrater reliability of Cohen's $\kappa = 0.87$ (95 % CI 0.79 to 0.95), $p < 0.001$.

Mock witness reports of what they had witnessed in the video clip were coded as accurate (the details reported by the participant matched the details in the video) or inaccurate (participant reported something not present in the video). The interrater reliability for the coding using 12 randomly selected reports using Kendall tau analysis was $\tau\_b = 0.985$, $p < 0.01$ for accurate details, and $\tau\_b = 0.977$, $p < 0.01$ for inaccurate details.

Besides the number of correct and incorrect responses produced by the avatars, we calculated the proportion of correct avatar responses relative to all potential correct avatar responses in the system.

2.5. Data analyses

The dependent measures for testing hypothesis 1 were the number of recommended and non-recommended questions, and the proportion of recommended questions. Free recall, cued recall and open questions were grouped as recommended questions; closed, forced-choice, suggestive and misleading questions were grouped as non-recommended questions. The proportion of recommended questions was calculated by dividing the number of recommended questions by the sum of recommended and non-recommended questions in each interview.

To evaluate how receiving feedback affected the questions asked, we conducted three separate 2 (group: feedback × control; between-subjects) × 4 (time: interviews 1–4; within-subjects) repeated measures analyses of covariances (ANCOVAs), controlling for participants' age, gender and education, with the number of recommended questions, the number of non-recommended questions and the proportion of recommended questions as dependents. When Mauchly's test of sphericity was significant, we used a Greenhouse-Geisser correction. We used SPSS (Statistical Package for the Social Sciences) version 27 with p level < 0.05. To test hypothesis 1, we were interested in both the group (feedback versus control) main effect and the group × time interaction.

When the interaction effect was statistically significant, we conducted pairwise comparisons with a Bonferroni correction.

To test hypotheses 2 and 3 we used t-tests for independent samples. To test hypothesis 4 we used correlation analyses (Pearson) and z-tests. The hypotheses were preregistered on AsPredicted ([15]).

3. Results

3.1. Descriptive analyses

The average length of the avatar interviews from the first to the last question was 548 seconds (SD = 80, range 89 to 612 seconds), with no differences between the feedback and control groups, $t(230) = -1.41$, $p = 0.161$ (feedback M = 540, SD = 80 versus control M = 555, SD = 79).

The average length of mock witness interviews from the first to the last question asked was 345 seconds (SD = 158, range 106 to 745 seconds), with no differences between the feedback and control groups, $t(56) = -1.41$, $p = 0.163$ (feedback M = 316, SD = 157 versus control M = 374, SD = 156).

([15]) https://aspredicted.org/blind.php?x=JQN_DX3.

The average realism rating of the AWAs was M = 5.41 (SD = 1.41, range 2–8) with a difference between feedback and control groups, $t(56) = 4.11$, $p = 0.001$, Cohen's $d = 1.245$ (95 % CI 0.524 to 1.628), with the former rating the AWAs to be more realistic than the latter (feedback M = 6.09, SD = 1.26 versus control M = 4.74, SD = 1.23).

The question categorisation task results did not differ between the control and feedback groups either before (feedback M = 9.79 (SD = 0.49) versus control M = 9.69 (SD = 0.71), $t(56) = 0.644$, $p = 0.261$) or after the avatar training session (feedback M = 9.83 (SD = 0.47) versus control M = 9.59 (SD = 0.63), $t(56) = 1.66$, $p = 0.051$). As four different versions were used, randomised to the participants, we tested whether there were differences between the versions. No significant differences were found ($p > 0.08$ in all versions).

3.2. Effect of feedback on question type while controlling for covariates

The descriptive statistics for the number of recommended and non-recommended questions and the proportion of recommended questions in both the feedback and control groups and for each interview are presented in Table 6.3.

Table 6.3. Differences in interview quality between feedback and control groups

Number of interviews		Recommended		Not recommended		Proportion of recommended	
		M	SD	M	SD	M	SD
1	Feedback	14.66	6.10	4.55	3.62	74.41	21.42
	Control	14.03	6.72	6.07	4.66	69.21	23.36
2	Feedback	17.24	5.24	3.48	3.52	83.07	17.72
	Control	16.03	6.01	6.00	4.30	72.34	18.67
3	Feedback	19.72	5.08	1.66	2.07	91.55	11.97
	Control	17.48	6.12	5.38	3.90	75.62	16.22
4	Feedback	21.48	5.65	1.72	2.42	92.28	11.98
	Control	19.72	6.93	5.45	4.20	77.24	18.00
Overall	Feedback	18.10	6.22	3.07	3.34	84.20	18.36
	Control	16.99	6.58	5.51	4.25	74.73	19.19

First, we analysed the results concerning the number of recommended questions. Mauchly's test of sphericity was not significant, so sphericity was assumed. The analysis showed that there was no main effect of group, $F(1, 53) = 1.06$, $p = 0.31$, $\eta p^2 = 0.02$, $1 - \beta$ ([16]) $= 0.17$, or of time, $F(3,159) = 0.47$, $p = 0.70$, $\eta p^2 = 0.01$, $1 - \beta = 0.14$. As shown in Figure 6.1,

([16]) $1 - \beta$ indicates statistical power.

participants in the feedback group did not use more recommended questions than participants in the control group. There was also no interaction between group and time, $F(3, 159) = 0.44$, $p = 0.73$, $\eta p^2 = 0.01$, $1 - \beta = 0.14$, and pairwise comparisons indicated no differences within interviews.

Next, we analysed the results concerning the number of non-recommended questions. As Mauchly's test of sphericity was significant, we used a Greenhouse-Geisser correction. The analysis showed that there was a significant main effect of group, $F(1, 53) = 11.72$, $p = 0.001$, $\eta p^2 = 0.18$, $1 - \beta = 0.92$, but not of time $F(2\ 543\ 134.781) = 0.25$, $p = 0.86$, $\eta p^2 = 0.01$, $1 - \beta = 0.10$. Figure 6.2 shows that the number of non-recommended questions was higher for the control group than for the feedback group. Lastly, there was no interaction between group and time, $F(2{,}543{,}134.781) = 2.70$, $p = 0.057$, $\eta p^2 = 0.048$, $1 - \beta = 0.596$. However, pairwise comparisons indicated differences between feedback and control groups within interviews 2, 3 and 4 ($p = 0.001$).

Finally, we analysed the results concerning the proportion of recommended questions. As Mauchly's test of sphericity was significant, we used a Greenhouse-Geisser correction. ANCOVA indicated a significant main effect of group, $F(1, 53) = 8.39$, $p = 0.005$, $\eta p^2 = 0.137$, $1 - \beta = 0.812$. There was no significant main effect of time, $F(1.969{,}104.352) = 0.43$, $p = 0.65$, $\eta p^2 = 0.01$, $1 - \beta = 0.12$, and no interaction between group and time, $F(1\ 969\ 104.352) = 2.52$, $p = 0.086$, $\eta p^2 = 0.045$, $1 - \beta = 0.491$. However, pairwise comparisons indicated differences between feedback and control groups within interviews 3 and 4 ($p = 0.001$). Figure 6.3 illustrates that the proportion of recommended questions increased in the feedback group with each subsequent interview, while in the control group it did not.

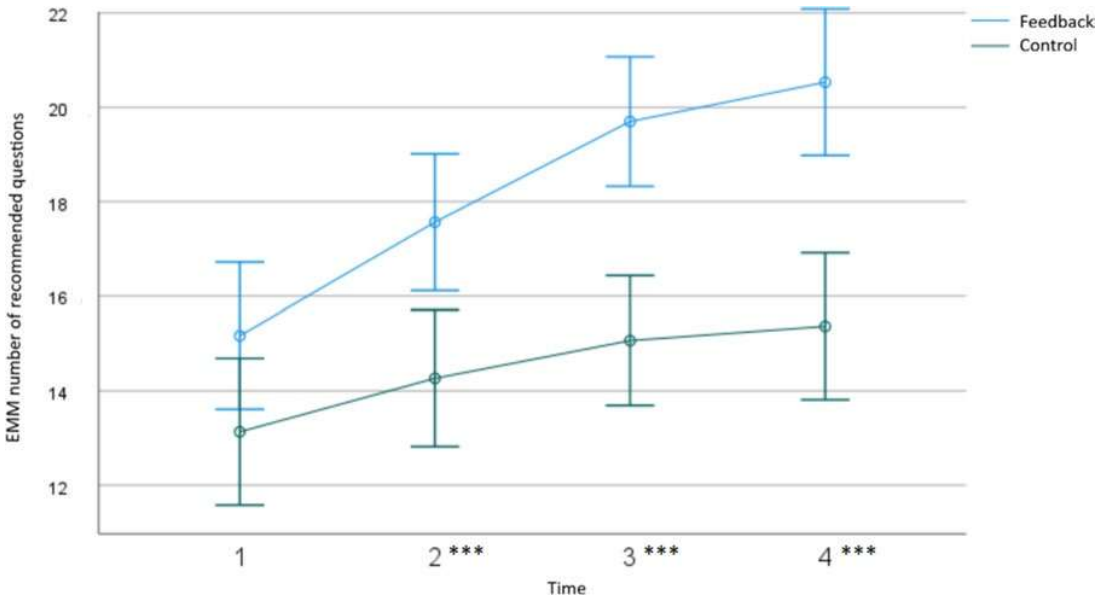

Figure 6.1. Estimated marginal means (EMM) of the difference between feedback and control groups in the number of recommended questions across four AWA interviews

NB: Error bars show 95 % confidence intervals. ***, $p < 0.001$.

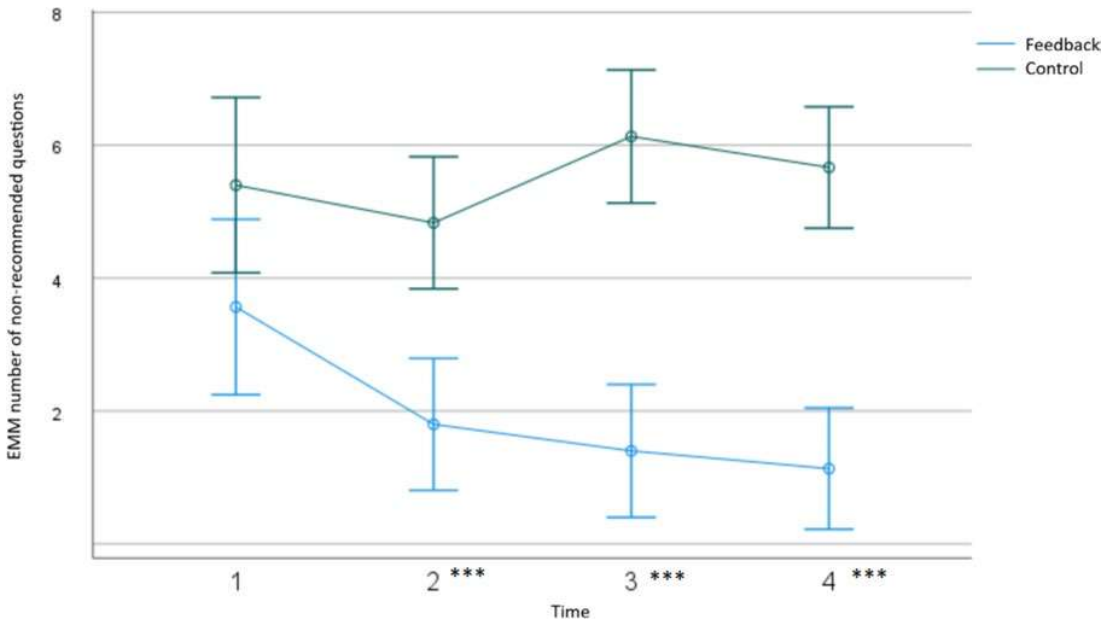

Figure 6.2. EMM of the difference between feedback and control groups in the number of non-recommended questions across four AWA interviews

NB: Error bars show 95 % confidence intervals. *** = $p < 0.001$.

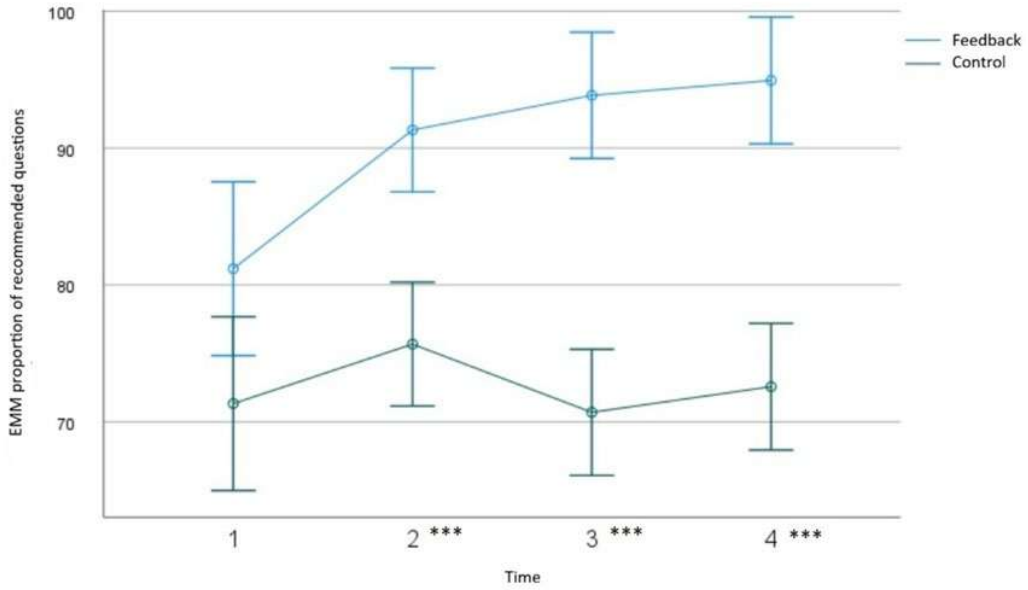

Figure 6.3. EMM of the difference between feedback and control groups in the proportion of recommended questions across four AWA interviews

NB: Error bars show 95 % confidence intervals. *** = $p < 0.001$.

Next, hypothesis 2 was tested. The feedback and control groups did not differ in the number of recommended questions in the interviews with human mock witnesses (see Table 6.4), $t(56) = -0.05$, $p = 0.96$, while the feedback group asked significantly fewer non-recommended questions than the control group, $t(56) = -3.32$, $p = 0.002$, $d = 0.872$. Also, the feedback group asked a higher proportion of recommended questions than the control group, $t(56) = -3.43$, $p = 0.001$, Hedge's $g = 0.89$.

Table 6.4. Differences between feedback and control groups in the number and proportion of question types in mock interviews with human mock witnesses

Category	Group	*n*	M	SD
Number of recommended questions in mock interview	Feedback	29	9.14	5.40
	Control	29	9.21	5.42
Number of non-recommended questions in mock interview	Feedback	29	5.14	5.48
	Control	29	10.38	6.50
Proportion of recommended questions in mock interview	Feedback	29	0.67	0.23
	Control	29	0.49	0.16

Next, hypothesis 3 was tested. The feedback and control groups did not differ in the number of correct details (see Table 6.5), $t(56) = -0.445$, $p = 0.658$, or the number of incorrect details, $t(56) = 0.049$, $p = 0.961$, or the proportion of correct details, $t(56) = -0.883$, $p = 0.381$, elicited from the human mock witnesses.

Table 6.5. Differences between feedback and control groups in the number and proportion of details elicited from human mock witnesses

Category	Group	*n*	M	SD
Number of correct details in mock interview	Feedback	29	80.62	43.04
	Control	29	85.90	47.09
Number of incorrect details in mock interview	Feedback	29	23.41	20.54
	Control	29	23.10	27.22
Proportion of correct details in mock interview	Feedback	29	0.78	0.16
	Control	29	0.80	0.12

Finally, hypothesis 4 was tested. The proportion of recommended questions in avatar interviews predicted the proportion of recommended questions in the mock witness interviews, $r(58) = 0.655$, $p = 0.001$. However, the proportion of correct details elicited from the human mock witnesses was not correlated with either the proportion of recommended questions in the interviews with the human mock witnesses, $r(58) = -0.091$, $p = 0.499$, or the proportion of recommended questions in the avatar interviews, $r(58) = -0.105$, $p = 0.434$. When feedback and control groups were analysed separately, only the proportion of recommended questions in avatar interviews was significantly and positively correlated with the proportion of recommended questions in interviews with human mock witnesses in the feedback group, $r(58) = 0.606$, $p = 0.001$, and in the control group, $r(58) = 0.620$, $p = 0.001$. The difference between the correlations was not significant using *z*-tests, $z = -0.081$, $p = 0.468$.

4. Discussion

We examined whether training with AWAs coupled with feedback would be effective in changing interviewer behaviour within the AWA training sessions and in subsequent interviews with human mock witnesses. We found that feedback was linked to a high proportion of recommended questions within the AWA training session itself. Importantly, we also demonstrated that interviewers who had received feedback during AWA training sessions asked a higher proportion of recommended questions in mock witness interviews. However, there were no significant differences between the feedback and control groups in terms of correct or incorrect details elicited from mock witnesses.

4.1. AWA training effects

In testing hypothesis 1, groups did not differ in the number of recommended questions. However, the feedback group asked fewer non-recommended questions and had a higher proportion of recommended questions than the control group, especially in interviews 2 to 4, immediately following feedback.

These results are largely in line with the previous experiment where AWA was used (Tohvelmann et al., 2025); however, the only major difference in present results is that there was no significant group or group × time interaction effect for the number of recommended questions. Overall, we can state that process feedback given to novice interviewers improved the overall proportion of recommended questions, as shown in previous similar studies where immediate and personalised feedback was given during avatar-based training sessions (Pompedda, 2018). These results confirm that serious gaming solutions contribute to improving the quality of interviews. The current study indicated that the overall proportion of recommended questions increased, though this was caused more by a decrease in the number of non-recommended questions than by an increase in the number of recommended questions.

4.2. Transfer effects

In hypothesis 2 testing, the feedback and control groups asked a similar number of recommended questions, but the feedback group asked fewer non-recommended and thus a higher proportion of recommended questions. These findings support previous notions that continuous feedback is crucial in creating a change in interviewer behaviour (Casey and Powell, 2021).

The third hypothesis was not supported – the feedback group did not elicit more correct, fewer incorrect or a higher proportion of correct details than the control group. The explanation for this finding could be that our sample was young and inexperienced in investigative interviewing; therefore, it may have been difficult for them to ask questions that are important in collecting evidence in real-life cases. Participants may have been concentrating more on the storyline and trying to find the potential offender rather than helping the avatars recall details otherwise necessary for police investigation. Therefore, more action-related questions (e.g. 'What happened after the conversation?'; 'Uh-uh, what happened next?') were asked even when the operator instructed the participants to ask for more details. Also, participants may have been pressured by the 10-minute time limit, which may have led them to concentrate on the storyline in order to get a complete report of the incident. On the other hand, they were interviewing mock witnesses from a distance (using Google Meet) and asked the mock witnesses about a short (one

minute long) video clip they had seen on average a week ago. As the video clip was short, there were not many details to elicit, resulting in a possible floor effect.

Another possible reason for the lack of significant differences between groups could be the relatively short delay period used with adult participants. Adults tend to show higher accuracy in testimony than children or older adults (Fitzgerald and Price, 2015) and are less affected by the delay period (Flin et al., 1992; Memon et al., 2003). The proportion of accurate details elicited from adult witnesses in the control group with a one-week delay period after viewing a mock event is generally around 80 % to 90 % (e.g. Gabbert et al., 2009; Hope et al., 2014). Therefore, when a short delay period like the one used in this study is implemented, it is likely that the control group who did not receive any intervention will demonstrate a ceiling effect in terms of the accuracy of the details they provide. Considering this, future research might benefit from using a longer delay period.

The proportion of recommended questions in avatar and human mock-witness interviews were positively correlated, but correct details were unrelated to question type. This correlation appeared in both feedback and control groups, with no significant difference between them. These results suggest that there were consistent individual differences between the participants. It may have been that the training sessions with avatars with the feedback group and the non-feedback group improved the use of recommended questions but did not train the participant to follow a traditional interviewing structure as the participants did not have any prior knowledge of that topic.

Participants in the feedback group rated AWAs as more realistic than those in the control group. AWAs use algorithms to generate responses based on the types of questions asked. Participants who received feedback may have learned that asking recommended questions would lead to a more logical storyline, whereas participants who did not receive any feedback may have been confused during the process due to the incorrect details the avatars recalled. Therefore, these participants may have considered that the AWA did not understand them properly and the programme needs to be improved.

There were no differences between the groups in question categorisation tasks before and after AWA training. As this task was given right after the participants read the description of the best practices of investigative interviewing of adults, which included answering two control questions correctly, their knowledge about the topic was tested using surface cues (i.e. recognition memory). However, for a deeper and more conscious learning process to emerge, a task where they must formulate the questions themselves (as they did in interviewing different avatars) could have resulted in different findings.

4.3. Limitations

The AWA training session was conducted by a single experimenter. To reduce the risk of categorising the question types in a similar manner, post-experiment interrater reliability analyses were conducted, indicating good reliability. Also, the number of correct pieces of information did not increase in mock interviews with human adult witnesses. This finding may be explained by the fact that our interviewers were not police investigators or police cadets, so they had very limited knowledge about the components of evidence in interviews (such as the time, place and manner of conducting the crime). As this study was conducted with student participants, the results cannot directly be generalised to practitioner samples.

4.4. Future research

Future research can focus on adding a modelling component to AWA training sessions, in addition to feedback, as previous studies have shown that the combination of feedback and modelling lead to greater improvement in interviewers' questioning style than feedback alone (Haginoya et al., 2021). In future, the human operator who currently codes question types and styles could be replaced by an artificial intelligence system that can be trained (machine learning) to code these question types and styles automatically (see Haginoya et al., 2023).

5. Conclusion

In the present study, we confirmed that training with AWAs paired with feedback had an effect on interviewer behaviour both within the AWA training sessions (replicating Tohvelmann et al., 2025) and now also in subsequent interviews with human mock witnesses. Our results demonstrate the potential to help interviewers use more recommended questions. Traditional training sessions are often time-consuming, costly and difficult to arrange (Pompedda, 2018), especially when they involve immediate and personalised feedback. The proposed solution offers a cost-effective means of helping practitioners train their question formulation skills, as this can currently be conducted remotely via the web in investigators' workplaces. However, before AWA can be fully implemented by law enforcement agencies, it should be tested with real police investigators to determine whether training effects transfer to their real interviews with victims and witnesses, who are an important source of information in criminal proceedings, especially in cases concerning high-risk criminal networks. These findings and tools could then be incorporated into initial academy training to set strong interviewing habits from the start. Agencies can then provide brief booster sessions at regular intervals throughout an officer's career to refresh skills and prevent drift. This approach has the potential to keep practice aligned with evidence.

References

Bjerknes, O. T. and Fahsing, I. A. (2018), *Etterforskning – Prinsipper, metoder og praksis* [*Investigation – Principles, methods and practice*], Fagbokforlaget, Bergen, https://hdl.handle.net/11250/4325149.

Casey, S. and Powell, M. B. (2021), 'Usefulness of an e-simulation in improving social work student knowledge of best-practice questions', *Social Work Education*, Vol. 41, No 6, pp. 1253–1271, https://doi.org/10.1080/02615479.2021.1948002.

Fisher, R. P. and Geiselman, R. E. (1992), *Memory Enhancing Techniques for Investigative Interviewing: The cognitive interview*, Charles C. Thomas, Springfield, Illinois, https://psycnet.apa.org/record/1992-98595-000.

Fitzgerald, R. J. and Price, H. L. (2015), 'Eyewitness identification across the life span: A meta-analysis of age differences', *Psychological Bulletin*, Vol. 141, No 6, pp. 1228–1265, https://doi.org/10.1037/bul0000013.

Flin, R., Boon, J., Knox, A. and Bull, R. (1992), 'The effect of a five-month delay on children's and adults' eyewitness memory', *British Journal of Psychology*, Vol. 83, Issue 3, pp. 323–336, https://doi.org/10.1111/j.2044-8295.1992.tb02444.x.

Graafland, M., Schraagen, J. M. and Schijven, M. P. (2012), 'Systematic review of serious games for medical education and surgical skills training', *British Journal of Surgery*, Vol. 99, Issue 10, pp. 1322–1330, https://doi.org/10.1002/bjs.8819.

Gabbert, F., Hope, L. and Fisher, R. P. (2009), 'Protecting eyewitness evidence: Examining the efficacy of a self-administered interview tool', *Law and Human Behavior*, Vol. 33, No 4, pp. 298–307, https://doi.org/10.1007/s10979-008-9146-8.

Griffiths, A. and Milne, R. (2006), 'Will it all end in tiers? Police interviews with suspects in Britain', in: Williamson, T. (ed.), *Investigative Interviewing – Rights, research, regulation*, Willan, London, pp. 167–189, https://doi.org/10.4324/9781843926337.

Griffiths, A. and Rachlew, A. (2018), 'From interrogation to investigative interviewing', in: Griffiths, A. and Milne, R. (eds), *The Psychology of Criminal Investigation: From theory to practice*, Routledge, London, pp. 154–178, https://doi.org/10.4324/9781315637211-9.

Halley, J., Walsh, D., Myklebust, T. and Bjerknes, O. T. (2023), 'Structured models of interviewing', in: Oxburgh, G. E., Myklebust, T., Fallon, M. and Hartwig, M. (eds), *Interviewing and Interrogation: A review of research and practice since World War II*, Torkel Opsahl Academic EPublisher, Brussels, pp. 257–281, https://www.legal-tools.org/doc/ln43ta/.

Haginoya, S., Yamamoto, S. and 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 and Neglect*, Vol. 115, https://doi.org/10.1016/j.chiabu.2021.105013.

Haginoya, S., Ibe, T., Yamamoto, S., Yoshimoto, N., Mizushi, H. et al. (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*, Vol. 14, https://doi.org/10.3389/fpsyg.2023.1133621.

Hope, L., Gabbert, F., Fisher, R. P. and Jamieson, K. (2014), 'Protecting and enhancing eyewitness memory: The impact of an initial recall attempt on performance in an investigative interview', *Applied Cognitive Psychology*, Vol. 28, Issue 3, pp. 304–313, https://doi.org/10.1002/acp.2984.

Kask, K., Baugerud, G. A. and Volbert, R. (2024), 'Editorial: Technological solutions helping to train specialists' interviewing skills of possible victims and witnesses', *Frontiers in Psychology,* Vol. 15, https://doi.org/10.3389/fpsyg.2024.1406867.

Lamb, M. E. (2016), 'Difficulties translating research on forensic interview practices to practitioners: Finding water, leading horses, but can we get them to drink?', *American Psychologist*, Vol. 71, No 8, pp. 710–718, https://doi.org/10.1037/amp0000039.

Memon, A., Bartlett, J., Rose, R. and Gray, C. (2003), 'The aging eyewitness: Effects of age on face, delay, and source-memory ability', *Journals of Gerontology – Series B*, Vol. 58, Issue 6, pp. 338–345, https://doi.org/10.1093/geronb/58.6.P338.

Milne, R. and Bull, R. (1999), *Investigative Interviewing: Psychology and practice*, Wiley, Chichester.

Nunan, J., Stanier, I., Milne, R., Shawyer, A. and Walsh, D. (2020), 'Source handler telephone interactions with covert human intelligence sources: An exploration of question types and intelligence yield', *Applied Cognitive Psychology*, Vol. 34, Issue 6, pp. 1473–1484, https://doi.org/10.1002/acp.3726.

Otgaar, H., Houben, S. T. L., Muris, P. and Howe, M. L. (2023), 'Does interviewing affect suggestibility and false memory formation?', in: Oxburgh, G. E., Myklebust, T., Fallon, M. and Hartwig, M. (eds), *Interviewing and Interrogation: A review of research and practice since World War II*, Torkel Opsahl Academic EPublisher, Brussels, pp. 193–209, https://www.legal-tools.org/doc/v3wphg/.

Oxburgh, G. E., Myklebust, T. and Grant, T. (2010), 'The question of question types in police interviews: A review of the literature from a psychological and linguistic perspective', *International Journal of Speech, Language and the Law*, Vol. 17, Issue 1, pp. 45–66, https://doi.org/10.1558/ijsll.v17i1.45.

Phillips, E., Oxburgh, G. E., Gavin, A. and Myklebust, T. (2011), 'Investigative interviews with victims of child sexual abuse: The relationship between question type and investigation relevant information', *Journal of Police and Criminal Psychology*, Vol. 27, pp. 45–54, https://doi.org/10.1007/s11896-011-9093-z.

Pompedda, F. (2018), *Training in Investigative Interviews of Children: Serious gaming paired with feedback improves interview quality*, Åbo Akademi University, Turku, https://doi.org/10.13140/RG.2.2.29825.07526.

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

Pompedda, F., Zhang, Y., Haginoya, S. and 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*, Vol. 37, pp. 485–498, https://doi.org/10.1007/s11896-022-09509-7.

'Principles on effective interviewing for investigations and information gathering' (2021), interviewingprinciples.com website, https://interviewingprinciples.com/.

Sternberg, K. J., Lamb, M. E., Esplin, P. W., Orbach Y. and Hershkowitz I. (2002), 'Using a structured protocol to improve the quality of investigative interviews', in: Eisen, M., Quas, J. A. and Goodman, G. S. (eds), *Memory and Suggestibility in the Forensic Interview*, Lawrence Erlbaum Associates Publishers, Mahwah, pp. 409–436, https://psycnet.apa.org/record/2001-05106-017.

Tohvelmann, M.-L. and Kask, K. (2022), 'From child to adult victims and witnesses: Ways of improving the quality of investigative interviews', *Juridica International*, Vol. 31, pp. 136–146, https://doi.org/10.12697/JI.2022.31.10.

Tohvelmann, M.-L., Kask, K., Palu, A., Haginoya, S. and Santtila, P. (2025), 'Providing feedback in simulated investigative interviews with adult witness avatars increases the use of free recall and open questions', *International Journal of Police Science and Management*, Vol. 27, Issue 2, pp. 182–198, https://doi.org/10.1177/14613557241310014.

UK Ministry of Justice (2022), *Achieving Best Evidence in Criminal Proceedings: Guidance on interviewing victims and witnesses, and guidance on using special measures*, https://assets.publishing.service.gov.uk/media/6492e26c103ca6001303a331/achieving-best-evidence-criminal-proceedings-2023.pdf.

United Nations Department of Peace Operations, United Nations Office of the High Commissioner for Human Rights and United Nations Office on Drugs and Crime (2024), *Manual on Investigative Interviewing for Criminal Investigation*, https://resourcehub01.blob.core.windows.net/$web/Policy %20and %20Guidance/corepeacekeepingguidance/Thematic %20Operational %20Activities/Police %20and %20Law %20Enforcement/2024.01 %20Manual %20on %20Investigative %20Interviewing %20for %20Criminal %20Investigation %20(2024).pdf.

Van der Zee, D.-J., Holkenborg, B. and Robinson, S. (2012), 'Conceptual modeling for simulation-based serious gaming', *Decision Support Systems*, Vol. 54, Issue 1, pp. 33–45, https://doi.org/10.1016/j.dss.2012.03.006.

Walsh, D., Areh, I., Barela, S., Boukalova, H., Bull, R. et al. (2025), 'A co-ordinated global initiative to enhance interview practice', *Investigative Interviewing: Research and practice*, Vol. 15, Issue 1, pp. 7–27, https://iiirg.org/wp-content/uploads/2025/09/II-RP-15-Special-Bilingual-Issue-2025-WEB.pdf.

Appendix 6.1. Examples of question categorisation pairs

1	Tell me what happened in the house.	Were there a lot of people?
2	Tell me what he/she was wearing.	Was he/she wearing a sweater or a hoodie?
3	What else do you remember?	Can you remember anything else?
4	You don't remember anything else, do you?	Tell me what else you remember about the incident.
5	Tell me more about this car.	Was the car red?
6	What colour was the car?	Was the car blue or green?
7	Tell me what happened in the house.	Did something happen?
8	Tell me who did what.	Was there a fight?
9	Were you afraid?	What did you feel?
10	Tell me what you were thinking when you saw this.	Were you afraid or not while all this happened?