The Effects of Feedback and Reflection on the Questioning Style of Untrained Interviewers in Simulated Child Sexual Abuse Interviews

Niels Krause[1]
Francesco Pompedda[2]
Jan Antfolk[2]
Angelo Zappalá[3]
Pekka Santtila[2]

[1] Humboldt-Universität zu Berlin, Berlin, Germany; Åbo Akademi University, Turku, Finland
[2] Åbo Akademi University, Turku, Finland
[3] CRIMELAB, IUSTO – Pontifical Salesian University, Turin, Italy; Åbo Akademi University, Turku, Finland

Correspondence to: Francesco Pompedda, Department of Psychology, Åbo Akademi University, Fabriksgatan 2, 20500 Turku, Finland.
E-mail: fpompedd@abo.fi

Due to a lack of eyewitnesses and corroborating evidence, investigative interviews with alleged victims are of central importance in child sexual abuse (CSA) investigation. In almost 70% of cases, the child's statement is the only evidence to rely on in court (Elliott & Briere, 1994; Herman, 2009). As international research has shown, interview quality re- mains quite poor worldwide (Cederborg, Orbach, Sternberg, & Lamb, 2000; Korkman, Santtila, Westeråker & Sandnabba, 2008b; Sternberg, Lamb, Davies, & Westcott, 2001). For example, a Joint Inspectorate report in England and Wales (Criminal Justice Joint Inspection, 2014) described a 'widespread tendency to also pose specific closed questions throughout the interview, which tended to elicit shorter and less detailed responses' (p. 22) and 'the use of leading questions was common where a more open style of questioning would have been appropriate' (p. 23), showing the continued need for training. Training programs, even if some promising results have been reported (Benson & Powell, 2015; Cederborg, Alm, Lima da Silva Nises, & Lamb, 2013; Yi, Jo, & Lamb, 2015), have generally failed in creating and maintaining improvements in the quality of these interviews. For example, a Norwegian follow up (Johnson et al., 2015) showed no improvement in interview quality over a time span of 22 years, in spite of considerable investment in training. The most promising research has shown that, together with a structured protocol, feedback on questions used must be provided in an immediate, continuous and detailed way (Lamb, Sternberg, Orbach, Esplin and Mitchell, 2002a; Smith, 2008). This is a problem for a number of reasons: CSA interviewers rarely get feedback on their use of question types outside of scientific studies. Organizing this type of training can result in high costs and logistical problems. Also, in most real CSA cases, it cannot be reliably known whether a child's statement, or parts of it, are actually true (Vrij, 2005), resulting in a lack of feedback on the conclusions drawn by the interviewers.

For this reason, we have applied the concept of serious gaming and response algorithms to training interviewers in the context of CSA cases. Serious gaming, that Ritterfeld, Cody, and Vorderer (2009)

defined as ‘any form of interactive computer-based game software for one or multiple players to be used on any platform and that has been developed with the intention to be more than entertainment’ (p. 6), has been proved to be effective in improving learning in different fields (for a review see Olszewski, 2016; Van Dijk, Spil, van der Burg, Wenzler, & Dalmolen, 2015; Wouters, Van Nimwegen, Van Oostendorp, & Van Der Spek, 2013), in improving complex skills, for example, in surgical skills training (Graafland, Schraagen, & Schijven, 2012) and in improving the use of open-ended questions in a group of teachers (Brubacher, Powell, Skouteris, & Guadagno, 2015) and students (Pompedda, Zappalà, & Santtila, 2015). However, the current study adds the concepts of probabilistic response algorithms and reflection that was not present in the previously mentioned work. In the simulation, virtual children (avatars) are displayed on a computer screen, and the trainees are asked to interview them about a specific sexual abuse allegation. The avatars possess pre-defined ‘memories’ that alongside other information either do or do not contain memories of sexual abuse. The avatars’ answers (contained in video clips) are determined by algorithms based on research about children’s responses to different kinds of questions in interviews. The task of the trainees is to conduct the interview and based on the avatar’s responses reach a conclusion about what has happened (Pompedda, Zappalà, & Santtila, 2015).

Best practice in interviewing children

According to research, the interview should comprise an introductory phase including rapport-building, an explanation of ground rules and a practice narrative (Lamb, Hershkowitz, Orbach, & Esplin, 2008; Lamb, La Rooy, Malloy, & Katz,2011). The following main rules are recommended during an interview: First, questions should be non-leading because leading questions can have a negative impact on children, creating less accurate statements and contaminated memories (Bruck & Ceci, 1999; Ceci & Bruck, 1993, 1995). 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 accurate answers (Lamb et al., 2003, 2008; Lyon, 2014; Rocha, Marche, & Briere, 2013; Waterman, Blades, & Spencer, 2000). Third, questions should be formulated using clear and easy language in order to be understood by the child, thereby avoiding misunderstandings on both sides(Korkman, Santtila, Drzewiecki & Sandnabba, 2008a; Lyon, 2014). In particular, questions concerning such complex cognitive domains as time (Friedman & Lyon, 2005; Wandrey,Lyon, Quas, & Friedman, 2012) or feelings (Pons, Harris,& de Rosnay, 2004; Pons, Lawson, Harris, & de Rosnay, 2003) should be avoided when interrogating very young children. Fourth, activating fantasy by asking children to imagine how something might have happened is viewed as a potentially harmful suggestive technique (Lamb, Sternberg, & Esplin, 1998; Lamb et al., 2008; Poole & Lamb, 1998).

Training investigative interviewers

Considerable effort has been invested in developing and testing training programs to improve interviewer’s ability to follow best practice. These programs are typically in the form of short and intensive courses including lectures, discussions and partnered exercises. Although improving theoretical knowledge, they have generally failed to transform improved theoretical knowledge into changed practice (see, for example, Cederborg & Lamb, 2008; Sternberg et al., 2001). As a possible solution to these challenges, we have developed a simulation of CSA investigative interviews to provide interviewers with appropriate feedback (Pompedda, Zappalà, & Santtila,2015). Two kinds of feedback can be given within this paradigm: (i) feedback can be provided on question types used by

the interviewers. Participants thus get information on which kind of questions they are supposed to use more and which questions to avoid. In contrast to real cases, we also (ii) give feedback on the correctness of the conclusion reached by the interviewer. That is, a sub-standard interview might lead to an erroneous conclusion, and when the ground truth is known, interviewers can receive feedback on both the poor questioning and exactly how the poor questioning style led to the wrongful conclusion. For example, if the interviewer asks an option-posing question, with this simulation it is possible to state if that particular question elicited a wrong detail. As stated before, feedback must be detailed (Smith, 2008); however, the feedback provided within other training programs or real interviews generally lacks this level of detail.

Enhancing feedback through a reflection task

Reflection on previous task-related behavior has been proposed as a tool for enhancing learning from experience and to help acknowledging what one has learned from feedback (Ellis, Carette, Anseel, & Lievens, 2014; Seibert, 1999). A variety of approaches to reflection have been studied in different fields, such as education (Espinet, Anderson, &Zelazo, 2013), military leadership (Matthew & Sternberg, 2009) and aircraft navigation (Ron, Lipshitz, & Popper, 2006), showing that reflection can affect motivational and cognitive processes as well as behavioral outcomes 'resulting in a prominent tool for learning from experience'(see Ellis et al., 2014, for a review of recent studies).

Anseel, Lievens, and Schollaert (2009) constructed a simple intervention to stimulate reflection after feedback, drawing on a paradigm from persuasion research: Stimulating deeper processing of arguments through the generation of examples. They focused on an understanding of reflection as 'aim[ing] to intensify cognitive elaboration of experiential data, leading to the necessary behavioral changes' (Anseel et al., 2009, p. 24). Through reflection tasks, participants can allocate the necessary cognitive resources into processing the feedback message (Anseel et al., 2009 p. 24). After giving feedback on four broad categories in an online work simulation, they asked participants to think back to their previous performance and generate examples of successful and unsuccessful behavior related to the feedback categories. Participants who received feedback and the reflection task performed better in a subsequent parallel task than those who only received feedback (Anseel et al., 2009).

While there is evidence of reflection tasks being effective in enhancing different training effects, there are (to our best knowledge) no studies that have investigated if reflection tasks can improve interview quality in the context of CSA cases beyond feedback alone.

Aims and hypotheses of the current study

A previous study, using a similar interview simulation, showed that feedback can increase the use of open-ended questions, while option-posing and suggestive questions become less frequent (Pompedda, Zappalà, & Santtila, 2015).

The aim of the present study was two-folded. The first aim was to replicate the effects of feedback on questioning style and correct conclusions using updated algorithms and an improved interview simulation. Whereas in previous studies, an operator had to select the avatar's answers manually while keeping track of the algorithms, here the algorithms were automated in the training software. The new version of algorithms works in a probabilistic way—simulating the possible answers that a child of a specific age would provide according to the question type asked. The appearance of the avatars was also improved, and new avatars were developed. Moreover, the current study employed a

larger sample, and training sessions lasted longer (Pompedda, Zappalà, & Santtila, 2015). Finally, participants conducted eight 10-min interviews in a row, twice as many as in the previous study (Pompedda, Zappalà, & Santtila, 2015).

The first hypothesis concerned the effect of feedback on the general quality of the interviews. With the term 'recommended questions' we intended all question types that, according to different studies, more probably elicit a reliable answer from the child; with the term 'not recommended questions', we intended all question types that, according to different studies, less probably elicit a reliable answer from the child. We expected that interview quality, measured as the proportion of recommended questions per interview and the total number of relevant and neutral details obtained from the avatars, would increase, whereas wrong details would decrease over time for participants who received feedback. As results of those improvements, the participants in the feedback groups would reach more correct conclusions.[1] In addition, we measured the proportion of participants reaching reliable change in the use of recommended questions (as defined in the introduction) during the interviews.

Method

Participants

Fifty-nine participants (35 female) aged 18 to 36 years (M= 24.4, SE= 0.2) were recruited mainly from a university campus and rewarded with two movie tickets for their participation. Most (n= 50) were university students; the remaining had gone through at least three years of higher education. None had work experience in CSA investigation, and only two were parents. The data from two participants out of the original sample (n= 61) were not used because of an error when determining the sequence of avatars they had to interview.

Design

The study used a mixed factorial design, with two independent factors, one between-subject factor with three levels and one within-subject factor with eight levels. Participants were randomly divided into three experimental groups [control (n = 19), feedback (n = 19), feedback plus reflection (n = 21)] and conducted eight interviews within one session. Procedures were equal for all participants apart from the feedback provided after each interview (both in the feedback and the feedback plus reflection groups) and the reflection ask following the feedback (only in the feedback plus reflection group). The procedure was approved by The Ethics Committee of the Department of Psychology and Logopedics at the Åbo Akademi University.

Materials

Simulation of investigative interviews

Simulated interviews with avatars were performed using the Empowering Interviewer Training software (EIT®, Version1.12.7, Åbo Akademi University, Turku, Finland). During the interviews, an operator listened to the interviewer's questions, categorized them and fed the categories into the software via a graphical interface. The avatar's answer was then selected by probabilistic algorithms derived from a broad range of empirical results on children's memory performance and suggestibility in investigative interview situations.

[1] Throughout the manuscript when we will mention interview quality, we will always refer to this definition, which is exclusively based on the questioning style used by the interviewer.

Avatars

The program comprised 16 child avatars differing in age (4vs. 6 years), gender (female vs. male), emotionality [emotional (crying) vs. neutral] and the presence or absence of a sexually abusive incident in the avatars' 'memory'. Avatars were created by the procedure described in (Pompedda, Zappalà, & Santtila, 2015), morphing real children's pictures and animating them with the SitePal (Oddcast, New York,NY) video engine.

Procedure

Question classification

The coding of question types was based on a scheme adapted from previous work of Korkman, Santtila, and Sandnabba (2006) and Sternberg et al. (1996) (see Table 1 for an overview). Each of the experimental sessions was conducted by one of two experimenters. One of the experimenters had conducted more than 300 interviews with this tool, and four pilots sessions were conducted together by both experimenters in order to reach agreement regarding the coding of the question types. Cases of disagreement were resolved by discussion after the end of the session. Moreover, we conducted analyses for interrater agreement. One psychology student, who was blind to the purposes of the research, coded 118 interviews (25% of the total sample) comprising a total of 4738 questions. It is known that Cohen's Kappa is sensitive to marginal homogeneity (Di Eugenio & Glass, 2004; Feinstein & Cicchetti, 1990; Gwet, 2002), and Gwet's *AC1* has been proposed to solve this bias (Gwet, 2008, 2010; Wongpakaran, Wongpakaran, Wedding, & Gwet, 2013). Because a Stuart-Maxwell test (Bickeböller & Clerget-Darpoux, 1995; Maxwell, 1970; Stuart, 1955) showed lack of overall marginal homogeneity in our sample, $\chi^2(7, N= 4738) = 377.262$, $p<.001$, $\varphi c = .11$, we decided to report both measures: Cohen's Kappa (Cohen, 1960) because of its wide use in the literature and Gwet's *AC1* (Gwet, 2002) as a solution to this bias. The overall percentage of agreement between raters was 80%, *CI* [79%, 81%], with $\kappa = .684$, $p<.001$, 95% *CI* [.669, .701], and Gwet's *AC1* = .785, $p<.001$, 95% *CI* [.773, .798] evidencing adequate interrater reliability.

Table 1 Description and examples of recommended and not recommended question types

Question types	Description	Examples
Recommended questions		
Facilitators	Non-suggestive utterances that encourage the child to continue with an answer. They may also be requests for clarification.	'What happened after that?' 'Continue' or 'Ok'
Invitations	Non-suggestive and general questions allowing the child to produce recollections	'What do you usually do with your dad?' 'Tell me what you did yesterday afternoon'
Directive	Questions that focus the child's attention on a detail that has previously been mentioned by the child and ask for further explanation.	'Where did you go with mum?' 'What game did you play with Philip?'

Not recommended questions		
Option posing	Closed questions that focus the child's attention on a detail not mentioned previously, but do not imply a particular type of response.	'Do you play with dad?'
Specific suggestive	Questions communicating an expected response and which assume details that the child has never mentioned before	'Did he do something that you did not like?' 'Is your dad a bad person?'
Unspecific suggestive	Questions communicating an expected response but do not assume details that have not been revealed by the child earlier.	'I know that you have something to tell me. Just talk about it!'
Repetitions	This category refers to questions that are asked more than once in a row.	
Too long	Questions containing more than one concept or several questions in a series within the same question.	'Did you go to the park and did you behave?'
Unclear	These questions contain words too difficult for the cognitive level of the child or the questions have been formulated in a haphazard manner.	'What is the relationship between your mum and your dad like?'
Multiple choice	These questions lead the child's focus towards certain answers or force him/her to choose between alternatives.	'Did you go to the training with Fabian or Matthew?'
Time	These types of questions rely on cognitive processes that are still unreliable in children up to six years of age.	'When did your mum leave the park?'
Fantasy	These type of questions can activate the child fantasy, yielding possible inaccurate answers.	'Pretend to be your father...What did he do?'
Feelings	These type of questions rely on cognitive processes that can be not completely developed in children up to six years of age.	'How do you feel about your grandfather?'

Answer selection

The EIT software contains probabilistic algorithms for each question type specifying the probabilities for a range of answers (see Table 2). For each question entered by the opera-tor, an answer was

chosen and played by the computer with the help of a random number generator. There are different sets of algorithms, both for 4- and 6-year-old avatars. Differences between the algorithms were based on empirical findings on children's memory and suggestibility in investigative interviews. (An example of an algorithm is displayed in Figure 1).

Narrative responses (details) were revealed only in reply to a recommended question. Every avatar held a set of nine relevant and nine neutral details in its 'memory' that would be presented one at a time in the form of one or two short sentences. Relevant details were related to the incident to be investigated, they were presented in a fixed order to all interviewers and only the last four relevant details provided information allowing a correct conclusion about the CSA allegation. Neutral details were presented in the same way as relevant details but provided unrelated information about the child's life, for example, what the child was eating on a specific day. However, those details were not useful in order to achieve a correct conclusion on the case. Every time the interviewers asked a recommended question, the probability of eliciting a relevant or a neutral detail was set to 12.5% each for 4 year old and to 25% for 6 year old, reflecting differing abilities to recollect from memory.

Table 2 Avatars' possible responses to the question types specified by the algorithms

Question type	Possible answers
Recommended	'Yes', 'No', 'I don't know', relevant detail, neutral detail
Option posing	'Yes', 'No', 'I don't know'
Specific suggestive	'Yes', 'No', 'I don't know'
Repetition of an option posing	'Yes', 'No', 'I don't know'
Multiple choice	'The first one you said', 'The last one you said', 'I don't know'
Unspecific suggestive	'I don't know how to say it'
Time/feelings/fantasy	'Have we finished already?'
Too long or unclear	'Yes', 'No', 'I don't understand', random detail
Repetition of a recommended question	'I don't know how to say it', 'Have we finished already?', 'I don't understand', 'Enough, I won't speak anymore!'

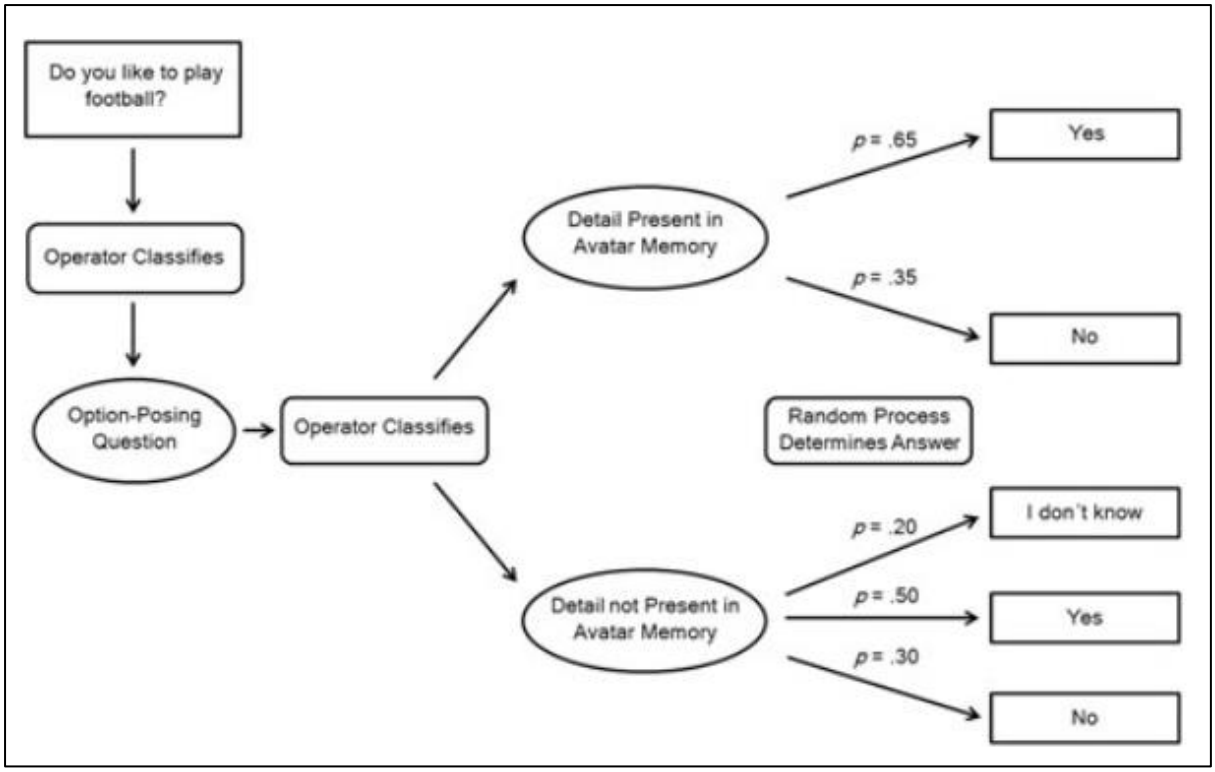

Figure 1 Example of an algorithm. Here, the algorithm is determining the possible answers to option-posing questions in 4-year-old avatars

To increase perceived realism of the simulation, there were varying numbers of additional 'side details' present in each child's repertoire, which could be chosen manually from a list by the operator. They were sorted into different topics and coded with one or two words to make them easily accessible for the operator. For example, when the interviewer asked 'What do you usually do with your grandmother?', instead of activating the 'recommended question' algorithm, the operator chose the first side detail in which a description of the activities was present. The operators provided answers of that type until all details of that category had been revealed or the interviewer changed topic. Side details were provided only in response to recommended questions. However, recommended questions related to the allegation or that did not match a side detail category (or one for which all details had been already provided), led to activating the standard algorithm.

Wrong details instead consisted of erroneous information that an interviewer can obtain using a not recommended questioning style. Because we had predefined the memory contents of each avatar, we know with certainty if a specific question elicited a wrong detail. For example, if the avatar based on the response algorithms answers 'No', we can easily check if this is a wrong detail or not.

Preparations

For each participant, eight out of 16 possible avatars were randomly selected for the participant to interview. The avatars were selected so that they would include all possible combinations of age, gender, abuse or non-abuse history and emotionality. The sequence of the selected avatars was then randomized.

Upon arrival, participants signed informed consent and confidentiality agreements. Subsequently, they received instructions about best practice in child interviewing (see Appendix 1). To make sure they had read and understood the brief, participants had to answer two questions about its contents. If they gave even one wrong answer, they were instructed to go through the brief again. Finally, they received oral instructions on the procedure to be followed within the study.

Interviews

Sessions lasted between 1.5 and 2.5 h in total, and the inter-views were videotaped. Before the start of each interview, participants were given a background sheet about the avatar. The sheet contained information about her or his home situation and on which grounds sexual abuse was suspected. An example is provided in Appendix 2. Two questions were asked before the interview; the interviewers' first impression regarding the presence or absence of abuse, together with a question regarding how sure they were about their response on a scale from 50% (guessing) to 100% (completely sure). The participants were free to conduct the interview as they preferred, but they were instructed to focus on the investigation of the alleged abuse. Interviews lasted a maximum of 10 min, but participants were instructed that they could finish in advance if they were satisfied with the information elicited from the child. The experimenter sat in a room next to the participant.

At the end of each interview, participants were asked for a conclusion about the case (had sexual abuse taken place or not). Additionally, they were asked to provide as much detail as possible regarding what they thought had happened based on their findings in the interview. In a sexual abuse case, they had to describe the abusive situation and the abuser in as much detail as they could; in a not abuse case, they had to give an alternative explanation for the allegation in as much detail as they could. A conclusion was coded as correct only if all the details related to the story were provided: who was the perpetrator, where and how the abuse happened or the alternative explanation of the story if no abuse has taken place.

Participants in the two treatment groups then received feedback. First, they were presented with the correct solution of the case. This type of feedback is hardly achievable in a real-life context. After that, they received feedback on the first two recommended and the first two not recommended question types they had used during the interview. This included the question they had used, a description of the question type and information on whether an answer to this kind of question could be considered reliable or not.

Subsequently, in the following interviews, priority was given to new types of question categories used by the interviewer. For example, if the first feedback on not recommended questions comprised option-posing and specific suggestive questions and the interviewer in the subsequent interview still used these two categories but also new ones (for example, too long questions and repetitions), priority was given to the new ones, otherwise feedback was provided on the old ones.

Participants also receiving the reflection task were finally asked to think back to the interview and provide one additional example for every question type they had received feedback on. If they did not remember a suitable question from the interview, they were allowed to make up a new one instead. If the question did not match the expected category, participants were asked to provide another example without further comments from the researcher.

A new trial then began with handing out the next background story. After the last interview, participants received their reward as well as debriefing about which experimental group they belonged to and had the possibility to ask questions.

Statistical analyses

Hypotheses were tested conducting Group (3) by Time (8) repeated measures analyses of variance (ANOVA) on percentages of recommended questions, number of relevant, neutral and wrong details and also for the percentages of correct conclusions. Except in four cases, the Mauchly's Test of Sphericity indicated that ANOVA assumptions had been violated: the value of Epsilon (ε) was always <.75. For this reason, a Greenhouse–Geisser correction for degrees of freedom was used (Girden, 1992). Confidence intervals for group means were corrected with a procedure proposed by Jarmasz and Hollands (2009). We also conducted post hoc group (2) × time (8) repeated measures ANOVAs comparing pairs of experimental groups to each other. For reasons of brevity, the statistical details from these comparisons are reported in the Appendix (Tables A1–A4). Reliable change indices (RCI) for the variable percentage of recommended questions were calculated to determine how many of the participants had reliably changed their questioning style. We applied the RCI proposed by Chelune, Naugle, Lüders, Sedlak,and Awad (1993), using standard deviations of the whole sample at baseline (the first interview) and the intraclass correlation coefficient from the control group across all time points for the test of the reliability and used a ± 1.645 change (90% of the confidence interval) as the criterion (Parsons, Notebaert, Shields, & Guskiewicz, 2009). For the calculation of the RCI, we decided to compare the first interview to the average value of the last seven, instead of comparing the first versus the last interview. In this way, we used a value free from biases related to the performance in a single interview.

Results

Descriptive statistics

Overall means for the first interview were: 33.4 (*SE*= 0.3) for percentage of recommended questions and 66.6 (*SE*= 0.2) for percentage of not recommended questions, 2.1 (*SE*= 0.3) for numbers of relevant details, 2.1 (*SE*= 0.2) for numbers of neutral details and 2.6 (*SE*= 0.4) for numbers of wrong details. In Table 3 instead are presented the correlations among questions, details and conclusions among all the interviews.

Participants overall reached the correct conclusion in 19% of the cases showing that the task was quite difficult. The participants' dichotomous conclusions about whether the avatar had been sexually abused or not were correct 62% of the time (chance expectation was 50% as half of the avatars were sexually abused) with no differences between the experimental groups.; however, only participants in the feedback groups were able to substantiate their conclusion with correct details.

Correlations between recommended questions, details and conclusions

The algorithms worked in the expected way with the proportion of recommended questions being positively correlated with the number of relevant and neutral details as well as with the probability of correct conclusions and negatively correlated with the number of wrong details. All these correlations were statistically significant (Table 3).

Baseline performance

One-way ANOVAs did not reveal any significant differences between the experimental groups in baseline performance; neither did any of the demographic variables Age, Gender or Education differ between the groups. We also found no differences regarding the dependent variables of interest. For Wrong Details, however, the Levene test for the Homogeneity of Variances was significant, but a subsequent Brown–Forsythe Robust test revealed no significance.

Interview quality

Over the eight interviews, an improvement in interview quality could be observed in both groups that received feedback, but not in the control group. The ANOVA on percentage of recommended questions revealed significant main effects for group (F[2, 56] = 22.29, p<.001, η_p^2= .44, 1−β>.99), time (F[4.15, 232.12] = 64.93, p<.001, η_p^2= .54, 1−β>.99), as well as a significant group × time interaction, (F[8.29, 232.12] = 12.32, p<.001, η_p^2= .31, 1−β>.99) (see Figure 2, Panel A). The results of the planned comparisons (Table A1) showed a significant effect of group, time and time × group interaction for control versus feedback groups and control versus feedback plus reflection groups.

Table 3 Means, standard error and correlations between recommended questions, details and conclusions

Variable	M	SE	1	2	3	4	5
1. Number of relevant details	4.23	0.14	-				
2. Number of neutral details	4.05	0.13	0.66**	-			
3. Number of wrong details	1.78	0.15	-0.30**	-0.19**	-		
4. Conclusion correct	0.19	0.02	0.53**	0.46**	-0.20**	-	
5. Percentage recommended questions	55.17	1.18	0.67**	0.63**	-0.56**	0.40**	-

*Note: The values of the percentage of not recommended questions are the reverse of the recommended questions. **p<.01.*

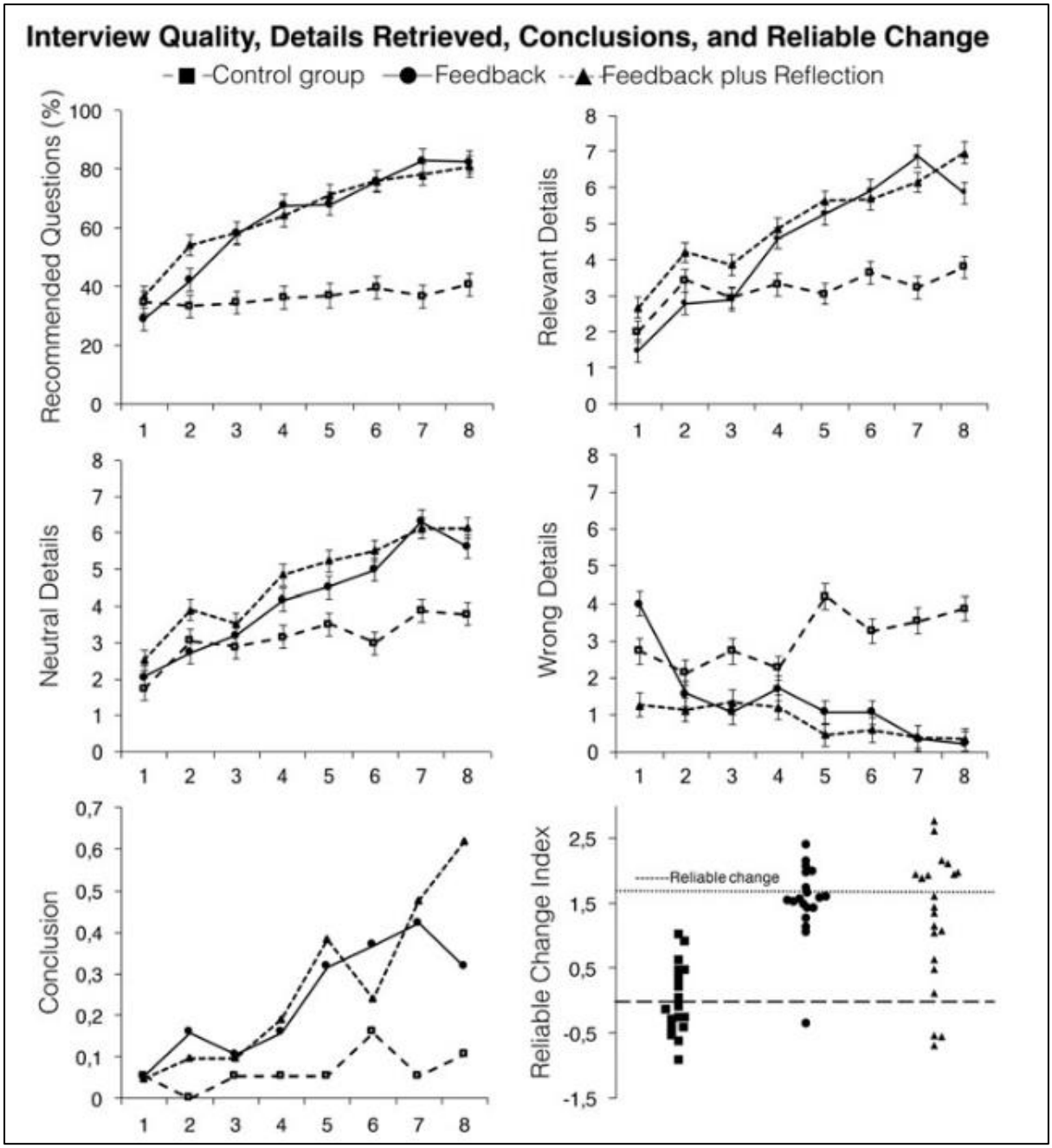

Figure 2 Interview quality, details retrieved, conclusions and reliable change by group. In panels A–E, the x-axis displays the interview number (1–8). Panel A displays the use of recommended question by group. Panels B–D display the details (relevant, neutral and wrong) retrieved by group. Panel E displays the probability of reaching a correct conclusion by group. Panel F displays the reliable change by participant and group

Apart from differences at some particular time points, no meaningful differences were found between the feedback and the feedback plus reflection group.

Number of details elicited from the avatars

For relevant details, the ANOVA revealed significant main effects for group, (F[2, 56] = 5.54, p<.001, η_p^2 = .17, 1−β>.84), time, (F[5.66, 317, 30] = 20.52, p<.001, η_p^2= .27, 1−β>.99), as well as a significant group × time interaction, (F[11.33, 317, 30] = 2.98, p = .001, η_p^2= .96,1−β>.98) (see

Figure 2, Panel B). The results of the planned comparisons (Table A2) showed a significant effect of group, time and time × group interaction for control versus feedback groups and control versus feedback plus reflection groups. Taken together, in the two feedback groups, but not in the control group, the number of elicited relevant details increased as a function of time. We only found differences at particular time points when comparing the feedback and the feedback plus reflection groups.

For neutral details, the ANOVA[2] also revealed significant main effects for group (F[2, 56] = 5.30, p= .008, η_p^2= .16, 1−β>.82) and time (F[7, 392] = 16.32, p<.001, η_p^2= .23, 1−β>.99), but not a significant group × time interaction (Figure 2, Panel C). The results of the planned comparisons (Table A3) showed a significant effect of group and time for control versus feedback groups and control versus feedback plus reflection groups. Again, in the two feedback groups, but not in the control group, the number of elicited neutral details increased as a function of time. Apart from differences at some particular time points, no meaningful differences were found between the feedback and the feedback plus reflection group.

For wrong details, we expected the reverse pattern. Also here, the ANOVA revealed significant main effects for group, (F[2, 41] = 6.54, p = .003, η_p^2= .24, 1−β>.88)and group × time interaction, (F[8.75, 179.40] = 3.02, p = .002, η_p^2= .13, 1−β>.96), but not for time (see Figure 2, Panel D). The results of the planned comparisons (Table A4) showed a significant effect of group, time and time × group interaction for control versus feedback groups and control versus feedback plus reflection groups. As expected, in the two feedback groups, but not in the control group, the number of elicited wrong details decreased as a function of time. Again, for feedback versus feedback plus reflection, we only found occasional differences at some particular time points.

Proportion of correct conclusions

A one-way ANOVA was conducted to compare the effect of feedback on the proportion of correct conclusions (see Figure 2, Panel E). The proportion of correct conclusions was calculated using all the interviews except for the first. The analyses showed significant differences between the groups on the proportion of correct conclusions (F[2,56] = 12.03, p≤.000, η_p^2= .30). A Tukey HSD post-hoc test revealed that the control group (M= 6.8, SE= 2.3) was significantly different compared to the feedback (M= 26.3, SE= 4.0) and feedback plus reflection groups (M= 29.9, SE= 4.1). However, no differences were found between feedback and feedback plus reflection groups (see Figure 2, Panel E).

Reliable change analysis

As expected, none of the participants from the control group reached a positive reliable change in the proportion of recommended questions used during the interviews with 53% (n= 10) of participants actually decreasing their performance compared to the first interview. In the feedback group, 95% (n= 18) of the participants increased their performance, and 32% (n= 6) reached a reliable change. In the feedback plus reflection group, 86% (n= 18) improved their performance, and 43% reached a reliable change (n= 9) (see Figure 2, Panel F).

[2] No Greenhouse–Geisser correction needed.

Discussion

In the present study, we simulated interviews using child avatars with answering behavior determined by research-based algorithms and pre-defined memories. Participants who received feedback improved the quality of their interviews, eliciting more correct details and reaching more accurate conclusions. In contrast, the overall performance in the control group did not improve. It should also be noted that participants in the two feedback groups were reluctant to guess the conclusion. When they did not find out enough details about the story, they were more likely to say 'I don't have enough information to provide a reliable conclusion' compared to the control group. The participants in the control group were as good as participants in the feedback group to correctly differentiate between abuse and not abuse without providing any details about the story. However, this is not surprising considering that one could guess 50% of the conclusions correctly by chance alone. In fact, only the participants in the feedback groups were able to substantiate their decision with correct details in 24% of cases, while participants in the no feedback group were able to do so only in 7% of cases. We also investigated the training effects on an individual level by calculating RCI. This is especially interesting as there is considerable variance between participants in the proportion of recommended versus not recommended questions. The findings show that almost all participants in the feedback groups increased the use of recommended questions, and that 32 and 43% reached a reliable change in the feedback and feedback plus reflection group, respectively. In the control group, none reached a reliable change. The first hypothesis stating that interview quality improves within the groups receiving feedback was thus corroborated.

These results stand in line with a range of studies that have established the beneficial effect of feedback on interviewer's performance in training and forensic practice (Benson &Powell, 2015; Cederborg et al., 2013; Lamb, Sternberg, Orbach, Esplin and Mitchell, 2002a; Yi et al., 2015), compared to the traditional classroom-based program (Aldridge & Cameron, 1999; Cederborg et al., 2000; Orbach et al.,2000; Warren et al., 1999). For example, Lamb, Sternberg, Orbach, Esplin and Mitchell (2002a) compared four different types of training; the results showed that only the groups that received continuous training and feedback improved significantly regarding the proportion of open-ended questions used during the interviews.

To date, all successful training programs have lasted at least a few days, resulting in proportions of recommended questions after training between 57 and 79% (Benson & Powell, 2015). Within one training session lasting only two and a half hours, we have been able to achieve the same level of effectiveness.

Contrary to our expectations, combining feedback with a simple reflection task did not enhance feedback effectiveness. No significant differences were found in the questioning style or in the proportion of correct conclusions between the standard feedback group and the group receiving the reflection task. The second hypothesis thus remains unsupported. It could be that reflection, in general, has no effect in this context or that the specific task failed to produce an effect. By asking participants to recall additional examples for question categories (or make up new examples if they did not remember any), the focus might have been put too much on the elaboration of question categories and less on critical examination of one's own performance and strategies, which is considered an important part of systematic reflection (Ellis et al., 2014).

Compared to the study by Anseel et al. (2009), from which our reflection intervention had been adapted, the feedback intervention used in the present study already included examples of successful

and unsuccessful behavior. The reflection intervention, asking for additional examples, might not have differed enough from feedback alone in order to stimulate deeper processing of the feedback message. A more useful way to implement reflection within the present training paradigm might be to review one's experiences and performance after a training session has ended (Ellis & Davidi, 2005). This might help the interviewer to remember the acquired questioning strategies in a follow-up training and actual investigative practice.

Limitations

Both experimenters conducted four pilot sessions together in order to achieve agreement upon question categories, and cases of disagreement were discussed at the end of each session. However, the evaluation of the inter-rater reliability was conducted only at the end of the study, showing anyway a good level of inter-rater agreement. In the current version of the simulation algorithm, no difference was made between subtypes of open-ended questions. The avatars' reaction pattern was the same regardless of whether the participant asked for specific information, further elaboration on a topic or changed topic completely. The next step could be to cluster avatars' memories around some macro-topics in the way used to provide additional neutral details in the current study. We asked our participants to conduct the interview as they preferred, and we asked them their first impression on the case. The use of those statements might have affected interviewers' questioning style and conclusions. Also, an obvious next step will be to test the effectiveness of this training with professional CSA investigators. Moreover, it would be interesting to test how long these effects will last and how frequently feedback should be provided.

Future development of the current training setup

The present study has shown EIT training setup to be effective in improving the quality of investigative interviews in a short period of time. However, a couple of issues deserve further attention. It is known, especially regarding younger children, that encouraging to elaborate further on topics already mentioned in the interview helps to elicit more complete accounts of events (Korkman, et al., 2008b; Lamb et al., 2008). This should be taken into account when developing the simulation for further use.

As a next step, it is important to investigate the transfer of the present study's impressive results into real-life interviewing situations. Professional CSA investigators should be trained within the same setup while pre- and post-training performance in real CSA interviews is assessed. Furthermore, the durability of training effects has to be demonstrated. If needed, follow-up sessions can be arranged easily using the EIT simulation. Building on the existing algorithm structure, more avatars can be developed in order to provide material for follow-up trainings. This is a preliminary training program, and we believe that we reached an acceptable level of realism, especially if com-pared to traditional training programs that include role-playing between two adults. The use of response algorithms and the possibility of providing detailed feedback make this training promising for future use in training CSA interviewers. Next, we can include speech recognition in our software; we can implement the NICHD protocol within the training in order to teach interviewers also how to, for example, build a rapport with the child before the interview. The use of serious gaming could be adapted not only to CSA interviews with professionals, but also to other situations in which interviews are conducted.

Conclusions

Performing a simple reflection task right after receiving feedback did not prove effective in enhancing the training effects in a group of students. However, the results of this study have confirmed previous findings showing that feedback on question types and conclusions is effective in improving interview's quality in a CSA investigative interview simulation with students. If these training effects can be shown to transfer onto professional interviewers' performance in real cases, the approach investigated here constitutes a major step ahead in developing cost-effective, fast and easily applicable training for child investigative interviewers.

References

Aldridge, J., & Cameron, S. (1999). Interviewing child witnesses: Questioning strategies and the effectiveness of training. *Applied Developmental Science*, *3*(2), 136–147. DOI:10.1207/s1532480xads0302_7

Anseel, F., Lievens, F., & Schollaert, E. (2009). Reflection as a strategy to enhance task performance after feedback. *Organizational Behavior and Human Decision Processes*, *110*(1), 23–35. DOI:10.1016/j.obhdp.2009.05.003

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

Bickeböller, H., & Clerget-Darpoux, F. (1995). Statistical properties of the allelic and genotypic transmission/disequilibrium test for multiallelic markers. *Genetic Epidemiology*, *12*(6), 865–870. DOI:10.1002/gepi.1370120656

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

Bruck, M., & Ceci, S. J. (1999). The suggestibility of children's memory. *Annual Review of Psychology*, *50*, 419–439. DOI:10.1146/annurev.psych.50.1.419

Ceci, S. J., & Bruck, M. (1993). Suggestibility of the child witness: A historical review and synthesis. *Psychological Bulletin*, *113*(3), 403–439. DOI:10.1037/0033-2909.113.3.403

Ceci, S. J., & Bruck, M. (1995). *Jeopardy in the courtroom: A scientific analysis of children's testimony*. Washington, DC, US: American Psychological Association. DOI:10.1037/10180-000

Cederborg, A.-C., & Lamb, M. E. (2008). Intensive training of forensic interviewers. In T. I. Richardsson, & M. V. Williams (Eds.), *Child abuse and violence*, (pp. 1–17). New York: Nova Science Publishers.

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

Cederborg, A.-C., Orbach, Y., Sternberg, K. J., & Lamb, M. E. (2000). Investigative interviews of child witnesses in Sweden. *Child Abuse and Neglect*, *24*(10), 1355–1361. DOI:10.1016/S0145-2134(00)00183-6

Chelune, G. J., Naugle, R. I., Lüders, H., Sedlak, J., & Awad, I. A. (1993). Individual change after epilepsy surgery: Practice effects and base-rate information. *Neuropsychology*, *7*(1), 41–52. DOI:10.1037/0894-4105.7.1.41

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*(1), 37–46. DOI:10.1177/001316446002000104

Criminal Justice Joint Inspection. (2014). Achieving best evidence in child sexual abuse case—A joint inspection. Retrieved from: https://www.justiceinspectorates.gov.uk/cjji/wp-content/uploads/sites/2/2014/12/CJJI_ABE_Dec14_rpt.pdf

Di Eugenio, B., & Glass, M. (2004). The kappa statistic: A second look. *Computational Linguistics*, *30*(1), 95–101. DOI:10.1162/089120104773633402.EIT (Version 1.12.7) [Computer software].

Åbo, Finland: Åbo Akademi University. Elliott, D. M., & Briere, J. (1994). Forensic sexual abuse evaluations of older children: Disclosures and symptomatology. *Behavioral Sciences & the Law*, *12*(3), 261–277. DOI:10.1002/bsl.2370120306

Ellis, S., & Davidi, I. (2005). After-event reviews: Drawing lessons from successful and failed experience. *The Journal of Applied Psychology*, *90*(5), 857–871. DOI:10.1037/0021-9010.90.5.857.

Ellis, S., Carette, B., Anseel, F., & Lievens, F. (2014). Systematic reflection: Implications for learning from failures and successes. *Current Directions in Psychological Science*, *23*(1), 67–72. DOI:10.1177/0963721413504106

Espinet, S. D., Anderson, J. E., & Zelazo, P. D. (2013). Reflection training improves executive function in preschool-age children: Behavioral and neural effects. *Developmental Cognitive Neuroscience*, *4*, 3–15.

Feinstein, A. R., & Cicchetti, D. V. (1990). High agreement but low kappa: I. The problems of two paradoxes. *Journal of Clinical Epidemiology*, *43*(6), 543–549. DOI:10.1016/0895-4356(90)90158-L.

Friedman, W. J., & Lyon, T. D. (2005). Development of temporal-reconstructive abilities. *Child Development*, *76*(6), 1202–1216.

Girden, E. R. (1992). ANOVA: Repeated measures (No. 84). Sage.

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

Gwet, K. (2002). Inter-rater reliability: Dependency on trait prevalence and marginal homogeneity. *Statistical Methods for Inter-Rater Reliability Assessment Series*,*2*,1–9.

Gwet, L. (2008). Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology*,*61*, 29–48. DOI:10.1348/000711006X126600

Gwet, L. (2010). *The handbook of inter-rater reliability*. Gaithersburg: Advanced Analytics.

Herman, S. (2009). Forensic child sexual abuse evaluations: Accuracy, ethics, and admissibility. In K. Kuehnle & M. Connell (Eds.),*The evaluation of child sexual abuse allegations: A comprehensive guide to assessment and testimony* (pp. 247–266). Hoboken, NJ, US: Wiley. doi:10.13140/2.1.3235.4082

Jarmasz, J., & Hollands, J. G. (2009). Confidence intervals in repeated-measures designs: The number of observations principle. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, *63*(2), 124. DOI:10.1037/a0014164

Johnson, M., Magnussen, S., Thoresen, C., Lønnum, K., Burrell, L. V., & 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. *Applied Cognitive Psychology*, *29*(5), 661-668. DOI:10.1002/acp.3147

Korkman, J., Santtila, P., Drzewiecki, T., & Sandnabba, N. K. (2008a). Failing to keep it simple: Language use in child sexual abuse interviews with 3–8-year-old children. *Psychology, Crime & Law,14* (1), 41–60. DOI:10.1080/10683160701368438

Korkman, J., Santtila, P., & Sandnabba, N. K. (2006). Dynamics of verbal interaction between interviewer and child in interviews with alleged victims of child sexual abuse. *Scandinavian Journal of Psychology, 47*, 109–119. DOI:10.1111/j.1467-9450.2006.00498.x

Korkman, J., Santtila, P., Westeråker, M., & Sandnabba, N. K. (2008b).Interviewing techniques and follow-up questions in child sexual abuse interviews. *The European Journal of Developmental Psychology, 5*,108–128. DOI:10.1080/17405620701210460

Lamb, M. E., Hershkowitz, I., Orbach, Y., & Esplin, P. W. (2008).*Tell me what happened*. Chichester, UK: John Wiley & Sons, Ltd.

Lamb, M. E., La Rooy, D. J., Malloy, L. C., & Katz, C. (2011). *Children's testimony: A handbook of psychological research and forensic practice.* (M. E. Lamb, D. J. La Rooy, L. C. Malloy, & C. Katz, Eds.) (2nd ed.).Chichester, UK: John Wiley & Sons, Ltd. DOI:10.1002/9781119998495

Lamb, M. E., Sternberg, K. J., & Esplin, P. W. (1998). Conducting investigative interviews of alleged sexual abuse victims. *Child Abuse and Neglect*, *22*(8), 813–823. DOI:10.1016/S0145-2134(98)00056-8

Lamb, M. E., Sternberg, K. J., Orbach, Y., Esplin, P. W., & Mitchell, S. (2002a). Is ongoing feedback necessary to maintain the quality of investigative interviews with allegedly abused children? *Applied Developmental Science*, *6*(1), 35–41.DOI:10.1207/S1532480XADS0601_04

Lamb, M. E., Sternberg, K. J., Orbach, Y., Esplin, P. W., Stewart, H., & Mitchell, S. (2003). Age differences in young children's responses to open-ended invitations in the course of forensic interviews. *Journal of Consulting and Clinical Psychology,71*(5), 926–934. DOI:10.1037/0022-006X.71.5.926

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

Lyon, T. D. (2014). Interviewing children. *Annual Review of Law and Social Science*, *10*(1), 73–89. DOI:10.1146/annurev-lawsocsci-110413-030913

Matthew, C. T., & Sternberg, R. J. (2009). Developing experience-based (tacit) knowledge through reflection. *Learning and Individual Differences*,*19*(4), 530–540.

Maxwell, A. E. (1970). Comparing the classification of subjects by two independent judges. *British Journal of Psychiatry*,*116*, 651–655. DOI:10.1192/bjp.116.535.651

Olszewski, A. E. (2016). Virtual simulation and serious games for medical education: A review of the literature and development of a virtual peritoneal dialysis simulator. (Doctoral dissertation). Retrieved from Harvard University's DASH repository. http://nrs.harvard.edu/urn-3:HUL.InstRepos:27007751

Orbach, Y., Hershkowitz, I., Lamb, M. E., Sternberg, K. J., Esplin, P. W., & Horowitz, D. (2000). Assessing the value of structured protocols for forensic interviews of alleged child abuse victims. *Child Abuse andNeglect, 24*(6), 733–752. DOI:10.1016/S0145-2134(00)00137-X

Parsons, T. D., Notebaert, A. J., Shields, E. W., & Guskiewicz, K. M.(2009). Application of reliable change indices to computerized neuropsy-chological measures of concussion. *International Journal of Neuroscience, 119*(4), 492–507. DOI:10.1080/00207450802330876

Pompedda, F., Zappalà, A., & Santtila, P. (2015). Simulations of child sexual abuse interviews using avatars paired with feedback improves interview quality. *Psychology, Crime & Law*, *21*(1), 28–52.

Pons, F., Harris, P. L., & de Rosnay, M. (2004). Emotion comprehension between 3 and 11 years: Developmental periods and hierarchical organization. *The European Journal of Developmental Psychology, 1*(2),127–152. DOI:10.1080/17405620344000022

Pons, F., Lawson, J., Harris, P. L., & de Rosnay, M. (2003). Individual differences in children's emotion understanding: Effects of age and language. *Scandinavian Journal of Psychology, 44*(4), 347–353. DOI:10.1111/1467-9450.00354

Poole, D. A., & Lamb, M. E. (1998). *Investigative interviews of children: A guide for helping professionals*. Washington, DC: American Psychological Association. DOI:10.1037/10301-000

Ritterfeld, U., Cody, M., & Vorderer, P. (Eds) (2009). *Serious games: Mechanisms and effects*. New York, NY: Routledge.

Rocha, E. M., Marche, T. A., & Briere, J. L. (2013). The effect of forced-choice questions on children's suggestibility: A comparison of multiple-choice and yes/no questions. *Canadian Journal of Behavioural Science, 45*(1), 1–11. DOI:10.1037/a0028507

Ron, N., Lipshitz, R., & Popper, M. (2006). How organizations learn: Post-flight reviews in an F-16 fighter squadron. *Organization Studies, 27*(8),1069–1089.

Seibert, K. W. (1999). Reflection-in-action: Tools for cultivating on-the-job learning conditions. *Organizational Dynamics,27* (3), 54–65.DOI:10.1016/S0090-2616(99)90021-9

Sitepal [Computer software]. (2004). New York, NY: Oddcast.

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

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

Sternberg, K. J., Lamb, M. E., Hershkowitz, I., Esplin, P. W., Redlich, A., & Sunshine, N. (1996). The relation between investigative utterance types and the informativeness of child witnesses. *Journal of Applied Developmental Psychology,17*(3), 439–451. DOI:10.1016/S0193-3973(96)90036-2

Stuart, A. A. (1955). A test for homogeneity of the marginal distributions in a two-way classification. *Biometrika, 42*, 412–416.

Van Dijk, T., Spil, T., van der Burg, S., Wenzler, I., & Dalmolen, S. (2015). Present or play: The effect of serious gaming on demonstrated behaviour. *International Journal of Game-Based Learning (IJGBL), 5*(2), 55–69.DOI:10.4018/ijgbl.2015040104

Vrij, A. (2005). Criteria-based content analysis: A qualitative review of the first 37 studies. *Psychology, Public Policy, and Law, 11*(1), 3. DOI:10.1037/1076-8971.11.1.3

Wandrey, L., Lyon, T. D., Quas, J. A., & Friedman, W. J. (2012). Maltreated children's ability to estimate temporal location and numerosity of placement changes and court visits. *Psychology, Public Policy, and Law, 18*(1), 79–104. DOI:10.1037/a0024812

Warren, A. R., Woodall, C. E., Thomas, M., Nunno, M., Keeney, J., Larson, S., & Stadfeld, J. (1999). Assessing the effectiveness of a training program for interviewing child witnesses. *Applied Developmental Science, 3*(2), 128–135. DOI:10.1207/s1532480xads0302_6

Waterman, A. H., Blades, M., & Spencer, C. (2000). Do children try to answer nonsensical questions? *British Journal of Developmental Psychology, 18*(2), 211–225. DOI:10.1348/026151000165652

Wongpakaran, N., Wongpakaran, T., Wedding, D., & Gwet, K. L. (2013). A comparison of Cohen's Kappa and Gwet's AC1 when calculating inter-rater reliability coefficients: A study

conducted with personality disorder samples. *BMC Medical Research Methodology, 13*(1), 1.DOI:10.1186/1471-2288-13-61

Wouters, P., Van Nimwegen, C., Van Oostendorp, H., & Van Der Spek, E.D. (2013). A meta-analysis of the cognitive and motivational effects of serious games. *Journal of Educational Psychology, 105*(2), 249.

Yi, M., Jo, E., & Lamb, M. E. (2015). Effects of the NICHD protocol training on child investigative interview quality in Korean police officers. *Journal of Police and Criminal Psychology*, Advance Online Publication. DOI: 10.1007/s11896-015-9170-9.

Appendix A

Appendix 1: Instructions about best practice interviewing

Guidelines for the correct questioning style

The behavior of children in interviews is considerably different from that of adults. Especially, young children are vulnerable to suggestion: They tend to perceive any information told by adults as true and do not negate wrong statements by the interviewer.

Because of that, there are established recommendations on which question types to use when interviewing children: In general, open-ended question types are recommended. This is because they most likely elicit a reliable answer. Open-ended questions can be:

Invitations. Open-ended utterances (questions, statements or imperatives) used to elicit free recall responses from the child. Invitations could be general ('Tell me everything that happened from the very beginning to the end') or relate to something just mentioned by the child ('Tell me more about that').

Facilitators. Non-suggestive encouragements to continue with a response. These include utterances like'ok,' restatements (echoing) of the child's previous utterance.

Directive utterances. These refocus the child's attention on details already mentioned by the child and request further elaboration (for example, 'Where were you when that happened?').On the other hand, closed and suggestive questions should be avoided, as the answers elicited by them are much less reliable and they might create false memories in children.

Closed and suggestive questions include:

Option-posing utterances. These focus the child's attention on issues that the child had not previously mentioned but do not imply that a particular response is expected. The answer to these types of questions is usually 'yes or no'. For example, the interviewer might ask 'Did he touch your penis?' or 'Did he do anything with his penis?'

Suggestive utterances. These are stated in such a way that the interviewer strongly communicated what response was expected (for example: 'He forced you to do that, didn't he?'), or assumed details that had not been revealed by the child (for example: Child: 'We laid on the sofa.' Interviewer: 'He laid on you or you laid on him?').

Questions before the interview: If the child doesn't provide any detail regarding the alleged situations of abuse, the interviewer should ask to the child questions related to the alleged situation. For example: Did your father touch you?

Yes	No

If the child provides a detail regarding the alleged situation, for example ‘he punched me’ which is the best question to ask?

1. Did it hurt?
2. Who punched you?
3. Was it your father?

Appendix 2: the story of Fabiana
Fabiana, 6 years old

She lives with her grandparents (she is orphan). She started to go to school this year and seems to have a good relationship with her classmates and teachers. Her grandparents Helen (70 years) and Albert (78 years) are always present during her activities and support her closely.

The case

The social worker that follows the child has reported to the police because the child said that she is afraid of her grandfather. No bruises or other visible injuries were found on the child.

Table A1. Post-hoc repeated measures ANOVAs on the percentage of recommended questions

Groups	Effect	df_{Effect}	df_{Error}	F	p	η^2_p	$1-\beta$
Control versus Feedback	Group	1	36	29.75	<.001	.45	>.99
	Time	4.49	161.78	34.48	<.001	.49	>.99
	Group x Time	4.49	161.78	23.02	<.001	.39	>.99
Control versus Feedback plus Reflection	Group	1	38	32.81	<.001	.46	>.99
	Time	3.76	142.90	22.97	<.001	.38	>.99
	Group x Time	3.76	142.90	13.56	<.001	.26	>.99
Feedback versus Feedback plus Reflection	Time	3.48	132.12	81.33	<.001	.68	>.99

Note: The three planned comparisons are group (2) by time (8) repeated measure ANOVAs.

Table A2. Post-hoc repeated ANOVAs on the number of relevant details

Groups	Effect	df_{Effect}	df_{Error}	F	p	η^2_p	$1-\beta$
Control versus Feedback	Group	1	36	5.60	.023	.04	>.63
	Time	7	252	11.55	<.001	.24	>.99
	Group x Time	7	252	5.27	<.001	.13	>.99
Control versus Feedback plus Reflection	Group	1	38	9.13	.004	.19	>.83
	Time	4.89	185.77	8.83	<.001	.19	>.99
	Group x Time	4.89	185.77	2.57	.029	.26	>.78
Feedback versus Feedback plus Reflection	Time	5.33	202.37	22.94	<.001	.38	>.99

Note: The three planned comparisons are group (2) by time (8) repeated measure ANOVAs.
No Greenhouse-Geiser correction needed for the pair comparison Control versus Feedback.

Table A3. Post-hoc repeated measures ANOVAs on the number of neutral details

Groups	Effect	df_{Effect}	df_{Error}	F	p	η^2_p	$1-\beta$
Control versus Feedback	Group	1	36	4.89	.034	.12	>.58
	Time	7	252	9.37	<.001	.21	>.99
Control versus Feedback plus Reflection	Group	1	38	8.96	.005	.19	>.83
	Time	7	266	8.66	<.001	.19	>.99
Feedback versus Feedback plus Reflection	Time	5.19	197.39	15.37	<.001	.29	>.99

Note: The three planned comparisons are group (2) by time (8) repeated measure ANOVAs.
No Greenhouse-Geiser correction needed for the pair comparison Control versus Feedback and Control versus Feedback plus Reflection.

Table A4. Post-hoc repeated measures ANOVAs on the number of wrong details

Groups	Effect	df_{Effect}	df_{Error}	F	p	η^2_p	$1-\beta$
Control versus Feedback	Group	1	27	4.81	.037	.15	>.56
	Group x Time	3.96	106.86	4.09	.004	.13	>.90
Control versus Feedback plus Reflection	Group	1	28	9.62	.004	.26	>.85
	Group x Time	4.22	118.16	2.50	.043	.13	>.71
Feedback versus Feedback plus Reflection	Time	3	80.99	5.09	.003	.16	>.90