Transfer of Simulated Interview Training Effects into Interviews with Children Exposed to a Mock Event

Francesco Pompedda[ab*] , Annegrete Palu[c] , Kristjan Kask[d] 4, Karolyn Schiff[d] 4, Anna Soveri[e] , Jan Antfolk[be] , Pekka Santtila[f]

[a]*School of Natural & Social Sciences, University of Gloucestershire, Cheltenham, UK*

[b]*Faculty of Arts, Psychology, and Theology, Åbo Akademi University, Turku, Finland*

[c]*Institute of Psychology, University of Tartu, Tartu, Estonia*

[d]*School of Natural Sciences and Health, Tallinn University, Tallinn, Estonia*

[e]*Turku Brain and Mind Center, Turku, Finland*

[f]*Faculty of Arts and Sciences, NYU Shanghai, Shanghai, China*

*** Correspondence:**

Francesco Pompedda, fpompedda@glos.ac.uk

"This is an Accepted Manuscript of an article published by Taylor & Francis in Nordic Psychology on 07/07/2020, available online:
http://dx.doi.org/10.1080/19012276.2020.1788417

Transfer of Simulated Interview Training Effects into Interviews with Children Exposed to a Mock Event

Research on students suggests that repeated feedback in simulated investigative interviews with avatars (computerized children) improves the quality of the interviews conducted in this simulated environment. It remains unclear whether also professional groups (psychologists) benefit from the training and if the effects obtained in the simulated interviews transfer into interviews with real children who have witnessed a mock event. We trained 40 psychologists (Study I) and 69 psychology students (Study II). In both studies, half of the participants received no feedback (control group) while the other half received feedback (experimental group) on their performance during repeated interviews with avatars. Each participant then interviewed two 4-6-year-old children who had each witnessed a different mock event without any feedback being provided. In both studies, interview quality improved in the feedback (vs. control) group during the training session with avatars. The analyses of transfer effects showed that, compared to controls, interview quality was better in the experimental group. More recommended questions were used in both studies, and more correct details were elicited from the children in Study I, during the interviews each participant conducted with two children (N = 76 in Study I; N = 116 in Study II) one week after the training. Although the two studies did not show statistically significant training effects for all investigated variables, we conclude that interview quality can be improved using avatar training and that there is transfer into actual interviews with children at least in the use of recommended questions.

Keywords: child sexual abuse (CSA), serious gaming, training with virtual reality, interview training, investigative interviewing, avatar

Introduction

Interviewing children in alleged child sexual abuse (CSA) cases is a complex task (Powell & Wright, 2008) that requires specialized training (Benson & Powell, 2015). Research has shown that there is a gap between theoretical knowledge of best-practice guidelines and the ability to follow these guidelines in real-life interviews (Sternberg, Lamb, Davies, & Westcott, 2001), and that this problem persists even after intensive training (Johnson et al., 2015).

Best-practice guidelines instruct interviewers to primarily use open-ended questions (hereafter, recommended questions) during an interview (Lyon, 2014), as they increase the reliability of the details elicited from children, and avoid closed and suggestive questions (hereafter, non-recommended questions) as they decrease the probability of eliciting reliable details (Lamb et al., 2018). In addition, multiple-choice questions, and questions concerning complex cognitive domains such as time (e.g., Wandrey, Lyon, Quas, & Friedman, 2012; Tillman et al., 2017) should be limited when interviewing young children (see Lamb et al., 2018, for a general review).

The use of structured protocols, such as the National Institute of Child Health and Human Development (NICHD) protocol (NICHD, 2011), and feedback during training (Powell, Fisher, & Hughes-Scholes, 2008) can increase the use of recommended questions. Recent studies (e.g., Powell, Guadagno, & Benson, 2016) have, however, shown that the overall proportion of open-ended questions after training is still low and, even if there are improvements, the effects of training usually disappear shortly after the training ends (Price & Roberts, 2011). It seems that multiple practice occasions and continuous (e.g., Lamb, Sternberg, & Orbach, 2002), immediate, and detailed (Smith, 2008) feedback are required to increase and maintain the use of recommended questions.

Besides in the context of investigative interviewing (e.g., Benson & Powell, 2015), studies in other contexts confirm the positive effect of feedback in increasing learning (e.g., Hattie &

Timperley, 2007). It is also clear that the type of feedback provided has an impact on learning, with procedural feedback showing larger effect sizes than other types of feedback (Van der Kleij et al., 2015). In a meta-analysis by Hatala and colleagues (2014), the authors highlighted how providing multiple sources of feedback compared to a single source elicited better learning results.

Currently, the most common type of training includes intensive theoretical face-to-face lectures and role-playing with an adult pretending to be a child or interviews with real children followed by feedback. In role-playing, the actors should provide responses reflecting the response patterns of actual children. However, actors tend to overestimate the number and frequency of details provided by children (Powell, Fisher, & Hughes-Scholes, 2008). On the other hand, the lack of detailed feedback is a problem in training based on interviews in real cases. With rare exceptions, it is impossible to know with certainty whether a detail provided by a child is factual or not, making it impossible to give feedback on the ability of the interviewers to elicit accurate information. Finally, mistakes in interviews in real cases can have serious consequences. Besides, such interview training programs are expensive, logistically challenging, and time-consuming making them difficult to implement (Powell, Guadagno, & Benson, 2016). Therefore, the successful development of feasible, cost-effective, and efficient training paradigms is of paramount importance. Serious gaming combined with feedback potentially provides such a solution (Benson & Powell, 2015; Pompedda et al., 2015, Powell et al., 2016).

Serious gaming can be defined as any game played within a safe environment with the aim of teaching and learning of especially complex, practical skills such as piloting a plane or conducting surgery (e.g., Wouters, van Nimwegen, van Oostendorp, & van der Spek, 2013). In the present studies, participants interviewed computerized avatars with whom they were able to interact as if they were interviewing real children. Given the particular cognitive demands of investigative interviewing, based on Cognitive Load Theory, the serious game we developed is ideal for increasing transfer effects, defined as the extent to which practice in one task affects

performance in another task (e.g., Blume, Ford, Baldwin, & Huang, 2010), in our case from avatar training to interviews with children.

According to Cognitive Load Theory (Ayres & Paas, 2009) permanent learning and transfer can be maximized by minimizing extraneous cognitive load while devoting the available cognitive resources in favour of the cognitive load (CL) related to learning, named germane cognitive load (Mugford, Corey, & Bennell, 2013). Effective learning in complex tasks requires the creation of schemata and their automation (Paas, Renkl, & Sweller, 2003). Schemata contain different pieces of information, but they are treated as a single unit in working memory. Creating schemata, given the limited capacity of the working memory, allows the processing of more information simultaneously. Bearing in mind the complexity of investigative interviewing, the avatar training was planned with guidance from the Cognitive Load Theory (see Ayres & Paas, 2009):

We increased the complexity of the task using mixed practice, which, even if it can reduce performance in the immediate future, is expected to promote better transfer (for a review see Helsdingen, van Gog, & van Merriënboer, 2011). Mixed practice was achieved by i) the avatars having different memory contents (abuse vs no abuse scenarios), ii) the avatars having probabilistic algorithms resulting in highly varied response patterns between the interviews, and iii) the avatars being presented in a randomized order in terms of age and abuse status. Also, we minimized passive learning (theoretical frontal lectures) and maximized active learning (practice) with the aim of enhancing schemata automatization (Sweller, van Merrienboer, & Paas, 1998). Instead of frontal lectures, the participants learned about the harmful effects of suggestive questions through their questioning and the feedback provided. To the best of our knowledge, this is the first interview training to apply minimal theoretical training. Further, transfer is more likely in conditions bearing a close resemblance to the training situation, both in terms of structural and surface similarity (e.g., Soveri et al., 2017). For this reason, the training simulation in our studies has been designed so that it would be similar to real interviews, both when it comes to surface

(i.e., using avatars that look and talk like children) and structural (the avatars have response algorithms that mimic those of a child of a specific age) features.

The present study

In previous experiments with students (Pompedda et al., 2015; Pompedda et al., 2017; Krause et al., 2017; Pompedda, 2018), interview quality in simulated investigative interviews with avatars was considerably improved in just one hour when the interviewers were given feedback on their performance after each interview. Importantly, there were improvements in all measures of interview quality used, and these were achieved without extensive theoretical instructions. However, two crucial questions remain unanswered. First, will similar improvements also be seen in a group of professionals, such as psychologists, and, secondly, will the improvements in interview quality, associated with the feedback within the avatar training, transfer into interviews with real children.

The main aims of the present studies were to replicate the effects of feedback manipulation in a group of psychologists and to test if the acquired skills transfer into interviews with actual children who witnessed a mock event, following previous studies (e.g., Lyon et al., 2014)

Study II was a close replication of Study I with minimal differences (e.g., the exclusion of the secret). All the differences between the two studies are shown in Appendix A.

Given the lack of previous experience in interviewing children in both samples (psychologists and students) we expected no differences between the groups after the training. Two hypotheses were formulated for both studies:

- Effect of feedback manipulation in simulated interviews with avatars: We expected the group that received feedback (vs the group that did not receive feedback) to use a higher proportion of recommended questions, elicit a higher number of correct and a

lower number of incorrect details, and to reach more correct conclusions during the interviews with the avatars.

- Transfer of the effect of the feedback manipulation in avatar interviews into interviews with real children who had witnessed a mock event: We expected the improvements achieved by the group that received feedback during avatar interviews (vs the group that did not receive feedback) to transfer into interviews with children who had witnessed a mock event. The interviewers who received feedback during the training with avatars were expected to use a higher proportion of recommended questions, elicit a higher proportion of correct details, compared to incorrect details, and reach more correct conclusions in interviews with children who had witnessed a mock event.

Materials and Methods

Participants

Study I

Forty psychologists who were randomly recruited in Italy using a social media advertisement (thirty-seven women and three men) participated (mean age 27 years) as interviewers. In Italy, you can be regarded as a psychologist only after passing a national examination and five years of studies. All psychologists present in the sample have passed the licensing exam before the experiment. Six participants have been working in alleged CSA cases, however none of them has neither conducted, nor received training in investigative interviews of children. Children were recruited from two different kindergartens (aged 4 to 5 years) and two different primary schools (aged 6 to 7 years) in Italy. 97 children were recruited to ensure that enough children would be present at school during the day of the interview. Of the 97 children, 76 were interviewed. Children were randomly divided into two groups. The first group participated in a mock event called “the pirate game” ($n_{kindergarten}$ = 16, mean age 56 months; n_{school} = 22, mean age 84 months),

and the second group participated in a mock event called "the paw patrol game" ($n_{kindergarten}$ = 16, mean age 55 months; n_{school} = 22, mean age 85 months). The interviewers received 50 euros for their participation.

Study II

Sixty-nine psychology students recruited in Estonia (forty-seven women and twenty-two men) participated (mean age 23 years) as interviewers. Children were recruited from eight different kindergartens in two major cities in Estonia, and, of the 197 recruited children 126 were interviewed. Children were randomly divided into two groups. The first group participated in a mock event called "the pirate game" (n = 100, mean age 70 months), and the second group participated in a mock event called "the clown game" (n = 97, mean age 70 months). The interviewers did not receive any reward for their participation.

Ethical approval

The board of research ethics at Åbo Akademi University for Study I, and Tallinn Medical Research Ethics Committee for Study II, approved the studies before the data collections commenced. All participants gave written informed consent and written informed parental consent was obtained from parents for all participating children in accordance with the Declaration of Helsinki.

Design

Participants in both studies were randomly divided into two groups (between-subject factor): an active control that did not receive feedback during avatar training ($n_{\text{study I}}$ = 20; $n_{\text{study II}}$ = 34) and a feedback group ($n_{\text{study I}}$ = 20; $n_{\text{study II}}$ = 35). Each participant conducted six avatar interviews (within subject factor) within a single training session. Each participant subsequently interviewed two children each of whom had participated in a different mock event. Two participants from the control group dropped out from the study after the training session (Study I), and five participants have been removed due to missing data in one or more interviews from Study II. The procedure

was identical for all participants apart from whether feedback was provided after each interview with the avatars (no feedback was provided during the interviews with the children).

Materials

Simulated investigative interviews using Empowering Interviewer Training

We performed simulated interviews with avatars using the same procedure as in previous studies (see Pompedda, 2018, for a detailed description). An operator listened to each question asked by the interviewer and categorized it (e.g., as option-posing), which will elicit a response (a video clip) based on the algorithms.

The training tool (Empowering Interviewers Training) consists of 16 different avatars (computerized children) with different scenarios to be investigated. Half of the avatars are emotional (i.e., crying) and half are neutral. Half of the stories include abuse while the other half do not, and half of the avatars are four years old while the other half are six years old. The avatars have predefined memories and details of the alleged CSA scenarios. The avatars respond to the interviewers' questions providing predefined details through probabilistic response algorithms that are derived from studies on children's responses to different question types. These algorithms increase the ecological validity compared to other training tools as they provide the interviewers with realistic response patterns (for an example of the algorithms see Pompedda, 2018). The realistic response pattern embedded in the avatars is also assumed to stimulate interviewers' immediate problem solving and emotion regulation abilities. For example, interviewers learn how to switch topic if the child is resistant or how to cope with the frustration after a series of irrelevant responses

Mock events

Study I

Two research assistants staged two different mock events in the schools of the children (the pirate game and the paw patrol game). The mock events were based on previous mock events presented in Roberts, Lamb, and Sternberg (1999). Events were constructed to include active involvement of children to increase ecological validity. Moreover, we included actions with higher forensic relevance as presented in Roberts et al. (1999). For example, the events included dressing and undressing moments, innocuous touching between both adult/child and child/child pairs, a secret, and the insertion of a cookie into the mouth. These activities were used successfully in previous studies; for example, dressing up (Roberts, Lamb, & Sternberg, 2004), innocuous touching (Davis & Bottoms, 2002), offering food (Finnilä et al., 2003), and a secret (Roberts et al., 1999). What was new in these two mock events was the direct insertion of food (here, the cookie was inserted into the child's mouth). Each of the mock events lasted about eight minutes and was videotaped. The structure of the two events was similar with some differences concerning the main character and some actions.

Study II

Based on the results in Study I, where relatively few children talked about the target event, we decided to make some changes to the mock events. We did not mention secrecy since it was one of the main reasons why children did not disclose in Study I, and decided to emphasize the name of the "character" to help children understand what the aim of the interview was without being suggestive. Additional secondary details were also changed (e.g., using a clown nose instead of a mask). Except for these differences, the mock events employed the same plots in both studies.

Procedure

Training with avatars

For each participant, six out of sixteen possible avatars were randomly selected. The selection of the avatars was randomized for the first four avatars, and the last two avatars were selected

among the remaining options to include all possible combinations of age (four/six years old) and scenarios (abuse/no-abuse) for each interviewer. The order of the obtained sequence was then randomized. Hence, each participant received three abuse and three no abuse situations. Because it was impossible to balance age by abuse situation across the six training interviews per participant (within six avatars), this was instead balanced within groups.

All participants signed informed consent and confidentiality agreements. Subsequently, they received instructions about best practices in child interviewing, and were given oral instructions on the different phases of the study and a paper with the background story of the first avatar they interviewed. Each interview was recorded and lasted a maximum of ten minutes (the total training lasted between 1 and 1.5 hours) and, at the end of each interview, participants were asked to explain what happened with as many details as they could, while being informed that "I do not have enough information to draw a conclusion" was an acceptable answer. After each interview, participants in the feedback group received a combination of outcome feedback (information about what had actually "happened" to the avatar) and process feedback, which consisted of an example of four questions (two recommended and two non-recommended questions) they used during the interview (e.g., "Do you play with daddy?"), with related description of the question type category and rationale for using it or otherwise. Feedback was provided in a way that covered as many question types as possible, prioritizing new question types used by the interviewer while keeping four examples as the maximum limit.

Mock events

Study I. Two research assistants went to the children's schools a week before the interviews and staged two different mock events (pirate game and paw patrol game). Children who were not present during the designated day were excluded from the study. The mock events were staged in different rooms in the schools, with 3 to 7 children participating in each mock event. A total of sixteen mock events were staged. The events were balanced for child age (younger-older), mock

event (type 1- type 2), and main actor (1-2). For each age group, eight mock events were staged (four of type 1 and four of type 2) with each research assistant performing two mock events per type. The mock events, which lasted eight minutes on average, were video recorded allowing us to consider any differences between the actual mock event and the script

Study II. Two pairs of research assistants, one pair in each city, went to the children's kindergartens a week before the interviews and staged two different mock events (pirate game and clown game). The number of children in each event varied from two to nine. Thirty-six mock events were staged. The events were balanced for mock event type (type 1- type 2), and children were divided into mock events so that there would be an equal number of younger and older children taking part in both types of events. The mock events were video recorded and lasted on average eight minutes.

Interviews with children

At the end of the training session with the avatars described before, all participants were provided with advice on how to conduct interviews with children, and instructions extracted from the NICHD protocol (National Institute of Child Health and Human Development, 2011), to help them create rapport with children. The participants were requested to read the instructions and bring them to the interviews with the real children. One week after (7±2 days) the training with avatars, each participant interviewed two children each of whom had witnessed one of the two mock events staged the week before (7±1 days). The order of the investigated mock events was balanced. Participants were informed that they were not allowed to investigate any personal information nor possible abuse experiences of the child, but their task was to find out as many details as possible about the mock event. Subsequently, participants were provided with preliminary information about the event the child had witnessed, and they were informed that some pieces of this information were true while some others were untrue. For both mock events,

this preliminary information contained two true elements, two that were untrue, and two that were inaccurate (a part was true and a part was untrue).

Rapport building. To help participants in building rapport with the child before the interview, we provided them with a protocol for interview and rapport building adapted from the NICHD protocol. We decided to have a pre-fixed maximum time (eight minutes) for this pre-substantive phase based on previous studies (e.g., Roberts et al., 2004, Teoh & Lamb 2010). The rapport building phase lasted on average between three and four minutes in both studies (Study I, *M* = 3.52, SD = 1.69; Study II, *M* = 3.65, SD = 1.26).

Main interview. After the rapport-building phase, the second part (free interview) lasted a maximum of 22 minutes, and after an initial standard question, the interviewer was free to conduct the interview. The free interview lasted on average nine minutes in Study I (*M* = 9.34, SD = 7.04), and five minutes in Study II (*M* = 5.17, SD = 3.01).

At the end of each interview with the child, children were provided with a questionnaire and a small toy as a reward. The questionnaire consisted of three questions “Did you like to play (name of the game)?” “Did you like to talk with the interviewer?” and “Did you say everything you know about the event to the interviewer?” The children answered the questions by crossing one of the two emoticons presented, a smiling one or a sad one.

Subsequently, the assistant asked the interviewer to explain what happened during the mock event with a standard question: “Describe in the most detailed way possible the event the child witnessed”. After this part, the researcher instructed all the participants (regardless of the experimental group they belonged to) to think about the question types they used during avatar interviews before interviewing the second child. In interviews with children we did not provide any type of feedback.

Coding

Both the interviews with avatars and the interviews with the children were coded for question 3types and for details elicited from the avatar/child. Question-type coding and the accuracy of details was based on a scheme used in previous studies (e.g., Pompedda et al., 2015) for interviews with avatars. In interviews with avatars, there is a maximum of nine correct details for each avatar. One detail was defined as a narrative phrase that the avatar provides in response to a recommended question (e.g., I was alone with daddy). Incorrect details were evaluated in the same way. As the interviewer's questioning style may create a varying number of incorrect details, there was no maximum number.

For coding the interviews with children, we used a method previously used in interviews with children by Roberts et al. (1999) (see Table 1). Only details that referred to the mock event were used in the analyses. If the nature of a detail was not possible to be determined with certainty, it was not included in the analyses. Details (narrative and yes/no answers) were divided into pieces of information, and the veracity of each piece of information was evaluated. Based on previous studies (Roberts et al., 1999), a piece of information was evaluated as any new logical element of the phrase (subject, verb, object, adjectives) for which veracity could be evaluated.

While avatars provided only correct information when presenting a narrative statement, children can introduce incorrect elements within the same narrative. For this reason, we evaluated the proportion of correct details (out of the total number of correct and incorrect details retrieved) in interviews with children. Correct information in response to the first fixed question (which is not related to the interviewer's skill) and correct information received as a rejection of a suggestive question (they do not reflect interviewer skills, but children’s proneness or otherwise to suggestibility), were not included as correct details.

Results

Due to the violation of normality and correlated data, we tested hypotheses for the avatar interviews using Generalized Estimating Equations (Group * Time) and Mann-Whitney We tested

the hypotheses regarding the transfer of effects into interviews with real children using multilevel modelling as the data did not violate the assumption of normality. Multilevel modelling was necessary because the two interviews with children were nested within the same participant. Multi-level analyses were conducted using the *lme4* package (Bates, Mächler, Bolker, & Walker, 2015) and the *lmerTest* package was used for computing *p*-values and model fit indices (Kunetsova, Brockhoff & Christensen, 2016). Linear mixed-effects models, fitted using a Maximum Likelihood procedure and with Satterthwaite approximation for degrees of freedom, was used for continuous variables (proportion of recommended questions and correct details), while a binomial, logit link, generalized linear mixed model was used for dichotomous variables (correct conclusions). We modelled interviewers as the random factor and experimental group as a fixed factor, while both Bayesian Information Criterion (BIC) and Akaike Information Criterion (AIC) were used for model selection (Lewis, Butler, & Gilbert, 2011). The model explaining the largest amount of variance was the one we used in our analyses.

We included only the interviews where the child provided at least one detail about the target event (Study I, n = 35; Study II, n = 103) in the analyses of the proportion of correct details and correct conclusions, as a way to differentiate between situations where the child did not talk about the event at all, and situations where the child talked about the event (during the first standard question or as correct rejection in response to a suggestive question), but the interviewer failed to elicit more correct details or conclusion.

Study I

Inter-coder reliability

Two research assistants who were blind to the experimental condition coded the question types in the interviews. Before the beginning of the study, the two assistants were trained until they reached 80% [76%, 85%] agreement, with $\kappa = .76$ [.71, .82], $p < .001$, and Gwet's AC1 = .79 [.73, .84], $p < .001$. After the experiment was concluded, they both coded same eight interviews with

children (roughly 10% of the total sample) with 87% [82%, 92%] agreement, with κ = .76 [.73, .88], p < .001, and Gwet's AC1 = .85 [.79, .91], p < .001. Regarding the details in child interviews, where disagreement was present, an agreement has been reached between coders and the first author. No formal inter-coder reliability was conducted for the variable conclusion.

Training sessions with avatars

Assessment of baseline demographics and performance differences between groups. There were no differences between the control and the experimental groups regarding interviewer age (F[1, 38] = .04, p = .835). For interviewer gender, Due to the low frequencies in one of the cells, we used Fisher's exact test. The result showed no difference regarding interviewer gender (Fisher's exact test, p = 1.0). The first interview of each participant was used for testing eventual differences in baseline performance between groups. One-way ANOVAs showed no differences between the two groups at baseline (Interview 1) for incorrect details (F[1, 38] = 3.86, p = .057), or correct conclusions (no correct conclusions were made during the first interview). For number of correct details a significant difference in favour of the feedback group was found (F[1, 38] = 7.96, p = .008). However, the correct details (as well as neutral and incorrect details) are not a direct expression of interviewer skills. They are subordinate to the use of recommended questions, which through the algorithms lead to correct details being elicited from the avatars with a predetermined probability. No differences between groups were found for the proportion of recommended questions, which are univocally related to the interviewer's skills (these are not affected by the algorithms). This suggests that the difference in the number of correct details was due to a random effect of the algorithms.

Effects of feedback manipulation. Receiving feedback (see Table 2 for the descriptive statistics) during the avatar training was associated with a higher proportion of recommended questions, Wald χ^2(11, N = 240) = 661.09, p < .001, a higher number of correct details, Wald χ^2(11, N = 240) = 262.04, p < .001, a lower number of incorrect details, Wald χ^2(11, N = 240) = 66.72, p < .001,

(Group*Time); and a higher proportion of correct conclusions between groups, Mann-Whitney U = 25, $p < .001$, within the avatar interviews. These results indicate that the feedback manipulation improved the quality of the avatar interviews on all indicators. Results of the pairwise comparisons between the groups in each of the six interviews, including the first interview, can be found in Appendix B.

Interviews with children

Questionnaire with children. We provided the children with a questionnaire to analyse their experience during the different phases of the experiment. Ninety-seven percent of the children reported that they liked the mock event they participated in, 94% reported that they liked to talk with the interviewer, and 53% (n = 31) claimed they told everything they remembered.

Assessment of demographic differences between groups. There were no differences between the 393 control and experimental groups regarding the children's age (F[1, 74] = 0.55, p = .460) or gender (χ2 394 [1] = .57, p = .450).

Transfer of feedback effects into child interviews. The means for the total sample (N = 76) were in favor of the feedback group for the variables proportion of recommended questions (see Table 3), and the proportion of correct details ($M_{Feedback}$ = 14.24, SD = 31.73 and $M_{Control}$ = 13.59, SD = 27.83) but not for the proportion of correct conclusions ($M_{Feedback}$ = 0.15, SD = 0.36 and $M_{Control}$ = 0.19, SD = 0.40). Including only the cases in which the children talked about the event (n = 35) showed that the means for all outcome variables were in the expected direction (see Table 3).

The proportions of correct details and correct conclusions are naturally influenced by the cases in which the child did not talk about the event. For this reason, the primary analyses are based on the whole sample for the proportion of recommended questions and restricted to the cases in which the child talked about the event for the proportion of correct details and correct conclusions. However, we ran an additional analysis for the proportion of recommended

questions, where we included only the cases in which the child talked about the event (Control, *n* = 23; Feedback, *n* = 12) to test if disclosure by the interviewed child influenced the questioning style of the interviewer.

Proportion of recommended questions. Receiving feedback during the avatar interviews significantly increased the proportion of recommended questions used by the interviewer in the child interviews compared to the control group in the whole sample (Figure 2 and Table 4) confirming transfer.

Proportion of correct details. Receiving feedback during the avatar interviews significantly increased the proportion of correct details elicited from the children compared to the control group when analyzing the cases in which the child talked about the event (Table 4) confirming that the changes in the use of questions also had the expected effect on eliciting information.

Correct conclusions. Receiving feedback during the avatar interviews did not significantly increase the number of correct conclusions. In this case, the full model including the interaction (Group x Time) failed to converge, suggesting that we added parameters with little or no explanatory value.

Supplementary analysis. We tested if the disclosure of the child affected the proportion of recommended questions used by the interviewer. Only including cases where the child talked about the event (*n* = 35) showed the same result as for the whole group (E = 31.59, t = 2.16, SE = 14.60, p = .039).

In sum, this study replicated previous findings on interviews with avatars in a group of psychologists. The psychologists, who received feedback in avatar interviews, employed a higher proportion of recommended questions, elicited a higher number of correct details, a lower number of incorrect details, and made a higher proportion of correct conclusions compared to the control group in the avatar interviews. Moreover, this effect transferred to interviews with children who had witnessed a mock event. The psychologists, who received feedback when

trained with avatars, employed a higher proportion of recommended questions and elicited a higher proportion of correct details in interviews with children compared to the control group. However, we did not find significant effects for the variable "conclusions" in interviews with children.

Study II

Inter-coder reliability

Two research assistants coded the question types in the interviews. Before the beginning of the study, the two assistants were trained until they reached 80% [76%, 85%] agreement, with κ = .76 [.70, .82], $p < .001$, and Gwet's AC1 = .78 [.73, .84], $p < .001$. After the experiment was concluded, they both coded the same twelve interviews with children (roughly 10% of the total sample) with 76% [69%, 83%] agreement, with κ = .70 [.62, .78], $p < .001$, and Gwet's AC1 = .73 [.66, .81], $p < .001$. ICC was calculated using a 2-way mixed-effects model with consistency for both the total number of correct details per interview (ICC = .98 [.95, .99]) and incorrect details per interview (ICC = .88 [.67, .96]). No formal inter-coder reliability was conducted for the variable conclusion.

Training sessions with avatars

Assessment of baseline demographics and performance differences between groups. There were no differences between the control and experimental groups for interviewers' age ($F[1, 67] = .14$, $p = .710$) or gender (χ^2 [1] = .073, $p = .787$). Due to the violation of normality, we used a Mann-Whitney test for analyzing differences at the baseline, that is, Interview 1. The test showed no differences between the two groups at baseline for the proportion of recommended questions ($U = 592$, $p = .996$) correct details ($U = 562$, $p = .701$), incorrect details ($U = 485$, $p = .158$). A chi-square test showed no differences for correct conclusions (χ^2 (1) = .16, $p = .900$).

Effects of feedback manipulation. Out of the 69 participants, five participants were deleted due to missing data in one or more interviews. The final sample included sixty-four participants ($n_{control}$

=32, $n_{feedback}$=32), see descriptive statistics divided by groups in Table 5. Importantly, we partially replicated the results of Study I. Receiving feedback during the training was associated with higher proportion of recommended questions, Wald $\chi^2(11, N = 384) = 222.38$, $p < .001$, eliciting a higher number of correct details, Wald $\chi^2(11, N = 384) = 59.47$, $p < .001$, and a lower number of incorrect details in the feedback group Wald $\chi^2(11, N = 384) = 32.23$, $p = .001$, (Group*Time). A Mann-Whitney test showed no differences between groups for the proportion of correct conclusions $U = 462$, $p = .495$. These results indicate that the feedback manipulation improved the quality of the avatar interviews on all but one indicator. Results of the pairwise comparisons between the groups in each of the six interviews can be found in Appendix C.

Interviews with children

Questionnaire with children. We provided the children with a questionnaire to analyze their experience during the different phases of the experiment. Ninety-five percent of the children reported that they liked the mock event they participated in, 94% reported that they liked to talk with the interviewer, and 57% claimed they said everything they remembered.

Assessment of demographic differences between groups. There were no differences between the control and experimental groups regarding the children's age ($F[1, 114] = 0.32$, $p = .572$) or gender (χ^2 [39] = 46.40, $p = .194$).

Transfer of feedback effects into child interviews. The means for the total sample ($n = 115$[1]) were in favor of the feedback group for the variables proportion of recommended questions (see Table 6), but not for the proportion of correct details ($M_{Feedback} = 48.90$, $SD = 35.97$ and $M_{Control} = 65.20$, $SD = 35.01$) nor the proportion of correct conclusions ($M_{Feedback} = 0.26$, $SD = 0.44$ and $M_{Control} = 0.28$, $SD = 0.45$) (see descriptive statistics in Table 6).

[1] Data from one interview have not been recorded

The main analyses are based on the whole sample for the proportion of recommended questions and restricted to the cases in which the child talked about the event for the proportion of correct details and correct conclusions. However, we ran the same supplementary analysis as in Study 1 here as well: For the proportion of recommended questions, we also ran an analysis including only the cases in which the child talked about the event (Control, *n* = 53; Feedback, *n* = 50) to test if the disclosure of the child affected the questioning style.

Proportion of recommended questions. Receiving feedback in the avatar interviews significantly increased the proportion of recommended questions used by the interviewer in the child interviews compared to the control group in the whole sample (Table 7) again confirming transfer.

Proportion of correct details. Surprisingly, receiving feedback significantly decreased the proportion of correct details elicited compared to the control group, when analyzing the cases in which the child talked about the event (Table 7).

Correct conclusions. Receiving feedback did not significantly increase the number of correct conclusions.

Supplementary analysis. We tested if the disclosure of the child affected the proportion of recommended questions used by the interviewer. Only including cases where the child talked about the event (*n* = 115) showed the same result as for the whole group (*E* = 11.74, *t* = 3.05, *SE* = 3.85, *p* = .003).

In sum, we partially replicated the results of Study I. The students, who received feedback during their training with avatars, employed a higher proportion of recommended questions in interviews with the children than the control group did. Surprisingly, in Study II, even if the participants in the feedback group employed more recommended questions and elicited a higher number of correct details, they also elicited a higher number of incorrect details. For this reason, we found that receiving no feedback was associated with a higher proportion of correct details.

Discussion

In the current studies, we investigated 1) whether psychologists and psychology students would improve their performance during simulated interview training with avatars and 2) whether the effects of training transferred to interviews with children exposed to a mock event. Receiving feedback improved the quality of both the avatar interviews during the training and the quality of subsequent interviews with real children for what concern the proportion of recommended questions.

The results are promising for several reasons, even if a perfect comparison with other training programs is not possible due to different coding schemes and designs used. The proportion of recommended questions used by interviewers in interviews with children (around 40% in both studies, one week after the training) is in the range of results reported in other training studies, which employed interviews with children about staged events (e.g., Aldridge & Cameron, 1999), and also field interviews. Benson and Powell (2015) showed 40% of open-ended questions immediately after training in their study and suggested an average of 25% in previous training evaluations conducted in the past 15 years. In previous studies, extensive theoretical training has been provided and the training period has lasted from a few days to several weeks. In the current study, we provided no extensive theoretical training to a group of non-expert interviewers, and the training lasted between one and two hours. A short intervention that is as effective as a longer one, carries inherent practical advantages in terms of reduced costs and time.

We found mixed results related to the variable proportion of correct details. Participants in the feedback group were able to obtain a higher proportion of correct details in interviews with children compared to the participants in the control group in Study I. These results were not replicated in Study II. Surprisingly, a higher proportion of recommended questions elicited more incorrect details from the children. Currently, it is not possible to fully explain this finding. Research into child interviews suggests that recommended questions elicit more reliable details,

and for this reason, the result is perplexing. A possible explanation is that psychologists in Study 1 used different types of recommended questions compared to the students in Study 2. Research emphasizes the importance of using invitations because they increase the probability of receiving a reliable statement, compared to, for example, directive questions. We checked for differences in the question types used by the interviewers in the two studies. Results showed that psychologists asked on average more recommended questions of each type. However, these differences were statistically significant only for facilitations (Mann–Whitney $U = 1368.5$, $n_1 = 103$ $n_2 = 35$, $p < 0.05$, two-tailed), and clarifications (Mann–Whitney $U = 1364.5$, $n_1 = 103$ $n_2 = 35$, $p < 0.05$, two-tailed). These differences do not provide a complete explanation for the finding.

Although the use of recommended questions is an important indicator of a well conducted interview; high-quality interviews should also include fewer non-recommended questions and elicit more correct details than incorrect details. An interviewer using forty recommended and forty non-recommended questions will have used 50% of recommended questions, whereas an interviewer that uses twenty recommended questions and five non-recommended questions will have used 80% of recommended questions. Therefore, the proportion of recommended questions provides important information, above and beyond the absolute number of recommended questions, about interview quality.

Taken together, the increased proportion of recommended questions and, for Study I, the higher proportion of correct details in interviews with children suggest that participants in the feedback group conducted interviews of higher quality than participants in the control group. This is indeed a remarkable result. When the ground truth is unknown as is the case in field studies, it is impossible to say anything about whether an interviewer has been successful in finding out the truth. Because of this, most previous research has focused mainly on the use of recommended questions (for a review, see Benson & Powell, 2015), with some studies also looking at the forensic relevance of details elicited (but not the veracity) as an indicator of interview quality (Lamb et al.,

2002a). On the other hand, the participants failed to reach correct conclusions. The poor ability to reach a correct conclusion can be explained by the strict operationalization used for coding a conclusion as a correct (the conclusion had to contain all correct details). As very few children provided enough details to allow the interviewer to reach a correct conclusion, future studies should try to address this problem.

Limitations

Some limitations of the current study must be acknowledged. Powell and colleagues (2008) highlight how interviews with children who have participated in mock events do not necessarily mimic interviews with children in alleged CSA cases. One such difference might be the child's willingness to disclose the truth. However, as pointed out above, the use of a mock event has some qualities that are not present in interviews in a real context, such as the possibility to monitor the veracity of the details provided by the child (Lamb, Hershkowitz, Orbach, & Esplin, 2018).

In Study I, out of the seventy-six children, only twenty-six children talked extensively about the mock event, and a further nine mentioned some details about the event. The induced secrecy (children were requested not to talk about the cookie they received during the mock event) may have negatively affected the children's willingness to report the event (Lyon et al., 2014). The similarity between the scripts, together with the lack of differences in the proportion of recommended questions asked, suggest that the exclusion of the secret and the emphasis on the name of the actor explain the differences in the number of children who talked about the event between Study I and Study II.

The reluctance to talk about the event could also be explained with the relatively short rapport-building phase used. Several studies (e.g., Hershkowitz, Lamb, Katz, & Malloy, 2013) highlight the importance of the rapport-building phase of the interview. Moreover, the revised version of the NICHD, emphasizes the importance of supportive techniques in the pre-substantive

phase (e.g., Hershkowitz et al., 2017). Our participants were not trained in recognizing or dealing with reluctance.

The psychologists and students, who participated in the present studies, were all relatively inexperienced in interviewing children and they were not provided with extensive theoretical knowledge before the interviews, therefore, the effects of training may be different with interviewers with more training.

The coders were trained extensively before the experiment. Even if the coding of the conclusion should be an easier task than the coding of question types and with no interpretation, the lack of a formal test of interrater reliability for this variable is another potential limitation of the research.

The actual training structure is not immune from the effects related to having the practice in a short period, which enhance faster learning but lower retention over time (Schmidt & Bjork, 1992). However, the setup can be used to provide training over longer periods without cost burden. The small sample and effect sizes show the need for other studies to confirm these results.

Future developments

Using simulated avatar training for investigative interviews with children makes it possible to address several practical problems related to other training forms. For example, it is cost-effective, flexible, and more realistic in terms of avatar response patterns. To further develop this training technique, the next step will be to replicate the findings of this experiment with another group of participants, and subsequently test this training with professionals who perform interviews in alleged CSA cases. Other essential steps are addressing the longevity of training effects, and including a training for rapport building, drawing from previous literature both with adults (e.g., Alison et al., 2013) and children (e.g., Lamb et al., 2018).

Conclusions and impact on the field

The results highlight the potential of serious gaming in improving the quality of investigative interviews with children. Solutions that decrease time and money demands are always a positive innovation. The failure of solely information-based training seems to suggest that investigative interviewing is more a practical task (like learning a new sport) than a purely knowledge-based task. When learning something practical, the knowledge of the rules is important, however, only practice with feedback can foster real improvement. Instead of learning in the field, serious gaming can help professionals master interviewing techniques before performing interviews in the real world. Practitioners also tend to highlight the gap between academic-based solutions (that might not be applicable in a real context) and the job in the field. This research aims to bridge this gap by providing a solution that has practical implications for different practitioners (e.g., forensic psychologists, police interviewers, and social workers). This training tool potentially remedies unresolved problems in the field of CSA interviews: a one-hour training with avatars it is at least comparably effective, performance-wise, to training interviewing techniques with actors for what concerns near transfer.

Acknowledgements

We want to give our gratitude to the school heads, the teachers, the school personnel, the families and least but not last all the children who participated in this research, both in Italy and Estonia, for their great attitude and support. A special thanks to Angelica Rizzo Scaccia and Antonella Palazzo for helping in the data collection in Italy and to Steven Saagpakk and Elisabeth Kendrali for helping in the data collection in Estonia.

Conflict of Interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

Aldridge, J., & Cameron, S. (1999). Interviewing child witnesses: Questioning techniques and the role of training. Applied Developmental Science 3(2), 136–147. https://doi.org/10.1207/s1532480xads0302_7

Alison, L. J., Alison, E., Noone, G., Elntib, S., & Christiansen, P. (2013). Why tough tactics fail and rapport gets results: Observing Rapport-Based Interpersonal Techniques (ORBIT) to generate useful information from terrorists. Psychology, Public Policy, and Law, 19(4), 411.

Ayres, P., & Paas, F. (2009). Interdisciplinary perspectives inspiring a new generation of cognitive load research. Educational Psychology Review, 21(1), 1–9. https://doi.org/10.1007/s10648-008-9090-7

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. Journal of Statistical Software, 67(1). https://doi.org/10.18637/jss.v067.i01

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

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

Brysbaert, M., & Stevens, M. (2018). Power analysis and effect size in mixed effects models: A tutorial. Journal of Cognition, 1(1), 1–20, https://doi.org/10.5334/joc

Davis, S. L., & Bottoms, B. L. (2002). Effects of social support on children's eyewitness reports: A test of the underlying mechanism. Law and Human Behavior, 26(2), 185–215. https://doi.org/10.1023/A:1014692009941

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. Journal of Experimental Child Psychology, 85(1), 32–49. https://doi.org/10.1016/S0022-0965(03)00025-0

Hatala, R., Cook, D. A., Zendejas, B., Hamstra, S. J., & Brydges, R. (2014). Feedback for simulation-based procedural skills training: a meta-analysis and critical narrative synthesis. Advances in Health Sciences Education, 19(2), 251-272.

Hattie, J., and Timperley, H. (2007). The power of feedback. Rev. Educ. Res. 77, 81–112. doi: 654 10.3102/003465430298487#

Helsdingen, A., van Gog, T., & van Merriënboer, J. (2011). The effects of practice schedule and critical thinking prompts on learning and transfer of a complex judgment task. Journal of Educational Psychology, 103(2), 383–398. https://doi.org/10.1037/a0022370

Hershkowitz, I., Ahern, E. C., Lamb, M. E., Blasbalg, U., Karni-Visel, Y., and Breitman, M. (2017). Changes in interviewers' use of supportive techniques during the revised protocol training. Appl. Cogn. Psychol. 31, 340–350. doi: 10.1002/acp.3333

Hershkowitz, I., Lamb, M. E., Katz, C., & Malloy, L. C. (2013). Does enhanced rapport-building alter the dynamics of investigative interviews with suspected victims of intra-familial abuse? Journal of Police and Criminal Psychology, 30(1), 6–14. https://doi.org/10.1007/s11896-664 013-9136-8

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. https://doi.org/10.1002/acp.3147

Krause, N., Pompedda, F., Antfolk, J., Zappalá, A., & Santtila, P. (2017). The effects of feedback and reflection on the questioning style of untrained interviewers in simulated child sexual abuse interviews. Applied Cognitive Psychology, 31(2), 187-198.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2016). Tests in linear mixed effects models. R package version, 2, 33.

Lamb, M. E., Brown, D.A., Hershkowitz, I., Orbach, Y., & Esplin, P. W. (2018). Tell me what happened: Questioning children about abuse (2nd ed). Chichester: John Wiley & Sons Ltd.

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. https://doi.org/10.1207/S1532480XADS0601_04

Lewis, F., Butler, A., & Gilbert, L. (2011). A unified approach to model selection using the likelihood ratio test. Methods in Ecology and Evolution, 2(2), 155–162. https://doi.org/10.1111/j.2041-210X.2010.00063.x

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

Lyon, T. D., Wandrey, L., Ahern, E., Licht, R., Sim, M. P. Y., & Quas, J. A. (2014). Eliciting maltreated and nonmaltreated children's transgression disclosures: Narrative practice rapport building and a putative confession. Child Development, 85(4), 1756–1769. https://doi.org/10.1111/cdev.12223

Mugford, R., Corey, S., & Bennell, C. (2013). Improving police training from a cognitive load perspective. Policing: An International Journal of Police Strategies & Management, 36(May), 312–337. https://doi.org/10.1108/13639511311329723

National Institute of Child Health and Human Development. (2011). The National Institute of Child Health and Human Development (NICHD) Protocol: Interview Guide. (M. E. Lamb, D. La Rooy, L. C. Malloy, & C. Katz, Eds.), Children's testimony: A handbook of psychological research and forensic practice (2nd ed), 431-448. John Wiley& Sons.

Paas, F., Renkl, A., & Sweller, J. (2003). Cognitive load theory and instructional design: Recent developments. Educational Psychologist. http://www.tandfonline.com/doi/pdf/10.1207/S15326985EP3801_1

Pompedda, F., (2018). Training in Investigative Interviews of Children: Serious Gaming Paired with Feedback Improves Interview Quality. Åbo Akademi University, Turku: Painosalama Oy. (Ph.D. thesis based on articles). Available at http://urn.fi/URN:ISBN:978-952-12-3679-2

Pompedda, F., Antfolk, J., Zappalà, A., & Santtila, P. (2017). A combination of outcome and process feedback enhances performance in simulations of child sexual abuse interviews using avatars. Frontiers in Psychology, 8, 1474.

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.

Powell, M. B., Fisher, R. P., & Hughes-Scholes, C. H. (2008). The effect of intra- versus post-interview feedback during simulated practice interviews about child abuse. Child Abuse and Neglect, 32(2), 213–227. https://doi.org/10.1016/j.chiabu.2007.08.002

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

Powell, M. B., & Wright, R. (2008). Investigative interviewers' perceptions of the value of different training tasks on their adherence to open-ended questions with children. Psychiatry, Psychology and Law, 15(2), 272–283. https://doi.org/http://dx.doi.org/10.1080/13218710802014493

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

Roberts, K., Lamb, M., & Sternberg, K. (1999). Effects of the timing of postevent information on preschoolers' memories of an event. Applied Cognitive Psychology 13(6), 541-559.

Roberts, K. P., Lamb, M. E., & Sternberg, K. J. (2004). The effects of rapport-building style on children's reports of a staged event. Applied Cognitive Psychology, 18(2), 189–202. https://doi.org/10.1002/acp.957

Schmidt, R. A., & Bjork, R. A. (1992). New conceptualizations of practice: Common principles in three paradigms suggest new concepts for training. Psychological Science, 3(4), 207–218. https://doi.org/10.1111/j.1467-9280.1992.tb00029.x

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

Soveri, A., Antfolk, J., Karlsson, L., Salo, B., & Laine, M. (2017). Working memory training revisited: A multi-level meta-analysis of n-back training studies. Psychonomic bulletin & review, 24(4), 1077-1096.

Sternberg, K., Lamb, M., Davies, G., & Westcott, H. (2001). The memorandum of good practice: Theory versus application. Child Abuse & Neglect.

Sweller, J., van Merrienboer, J. J. G., & Paas, F. G. W. C. (1998). Cognitive architecture and instructional design. Educational Psychology Review, 10(3), 251–296. https://doi.org/10.1023/A:1022193728205

Teoh, Y.-S., & Lamb, M. E. (2010). Preparing children for investigative interviews: Rapport-building, instruction, and evaluation. Applied Developmental Science, 14(3), 154–163. https://doi.org/10.1080/10888691.2010.494463

Van der Kleij, F. M., Feskens, R. C. W., and Eggen, T. J. H. M. (2015). Effects of feedback in a computer-based learning environment on students' learning outcomes: a meta-analysis. Rev. Educ. Res. 85, 475–511. doi: 10.3102/0034654314564881

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.

Westfall, J., Kenny, D. A., & Judd, C. M. (2014). Statistical power and optimal design in experiments in which samples of participants respond to samples of stimuli. Journal of Experimental Psychology: General, 143(5), 2020–2045. DOI: https://doi.org/10.1037/xge0000014

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.

Tables and Figures

Table 1. *Example of the procedure applied for scoring details.*

Interviewer Question	Child Answer	Correct Details	Incorrect Details	Question Type Used by the Interviewer
What did you do in the pirate game?	We dressed up and made a cake	We dressed up (2)	Made, cake (2)	Directive (child already mentioned the pirate game
A magician came to visit you, didn't he?	Yes	--	A magician, That was a "he", visited the child (4)	Suggestive (child never mentioned any magician)

Table 2. *Descriptive statistics divided by group in avatar interviews in Study I.*

Dependent Variable	Control			Feedback		
	N	*M (Min/Max)*	*SD*	*N*	*M (Min/Max)*	*SD*
% Recommended Questions	120	30.83 (0/65)	15.22	120	63.19 (14/100)	20.18
Correct Details	120	2.75 (0/9)	2.25	120	6.39 (0/9)	2.63
Incorrect Details	120	4.03 (0/17)	3.92	120	0.71 (0.8)	1.31
Correct Conclusions	120	0.04 (0/1)	0.20	120	0.35 (0/1)	0.48

Note: N is based on the total number of interviews. Total participants 40 (20 per group)

Table 3. *Descriptive characteristics divided by group in child interviews in Study 1.*

Dependent Variable	Control			Feedback		
	N	*M (Min/Max)*	*SD*	*N*	*M (Min/Mix)*	*SD*
% Recommended Questions	36	25.57 *(0/55)*	14.62	40	40.44 *(0/95)*	22.78
Correct Conclusions[1]	23	0.30 *(0/1)*	0.47	12	0.50 *(0/1)*	0.52
Proportion of Correct Details[1,2]	23	21.28 *(0/100)*	32.57	12	47.45 *(0/100)*	43.01

Note: [1]Includes only the interviews where the child talked about the event [2]Correct details elicited from the first standard question and as correct rejection have been excluded. *N* = number of interviews.

Table 4. *Impact of group on dependent variables (Child interviews in Study I).*

Dependent Variable		*Estimate*	*t*	*95% CI*	*d*		*Variance (SD)*
	Fixed Effects					*Random Effects*	
% Recommended questions	Intercept	25.57	6.57***	(17.94, 33.20)		Subjects	180.2 (13.42)
	Group	14.87	5.36**	(4.36, 25.38)	.04	Residuals	184.5 (13.58)
Proportion of Correct Details	Intercept	22.59	2.51*	(4.26, 40.96)	.	Subjects	1159 (34.06)
	Group	31.59	2.16*	(1.94, 61.37)	02	Residuals	179.4 (13.39)

*Note. *p < .05 ** p < .01 *** p < .001. d* has been calculated based on Westfall et al., 2014, cited in Brysbaert et al., 2018

Table 5. *Descriptive statistics divided by group in avatar interviews in Study II.*

Dependent Variable	Control			Feedback		
	N	*M (Min/Max)*	*SD*	*N*	*M (Min/Max)*	*SD*
% Recommended Questions	192	52.63 (14/94)	16.55	192	69.44 *(26/100)*	17.41
Correct Details	192	3.37 (0/9)	2.56	192	3.98 *(0/9)*	2.55
Incorrect Details	192	1.67 (0/10)	1.96	192	0.89 *(0/8)*	1.30
Correct Conclusions	192	0.32 (0/1)	0.67	192	0.40 *(0/1)*	0.76

Note: N is based on the total number of interviews. Total participants were 64 (32 per group)

Table 6. *Descriptive statistics divided by group in child interviews in Study II.*

Dependent Variable	Control			Feedback		
	N	*M (Min/Max)*	*SD*	*N*	*M (Min/Mix)*	*SD*
% Recommended Questions	57	28.43 *(0/71)*	18.32	58	40.80 *(0/82)*	15.87
Correct Conclusions[1]	53	0.30 *(0/1)*	0.46	50	0.30 *(0/1)*	0.46
Proportion of Correct Details[1,2]	53	70.12 *(0/100)*	31.11	50	56.73 *(0/100)*	32.44

Note: [1]Includes only the interviews where the child talked about the event [2]Correct details elicited from the first standard question and as correct rejection have been excluded. *N* = number of interviews.

Table 7. *Impact of group on dependent variables (Child interviews in Study II).*

Dependent Variable		*Estimate*	*t*	*95% CI*	*d*		*Variance (SD)*
	Fixed Effects					*Random Effects*	
% Recommended questions	Intercept	28.45	10.46***	(23.03, 33.87)		Subjects	135.4 (11.63)
	Group	11.96	3.15**	(4.30, 19.60)	.04	Residuals	154.6 (12.43)
Proportion of Correct Details	Intercept	70.14	16.03*	(5.07, 40.43)	.	Subjects	27.4 (5.23)
	Group	-13.46	-2.14*	(-26.07, -1.00)	-.01	Residuals	962 (31.01)

*Note. *p < .05 ** p < .01 *** p < .001. d* has been calculated based on Westfall et al., 2014, cited in Brysbaert et al., 2018

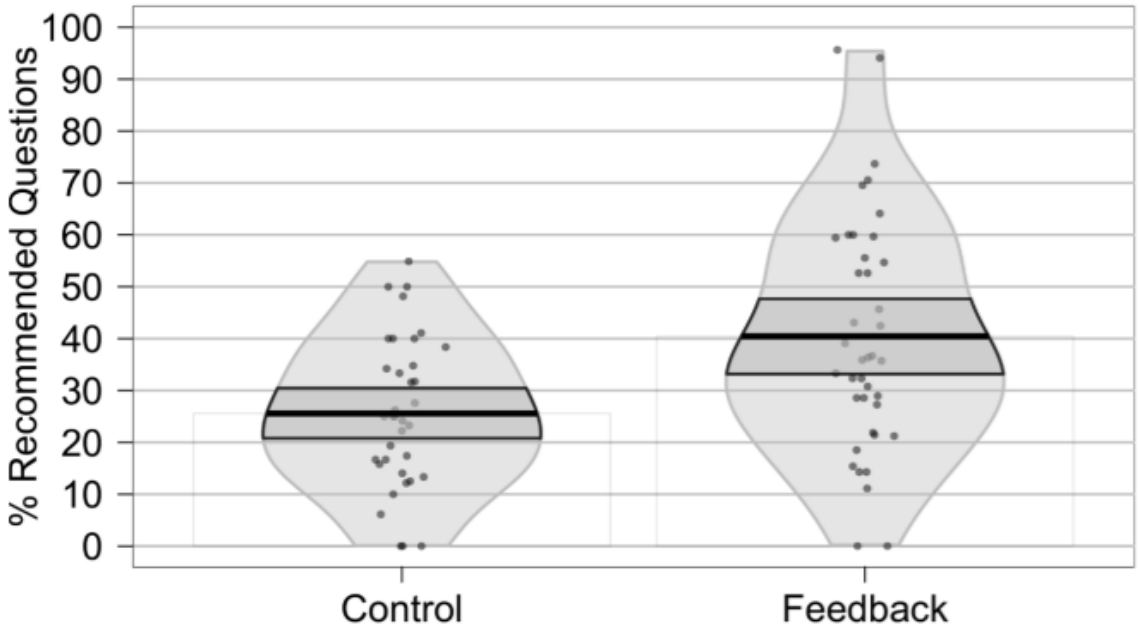

Figure 1. RDI (Raw data, Descriptive and Inferential statistic) plot of the proportion of recommended questions, out of all questions, asked by the interviewers in interviews with children and divided by group (Control, *n* = 36; Feedback, *n* = 40). A Point represents a single interview within each group. The black line represents the average value, highlighted in dark grey is the 95% CI and in light grey the smoothed density.