Journal of Police and Criminal Psychology
https://doi.org/10.1007/s11896-022-09509-7

A Mega-Analysis of the Effects of Feedback on the Quality of Simulated Child Sexual Abuse Interviews with Avatars

Francesco Pompedda[1] · **Yikang Zhang**[2] · **Shumpei Haginoya**[3,5] · **Pekka Santtila**[4]

Accepted: 23 March 2022

Abstract

The present study aimed to test the effectiveness of giving feedback on simulated avatar interview training (Avatar Training) across different experiments and participant groups and to explore the effect of professional training and parenting experience by conducting a mega-analysis of previous studies. A total of 2,208 interviews containing 39,950 recommended and 36,622 non-recommended questions from 394 participants including European and Japanese students, psychologists, and police officers from nine studies were included in the mega-analysis. Experimental conditions were dummy-coded, and all dependent variables were coded in the same way as in the previously published studies. Professional experience and parenting experience were coded as dichotomous variables and used in moderation analyses. Linear mixed effects analyses demonstrated robust effects of feedback on increasing recommended questions and decreasing non-recommended questions, improving quality of details elicited from the avatar, and reaching a correct conclusion regarding the suspected abuse. Round-wise comparisons in the interviews involving feedback showed a continued increase of recommended questions and a continued decrease of non-recommended questions. Those with (vs. without) professional and parenting experience improved faster in the feedback group. These findings provide strong support for the efficacy of Avatar Training.

Keywords Child sexual abuse (CSA) · Investigative interviewing · Simulation training · Feedback · Serious gaming

Child sexual abuse (CSA) is prevalent in all societies, with prevalence estimates ranging from 8 to 31% for girls and 3 to 17% for boys (Barth et al. 2013). It is also clear that CSA is associated with a plethora of negative psychological, relational, and somatic health consequences (Hailes et al. 2019). It is, therefore, important to investigate suspected CSA cases effectively. Unfortunately, these investigations present special challenges. In almost 70% of suspected CSA cases, the child's statement is the only available evidence in these cases (Elliott and Briere 1994; Herman, 2009). Due to the usual lack of corroborating evidence, investigative interviews with alleged victims are of central importance in CSA investigations. While most experts agree that children can in principle provide accurate reports, there is also little doubt that accounts can be distorted both by improper interviewing and by normal memory decay (Ceci and Bruck 1995). It is, therefore, worrying that the quality of investigative interviews in these cases is still poor in many regions across the world: Closed questions are still the most common type of questions, regardless of expert warnings about the use of this type of questions (Cederborg et al. 2000; Korkman et al. 2008; Sternberg et al. 2001).

Francesco Pompedda and Yikang Zhang are shared first authors.

✉ Yikang Zhang
kang.y.zhang@outlook.com

Francesco Pompedda
fpompedda@glos.ac.uk

Shumpei Haginoya
haginoya@psy.meijigakuin.ac.jp

Pekka Santtila
pekka.santtila@nyu.edu

1 School of Natural & Social Sciences, University of Gloucestershire, Cheltenham, UK

2 Faculty of Psychology and Neuroscience, Maastricht University, Maastricht, the Netherlands

3 Faculty of Psychology, Meiji Gakuin University, Tokyo, Japan

4 NYU Shanghai and NYU-ECNU Institute for Social Development, Shanghai, China

5 Mykolas Romeris University, Vilnius, Lithuania

Published online: 01 April 2022

Recommendations for Investigative Interview

To improve interview quality, a lot of effort has been directed at training interviewers. The following main rules are recommended when interviewing children: First, questions should be non-leading. Leading questions can have a negative impact on children, creating less accurate statements and contaminated memories (Bruck and Ceci 1999; Ceci and Bruck 1993, 1995). For example, using a realistic mock event, Finnilä et al. (2003) showed that children who had been visited at daycare by a clown gave false affirmative responses at alarming rates (e.g., a 20% false-positive rate to the question "He told you that what you did together was a secret and that you couldn't tell anyone, didn't he?"). Second, open-ended rather than option-posing questions (i.e., asking for a yes/no-response or providing a list of alternatives) should be used. Whereas option-posing questions tap less accurate recognition processes, open-ended questions rely on recall memory and are, therefore, more likely to elicit useful answers (Lamb et al. 2003; 2008; Lyon, 2014). Even though they have been widely disseminated over the years in theoretical training programs targeting professionals in the field, the recommendations are not always followed in practice (e.g., Johnson et al. 2015).

Implementation Through Training Programs

Despite the lack of proven efficacy of most training programs, Lamb et al. (2002) have shown that the use of a structured protocol giving the interviewers clear guidance on the questions to use in the different phases of the interview coupled with extensive feedback has resulted in improvements. Unfortunately, performance deteriorates rapidly after feedback is discontinued (Lamb et al. 2002a, b). Another limitation is that, as it is usually impossible to know what actually happened in real cases, feedback can only be given on the questions used by the interviewer but not on whether the elicited responses are true or false, that is, there is no outcome feedback. This means that interviewers are not alerted to when their questions have resulted in a false allegation of CSA; therefore, they may not experience the need to change their interviewing behaviors. The training program *Specialist Vulnerable Witness Forensic Interview Training* (Benson and Powell 2015; Powell et al. 2016) has also shown promising results in improving CSA interview quality in terms of question use. However, it is a comprehensive training project that includes 15 modules and takes months to be implemented. The feedback provided usually consists of process feedback on questions used and behaviors employed during the interviews, with no feedback on whether the interviewer reached the correct conclusion. Some trainings employ actors who play the role of the allegedly abused child, which, while introducing the interactive components of the training, may not be optimal to mimic the behavior of actual children in interview situations in terms of memory recall and suggestibility. The reliance on actors and experts also poses difficulties for the scalability of the program.

The Simulated Avatar Interview with Feedback Approach

To address the situation illustrated above, a series of experiments exploring the efficacy of simulated avatar interview training programs have been conducted, where individuals have interviewed child avatars and received feedback on the questions used and the correctness of elicited information (Haginoya et al. 2020, 2022b; Pompedda et al. 2020; Krause et al. 2017; Pompedda et al. 2015). The algorithms in the program were set to mimic the behavioral pattern of real children during interviews. That is, the avatars have predefined "memories" of an event of interest and respond to questions in a way that is consistent with research on suggestibility of children of different ages (4- and 6-year-olds). For example, if the interviewer asks a question about a detail that is absent from the avatar's memory, the avatar responds "No." But if the question is repeated, a 4-year-old avatar will change the response to "Yes" with a probability of 0.50. This way, suggestive questions (and other types of non-recommended questions) can lead to inaccurate details being contained in the avatars' responses, similar to what may happen in actual CSA interviews. The correlations between question types and types of details elicited and the correctness of the conclusions are used to inspect whether the algorithms function as expected. Consistent with analyses of real-life interviews in previous studies (e.g., Lamb et al. 2007), correct conclusions were positively predicted by recommended questions as well as relevant details and negatively predicted by non-recommended questions and wrong details in the avatar interviews. Evidence has repeatedly shown that CSA Avatar Training coupled with feedback on the interviewers' performance results in improvement of interview quality compared to controls who receive no feedback (Haginoya et al. 2020, 2021; Krause et al. 2017; Pompedda et al. 2015, 2020). Subsequently, additional studies have focused on further factors that have been expected to improve the training effect by incorporating new features in the program.

Specifically, Pompedda et al. (2017) compared the effects of types of feedback on interview quality improvement. Compared with outcome feedback (i.e., feedback on the correctness of the elicited information from the avatar) and process feedback (i.e., feedback on the usage of recommended and non-recommended questions), the combined feedback with both outcome and process information produced the largest improvement. Krause et al. (2017) further examined whether instructions of reflection could contribute to the improvement above and beyond combined feedback. Though empirical evidence supports the efficacy of reflection on task performance in other contexts such as education (Espinet et al. 2013), military leadership (Matthew and Sternberg 2009) and aircraft navigation (Ron et al. 2006), Krause et al. (2017) did not find clear evidence favoring the reflection design in CSA interview training.

More recently, Haginoya et al. (2020) extended this line of research to an Asian population and online context with results showing that Avatar Training with feedback improves interview quality across cultural contexts and implementation settings. After establishing the effectiveness of the approach, Haginoya et al. (2021) further examined the effect of behavioral modeling and its combination with feedback. Behavioral modeling training (BMT) originates from social learning theory (Bandura and McClelland 1977). This approach includes five components: (1) identifying well-defined behaviors, (2) showing the effective use of those behaviors through model(s), (3) giving opportunities to practice those behaviors, (4) providing feedback and social reinforcement, and (5) taking measures of maximizing the transfer of those behaviors to practical tasks (Taylor et al. 2005), the latter three of which have been the integral part of the Avatar Training approach used in the earlier studies. By also incorporating the first and second component into the Avatar Training and providing the participants both negative and positive models and the consequences of these behaviors, Haginoya et al. (2021) showed that the combination of feedback and modeling improves interview quality more than feedback alone. In addition to improving interview quality directly, Haginoya et al. (2022a) also tested whether adding feedback on supportive statements could be done while also improving the use of appropriate question types. The use of supportive statements would be helpful in enhancing rapport-building and, consequently, abuse disclosure by reluctant children (Blasbalg et al. 2018). The results confirmed that it is possible to improve the use of recommended questions while also improving the use of supportive statements by providing feedback.

Potential Moderators of the Training Effect

Whether experience of interacting with children, either as a parent, babysitter, or as a professional child interviewer can have an impact on training, has not been exhaustively examined before. In many of the situations in which children interact with adults, closed questions are used to ask children about topics that adults are already aware, for example, a teacher assessing knowledge after teaching (e.g., Pate 2012). It has been proposed that past experience may thus negatively affect training outcomes as the interviewers could have trouble refrain from using non-recommended questions or be reluctant to change (Pompedda 2018). According to the proactive interference theory, previously learned information might interfere with newly learned information. In the case of investigative interviews, the previously learned use of closed questions might interfere with the use of open questions (Powell et al. 2014). In addition, the frequent lack of knowledge regarding the ground truth in alleged CSA cases, and the use of the judicial outcome as proof of the quality of the interview, might exacerbate the effects of proactive interference (Jacoby et al. 2001). However, the literature shows mixed results between field studies where there is no association between experience and use of open questions (e.g., Wolfman et al. 2016), with some exceptions (e.g., Lafontaine and Cyr 2017), and simulated interviews, where there is evidence of a negative association between experience and the use of open questions (e.g., Powell et al. 2014), with some studies showing no differences after training (e.g., Benson and Powell 2015). Moreover, experienced interviewers, while potentially more incline to use closed questions, could also have other skills that can help them during the interview, for example, better ability in creating rapport (Hershkowitz et al. 2017) or better communication skills (e.g., MacDonald et al. 2017).

Rationale for Conducting a Mega-Analysis

Though each single study has already provided evidence for the efficacy of the program, the current research intended to offer more insight through a systematic mega-analysis of all the studies conducted so far. Mega-analysis integrates raw data from individual studies followed by new analyses using multilevel models accounting for the heterogeneity among studies to reach robust conclusions. Compared with meta-analyses where summary statistics from the original studies are integrated, mega-analyses allow evidence synthesis using the raw data, avoiding potential bias and errors related to the analyses in the reports of the original studies. Moreover, as the original studies employing the avatars were limited by the individually low sample sizes and only including particular participant types, previous examinations of the effect of potential moderators of the reported training effects have been underpowered. In the present analyses, we were able to explore the effects of individual differences, research design, and relevant demographic variables such as professional training or parenting experience on the outcomes

with the pooled data providing adequate power to do so. Importantly, this research also allows a better estimate of the effects of providing a combination of outcome and process feedback across different professional groups and countries.

Method

Participants

The nine studies included in this mega-analysis collected training data using participant samples of European and Japanese students, psychologists, and police officers. Detailed information regarding all of the samples can be found in Table 1.

Materials

Avatar Training

Simulated interviews with avatars were conducted using different languages based on the country where the interviews took place. The different language versions were identical with the exception of a small number of cultural adaptations (e.g., religious settings and some games played by the child avatar were changed). The simulation comprised 16 different avatars equally divided between age (4 vs 6), ground truth abused vs not abused, and emotions displayed (crying vs no crying). For each case, a series of details both related to the alleged abuse (e.g., details that describe the abuse or that provide an alternative explanation), but also details that are not relevant for the investigation (e.g., favorite toy or other activities the avatar had experienced) were created. Interviewers faced a screen where one of the avatars was presented and vocally asked a question to the avatar like in a real interview. Meanwhile, an operator listened to the questions asked and categorized them in real time by clicking the appropriate button in the simulation interface. The categorization triggered the algorithms (different between 4-year-old and 6-year-old avatars) that were based on research in children's memory and suggestibility, as well as the available details in the memory of the avatar and resulted in the launch of the appropriate video clip containing the avatar's response. In each study, a randomized selection of avatars was used (providing an equal balance between abuse vs not abused avatars and 4 years old vs 6 years old) and provided in a random order.

Data Coding

Experimental Conditions All experimental conditions were dummy-coded (see Table S1). Feedback referred to the

Table 1 Descriptive statistics of the nine data sets included in the mega-analysis

Study	Experimental conditions	Rounds of interviews	Participant			
			population	n_{Female}	M_{age}	SD_{age}
Pompedda et al. (2017)	Control ($n=12$) Outcome feedback ($n=12$) Process feedback ($n=12$) Feedback ($n=12$)	4	Europe; students	*38*	27.9	9.1
Krause et al. (2017)	Control ($n=19$) Feedback ($n=19$) Feedback + reflection ($n=21$)	8	Europe; students	35	24.4	3.7
Haginoya et al. (2020)	Control ($n=15$) Feedback ($n=17$)	6	Japan; students	23	20.5	0.6
Haginoya et al. (2022a) Study 1	Control ($n=20$) Feedback ($n=20$)	6	Europe; psychologists	37	27.4	2.2
Haginoya et al. (2022a) Study 2	Control ($n=32$) Feedback ($n=32$)	6	Europe; students	44	23.1	3.6
Haginoya et al., (2021)	Modeling ($n=11$) Feedback ($n=10$) Feedback + modeling ($n=11$)	5	Japan; psychologists	22	35.1	8.7
Haginoya et al., (2022b)	Control ($n=10$) Control + feedback ($n=11$)	4/8	Japan; police	8	35.5	5.4
Kask et al. (2022)	Control ($n=11$) Control + feedback ($n=11$)	4/8	Europe; police	3	41.2	6.2
Haginoya et al. (2022a)	Control ($n=20$) Feedback ($n=20$) Supportive ($n=20$) Feedback + supportive ($n=20$)	4	Japan; mixed	53	35.6	9.9

condition where participants received both process feedback (i.e., which of the questions they had used were appropriate and which were inappropriate and why) and outcome feedback (whether they had reached the correct conclusion after the interview; they were explained what had really happened in the case, i.e., which memory contents did the avatars have). The supportive statement manipulation was not aimed at improving recommended question use or conclusion accuracy, and it was therefore ignored in current analyses.

Interview Round In all but two police studies (Kask et al. 2022b, Haginoya et al. 2022a), participants were assigned either to a feedback condition or a no feedback condition, with interview rounds ranging from 4 to 8. Instead, to maximize the utility of simulation training in the police force, in the two studies using police samples, participants were either assigned to a condition where they finished four rounds of interviews with feedback or a condition where they first finished four rounds of interviews without feedback and then finished four rounds of interviews with feedback. In current analyses, we recoded the latter condition so that the first four rounds without feedback were coded as belonging to the no feedback condition, and the second four rounds with feedback were coded as belonging to the feedback condition. The 5–8 rounds of interviews in the control + feedback condition were thus coded as rounds 1–4 in the feedback condition.

Recommended and Non-Recommended Questions Recommended and non-recommended questions were coded as continuous variables with the value indicating number of questions asked in an interview. The current analysis coded the data in the same way as in the previously published studies (reference eliminated due to blind peer review; see Table 2).

Relevant, Neutral, and Wrong Details Relevant, neutral, and wrong details were coded as continuous variables with the value indicating number of details elicited from the avatar in an interview. Relevant details were the forensically relevant details that related to the alleged abusive situation (e.g., details that would clarify if the abuse happened or not). Neutral details were details related to other situations that the avatar had experienced but that were not forensically relevant to the investigation of the alleged abuse (e.g., games played with other persons). Finally, wrong details were details that contradicted the pre-defined memories of each avatar (e.g., using repeated suggestive questions, the interviewer found that the dad would have touched the child, while in reality it was the uncle) There were one hundred and twenty cases missing the number of wrong details in the combined data set.

Conclusion Content Correctness In all the studies, participants were deemed as having reached a correct conclusion only if they (1) first provided a correct answer regarding the presence (or absence) of an abuse and then (2) offered a correct account of the sexual abuse (who, when, where, and what transpired) in the former case or an explanation of what happened instead of an abuse in the latter case. Conclusion

Table 2 Question-type coding used in the studies

Category	Definition
Recommended questions	
Facilitators	Open-ended and non-suggestive questions that encourage the child to continue with the previous answer
Invitations	Open-ended and non-suggestive questions. They are broad and let the child talk freely
Directive	Open-ended and non-suggestive questions that focus the child attention on a previously mentioned detail asking for a specific explanation
Not recommended questions	
Option-posing	Closed-ended questions that focus on unmentioned detail (without implying a particular type of response) or on a mentioned detail asking the child to provide a yes/no answer
Specific suggestive	Open-ended or closed-ended questions that are based on an unmentioned detail and express the expected response
Unspecific suggestive	Open-ended or closed-ended questions that are not based on an unmentioned detail but express the expected response
Repetitions	Repetitions of a previous recommended or non-recommended question
Too-long/unclear	Questions that use a logical structure that is too complicated for the cognitive level of the child and/or are formulated in a haphazard manner and/or contains more than one concept at the time
Multiple choice	Questions that provide a predetermined list which the child is requested (explicitly or implicitly) to pick from
Time	Open-ended or closed-ended questions that require the child to provide or recollect precise time-related information
Fantasy	Open-ended or closed-ended questions that move the discussion from the reality to the fantasy level
Feelings	Open-ended or closed-ended questions that require the child to provide accounts regarding own or other's feelings

content correctness was coded as dichotomous variable in all except for two data sets (Study 2 from Haginoya et al. 2022b; Kask et al. 2022). In these two studies, conclusion content correctness was coded with three categories: correct, incorrect, and not enough information to reach a conclusion. Therefore, in the current mega-analysis, we coded conclusion content correctness with two categories: correct and not correct, with the latter including both incorrect and fail-to-reach-conclusion cases. As the data set containing the Japanese police response did not record conclusions, there were one hundred and twenty-four cases missing the information on conclusion correctness in the combined data set.

Professional Experience Regarding Child Interview The data sets included in the current mega-analysis employed several non-identical measures to assess participants' professional training and/or experiences with child interview. Haginoya et al. (2021) and Haginoya et al. (2022a) employed three questions about child interview training, experience, and child sexual abuse interview experience, specifically. Krause et al. (2017), Pompedda et al. (2020), and Haginoya (2022a) used one item to assess child sexual abuse interview experience. Kask et al. (2022b) documented years of conducting child interview, years of conducting child sexual abuse interview, and number of interviews conducted in a continuous manner. Pompedda et al. (2017) did not report child interview experience data in their published manuscript nor documented the information in the data set available to us. We obtained the child interview experience information through private communication with the authors. Therefore, in current analysis, we coded child interview experience as a dichotomous variable, with yes indicating having interviewing experience with children. The combined data set contained 33 participants (resulting in 149 interviews) in the no feedback condition and 62 participants (resulting in 276 interviews) in the feedback condition who had child interview experience before participating in one of the included studies.

Parenting Experience All data sets except for the data set of Kask et al. (2022b) contained parenting experience, though the operationalization was not always the same. Pompedda et al. (2017), Krause et al. (2017), and Pompedda et al. (2020) asked participants whether they had children or not. Haginoya et al. (2020; 2021) and Haginoya et al. (2022a, b) asked participants whether they had child-rearing experience. In current mega-analysis, we coded parenting experience as dichotomous variable, with yes either indicating having children or having child-rearing experience. There were 32 participants with parenting experience (150 interviews) in the no feedback condition and 45 participants with parenting experience (158 interviews) in the feedback condition in the combined data set.

Statistical Analyses

All statistical analyses were conducted in R (version 4.05). We first employed correlational analysis to investigate the validity of the algorithms used in the studies. Then we used lme4 (Bates et al. 2014) to perform a series of linear mixed effects analyses to examine the efficacy of simulation training on CSA interview quality. As fixed effects, we entered interview round, feedback, process feedback, outcome feedback, reflection, modeling, and the interaction term between interview round and feedback into the model. As random effects, we had intercepts for participants and studies. In subsequent analyses, we also examined potential moderating effects of the demographic variables, professional training and experience, and parenting experience by including their interaction terms with interview round as fixed effects and study, feedback condition, and participants as random effects. Since professional experience and parenting experience were correlated, χ^2 (1) = 238.04, $p < 0.001$, separate models were run for the two potential moderators. Confidence intervals (95%) of the parameters in the linear mixed models were calculated using bootstrap method with 5,000 draws. The 95% confidence intervals of the parameters in the generalized linear mixed model were calculated using the Wald method.

To make the results more interpretable, interview round in the analyses was coded starting from zero, that is, the first round was designated with the value of 0, the second round with value of 1, and so on. This way, the intercepts of the models represented the estimates of the first-round performance in the baseline conditions.

In addition, we calculated a Reliable Change Index (RCI) for each participant in the feedback condition for question use and details elicited to provide more nuanced information regarding individual differences in the training effect as well as how training design could have an impact on reliable change. As there is no established norm for these measurements, the reliability of the measurement (r) was operationalized as the correlation between the first-round performance and the last-round performance in the no feedback condition while excluding cases who received modeling instructions or process feedback instructions. The standard deviation (SD) of the measure was operationalized as the standard deviation of the first-round performance in the feedback group, which was used to calculate the standard error of the difference. The RCI formula is as follows:

$$\mathrm{RCI} = (\mathrm{Performance}_{\text{last round}} - \mathrm{Performance}_{\text{1st round}}) / [2 \times (1 - r) \times SD^2]^{1/2}$$

RCI greater than 1.96 (the z score corresponding to a distance of 2 standard deviations from the mean) in the case

of recommended questions or smaller than − 1.96 in the case of non-recommended questions would indicate that the participant had a reliable change in their interview quality.

Results

Correlations Between Questions, Details, and Conclusion Correctness

Descriptive statistics for interview quality indicators are presented in supplementary materials (Table S2). The number of recommended questions was positively correlated with number of relevant details elicited ($r = 0.79$, $p < 0.001$), number of neutral details elicited ($r = 0.79$, $p < 0.001$), and negatively correlated with number of wrong details elicited ($r = -0.25$, $p < 0.001$). The number of recommended questions was also positively correlated with conclusion correctness ($r = 0.36$, $p < 0.001$). The number of non-recommended questions was negatively correlated with number of relevant details elicited ($r = -0.69$, $p < 0.001$), number of neutral details elicited ($r = -0.20$, $p < 0.001$), and positively correlated with number of wrong details elicited ($r = 0.57$, $p < 0.001$). The number of non-recommended questions was negatively correlated with conclusion correctness ($r = -0.13$, $p < 0.001$). The correlational structure provided robust evidence that the algorithms used in this series of studies functioned as expected (for the correlation matrix, see Table S3 in supplementary materials).

The Effect of Simulation Training with Feedback on Interview Quality

The complete results of the linear mixed models for question use, details elicited, and conclusion correctness can be accessed in supplementary materials (Tables S4, S5, and S6, respectively). There were considerable individual differences and within-study variation of interview quality as indicated by the random effects and intraclass correlations (ICCs) of the models. Simulation with feedback had robust effects on increasing recommended question use and decreasing non-recommended question use, improving details retrieval from the avatar, and finally reaching a correct conclusion regarding the suspected abuse. For models predicting question use, our main interest, the interaction term between feedback condition and round significantly predicted increased interview quality (recommended questions: $B = 2.03$, $SE = 0.16$, 95% *CI* [1.71, 2.34]; non- recommended questions: $B = -2.37$, $SE = 0.17$, 95% *CI* [− 2.68, 2.05]; percentage of recommended questions: $B = 5.34$, $SE = 0.32$, 95% *CI* [4.74, 5.95]). Similar patterns emerged for the details elicited during interviews, with increasing improved interview quality in the feedback condition (relevant details: $B = 0.40$, $SE = 0.05$, 95% *CI* [0.30, 0.50]; neutral details: $B = 0.30$, $SE = 0.05$, 95% *CI* [0.20, 0.40]; wrong details: $B = -0.40$, $SE = 0.05$, 95% *CI* [− 0.50, − 0.30]). In the generalized linear mixed model predicting conclusion correctness, the significant interaction between feedback and round showed increased correct rate in the feedback condition as the training progressed (interaction term: *odds ratio* = 1.39, $SE = 0.11$, 95% *CI* [1.20, 1.62]). The trends of interview quality improvement in the feedback condition can be seen in Figs. 1, 2, and 3. Note that these plots are not estimates from the mixed models but the actual data.

As for the effects of the other experimental conditions on interview quality, outcome feedback did not seem to have a robust influence on question use, details elicited, or conclusion correctness and neither did reflection (see Table S4, S5, and S6 in supplementary materials). All eight 95% CIs of outcome feedback indicated that no significant effect was found. Reflection only had a positive effect on neutral detail elicitation. Process feedback had positive effects on recommended question use ($B = 6.97$, $SE = 2.58$, 95% *CI* [1.88, 11.92]), relevant detail elicitation ($B = 1.61$, $SE = 0.65$, 95% *CI* [0.35, 2.85]), percentage of recommended questions ($B = 15.56$, $SE = 5.26$, 95% *CI* [5.05, 25.99]), and percentage of relevant details ($B = 26.98$, $SE = 7.40$, 95% *CI* [12.73, 41.40]), but no significant effect was detected on the conclusion correctness (*odds ratio* = 1.44, $SE = 0.78$, 95% *CI* [0.49, 4.19]). More importantly, modeling had significant effects on all interview quality indicators except for number of wrong details elicited. Modeling increased recommended question use ($B = 14.70$, $SE = 2.65$, 95% *CI* [9.47, 20.04]) while decreasing non-recommended question use ($B = -6.46$, $SE = 2.89$, 95% *CI* [− 12.22, − 0.82]), leading to a higher percentage of recommended questions ($B = 27.06$, $SE = 5.35$, 95% *CI* [16.34, 37.45]). Modeling also significantly increased the number of relevant ($B = 2.72$, $SE = 0.63$, 95% *CI* [1.49, 3.99]) and neutral details ($B = 2.86$, $SE = 0.61$, 95% *CI* [1.64, 4.04]), without increasing the number of wrong details in the meantime, resulting in higher percentage of relevant details ($B = 22.15$, $SE = 7.17$, 95% *CI* [8.07, 36.38]). As for the conclusions, modeling significantly increased conclusion correctness above and beyond the provision of feedback (*odds ratio* = 5.05, $SE = 2.81$, 95% *CI* [1.69, 15.05]).

We also did round-wise comparisons (i.e., compare each round's performance with the performance of the previous round) in the interviews that received either combined feedback or process/outcome feedback to examine the trend of training. The data contained 247 participants and 1307 interviews in total, and the results are presented in supplementary materials (Table S7, S8, and S9). Overall, round-wise increase of recommended question and decrease of non-recommended question were, for several comparisons, significant, suggesting continued improvement. The training

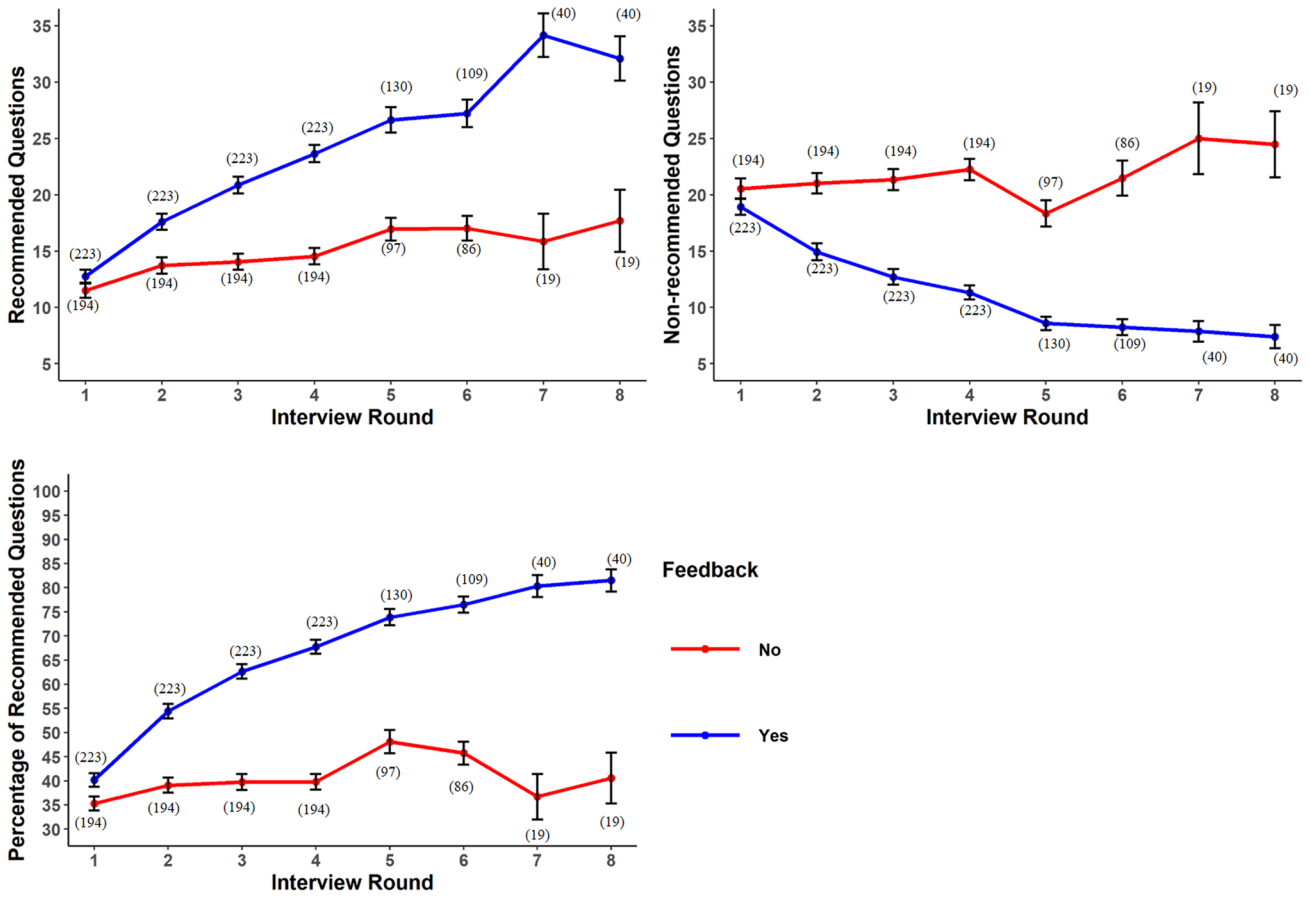

Fig. 1 Effects of simulated interviews with feedback on recommended and non-recommended question use. *Note*. Numbers in parentheses refer to the numbers of interviews in each condition. Error bars show standard error

effect did not seem to continue to improve in terms of details elicited, especially for wrong details. Round-wise difference of conclusion correctness was only significant in the first comparison (round 1 vs. round 2). Notably, all the comparisons between the 8th round and the 7th round were not significant. Whether this is an indicator of reaching plateau or a result of insufficient power demands further investigation.

Individual Differences in Interview Training: Reliable Change

The RCI results showed that only a minority of participants in the feedback group exhibited reliable change at the end of the training (see Table 3). For recommended question, 41.7% (93/223) of participants had a RCI greater than 1.96. But only 18.8% (42/223) had an RCI smaller than – 1.96 in the case of non-recommended questions. Similar patterns emerged when using the details elicited to examine reliable change: 26.0% (58/223) of participants achieved reliable change in terms of relevant detail elicitation and 30.5% (68/223) for neutral detail elicitation. As for wrong detail elicitation, only 8.5% (19/223) of participants had a RCI smaller than – 1.96.

A closer examination between training design and RCI showed that number of interviews had an impact on reliable change percentage. As shown in Table 3, participants who participated in a greater number of interviews were more likely to achieve a reliable change. These results corresponded to the round-wise comparison analyses, showing continuous improvement at the individual and the group level. Note that the higher percentage in the 5-round design from Haginoya et al. (2021) could be a result of small sample size and the added feature of modeling instead of indicating a non-linear trend for improvement. Combined, these results suggest that there are great individual differences in training effect, but with more practice, it is possible to improve the interview quality even among those who learn at a relatively slow pace.

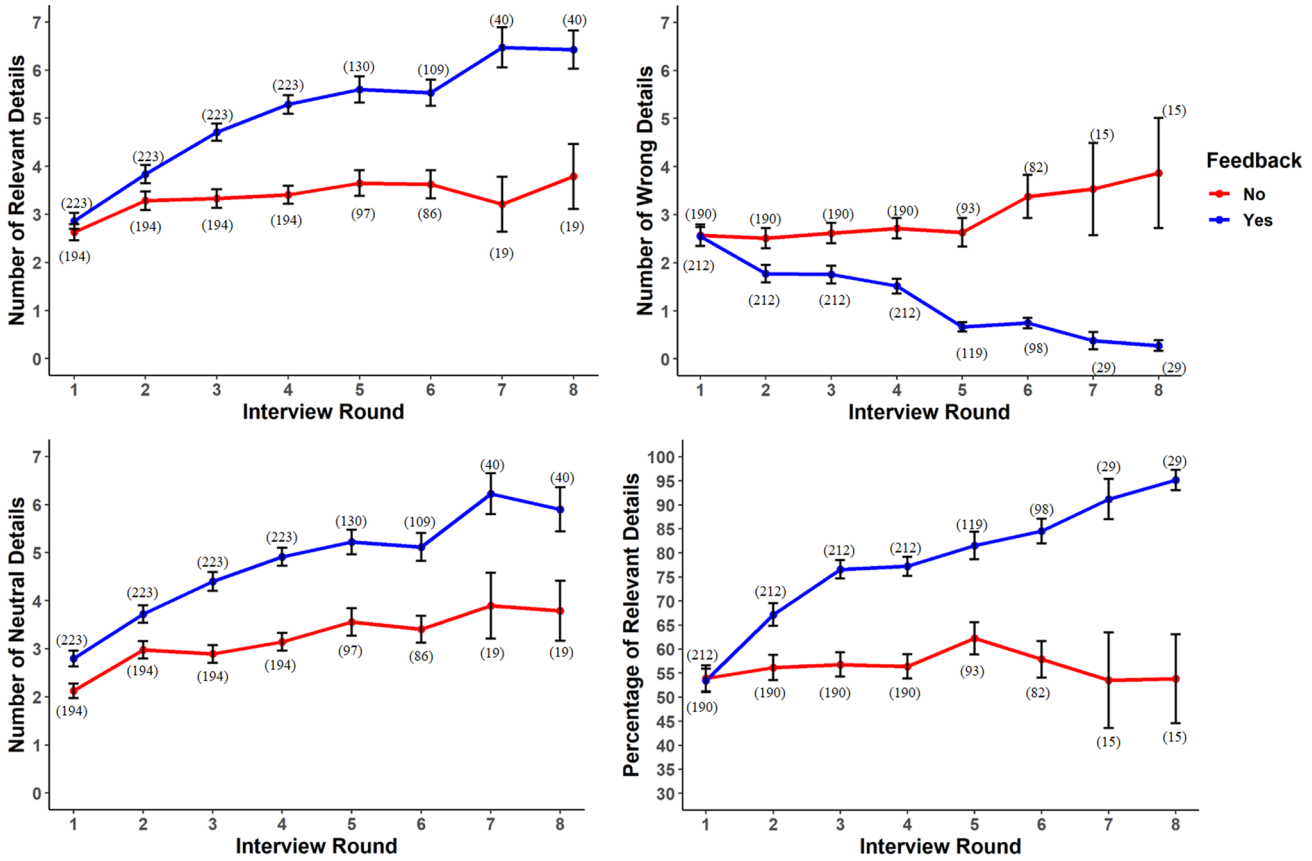

Fig. 2 Effects of simulated interviews with feedback on details elicited. *Note*. The percentage of relevant details elicited is calculated using the following formula: number of relevant details / (number of relevant details + number of wrong details). Numbers in parentheses refer to the numbers of interviews in each condition. Error bars show standard error

Professional and Parenting Experience Moderates Improvements over Interviews

To examine whether professional experience and parenting experience moderated the improvement, additional mixed models with professional experience, parenting, and the interaction term with round were run for all quality indicators. In terms of professional experience (for detailed results, see Table S10, S11, and S12 in supplementary materials), having professional experience positively predicted relevant details ($B = 0.82$, $SE = 0.35$, 95% CI [0.12, 1.53]) and neutral details elicited ($B = 0.72$, $SE = 0.34$, 95% CI [0.05, 1.40]), that is, all else being equal, individuals with professional experiences were better at eliciting information from the avatars in the first round of the interview. More importantly, the interaction term between professional experience and round was a significant predictor of number of recommended questions ($B = 0.96$, $SE = 0.25$, 95% CI [0.43, 1.45]), number of non-recommended questions ($B = -0.68$, $SE = 0.27$, 95% CI [−1.21, −0.17]), percentage of recommended questions ($B = 2.47$, $SE = 0.52$, 95% CI [1.40, 3.49]), relevant details elicited ($B = 0.18$, $SE = 0.08$, 95% CI [0.02, 0.33]), and neutral details elicited ($B = 0.26$, $SE = 0.08$, 95% CI [0.11, 0.42]). Professional experiences also predicted higher correct rate (*odds ratio* = 2.59, $SE = 0.99$, 95% CI [1.22, 5.49]), but the interaction with round was not significant (*odds ratio* = 0.88, $SE = 0.10$, 95% CI [0.70, 1.09]). After controlling for study-level, condition-level, and individual level variances, compared with those who had no experience in interviewing children, individuals with professional experience improved more over rounds of practice, as suggested by the significant interaction terms between professional experience and practice round.

Individuals with parenting experiences asked more non-recommended questions ($B = 3.13$, $SE = 1.28$, 95% CI [0.63, 5.68]) and obtained more wrong details at the beginning of the training ($B = 0.97$, $SE = 0.30$, 95% CI [0.38, 1.55]). Parenting experience also interacted with practice round to predict interview quality (for detailed results, see Table S13, S14, and S15 in supplementary materials). The 95% CI of the estimate of interaction term between parenting experience and round did not include zero for, number of non-recommended questions ($B = -0.73$, $SE = 0.29$, 95% CI [−1.30, −0.15]), relevant details elicited ($B = 0.20$, $SE = 0.09$, 95% CI [0.02,

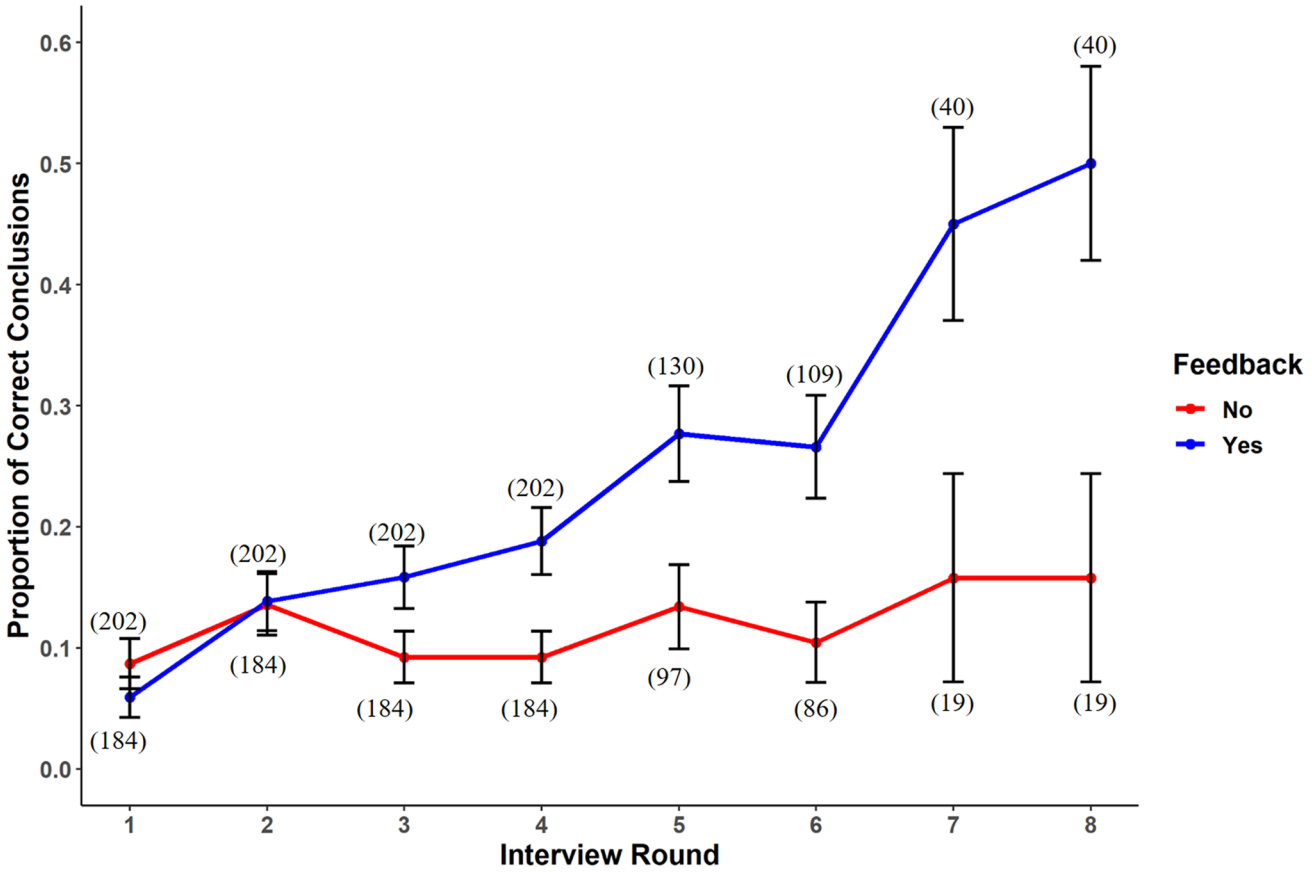

Fig. 3 Effects of simulated interviews with feedback on conclusion correctness. *Note.* Numbers in parentheses refer to the numbers of interviews in each condition. Error bars show standard error

0.36]), neutral details elicited ($B = 0.24$, $SE = 0.08$, 95% *CI* [0.07, 0.40]), wrong details elicited ($B = -0.30$, $SE = 0.08$, 95% *CI* [−0.46, −0.14]), and percentage of relevant details elicited ($B = 3.16$, $SE = 1.09$, 95% *CI* [1.02, 5.28]). All significant effects were in the expected direction. Parental experience had no impact on conclusion correctness regardless of training rounds. While all experimental conditions were controlled for, compared with those without parenting experience, individuals with parenting experiences showed greater improvement of interview quality over rounds of practices, using less non-recommended questions, eliciting more relevant and neutral but less wrong details during interviews. Importantly, though professional and parenting experiences interacted with interview rounds to positively predict interview quality, the effect sizes were smaller compared to those of experimental manipulations, as suggested by the smaller estimates of the fixed effects of the moderation models compared with models having all experimental manipulations as fixed effects.

Discussion

Through a systematic mega-analysis, the current research first examined the efficacy of the avatar training with feedback in terms of recommended question use, details

Table 3 Percentage of reliable change in designs with different rounds

Training design	Recommended questions	Non-recommended questions	Relevant details	Neutral details	Wrong details
4 rounds	22.6% (21/93)	12.9% (12/93)	17.2% (16/93)	14.0% (13/93)	9.7% (9/93)
5 rounds	71.4% (15/21)	14.3% (3/21)	47.6% (10/21)	71.4% (15/21)	9.5% (2/21)
6 rounds	36.2% (25/69)	17.4% (12/69)	23.2% (16/69)	36.2% (25/69)	2.9% (2/69)
8 rounds	80.0% (32/40)	37.5% (15/40)	40.0% (16/40)	37.5% (15/40)	15% (6/40)
χ^2 statistics	$\chi^2(3) = 46.60$, $p < .001$	$\chi^2(3) = 11.64$, $p = .009$	$\chi^2(3) = 13.20$, $p = .004$	$\chi^2(3) = 30.57$, $p < .001$	$\chi^2(3) = 5.14$, $p = .162$

elicited, and most of all, conclusion correctness confirming a strong and clear effect of feedback. Round-wise comparisons suggested continuous improvement in the use of recommended questions, while plateaus were reached in accurate information elicitation. Reliable change analyses offered insights on individual difference in training efficacy but also pointed out the potential of achieving reliable change by more practice. Moderation analysis also revealed that both professional experiences and parenting experiences may be conducive to the learning process.

The Efficacy of Simulated Interviews with Feedback

Interviewers, both legal and psychological professionals and university students, showed significant improvement in interview quality when undergoing the simulated training combined with feedback on their question use and interview outcome. Interviewers increased the use of recommended questions and reduced the use of non-recommended questions, which then led to better information gathering, that is, more relevant and neutral details and fewer wrong details being elicited from the avatars. Importantly, even though the program did not include explicit instructions on how correctly utilizing the elicited details to draw a conclusion, these variables had robust associations with reaching correct conclusion (see Table S2 in the supplementary materials). Note that in all the studies included in this mega-analysis, the criteria for a correct conclusion were very strict. The interviewers not only had to reach a correct judgment of whether an abuse had taken place but also had to provide a coherent account of how the abuse took place that was consistent with the ground truth of the case. Under this stringent standard, interviewers in the feedback condition were able to reach a correct conclusion at 22% rate on average and 50% rate in the 8th round.

The results are in line with previous research on the beneficial effect of feedback on the outcomes of child sexual abuse interviews (Benson and Powell 2015; Cederborg et al. 2013; Lamb et al. 2002a, b). While a perfect comparison is not possible, Benson and Powell (2015) showed on average between 57 and 79% of recommended question use. Within one training session that took 1–2 h, the avatar training program achieved improvement of interview quality comparable to other successful programs that last at least for a few days in terms of proportions of recommended questions.

How Many Interviews Will Be Enough and What Is the Goal?

Round-wise comparisons showed different patterns of development in question use, details elicited, and conclusion correctness. The recommended (vs. non-recommended) question use continued to improve, while the improvement of details elicited reached plateau earlier. This is not surprising as there is no limit to the number of questions a participant can ask (within the 10-min timeframe), while relevant and neutral details are finite within each avatar. As shown in Fig. 2, the average numbers of relevant and neutral details are close to 7 (the maximum is 9). The average number of wrong details was close to 1 in the later rounds of the training in the feedback condition suggesting a floor effect. This means that the interviewers used very few non-recommended questions that could have elicited wrong details at this point. Most of the round-wise comparisons for conclusion correctness were also not significant, which may suggest there is room for improvement in terms of using elicited information appropriately. However, it is also of note that all the estimates of the odds ratio were greater than 1, indicating a trend for continued improvement over interviews.

Though none of the round-wise comparisons between the 7th and 8th interview were significant, we should be cautious to draw the conclusion that the training reached plateau given the small number of interviews available in these comparisons. Also, it is important to note that even if the training effect did not reach plateau, it may not be necessary nor optimal to add more interviews to a training session. Instead, a more appropriate next step would be to focus on the relationship between learning during the simulated interviews and how this relates to transfer to actual interviews and then develop training plans also possibly including refresher training sessions (Cyr et al. 2021).

From the reliable change analyses, it is clear that there are large individual differences in the training effect. Only a minority of participants in the feedback group achieved reliable change. But it is also important to note that as the number of interviews increase, so do the percentages of reliable change. Therefore, by incorporating RCI into the design and offer individuals trainings with different lengths and intensities, future training programs can have greater impact.

Experience with Children on the Training Effect

The use of questions, both recommended and non-recommended, did not differ between those who had previous experience in interviewing children and those who did not in the first round of the simulated interview. However, individuals with interview experience were better at eliciting relevant and neutral information at the beginning. Individuals with parenting experience were more likely to use non-recommended questions and elicited more wrong details at the beginning. More importantly, both professional experience of child interview and parenting experience interacted with interview round to predict interview quality. Interviewers who had experience with children improved

faster compared with those who did not in terms of question use or information elicitation. Additional analyses were also run to probe if there were three-way interactions between experience, feedback, and interview round, but the results did not support the existence of a three-way interaction (see Table S16–S21 supplementary materials). This is a surprising result and goes against previous literature suggesting that professional experience is negatively associated with the use of open questions in simulated interviews and also goes against field studies that shows no relationship (for a review, see Lamb et al. 2018). However, this is in line with other studies that show how the type of training can overcome the effect of some a priori characteristics (e.g., Benson and Powell 2015). A possible explanation for this result is that parents and experienced interviewers, while might not have better ability in using open questions, might possess superior communication skills. The interactive nature of the training might have had a role in boosting the improvements in these groups of participants.

Limitations and Future Directions

Notwithstanding the strength of the mega-analysis approach, the current study has several limitations. First, we only examined the training efficacy within the Avatar Training system. That is, the scope of the current analysis did not include training efficacy in improving interview outside of the simulated environment. The reasons for not analyzing the transfer effect are, first, to keep the analysis concise and focused, and, second, as a result of lack of enough data for providing reliable results. At the moment of this analysis, only two studies have examined the transfer effect (Kask et al. 2022; Haginoya et al. 2022a). Secondly, also out of the scope of this article is how interview performance of previous rounds can influence learning of the subsequent rounds, and whether professionals and lay people respond to interview failures in the same way. Despite these limitations, the current study not only re-examined previous conclusions with a large sample offering more reliable estimates of the effects of combining process and outcome feedback, but also advanced our understanding of CSA interview training by its novel round-wise, RCIs, and moderation analyses.

The focus of this research has been on testing its efficacy in improving interview quality in a variety of different samples. However, this approach has the potential to be applied in other research areas in investigative psychology given that its algorithms are based on empirical research and have been proven to work as intended. The effects of contextual factors such as time pressure and fatigue as well as individual difference factors on interview quality can be examined within this training. That is, when not providing the interviewers with feedback, the avatar system could also function as a standardized assessment tool for the impact of contextual factors and individual's interview style and quality.

An interview is a dynamic interactive process between the interviewer and the interviewee; therefore, one other direction is to further develop the avatar training by tailoring the response patterns of the avatars based on family background, mental ability, and other factors that could make the children more or less vulnerable to suggestibility or compliance.

Conclusion

The present research demonstrated the robustness of the Avatar Training program in improving interview quality in interviewers with different backgrounds (e.g., working experience and specialty) and different training environments (face-to-face and remote online). This allows trainers in various fields to integrate Avatar Training into their interviewer training program flexibly. Moreover, this flexibility may imply successful training even when all procedures of the Avatar Training are automated to scale it to a large number of potential trainees such as police officers, clinical psychologists, child support center staff, and even school teachers.

Findings regarding interviewer background provided encouraging knowledge for interviewers who have experience with children. Experienced interviewers may be more likely improve faster than those without experience under the interactive training environment. Although potentially relevant factors (e.g., motivation to improvement) need to be investigated, this suggests that providers of training programs may need to consider an environment that promotes trainees to make the best use of their abilities.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11896-022-09509-7.

Funding Yikang Zhang's work is supported by the China Scholarship Council (No. 202106140025). Shumpei Haginoya's work is partially funded by the European Regional Development Fund (No 01.2.2-LMT-K-718-03-0067) under grant agreement with the Research Council of Lithuania (LMTLT).

Data and Code Availability Data are available upon request to respective authors. Code can be access at Open Science Framework (https://osf.io/hx2dr/).

Declarations

Ethics Approval The study utilized published data. All studies included in this mega-analysis obtained ethical approval from respective institutions.

Consent to Participate All studies included in this mega-analysis obtained informed consent from their participants before collecting data.

Conflict of Interest The authors declare no competing interests.

References

Bandura A, McClelland DC (1977) Social learning theory, vol 1. Prentice Hall, Englewood cliffs

Barth J, Bermetz L, Heim E, Trelle S, Tonia T (2013) The current prevalence of child sexual abuse worldwide: a systematic review and meta-analysis. Int J Public Health 58(3):469–483

Bates D, Mächler M, Bolker B, Walker S (2014) Fitting linear mixed-effects models using lme4. arXiv preprint arXiv:1406.5823

Benson MS, Powell MB (2015) Evaluation of a comprehensive interactive training system for investigative interviewers of children. Psychol Public Policy Law 21(3):309–322

Blasbalg U, Hershkowitz I, Karni-Visel Y (2018) Support, reluctance, and production in child abuse investigative interviews. Psychol Public Policy Law 24(4):518–527

Bruck M, Ceci SJ (1999) The suggestibility of children's memory. Annu Rev Psychol 50(1):419–439

Ceci SJ, Bruck M (1993) Suggestibility of the child witness: a historical review and synthesis. Psychol Bull 113(3):403–439

Ceci SJ, Bruck M (1995) Jeopardy in the courtroom: a scientific analysis of children's testimony. American Psychological Association, Washington, DC, US

Cederborg A-C, Alm C, da Silva L, Nises D, Lamb ME (2013) Investigative interviewing of alleged child abuse victims: an evaluation of a new training programme for investigative interviewers. Police Pract Res 14(3):242–254

Cederborg AC, Orbach Y, Sternberg KJ, Lamb ME (2000) Investigative interviews of child witnesses in Sweden. Child Abuse Negl 24(10):1355–1361

Cyr M, Dion J, Gendron A, Powell M, Brubacher S (2021) A test of three refresher modalities on child forensic interviewers' post-training performance. Psychol Public Policy Law 27(2):221–230

Elliott DM, Briere J (1994) Forensic sexual abuse evaluations of older children: disclosures and symptomatology. Behav Sci Law 12(3):261–277

Espinet SD, Anderson JE, Zelazo PD (2013) Reflection training improves executive function in preschool-age children: behavioral and neural effects. Dev Cogn Neurosci 4:3–15

Finnilä K, Mahlberg N, Santtila P, Sandnabba K, Niemi P (2003) Validity of a test of children's suggestibility for predicting responses to two interview situations differing in their degree of suggestiveness. J Exp Child Psychol 85(1):32–49

Hailes HP, Yu R, Danese A, Fazel S (2019) Long-term outcomes of childhood sexual abuse: an umbrella review. The Lancet Psychiatry 6(10):830–839

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: Forensic and Legal Psychology 11:998

Haginoya S, Yamamoto S, Mizushi H, Yoshimoto N, Santtila P (2022a) Improving supportiveness and questioning skills using online simulated child sexual abuse interviews with feedback. [Unpublished manuscript]

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 Negl 115:105013

Haginoya S, Yamamoto S, Santtila P (2022b) Improvement of interview quality in police officers using simulated child sexual abuse interviews with feedback. [Unpublished Manuscript]

Herman S (2009) Forensic child sexual abuse evaluations: accuracy, ethics and admissibility. In: Kuehnle K, Connell M (eds) The evaluation of child sexual abuse allegations: a comprehensive guide to assessment and testing. Wiley, Hoboken, NJ, pp 247–266

Hershkowitz I, Ahern EC, Lamb ME, Blasbalg U, Karni-Visel Y, Breitman M (2017) Changes in interviewers' use of supportive techniques during the revised protocol training. Appl Cogn Psychol 31(3):340–350

Jacoby LL, Debner JA, Hay JF (2001) Proactive interference, accessibility bias, and process dissociations: valid subject reports of memory. J Exp Psychol Learn Mem Cogn 27(3):686–700

Johnson M, Magnussen S, Thoresen C, Lonnum K, Burrell LV, Melinder A (2015) Best practice recommendations still fail to result in action: a national 10-year follow-up study of investigative interviews in CSA cases. Appl Cogn Psychol 29(5):661–668

Kask, K., Pompedda, F., Palu, A., Schiff, K., Mägi, M., & Santtila, P. (2022). Avatar training effects transfer to investigative field inter-views of children conducted by police officers [Manuscript submitted for publication]

Korkman J, Santtila P, Westeråker M, Sandnabba NK (2008) Interviewing techniques and follow-up questions in child sexual abuse interviews. European Journal of Developmental Psychology 5(1):108–128

Krause N, Pompedda F, Antfolk J, Zappalá A, Santtila P (2017) The effects of feedback and reflection on the questioning style of untrained interviewers in simulated child sexual abuse interviews. Appl Cogn Psychol 31(2):187–198

Lafontaine J, Cyr M (2017) The relation between interviewers' personal characteristics and investigative interview performance in a child sexual abuse context. Police Pract Res 18(2):106–118

Lamb ME, Brown DA, Hershkowitz I, Orbach Y, Esplin PW (2018) Tell me what happened: questioning children about abuse. John Wiley & Sons

Lamb ME, Hershkowitz I, Orbach Y, Esplin PW (2008) Tell me what happened: structured investigative interviews of child victims and witnesses. Hoboken, NJ: JohnWiley and Sons

Lamb ME, Orbach Y, Hershkowitz I, Horowitz D, Abbott CB (2007) Does the type of prompt affect the accuracy of information provided by alleged victims of abuse in forensic interviews? Applied Cognitive Psychology: the Official Journal of the Society for Applied Research in Memory and Cognition 21(9):1117–1130

Lamb ME, Sternberg KJ, Orbach Y, Esplin PW, Mitchell S (2002a) Is ongoing feedback necessary to maintain the quality of investigative interviews with allegedly abused children? Appl Dev Sci 6(1):35–41

Lamb ME, Sternberg KJ, Orbach Y, Esplin PW, Stewart H, Mitchell S (2003) Age differences in young children's responses to open-ended invitations in the course of forensic interviews. J Consult Clin Psychol 71(5):926–934

Lamb ME, Sternberg KJ, Orbach Y, Hershkowitz I, Horowitz D, Esplin PW (2002b) The effects of intensive training and ongoing supervision on the quality of investigative interviews with alleged sex abuse victims. Appl Dev Sci 6(3):114–125

Lyon TD (2014) Interviewing children. Annual Review of Law and Social Science 10(1):73–89
MacDonald S, Snook B, Milne R (2017) Witness interview training: a field evaluation. J Police Crim Psychol 32(1):77–84
Matthew CT, Sternberg RJ (2009) Developing experience-based (tacit) knowledge through reflection. Learn Individ Differ 19(4):530–540
Pate R (2012) Open versus closed questions: what constitutes a good question. Educational research and innovations pp. 29–39
Pompedda F (2018) Training in investigative interviews of children: serious gaming paired with feedback improves interview quality. (Doctoral dissertation, Abo Akademi University, Finland). Retrieved from 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. Front Psychol 8:1474
Pompedda F, Palu A, Kask K, Schiff K, Soveri A, Antfolk J, Santtila P (2020) Transfer of simulated interview training effects into interviews with children exposed to a mock event. Nordic Psychology 73(1):43–67
Pompedda F, Zappalà A, Santtila P (2015) Simulations of child sexual abuse interviews using avatars paired with feedback improves interview quality. Psychology, Crime and Law 21(1):28–52
Powell MB, Guadagno B, Benson M (2016) Improving child investigative interviewer performance through computer-based learning activities. Polic Soc 26(4):365–374
Powell MB, Hughes-Scholes CH, Smith R, Sharman SJ (2014) The relationship between investigative interviewing experience and open-ended question usage. Police Pract Res 15(4):283–292
Ron N, Lipshitz R, Popper M (2006) How organizations learn: post-flight reviews in an F-16 fighter squadron. Organ Stud 27(8):1069–1089
Sternberg KJ, Lamb ME, Davies GM, Westcott HL (2001) The memorandum of good practice: theory versus application. Child Abuse Negl 25(5):669–681
Taylor PJ, Russ-Eft DF, Chan DW (2005) A meta-analytic review of behavior modeling training. J Appl Psychol 90(4):692–709
Wolfman M, Brown D, Jose P (2016) Talking past each other: interviewer and child verbal exchanges in forensic interviews. Law Hum Behav 40(2):107–117