2010

Reality-monitoring characteristics in confirmed and doubtful allegations of abuse

Kim P. Roberts
Wilfrid Laurier University, kroberts@wlu.ca

Michael E. Lamb
University of Cambridge, mel37@cam.ac.uk

Recommended Citation

Roberts, K.P., & Lamb, M.E. (2010). Reality-monitoring characteristics in confirmed and doubtful allegations of abuse. Applied Cognitive Psychology, 24, 1049-1079. First published online 1 Sep 2009. DOI: 10.1002/acp.1600

Reality-monitoring characteristics in confirmed and doubtful allegations of child sexual abuse

Kim P. Roberts

Wilfrid Laurier University

Michael E. Lamb

University of Cambridge

Address for correspondence: Kim P. Roberts, Ph.D., Wilfrid Laurier University, Department of Psychology, Waterloo, Ontario, Canada, N2L 3C5. Tel: (519) 884 0710 ext. 3225#; Fax: (519) 746 7605; E-mail: kroberts@wlu.ca

Acknowledgements: We are grateful to Phillip W. Esplin, David C. Raskin, Ron Craig, and Irit Hershkowitz for providing transcripts and information about the cases. We are also grateful to Ron Craig, Sara Duncanson, Eva Guterman, Julie Harrison, Heather Lillemoe, Janelle Lynn, Brian Mainland, Nicole K. Sirrine, David W. Randall, Sean Cameron, Valerie Vorstenbosch, and especially Jennifer L. Zale for their assistance during the course of this study.

Abstract

According to reality-monitoring theory, memories of experienced and imagined events are qualitatively different, and can be distinguished by children from the age of 3. Across three studies, a total of 119 allegations of sexual abuse by younger (aged 3-8) and older (aged 9-16) children were analyzed for developmental differences in the presence of reality-monitoring criteria, which should characterise descriptions of experienced events. Statements were deemed likely or unlikely to be descriptions of actual incidents using independent case information (e.g., medical evidence). Accounts by older children consistently contained more reality-monitoring criteria than those provided by younger children, and age differences were particularly strong when the cases were deemed doubtful (Studies 1 and 2).

Reality-monitoring characteristics in confirmed and doubtful allegations of child sexual abuse

Many studies of child sexual abuse are hindered by the lack of both physical evidence and eyewitnesses, so children's accounts are of central importance to investigators. Full and accurate accounts of actually-experienced events can lead to successful prosecution or child protection, whereas convincing accounts of fictitious events can lead to false incarceration. Are there qualitative differences between accurate and fictitious accounts of sexual abuse, and do these differences vary depending on the age of child witnesses? Criteria-based content analysis (CBCA) was designed to discriminate between descriptions of actual and fictitious experiences of sexual abuse (Raskin & Esplin, 1991b; Undeutsch, 1982; Yuille, 1988), but some researchers have reported only moderate sensitivity with this scale (Boychuk, 1991; Craig, Scheibe, Raskin, Kircher, & Dodd, 1999; Lamb et al., 1997; Raskin & Esplin, 1991a; Vrij, 2005, 2008). The purpose of the present study was to a) explore the utility of another technique, inspired by reality-monitoring theory (RMT), for distinguishing between accurate and fictitious accounts and b) to investigate developmental differences in the presence of reality-monitoring criteria across a broad range of ages.

According to RMT, memories of actually-experienced and imagined events differ qualitatively (e.g., Johnson, Hashtroudi, & Lindsay, 1993; Johnson & Raye, 1981), and the distinctiveness of these profiles has been amply demonstrated in adults' reports of both experimentally-induced and autobiographical memories (e.g., Johnson, Foley, Suengas, & Raye, 1988; McGinnis & Roberts, 1996; Porter & Yuille, 1996; Schooler, Gerhard, & Loftus, 1986; Sporer & Küpper, 1995). These researchers suggest that RMT may indeed facilitate the evaluation of statements or allegations in forensic contexts (Alonso-Quecuty, 1992; Pezdek &

Taylor, 2000; Sporer, 1997), although the reality-monitoring characteristics of children’s allegations have not yet been investigated. In this study, we determined whether allegations that had been confirmed or rendered doubtful by independent evidence differed qualitatively from each other.

According to RMT, memories of actually experienced events contain more perceptual, contextual, sensory, and affective information than memories of non-experienced, imagined events. Memories of imagined events should also contain more information about the cognitive operations that took place at the time of the “event” (i.e., during the imaginative activity) than should memories of actual experiences (Johnson et al., 1993; Johnson & Raye, 1981; Suengas & Johnson, 1988). Johnson and colleagues developed the Memory Characteristics Questionnaire (MCQ) to allow participants to rate the qualitative characteristics of their memories (Johnson et al., 1988). As predicted by RMT, descriptions of experienced events rated using the MCQ contained more visual detail, more details about sounds, smells, tastes, the setting, location, spatial arrangements of objects and people, or time references, and more descriptions of memories from before and after the target event than did memories of imagined events. In addition, memories of experienced events were more positive in tone than memories of imagined events (Johnson et al., 1988). Manipulating the qualitative characteristics of memories of imagined events to resemble memories of experienced events (by asking the participant to focus on features typical of actually-experienced events such as perceptual information) makes it more difficult to distinguish between memories of imagined and experienced events (Johnson & Suengas, 1989; Suengas & Johnson, 1988). Hence, RMT provides a top-down, theoretical approach to understanding qualitative differences between truthful and false reports, and thus offers an advantage over other bottom-up approaches such as CBCA (Masip, Sporer, Garrido, & Herrero, 2005).

The predictions of RMT are also supported by studies of children. For example, Fremouw, Miller, and Nangle (1995) asked 11- to 12-year-olds to rate their memories of actual and imagined simple events (e.g., drinking soda and eating pretzels) using the MCQ. As predicted by RMT, the children's memories of actually-experienced events were clearer, more contextually descriptive, and contained more references to thoughts/feelings than their memories of imagined events.

The MCQ is a self-report scale, however, whereas forensic investigators need to know whether they can distinguish between children's *reports* of experienced and imagined events. The MCQ has thus been modified to permit independent ratings of others' reports. Few researchers have asked whether children's reports of actually-experienced and fictitious events have distinct qualitative profiles when examined in this fashion, however, and not all of the results have been in the predicted direction (for reviews, see Masip et al., 2005; Sporer, 2004). Alonso-Quecuty (1995) found that children's reports of a staged event contained more sensory information than deliberately fabricated accounts of such an event, as predicted, but the true statements also contained *fewer* contextual and semantic details, contrary to prediction. Similarly, Santilla, Roppola, and Niemi (1999) found that reports of personally-experienced events (e.g., getting an injection) contained more sensory and temporal information, but also *less* affective information than did reports of imagined events, and children in Joffe's (1994) study who were heavily coached to lie about an event that they had never experienced reported *more* spatial details than did children who actually experienced the event. Finally, Strömwall and Granhag (2005) found that 11-year-olds' reports of imagined events contained less auditory, affective, and temporal but *more* visual information than did reports of witnessed events. Such findings preclude conclusions about the extent to which children's reports of actual and fictitious events approximate the profiles predicted by the RMT, although further research is clearly warranted.

Because there are considerable developments in children's reality-monitoring abilities between the ages of about 3 and 10 (e.g., Foley & Johnson, 1985; Roberts & Blades, 1995), it is surprising that few researchers have studied developmental differences in reports of true and fictitious events. Santtila et al. (1999) and Joffe (1994) found that reports from children aged 7-8 contained fewer reality-monitoring criteria than did reports from older children, *regardless of truth status*. All of the previous studies have focused on children aged 7 and older (Alonso-Quecuty, 1995; Joffe, 1994; Santtila et al., 1999; Strömwall & Granhag, 2005), so more research with younger children is clearly warranted. In the current study, we included reports by children younger than age 7 to explore the extent to which the development of reality-monitoring skills might affect the quality of young children's reports, just as there are as age differences in the presence of CBCA criteria (Lamb et al., 1997).

Alonso-Quecuty (1995), Joffe (1994), and Strömwall and Granhag (2005) studied children's descriptions of artificially-created, non-traumatic events whereas many sexual abuse incidents are traumatic, children are often actively involved in the events, they may have ongoing relationships with the perpetrators, and the events may take place in familiar settings. In addition, the children studied by Alonso-Quecuty and Joffe were interviewed immediately after the events whereas disclosure of abuse is often delayed (the delay was not reported in Strömwall & Granhag's study). Because the quality of children's reports are largely dependent on their memories of the events, children’s reports of both experienced and non-experienced events should be compared after some delay (i.e., when memories of actual events have decayed).

Unlike Alonso-Quecuty (1995) and Joffe (1994), Santtila et al. (1999) studied children’s memories of complex, personally-relevant, and “mildly traumatic” events such as receiving an injection, but researchers have yet to examine younger and older children’s true and false reports of *sexual abuse* in forensic contexts (Masip et al., 2005). Each of the three studies described here involved a sample of forensic cases containing investigative interviews of 3- to 16-year-old children alleging sexual abuse. Each allegation had previously been judged as plausible or implausible using independent evidence gathered at the time of the investigation (e.g., medical findings, physical evidence). In the current studies, we compared reports of events that appeared to have happened with reports of events that appeared unlikely to have happened.

The MCQ was modified in several ways to permit the evaluation of sexual abuse disclosures and the resulting scale was named the “Report Characteristics Questionnaire” (RCQ; see Appendix 1). The RCQ comprised three parts – General characteristics (e.g., clarity and complexity of the account), Specific characteristics (e.g., amount of perceptual and contextual information), and non-MCQ Account characteristics (e.g., number of self-references, verbal hedges). The RCQ was extensively tested and revised to ensure that it could be used with high levels of inter-rater reliability.

The interviews were independently rated for the presence of reality-monitoring criteria by trained coders who were blind with respect to the independent case facts used to determine whether or not the incidents were likely to have happened. Based on RMT and the results of previous research, we expected that the Confirmed cases would receive higher RCQ scores than the Doubtful cases, as well as higher scores on the subscales of the RCQ. Thus, we expected that the Confirmed cases would receive higher ratings for all criteria associated with clarity and richness of detail (i.e., Clarity, Complexity, Realism, event Order, richness of Event Detail, Perceptual-People, Perceptual-Objects, Spatial, Temporal, and Affective information, and Supporting Memories), but would receive lower scores for rehearsal and information about cognitive operations. Although actions have never been rated separately before, we expected more actions to be reported in the Confirmed statements than in the Doubtful statements because action sequences provide richness and clarity about the event. We expected that older children's reports would contain more criteria relevant to status (e.g., more perceptual information for confirmed cases, more information about cognitive operations for doubtful cases) than younger children's reports. Further, because young children are less proficient at reality monitoring and less metacognitively skilled than older children, we expected that there would be fewer differences between older than younger children's reports of confirmed and doubtful incidents because older children would be more aware of the kinds of information that lend credibility to event descriptions.

STUDY 1

Method

Sample

Transcripts of forensic interviews initially included in Raskin and Esplin's (1991a) CBCA validation study were selected for study. These cases had been classified as Confirmed or Doubtful based on "ground truth" information (e.g., medical evidence, polygraph results, witness statements). The initial allocation of cases to the Doubtful group were criticized by Wells and Loftus (1991) because failure to prosecute was used as an index of doubtfulness, so we only included cases in which there was evidence independent of the child's account (i.e., lack of medical evidence, recantation, and polygraph evidence) that the allegation was doubtful (Raskin & Esplin, 1991b). We had access to 29 fully documented interviews which contained, on average, 31 interviewer prompts that elicited spontaneous descriptions from the child witnesses.

The alleged victims had made clear allegations which could be independently verified or falsified using independent case facts. Cases were previously classified as Confirmed if the perpetrator confessed to the alleged incidents before plea bargaining, and/or there was medical or physical evidence. If substantiating evidence was lacking, the perpetrator persistently denied the allegations, a polygraph test suggested that the alleged perpetrator was truthful, the case was not prosecuted, or a court concluded that no abuse had occurred, the allegations were considered Doubtful (Raskin & Esplin, 1991a).

The present sample consisted of 29 statements (15 Confirmed, 14 Doubtful) and all but two of the Confirmed and three of the Doubtful cases were interviews of girls. The statements were grouped according to age. Children's ages in the Younger group ranged from 3 to 8 years (8 Confirmed, 9 Doubtful cases), *M* = 5.29, *SD* = 1.72. In the Older group, children's ages ranged from 9 to 15 years (7 Confirmed, 5 Doubtful), *M* = 11.58, *SD* = 2.23. Of the 29 cases, 16 (7 Confirmed, 9 Doubtful) involved intrafamilial (step-fathers, mothers, or fathers) and 13 (8 Confirmed, 5 Doubtful) extrafamilial alleged perpetrators. None involved persons unfamiliar to the children. Fifteen (8 Confirmed, 7 Doubtful) of the cases contained clear allegations of anal or genital penetration and the remaining 14 (7 Confirmed, 7 Doubtful) allegedly involved non-penetrating abuse (e.g., fondling, sexualized kissing, exposure). Eighteen (10 Confirmed, 8 Doubtful) of the children claimed that they had been abused on only one occasion, and the remaining 11 children (5 Confirmed, 6 Doubtful) reported multiple incidents.

Procedure

Data preparation

The children's statements were coded from the time that the first "substantive" question referring to the alleged incidents was asked (e.g., "You told Dr. S. that something had happened between you and C. Can you tell me what happened?"). If children provided substantive information about the alleged incidents before the first substantive question, then the account was rated from the beginning of the children's first substantive utterances. The transcripts were coded up to the point where the interviewer terminated the interview or switched to a non-substantive topic shortly before the end of the interview. Although most of the interviews contained a standardized "truth/lie ceremony" during a rapport-building phase at the beginning of the interview, some of them contained additional references to truth-telling during the substantive phase. For example, in some cases, interviewers asserted that alleged victims had previously lied about abusive events. All references to the veracity of the child's current or previous accounts were removed from the transcripts before rating.

Only information that was spontaneously provided by the children was used in the RCQ ratings. For example, an affirmative response to the question "was his car red?" was not coded as evidence of the criteria Perceptual – Objects because the child could merely be acquiescing to the question rather than recalling perceptual (color) information. By contrast, a detailed description of the alleged perpetrator's car without focused prompting by the interviewer (e.g., in response to the probe "tell me what happened") would be scored as evidence of visually-encoded details in the child's memory.

Characteristics of the RCQ

The RCQ was developed using most of the original items from the MCQ (Johnson et al., 1988), although some of the MCQ items. were revised The following items were unchanged: Clarity, Complexity, Realism, Order of Event, Event Detail, Event Duration, Tone, Setting, Temporal, Supporting Memories, Affect, Rehearsal, Cognitive Operations. For purposes of the present study, Colour, Visual detail, Sound, Smell, Touch, and taste were combined into Perceptual-People and Perceptual-Objects categories; Location and Setting information were combined into Spatial-People and Spatial-Objects categories; Time, Year, Season, Day and Hour were combined into a Temporal category; Remembered Feeling, Positive/Negative affect, Intensity were combined into Affect; Events before and Events after were combined into Supporting memories; Covert and Overt rehearsal were combined into Rehearsal.

The MCQ was a self-report measure and so, in the present study, we also modified the MCQ criteria so that we could rate *other* peoples' reports of events and removed MCQ items that could not be rated in this way (e.g., the intensity of the feelings of the person remembering the events). As described above, the RCQ comprised two parts (see Appendix 1 for a summary and examples) – the General Characteristics and the Specific Characteristics.

The General Characteristics

In the first part of the RCQ, the Clarity, Complexity, Realism, Order, and richness of Detail of the descriptions of the alleged incident(s) were rated on 3-point scales with 0 indicating the weakest and 2 the strongest presence of each criterion. If an account was very clear, for example, it was rated '2' for Clarity, but if it was vague it was rated 0. The Duration (short, long), Tone (negative, positive, mixed, neutral), and Setting (familiar, unfamiliar) of the alleged incident(s) were rated categorically. Instances where no judgment was possible were recorded as such.

The Specific Characteristics

In the second part of the RCQ, any perceptual, contextual, affective, and cognitive information contained in the account was coded. Perceptual information (visual detail, sound, smell, physical sensation, and taste) was coded separately depending on whether it referred to people (Perceptual-People criterion) or objects (Perceptual-Objects criterion). Also coded were descriptions of Actions, Spatial information (descriptions of the location of the alleged incident[s], the environment, or spatial arrangements of people or objects), Affective information, evidence of covert or overt Rehearsal of the alleged incident(s), and any information about the Cognitive Operations that took place at the time of the alleged incident(s) such as what the child was thinking. Two criteria tapped descriptions of the temporal context: Temporal information (e.g., the day or month in which the events allegedly occurred) and Supporting Memories (i.e., descriptions of events that happened before, after, or [in cases with multiple allegations] in between the alleged incidents).

The relevant line numbers were noted each time an individual criterion was present in the account. The line numbers were then transformed to a “degree of presence” score such that if the criterion was absent, a score of ‘0’ was assigned; if there were 1 to 5 lines where the criterion was fulfilled, a score of ‘1’ was assigned; if the criterion was present in 6 to 30 lines, a score of ‘2’ was assigned; and the presence of a criterion in 31 or more lines resulted in a score of ‘3’. Inter-rater reliability, however, was calculated using the initial line numbers. After conversion to these rates, histograms were created for each criterion to ensure that all distributions of scores were approximately normal.

Coding

Only "substantive" information directly pertaining to the alleged incidents or events surrounding the incidents was coded using the RCQ. Details were only coded the first time that they were mentioned. Utterances could be coded for more than one criteria, so, for example, the utterance "He took his pants off. He was only wearing his shirt" would be considered as containing the criteria Actions and Perceptual – People (description of a person's appearance).

Extensive descriptions of each criteria were developed, tested, and revised using interviews not included in the present study. A pilot study was then conducted (see Roberts, Lamb, & Randall, 1997) and further modifications were made as necessary. This provided detailed, objective descriptions for each criterion. Copies of the complete code book can be obtained from the authors.

Raters were blind to the ground truth status of all interviews in training and in the actual samples studied. Raters were trained to employ the RCQ using interviews that were not included in the study. For the General and Specific Characteristics, a research assistant (RA) and the first author coded non-sample transcripts together until the RA was familiar with the objective descriptions of the criteria and the scoring method. The primary author and the RA then each independently rated five sets of five transcripts. Inter-rater reliability was calculated as each set was completed and disagreements were discussed. After rating the five sets, disagreements for the General Characteristics ratings were usually within one increment (e.g., one coder gave a rating of '1' and the other rated the same transcript as '2'), and reliability was calculated by dividing the number of agreements by eight (i.e., the number of General Characteristics to be rated per transcript). For the Specific Characteristics, inter-rater reliability was calculated separately for each criterion and for each sub-scale if the criterion had several components (e.g., reliability was calculated separately for the Visual, Sound, Smell, Physical Sensation, and Taste sub-scales of the criterion ‘Perceptual – People’). Reliability was calculated conservatively on a line-by-line basis. For example, both raters could agree that an utterance contained a particular criterion but if one rater considered it two lines and another one line this was counted as one agreement and one disagreement. Reliability was computed by dividing the number of agreements by the total number of lines recorded by the raters (i.e., total possible number of agreements). This process was continued until inter-rater reliability was at least 90% for each of the General Characteristics and 88% for each of the Specific Characteristics for each transcript in the last set coded.

After all of the transcripts included in the study had been coded for the General and Specific Characteristics, every fifth transcript in the sample was rated by the first author. Inter-rater reliability was 96% for the General Characteristics and 81% for the Specific Characteristics. The high levels of agreement are likely due to the extensive piloting and training.

Results

Data manipulation

As noted above, 11 of the cases involved multiple alleged incidents (range: 2-5). In three of the 11 cases, the individual incidents received the same rating for Duration, and in seven of the cases, there were identical Setting ratings for each incident in that case. These ratings were thus assigned to the statements. In the remaining cases, all but one of the incidents were rated the same, and the other incident in that case was rated as ‘no judgment possible’. Hence, a single aggregate rating was taken for each of these cases and that rating reflected the dominant rating. Only one case showed an exception to this pattern: the Setting rating was ‘long’ for the first incident and ‘short’ for the second. As the account of the first incident was more extensive than the second, the rating for the first incident was used.

Preliminary analyses

To ensure that neither the interviewers' styles nor the children's talkativeness affected the results, separate 2 (Status: Confirmed, Doubtful) independent groups *t*-tests were carried out on the total number of words in the child's account, the number of interviewer utterances eliciting spontaneous descriptions by the child, and the number of questions asked by the interviewer, but there were no significant effects of status, $0.24 \leq ts(27) \leq 1.68$, $ps > 0.10$.

Total RCQ score

The individual criteria in the General and Specific Characteristics sections were each weighted so that criteria that should be found in reports of experienced events were given positive scores, and criteria that should be found in reports of imagined events were given negative scores. A total score for each transcript was then calculated by summing the individual criteria (the three categorical variables of the General Characteristics section – Duration, Setting, Tone – were excluded), Cronbach's $\alpha = .91$. Scores could range from 0 to 31.

The total scores were entered into a 2 (Status: Confirmed, Doubtful) x 2 (Age Group: Younger, Older) analysis of variance (ANOVA). There were main effects of status, $F(1, 28) = 4.02$, $p = .05$, , $\eta_p^2 = .14$, age, $F(1, 28) = 28.44$, $p < .001$, $\eta_p^2 = .53$, and an interaction between them, $F(1, 28) = 5.79$, $p < .05$, $\eta_p^2 = .19$. Reports in the Confirmed group received higher ratings ($M = 19.29$, $SE = 1.06$) than reports in the Doubtful group ($M = 16.16$, $SE = 1.16$), and older children ($M = 21.89$, $SE = 1.20$) had higher scores than younger children ($M = 13.56$, $SE = 1.00$). To explore the interaction (displayed in Figure 1), 2 (age group) independent groups *t*-tests were run separately for the confirmed and doubtful cases. There were age differences in both groups, $t_{\text{confirmed}}(13) = -2.43$, $p = .03$ (Cohen's $d = 2.80$), and $t_{\text{doubtful}}(12) = -4.75$, $p < .001$ (Cohen's $d = 8.22$). Inspection of the means, however, showed that the age difference was larger for the

Doubtful cases (Younger: $M = 10.11$, $SE = 1.75$; Older: $M = 22.20$, $SE = 1.20$,) than the Confirmed cases (Younger: $M = 17.00$, $SE = 1.63$; Older: $M = 21.57$, $SE = 1.99$).

To determine how the Confirmed and Doubtful cases differed with respect to the individual criteria, the criteria in each section of the RCQ were then examined.

General Characteristics

[Insert Table 1]

The rates (which ranged from 0 to 2 and are presented in the General Characteristics section of Table 1) for the criteria Clarity, Complexity, Realism, Order of Event, and Detail were summed to give a total General Characteristics score, Cronbach's $\alpha = .88$. The scores were then entered into a 2 (status) x 2 (age group) ANOVA which revealed main effects of status, $F(1, 28) = 8.34$, $p = .008$, $\eta_p^2 = .25$, age, $F(1, 28) = 17.54$, $p < .001$, $\eta_p^2 = .41$, and an interaction between them, $F(1, 28) = 8.34$, $p = .008$, $\eta_p^2 = .25$. Reports in the Confirmed group received higher ratings ($M = 7.50$, $SE = 0.52$) than reports in the Doubtful group ($M = 5.28$, $SE = 0.56$), and older children ($M = 8.00$, $SE = 0.59$) had higher scores than younger children ($M = 4.78$, $SE = 0.49$). To explore the interaction, 2 (status) independent groups *t*-tests were run separately for each age group. Confirmed reports from children in the younger age group ($M = 8.88$, $SD = 2.42$) received higher total General scores than did Doubtful reports from their counterparts ($M = 3.00$, $SD = 3.00$), $t(15) = 4.41$, $p = .001$ (Cohen's $d = 2.28$), whereas Confirmed and Doubtful reports from older children did not differ (*M*s = 10.00, *SD*s = 0.00, respectively).

Because they were correlated, the individual scores for the Clarity, Complexity, Realism, Order of event, and Event detail criteria were entered into a 2 (status) x 2 (age group) multivariate analysis of variance (MANOVA). There were significant multivariate effects of status, Wilk's $\lambda = .55$; $F(5, 21) = 3.42$, $p = .02$, $\eta_p^2 = .45$; age, Wilk's $\lambda = .37$; $F(5, 21) = 7.14$, $p <$

.001, η_p^2 = .63; and an interaction between them, Wilk's λ = .55; $F(5, 21) = 3.42$, $p = .02$, η_p^2 = .45.

Confirmed reports were clearer, $F(1, 28) = 11.73$, $p = .002$, η_p^2 = .32, more complex, $F(1, 28) = 9.15$, $p = .006$, η_p^2 = .27, and richer in event detail, $F(1, 28) = 11.56$, $p = .002$, η_p^2 = .32, than were Doubtful reports (see Table 2). Older children's reports were clearer, $F(1, 28) = 17.69$, $p < .001$, η_p^2 = .41, more complex, $F(1, 28) = 22.86$, $p < .001$, η_p^2 = .48, and richer in event detail, $F(1, 28) = 21.15$, $p < .001$, η_p^2 = .46, than reports from younger children. There were significant Status x Age interactions for clarity, $F(1, 28) = 11.73$, $p = .002$, η_p^2 = .32, complexity, $F(1, 28) = 9.15$, $p = .006$, η_p^2 = .27, and event detail, $F(1, 28) = 11.56$, $p = .002$, η_p^2 = .32.

The interactions were followed up by conducting 2 (status) independent groups *t*-tests separately for each age group. Confirmed reports from the younger children were clearer, more complex, and richer in event detail than Doubtful reports, *t*s(15) ranged from 3.67 to 4.16, *p*s < .002 (Cohen's *d*s = 2.15, 1.91, 2.13, respectively). Reports from the older children did not vary depending on whether they were Confirmed or Doubtful.

To determine whether the Confirmed and Doubtful accounts from younger and older children differed with respect to Duration of Event (short vs. long), the cases that had 'no judgment possible' ratings were excluded (n = 11). Interestingly, all 11 "no judgment possible" cases came from younger children (2 confirmed, 9 doubtful). Only one incident was judged to be 'short'. Only one child (an older child who gave a confirmed report) was judged to have reported an unfamiliar setting and 11 children's reports received "no judgment possible" ratings for Setting. Again, these reports were all from younger children (3 confirmed, 8 doubtful). The frequencies of 'familiar' and 'no judgment possible' ratings were entered into 2 (age group) chi-square tests, separately for the confirmed and doubtful cases. Both tests gave significant results

(Confirmed: χ^2 [1, N = 14] = 2.86, p = .045 (1-tailed); Doubtful: χ^2 [1, N = 14] = 10.37, p = .001). In each analysis there were fewer familiar and more no judgment possible ratings for the younger children than would be expected by chance, but the reverse was true for the older children.

The tone (negative, neutral, mixed, positive) of the allegations was rated but none of the allegations were coded as positive or mixed. Most of the children in each age group (over 85%) provided reports that were judged to be negative in tone and a 2 (age) chi-square test was not significant. Proportionally fewer children in the Doubtful cases gave negative reports (55% and 60% of the younger and older children, respectively) but again the test was not significant. Chi-square tests conducted separately for each age group to compare the tone ratings of confirmed and doubtful cases, however, revealed non-significant tendencies for Confirmed cases in both age groups to have a more negative tone than would be expected by chance (Younger: χ^2 [1, N = 17] = 2.88, p = .09; Older: χ^2 [1, N = 12] = 3.00, p = .083).

In sum, general characteristics typical of actual events had a stronger presence in older children's reports than younger children's reports, even when the veracity of older children's reports was deemed doubtful by independent case evidence. The RCQ discriminated between younger children's confirmed and doubtful reports, however, because confirmed statements were clear, or more complex, and contained more event detail than did doubtful reports. The lack of clarity of the doubtful statements from younger children was also reflected in the absence of ratable information about the Duration and Setting of alleged incidents.

Specific Characteristics

The ratings (which ranged from 0 to 3) for the Perceptual–People, Perceptual–Objects, Actions, Spatial, Temporal, Supporting Memories, Rehearsal, Affect, and Cognitive Operations criteria were summed (with Rehearsal and Cognitive Operations negatively weighted), Cronbach's $\alpha = .93$. The full set of means is presented in the Specific Characteristics section of Table 1.

The total Specific Characteristics scores were then entered into a 2 (status) x 2 (age group) ANOVA. There was a main effect of age because reports from older children ($M = 11.89$, $SE = 0.80$) had higher scores than those from younger children ($M = 6.45$, $SE = 0.66$), $F(1, 28) = 27.58$, $p < .001$, $\eta_p^2 = .53$. Age also interacted with status, however, $F(1, 28) = 3.69$, $p = .033$ (1-tailed), $\eta_p^2 = .13$. In the younger age group, Confirmed reports had higher scores ($M = 8.13$, $SD = 2.90$) than did Doubtful reports ($M = 4.78$, $SD = 3.03$), $t(15) = 2.32$, $p = .035$ (Cohen's $d = 1.20$), while the older children's reports did not vary by status (*M*s = 11.57, 12.20, *SD*s = 1.99, 2.68, respectively).

To analyze the individual characteristics, and because they were correlated, the ratings for the Perceptual–People, Perceptual–Objects, Actions, Spatial, Temporal, Supporting Memories, Rehearsal, Affect, and Cognitive Operations criteria were each entered into a 2 (status) x 2 (age group) MANOVA. There was a significant multivariate effect of age, Wilk's $\lambda = .23$; $F(9, 17) = 6.23$, $p = .001$, $\eta_p^2 = .77$, and an interaction between status and age, Wilk's $\lambda = .45$; $F(9, 17) = 2.28$, $p = .035$ (1-tailed), $\eta_p^2 = .55$.

Follow-up univariate ANOVAs showed that the reports from older children contained more of all kinds of information than did reports from younger children, *F*s(1, 28) ranged from 6.11 to 43.60, *p*s $\leq .02$, η_p^2s ranged from .20 to .64.

The univariate ANOVAs showed significant Status x Age interactions for People-Perceptual, $F(1, 28) = 5.39$, $p = .029$, $\eta_p^2 = .18$, Actions, $F(1, 28) = 4.39$, $p = .047$, $\eta_p^2 = .15$, Spatial details, $F(1, 28) = 7.07$, $p = .014$, $\eta_p^2 = .22$, Temporal details, $F(1, 28) = 2.83$, $p = .05$ (1-

tailed), $\eta_p^2 = .10$, and Cognitive Operations, $F(1, 28) = 7.33$, $p = .012$, $\eta_p^2 = .23$. Confirmed reports from the younger children contained more of each kind of detail (except Temporal) than did Doubtful reports, *t*s(15) ranged from 2.31 to 3.66, *p*s < .03 (Cohen's *d*s = 1.26, 1.39, 1.89, 1.70, 0.61, respectively). Reports from the older children did not vary by status (see Table 1).

In sum, the older children's reports contained more details typical of actual events (i.e., more perceptual and contextual information), more actions, and more internal (non-observable) information (i.e., affect and cognitive operations) than did younger children's reports. As with the general characteristics, characteristics typical of actual events were present more often in the younger children's confirmed as opposed to doubtful reports. Means for all other criteria except rehearsal were in the same direction. Unexpectedly, there was also more mention of cognitive operations in the confirmed than in the doubtful reports of the younger children.[2]

Discussion – Study 1

The analyses revealed age differences in the reality-monitoring quality of children's allegations of abuse. Raters blind to the veracity of younger children's accounts gave higher RCQ scores to descriptions of confirmed events than to statements concerning events that may not have happened. Statements from younger children that were confirmed by independent evidence were clearer, more complex, and contained more event, perceptual, and contextual detail than did reports of doubtful incidents. Older children's reports contained such reality-monitoring properties whether or not the events were likely to have happened, but generally contained more details, and were clearer and more complex than younger children's reports.

This was the first study to show that the reports of younger children, who are still developing the capacity to accurately monitor their memories, contain characteristics predicted by RMT. This suggests that younger children's reality-monitoring difficulties may lie in an underdeveloped ability to draw inferences from qualitative differences in their memories, rather than the absence of useful source-specifying information. It is also possible that the older children were more aware of the explicit connection between the specific types of knowledge that are associated with different kinds of experiences (see O'Neill & Gopnik, 1991). Thus, some of the older children in the doubtful group may have inadvertently or intentionally emphasized characteristics more typical of actual events when giving their statements (e.g., vivid, perceptual information).

Contrary to the predictions of RMT, however, confirmed reports also contained significantly more references to Cognitive Operations than the doubtful reports. According to RMT, memories of actual events typically contain few references to cognitive operations. Although this result was found in early work on reality-monitoring characteristics (e.g., Johnson et al., 1988) and in one study with children (Alonso-Quecuty, 1992), other more recent studies have also reported that truthful statements contain more information about cognitive operations than do deceptive statements (with children: Strömwall & Granhag, 2005; with adults: Vrij, Edward, Roberts, & Bull, 2000). Much of the empirical work verifying this result has been based on reports of staged events (e.g., a videotape of a crime, Alonso-Quecuty, 1992), and no studies have assessed the reality-monitoring characteristics of events such as sexual abuse. It is possible that, more cognitive operations are mentioned when children are describing events with a high degree of personal significance as opposed to contrived events. Perhaps the increased number of references to cognitive operations in the confirmed reports reflects children's attempts to process what was happening to them (Fivush, Bohanek, Marin, & Sales, in press).

STUDY 2

Study 2 was conducted to see whether the results of Study 1 could be replicated with a completely different sample. Transcripts of forensic interviews used in Craig et al.'s (1999) study were coded using the RCQ. The children's ages ranged from 3-16 years. Case information relevant to each transcript had previously been analyzed independent of the transcript and used to judge the transcript as describing a "Confirmed" or a "Doubtful" event. Confirmed cases were so categorized on the basis of confession by the accused prior to plea bargaining, a failed polygraph (n = 1), or medical evidence. The Doubtful cases were so categorized on the basis of a later detailed and credible recantation by the child (e.g., indicating that sex was consensual or with a person other than the accused), the accused passed a polygraph, or medical evidence exonerated the accused.

Method

The sample consisted of 48 statements (35 Confirmed, 13 Doubtful), mostly from girls (37 girls, 11 boys). The statements were grouped according to age. Children's ages in the Younger group ranged from 3 to 8 years (12 Confirmed, 8 Doubtful cases), *M* = 5.75, *SD* = 1.55. In the Older group, children's ages ranged from 9 to 16 years (23 Confirmed, 5 Doubtful), *M* = 11.14, *SD* = 1.96.

As reported by Craig et al. (1999), allegations of intrafamilial (e.g., stepfathers) and extrafamilial (e.g., male friends) abuse were made in both confirmed and doubtful groups. There were 47 allegations of penetration (digital or penile) or fondling under clothes, and 10 allegations of non-penetrating abuse (e.g., fondling outside clothes, taking pornographic photographs). About half (n = 25, 52%) of the child complainants alleged multiple incidents which contained, on average, 24 interviewer prompts eliciting spontaneous descriptions from the children.

The procedure was the same as in Study 1. Inter-rater reliability ranged from 89-100% for each characteristic and remained high throughout coding.

Results – STUDY 2

Preliminary analyses

As in Study 1, aggregate ratings were taken whenever children reported multiple experiences. To ensure that neither the interviewers' styles nor the children's talkativeness affected the results, separate 2 (Status: Confirmed, Doubtful) independent groups *t*-tests were carried out on the total number of words in the child's account, the number of interviewer utterances eliciting spontaneous descriptions by the child, and the number of questions asked by the interviewer, but there were no significant effects of status, $0.29 \leq ts(46) \leq 0.64$, $ps > 0.50$.

Total RCQ score

As in Study 1, the individual criteria in the General and Specific Characteristics sections (excluding the categorical variables of Duration, Setting, and Tone) were summed after weighting appropriately, Cronbach's α = .92.

The total scores were entered into a 2 (Status: Confirmed, Doubtful) x 2 (Age: Younger, Older) ANOVA. There was a main effect of age, $F(1, 46) = 45.61$, $p < .001$, $\eta_p^2 = .52$, because older children's reports ($M = 21.44$, $SE = 1.03$) had higher scores than reports from younger children ($M = 11.82$, $SE = 1.00$). Age also interacted with status, $F(1, 46) = 6.63$, $p = .014$, $\eta_p^2 = .13$ (see Figure 1). Independent groups *t*-tests were carried out to compare age differences separately for the Confirmed and Doubtful groups. There were age differences in each group, $t_{\text{confirmed}}(33) = -4.02$, $p < .001$ (Cohen's $d = 5.65$), and $t_{\text{doubtful}}(12) = -4.75$, $p < .001$ (Cohen's $d = 8.82$). Inspection of the means, however, showed that the age difference was larger for the Doubtful cases (*M*s = 9.71, 23.00, and *SEs* = 1.82, 1.26, for the younger and older children, respectively) than the Confirmed cases (*M*s = 13.92, 19.87, and *SE*s = 1.64, 0.65, for the younger and older children, respectively).

General Characteristics

[Insert Table 2]

The rates (which ranged from 0 to 2 and are presented in the General Characteristics section of Table 3) for the criteria Clarity, Complexity, Realism, Order of Event, and Detail were summed to give a total General Characteristics score, Cronbach's $\alpha = .91$.

The scores were then entered into a 2 (status) x 2 (age group) ANOVA which revealed a main effect of age, $F(1,47) = 24.34$, $p < .001$, $\eta_p^2 = .36$, and a Status x Age interaction, $F(1, 47) = 2.91$, $p = .05$ (1-tailed), $\eta_p^2 = .06$. Older children ($M = 9.45$, $SE = 0.62$) had higher scores than younger children ($M = 5.29$, $SE = 0.57$). To explore the interaction, 2 (status) independent groups *t*-tests were run separately for each age group. Confirmed reports from children in the younger age group ($M = 6.58$, $SD = 3.34$) received higher scores than did Doubtful reports from their counterparts ($M = 4.00$, $SD = 3.34$), $t(18) = 1.69$, $p = .05$ (Cohen's $d = 0.81$), but Confirmed and Doubtful reports from older children did not differ in their assigned scores (*M*s = 9.30, 9.60, *SD*s = 1.84, 0.55, respectively), $t(26) = -.35$, *ns*.

Because they were correlated, the scores for Clarity, Complexity, Realism, Order of event, and Event detail were entered into a 2 (status) x 2 (age group) MANOVA. There were significant multivariate effects of status, Wilk's $\lambda = .78$; $F(5, 40) = 2.25$, $p = .034$ (1-tailed), $\eta_p^2 = .22$; age, Wilk's $\lambda = .60$; $F(5, 40) = 5.25$, $p = .001$, $\eta_p^2 = .40$; and an interaction between them, Wilk's $\lambda = .78$; $F(5, 40) = 2.32$, $p = .03$ (1-tailed), $\eta_p^2 = .23$. Confirmed reports were clearer than Doubtful reports, $F(1, 47) = 6.61$, $p = .014$, $\eta_p^2 = .13$. Older children's reports were clearer, $F(1, 47) = 20.89$, $p < .001$, $\eta_p^2 = .32$, more complex, $F(1, 47) = 23.91$, $p < .001$, $\eta_p^2 = .35$, orderly, *F*(1, 47) = 17.54, $p < .001$, $\eta_p^2 = .29$, and richer in event detail, *F*(1, 47) = 23.39, $p < .001$, $\eta_p^2 = .35$, than reports from younger children. There were significant Status x Age interactions for clarity, *F*(1, 47) = 4.88, $p = .032$, $\eta_p^2 = .10$, complexity, *F*(1, 47) = 4.74, $p = .035$, $\eta_p^2 = .10$, and event detail, *F*(1, 47) = 3.59, $p = .03$ (1-tailed), $\eta_p^2 = .08$.

The interactions were followed up by conducting 2 (status) independent groups *t*-tests separately for each age group. Confirmed reports from the younger children were clearer, *t*(18) = 2.85, $p = .01$ (Cohen's *d* = 1.36), more complex, *t*(18) = 2.00, $p = .03$ (Cohen's *d* = 0.95), and richer in event detail than Doubtful reports, although the analysis for the latter just failed to reach significance, *t*(18) = 1.64, $p = .06$ (Cohen's *d* = 0.79). Reports from the older children did not vary depending on whether they were confirmed or doubtful.

The accounts were rated with respect to Duration of Event (short vs. long). In the youngest age group, 14 accounts (70% of sample) were rated "no judgment possible" (7 confirmed, 7 doubtful); there were 5 such ratings (18%) in the older group (all confirmed cases). Most of the older children's accounts (75%) were judged to be about events of long duration (16 confirmed, 5 doubtful). The ratings of short, long, and no judgment possible were entered into 2 (age) chi-square tests separately for the confirmed and doubtful cases, and both tests were significant (Confirmed: χ^2 [2, *N* = 35] = 6.40, $p = .041$; Doubtful: χ^2 [1, *N* = 13] = 9.48, $p = .002$). In both analyses, fewer younger children received 'long' ratings and more received 'no judgment possible' ratings than would be expected by chance, while the reverse was true for the older children.

The setting was rated as unfamiliar in only three reports (1 older Confirmed, and a younger and an older child in the Doubtful group). While a 2 (age group) chi-square test on the remaining ratings for the Confirmed group did not reveal a significant pattern, the analysis for the Doubtful cases was significant, χ^2 (1, *N* = 11) = 3.59, *p* = .03 (1-tailed). Fewer younger children's reports were judged to be in familiar settings and more of their reports received "no judgment possible" ratings than would be expected by chance, whereas the reverse was true for the older children.

The tone (negative, neutral, mixed, positive) of the allegations was rated but none of the allegations were coded as positive and only one (from the Confirmed, older age group) was judged to be mixed. A 2 (age group) chi-square test was conducted on the remaining ratings of reports in the Confirmed group and it was significant, χ^2 (2, *N* = 34) = 7.29, *p* = .026. Fewer of the younger children's reports were judged to be negative and more were judged 'neutral' than would be expected by chance, while the reverse was true for the older children's reports. The analysis for the Doubtful reports was not significant.

As in Study 1, general characteristics typical of actual events had a stronger presence in younger children's Confirmed than Doubtful reports. Confirmed statements were more clear and complex than were Doubtful reports. The lack of clarity of the Doubtful statements from younger children was also reflected in the absence of ratable information about the Duration and Setting of alleged incidents. Older children's reports were clearer, more complex, more orderly, richer in event detail, and more negative in tone than younger children' reports.

Specific Characteristics

The ratings (which ranged from 0 to 3) for the Perceptual–People, Perceptual–Objects, Actions, Spatial, Temporal, Supporting Memories, Rehearsal, Affect, and Cognitive Operations criteria were summed (with Rehearsal and Cognitive Operations negatively weighted), Cronbach's α = .88. The full set of means is presented in the Specific Characteristics section of Table 2.

The total Specific Characteristics scores were then entered into a 2 (status) x 2 (age group) ANOVA. There was a main effect of age because reports from older children (M = 11.98, SE = 0.64) had higher scores than those from younger children (M = 6.74, SE = 0.62), F(1, 46) = 34.93, p < .001, η_p^2 = .45. Age also interacted with status, however, F(1, 46) = 5.14, p = .028, η_p^2 = .11. Although the reports of the older children had higher scores than those from younger children in both the Confirmed (t[33] = -3.47, p = .001; Ms = 10.57, 7.33, SDs = 2.33, 3.11, respectively) and Doubtful groups (t[10] = -4.98, p = .001; Ms = 13.40, 6.14, SDs = 2.88, 2.19, respectively), the difference was larger in the Doubtful group (Cohen's ds = 1.28, 3.20, for the Confirmed and Doubtful groups analyses, respectively).

To analyze the individual characteristics, and because they were correlated, the individual ratings for the Perceptual–People, Perceptual–Objects, Actions, Spatial, Temporal, Supporting Memories, Rehearsal, Affect, and Cognitive Operations criteria were entered into a 2 (status) x 2 (age group) MANOVA. There was a significant multivariate effect of age, Wilk's λ = .46; F(9, 35) = 4.63, p < .001, η_p^2 = .54. Follow-up univariate ANOVAs showed that the reports from older children contained more of all kinds of information except rehearsal than did reports from younger children, Fs(1, 46) ranged from 6.53 to 33.74, ps ≤ .01, η_p^2s ranged from .07 to .44. There was also a similar trend for rehearsal, F(1, 46) = 3.39, p = .07, η_p^2 = .16.

As with Study 1, the older children's reports contained more details typical of actual events (i.e., more perceptual and contextual information), more actions, and more internal information (i.e., affect and cognitive operations) than did younger children's reports. Although younger children's doubtful reports did not receive lower scores than confirmed reports, contrary to what we found in Study 1, there was some evidence of low levels of reality-monitoring criteria in their doubtful reports. Specifically, age differences were greater in the Doubtful than

Confirmed group because the reality-monitoring criteria consistent with memories of experienced events were less frequently present in doubtful reports by younger children. [1]

Discussion – STUDY 2

In general, the pattern of results in Study 1 and 2 was similar, with younger children's doubtful statements containing fewer reality-monitoring criteria typical of actual events than did their confirmed reports. These differences were evident in analyses of the total General characteristics, total Specific characteristics, and the individual general characteristics scores for clarity and complexity, but not for any of the individual specific characteristics scores. As before, there was also little difference between the older children's confirmed and doubtful reports.

Although the results were weaker than in Study 1, in sum, the results again showed that reality-monitoring criteria were less frequently evident in younger children's doubtful reports.

STUDY 3

Study 3 was conducted to see whether the RCQ revealed differences in younger and older children's descriptions of abuse that were confirmed or rendered doubtful when strict criteria for assessing ground truth were used. The 42 transcripts from Lamb et al.'s (1997) study were coded using the RCQ. The children's ages ranged from 3-16 years and all interviews were conducted in Hebrew. As in Studies 1 and 2, case information (e.g., medical, physical, and material evidence; witness and suspect statements) was independently used to clarify the statement as "confirmed" or "doubtful" (referred to as "plausible" and "implausible" by Lamb et al.). A conservative coding system that took into account, for example, the likely prevalence of different kinds of evidence, was used to deem cases confirmed or doubtful (full details are given by Lamb et al.).

Method

The sample consisted of 42 statements (20 Confirmed, 22 Doubtful) from 34 girls and 8 boys (2, 3, 2, and 1 were from the Younger/Confirmed, Younger/Doubtful, Older/Confirmed, and Older/Doubtful groups, respectively). The statements came from the Lamb et al. (1997) CBCA study. The Doubtful and Confirmed cases were matched on age until we had a sample of at least 20 per truth status condition. Children's ages in the Younger group ranged from 4 to 8 years (11 Confirmed, 12 Doubtful cases), *M* = 6.43, *SD* = 1.40. In the Older group, children's ages ranged from 9 to 13 years (9 Confirmed, 10 Doubtful), *M* = 10.37, *SD* = 1.21. Allegations of intrafamilial (step-fathers, mothers, or fathers) and extrafamilial abuse were present in all Age x Status cells, as were allegations involving anal or genital penetration and non-penetrating abuse (e.g., fondling, sexualized kissing, exposure). About two-thirds of the allegations involved repeated incidents.

The procedure and reliability checks were the same as in Studies 1 and 2, except that a native Hebrew speaker coded the transcripts for all criteria.

Results – STUDY 3

Preliminary analyses

As in Studies 1 and 2, aggregate ratings were used whenever children reported multiple experiences. To ensure that neither the interviewers' styles nor the children's talkativeness affected the results, separate 2 (Status: Confirmed, Doubtful) independent groups *t*-tests were carried out on the total number of words in the children's accounts, the number of interviewer utterances eliciting spontaneous descriptions by the child, and the number of questions asked by the interviewers, but there were no significant effects of status, *t*s(40) < 1.00, *p*s > 0.30.

Total RCQ score

As in Studies 1 and 2, a total score for each transcript was calculated by weighting the individual criteria appropriately and summing them (the three categorical variables in the General Characteristics section – Duration, Setting, Tone – were excluded), Cronbach's $\alpha = .87$. Scores could range from 0 to 31, and are displayed in Figure 1. The total scores were entered into a 2 (Status: Confirmed, Doubtful) x 2 (Age Group: Younger, Older) ANOVA. There was a main effect of age, $F(1, 39) = 12.27$, $p = .001$, $\eta_p^2 = .24$, because older children ($M = 17.07$, $SE = 0.99$) had higher scores than younger children ($M = 12.41$, $SE = 0.90$). Scores did not vary by status, however, $F(1, 39) = 1.97$, $p = .17$, $\eta_p^2 = .05$ (*M*s = 15.68, 13.81, and *SE*s = .96, .92, for the Confirmed and Doubtful groups, respectively).

To determine whether there were any differences between the Confirmed and Doubtful cases with respect to the individual criteria, the criteria in each section of the RCQ were then examined.

General Characteristics

[Insert Table 3]

The rates (which ranged from 0 to 2 and are presented in the General Characteristics section of Table 4) for the criteria Clarity, Complexity, Realism, Order of Event, and Detail were summed to give a total General Characteristics score, Cronbach's $\alpha = .90$. The scores were then entered into a 2 (status) x 2 (age group) ANOVA which revealed a main effect of age, $F(1, 41) = 15.44$, $p < .001$, $\eta_p^2 = .29$, because older children ($M = 8.12$, $SE = 0.62$) had higher scores than younger children ($M = 4.83$, $SE = 0.56$). Although Confirmed reports received higher scores ($M = 7.12$, $SE = 0.61$) than did Doubtful reports ($M = 5.83$, $SE = 0.58$), this difference just failed to reach significance, $F(1, 41) = 2.40$, $p < .06$ (1-tailed), $\eta_p^2 = .06$.

Because they were correlated, the individual Clarity, Complexity, Realism, Order of Event, and Detail scores were entered into a 2 (status) x 2 (age group) MANOVA. There was a significant multivariate effect of age, Wilk's $\lambda = .50$; $F(5, 34) = 6.70$ $p < .001$, $\eta_p^2 = .49$. Older children's reports were clearer, $F(1, 41) = 21.36$, $p < .001$, $\eta_p^2 = .36$, more realistic, $F(1, 41) = 15.12$, $p < .001$, $\eta_p^2 = .29$, more orderly, $F(1, 41) = 23.95$, $p < .001$, $\eta_p^2 = .39$, and richer in event detail, $F(1, 41) = 6.91$, $p = .012$, $\eta_p^2 = .15$, than were younger children's reports.

Most ratings of duration were coded as "no judgment possible": In the Confirmed group, 78% of the younger children's and 56% of the older children's reports received this rating; in the Doubtful group, 92% and 70% of the younger and older children's reports were thus rated. Chi-square tests confirmed that there were no significant differences.

None of the Confirmed cases were judged to have taken place in unfamiliar settings. The remaining ratings were entered into 2 (age group) chi-square tests, separately for confirmed and doubtful cases. With respect to the Confirmed cases, fewer younger children described familiar settings and more received 'no judgment possible' ratings than would be expected by chance, whereas this was reversed for the older children, χ^2 $(1, N = 20) = 10.48$, $p = .001$.

As before, mixed or positive tones in children's reports were rare. One younger child described the event positively and one older child had a mixed tone (both in the Confirmed group), and one younger child in the Doubtful group gave a mixed tone report. Separate 2 (age group) chi-square tests were run on the ratings of reports in the Confirmed and Doubtful groups. There was a significant result for the Doubtful cases, χ^2 $(2, N = 21) = 8.97$, $p = .011$. Fewer younger children gave negative reports and more gave neutral reports than would be expected by chance, but the reverse was true for the older children.

In sum, older children's reports again contained more of the General Characteristics consistent with reality-monitoring theory and were more negative in tone than younger children's reports, whose reports were again characterized by more vagueness. Although confirmed reports received higher total General Characteristics scores, there were no statistically significant differences between these two types of reports, as there was in the two previous studies.

Specific Characteristics

Cronbach's α for the nine specific characteristics was .71, but a more acceptable .74 with Perceptual-Objects removed. Thus the ratings (which ranged from 0 to 3) for the Perceptual–People, Actions, Spatial, Temporal, Supporting Memories, Rehearsal, Affect, and Cognitive Operations criteria were summed (with Rehearsal and Cognitive Operations negatively weighted). The full set of means is presented in the Specific Characteristics section of Table 3.

The total Specific Characteristics scores (excluding Perceptual-Objects) were then entered into a 2 (status) x 2 (age group) ANOVA. There was a main effect of age because reports from older children (*M* = 7.96, *SE* = 0.46) had higher scores than those from younger children (*M* = 6.63, *SE* = 0.42), *F*(1, 41) = 4.51, *p* = .04, η_p^2 = .11. Scores did not vary by status, *F*(1, 41) = 1.28, *p* = .265, η_p^2 = .03 (*M*s = 7.65, 6.94, and *SE*s = .45, .43, for the Confirmed and Doubtful groups, respectively). The analysis was repeated including Perceptual-Objects in the total scores and the results were the same: Status *F*(1, 41) = 5.01, *p* = .03, η_p^2 = .12.

To analyze the Specific Characteristics individually, the ratings for the Perceptual–People, Perceptual–Objects , Actions, Spatial, Temporal, Supporting Memories, Rehearsal, Affect, and Cognitive Operations criteria were entered into a 2 (status) x 2 (age group) MANOVA. There was a significant multivariate effect of age, Wilk's λ = .62; *F*(9, 30) = 2.09, *p* = .03 (1-tailed), η_p^2 = .39. Follow-up univariate ANOVAs showed that the reports by older children contained more references to actions, $F(1, 41) = 7.47$, $p = .009$, $\eta_p^2 = .16$, spatial information, $F(1, 41) = 2.79$, $p = .05$ (1-tailed), $\eta_p^2 = .07$, temporal information, $F(1, 41) = 10.77$, $p = .002$, $\eta_p^2 = .22$, and cognitive operations, $F(1, 41) = 7.03$, $p = .012$, $\eta_p^2 = .16$, than did younger children's reports.

In sum, the older children's reports contained more specific characteristics typical of actual events (i.e., more actions, contextual, and cognitive operations) than did younger children's reports. Unlike Studies 1 and 2, though, there were few differences between Confirmed and Doubtful cases. [1]

Discussion – STUDY 3

The age differences found in Studies 1 and 2 were replicated in Study 3. Older children's reports contained more individual criteria, as well as higher total General and Specific characteristics scores. There was little evidence in this sample, however, of qualitative differences between confirmed and doubtful reports.

The criteria applied in Lamb et al.'s (1997) study to distinguish between confirmed and doubtful cases were the strictest yet reported in the literature and thus the results were disconcerting. It could be that the predictions of RMT were not born out when the cases were more strictly and accurately classified. Alternatively, it could be that the RCQ criteria are more difficult to identify in Hebrew (the language used in the interviews conducted in Study 3) than in English (the language used in Study 1 and 2 interviews). It is unlikely that the differences were related to the quality of the interviews, because the Israeli interviews studied by Lamb et al. (1997) were reportedly as poor as those studied by Raskin and Esplin (1991) and Craig et al. (1999).

General Discussion

In the three Studies reported here, we examined the qualitative characteristics of children's allegations that had previously been confirmed or judged doubtful using independent evidence (Raskin & Esplin, 1991a; Craig et al., 1999; Lamb et al., 1997). Our analyses yielded the first evidence of consistent developmental differences in the presence of reality-monitoring criteria in children's allegations of personal experiences (specifically, sexual abuse), with more criteria present in older than younger children's reports. These qualitative differences were not an artifact of longer reports from older children because there were no age differences in the number of words in their accounts, and analyses of the presence of reality-monitoring criteria per 100 words found identical results. Second, in Study 1 and 2 (but not Study 3), younger children's confirmed allegations were qualitatively different in expected ways from doubtful reports when rated for a variety of characteristics known to differentiate between memories of experienced and imagined events. As predicted by RMT, reports of presumably experienced incidents contained more of the details expected to be present in memory-based accounts of experienced events (Johnson & Raye, 1981). Like adults' accounts, the reports of doubtful events lacked the characteristics typical of memories of experienced events (Johnson et al., 1993; Johnson & Raye, 1981; Suengas & Johnson, 1988). RMT may thus provide a useful guide to a theoretically-driven evaluation of children's accounts in forensic contexts, not only in the laboratory.

Consistent developmental differences in the qualitative characteristics of children's reports also provide important information about the kinds of details children might be expected to report at different ages. The older children explicitly acknowledged internal states that they experienced at the time of the alleged incidents, more often providing information about what they were feeling and thinking at the time. Younger children's accounts contained less of this information, perhaps because that they did not have reflective awareness of their affective or cognitive internal states. This is consistent with recommendations from forensic interviewing experts to avoid using “why?” questions with young children, that is, questions that require reflection on the part of the witness (e.g., Poole & Lamb, 1998). Young children did, however, tend to provide critical incident-relevant information about actions, perceptual, and contextual information when they were describing events that seemed likely to have happened, although these differences were not significant in all three studies

The reports provided by young children about incidents that appeared unlikely to have happened were more vague and less detailed than confirmed accounts in Studies 1 and 2. This raises the possibility that independent, confirmatory evidence was not obtained by law enforcement because the allegations were initially less clear. In the three studies reported here, though, the case information was rated entirely independently of the children’s statements – different coders rated the case information and the transcripts, the case coders did not see the transcripts, and the coders who used the RCQ did not have access to the case information or judgments about the case information. Also, as discussed earlier, Raskin and Esplin (1991b) re-analyzed their results after removing cases judged doubtful by “negative evidence” such as lack of prosecution, and their results did not change. Further, in the present study, there were no group differences in the number of interviewer utterances, interviewer utterances eliciting spontaneous details, or the length of the reports suggesting that these children were given as much opportunity to describe their experiences as children whose allegations were confirmed. It seems, then, that the doubtful allegations were less articulate and detailed than the confirmed allegations.

Because older children are more proficient at reality-monitoring, we expected that there would be less difference between confirmed and doubtful accounts in this age group than with younger children, and this was confirmed. Older children's confirmed and doubtful accounts were indistinguishable using the RCQ. These results suggest that by ages 9-10, children may know what kinds of information are consistent with memories of actual experiences and, thus, include more of this information when attempting to describe a fictitious event convincingly (e.g., including vivid, perceptual information). As no previous research had included a comparison group of children younger than 7/8, it was unknown whether older children were better able to provide an account that was consistent with the reality-monitoring criteria found in memories of experienced events. The current results show that they can, although it is unknown whether these children were aware of the association between reality-monitoring characteristics and the truth.

The development of language skills might also play some part in the developmental differences we observed in the reality-monitoring characteristics of children's reports. Older children tend to report more information than younger children (see Poole & Lamb, 1998), although even younger children can provide detailed reports if trained (e.g., Lamb et al., 2000). Our results were also analyzed by using rates per 100 lines to compensate for verbose reports, but the results remain the same (available from the first author). Thus, while language undoubtedly contributes to the quality and length of children's testimony, the older children (as a group) consistently provided reports richer in reality-monitoring characteristics than the group of younger children. Although the groups were split by age, further research could track the presence of reality-monitoring criteria in children's descriptions as a function of age in years.

As discussed in the Introduction, the CBCA technique has been used to distinguish between truthful and fabricated accounts of child sexual abuse (Boychuk, 1991; Craig et al., 1999; Lamb et al., 1997; Raskin & Esplin, 1991a; Undeutsch, 1982; Yuille, 1988), and because the RCQ developed out of a theory for which there is extensive empirical support, the reality-monitoring approach provides a theoretical basis for the CBCA technique. As noted by Sporer (1997), many MCQ items are similar to CBCA criteria: Both analyze the amount of detail in the child's account, for example, and total RCQ scores are significantly correlated with total CBCA scores (Roberts et al., 1997). A direct comparison of reality monitoring and CBCA approaches to the evaluations of children's allegations has yet to be attempted, however, although it is noteworthy that the MCQ appears more useful at distinguishing between plausible and doubtful accounts by younger rather than older children, whereas the CBCA technique distinguishes between plausible and doubtful allegations by older children better than those made by younger children (Lamb et al., 1997). This difference may reflect some fundamental but not obvious differences between the two approaches.

Whereas most of the results reported here were consistent with the predictions of RMT, there were some exceptions and these were all with the younger children's reports. Contrary to expectations, the confirmed reports contained more information about cognitive operations (Study 1) and filled pauses (e.g., "umm"; Study 2) than did the doubtful reports. The higher rate of cognitive operations information may not be surprising when the nature of the events described by the children is considered. Children may be actively involved, the events and consequences of the events were often anchored in their everyday lives, and they have enormous personal significance. For these reasons, frequent references to thoughts at the time of the event may reflect the extent of personal cognitive, affective and/or motor involvement in the events, as well as attempts to make sense of the events (Fivush, 1998). By contrast, previous research on the amount of cognitive operations in accounts has included memory for more neutral stimuli such as slides of traffic signs (Schooler et al., 1986) or a video (Alonso-Quecuty, 1992). Because the presence of cognitive operations in imagined events has not been evident in recent work (e.g., Strömwall & Granhag, 2005), the role of cognitive operations in reality-monitoring theory needs further exploration. It is less clear why there were more filled pauses in the Confirmed than Doubtful accounts, however, but because the Confirmed cases tended to be more negative in tone, it could be that the children found it difficult to describe their experiences and thus paused more often to collect their thoughts.

Although RMT helped us to target differences between the characteristics of confirmed and doubtful accounts, we caution that this technique should not be used to detect deception in the field. First, reality-monitoring characteristics differed for confirmed and doubtful cases only for children aged 8 and below. By contrast, there were few differences between the accounts of confirmed and doubtful statements by older children aged 9-16. Second, RCQ score differences between confirmed and doubtful statements were significant in only two of three samples.

The studies reported here suggest that there are interesting, theoretically-based differences in accounts of experienced and probably not experienced sexual abuse. Although many researchers have begun to study children's allegations of abuse in forensic rather than analogue contexts, we still know relatively little about many aspects of children's accounts, particularly those that may be false. Further research is needed to understand what other factors might affect the qualitative characteristics of children's abuse allegations. For example, the quality of allegations may be affected by the kinds of interviewer utterances used (e.g., the extent to which they are leading or open-ended). The present data have provided developmentally-appropriate expectations of the kinds of information that child witnesses of different ages provide when interviewed in the inadequate style that was typical in the 1980's and 1990's. The finding that even young children were able to provide forensically-relevant information such as the temporal and spatial context of events is consistent with other recent research showing that children have such capabilities (e.g., Orbach & Lamb, 2007) when the interviews are conducted in line with professional guidelines. It would thus be valuable to conduct further research of this kind using interviews conducted in accordance with the guidelines incorporated in the NICHD Protocol, for example, because this gives priority to open-ended prompts designed to elicit information from free recall (Lamb, Orbach, Hershkowitz, Esplin, & Horowitz, 2007; Lamb, Hershkowitz, Orbach, & Esplin, 2008).

Footnotes

[1] The interviews were also coded for a variety of characteristics identified in subsequent research: Doubt, Hedges, Pauses, Functional utterances, Faulty logic, Self-references, Diversions. These data are available from the primary author on request.

[2] The data were also analyzed in a variety of other ways. Instead of using 2 (status) x 2 (age) (M)ANOVAs, age was entered as a covariate in a series of 2 (status) ANOVAs. The pattern of results was the same as reported above, in that effects of status emerged only as reported above. Regressions were also run for each of the total RCQ, total General Characteristics, total Specific Characteristics, and total Account Characteristics scores with age and status entered as independent variables. All of the models except that on the total Account Characteristics in Study 3 were significant, and age and status effects mirrored those reported above. A detailed report of these analyses can be obtained from the first author.

References

Alonso-Quecuty, M. L. (1992). Deception detection and reality monitoring: A new answer to an old question? In F. Lösel, D. Bender, & T. Bliesener (Eds.), *Psychology and law: International perspectives* (pp. 328-332). Berlin: Walter de Gruyter.

Alonso-Quecuty, M. L. (1995). Detecting fact from fallacy in child and adult witness accounts. In G. Davies, S. Lloyd-Bostock, M. McMurran, & C. Wilson (Eds.), *Psychology, law, and criminal justice: International developments in research and practice* (pp. 74-80). Berlin: Walter de Gruyter.

Boychuk, T. (1991). *Criteria-based content analysis of children's statements of sexual abuse: A field based validation study*. Unpublished doctoral dissertation, University of Michigan, Ann Arbor.

Craig, R. A., Scheibe, R., Raskin, D. C., Kircher, J. C., & Dodd, D. H. (1999). Interviewer questions and content analysis of children's statements of sexual abuse. *Applied Developmental Science, 3*, 77-85.

Esplin, P. W., Houed, T., & Raskin, D. C. (1988, June). *Application of statement validity assessment*. Paper presented at NATO Advanced Study Institute on Credibility Assessment, Maratea, Italy.

Fivush, R., Bohanek, J.G., Marin, K., & Sales, J.M. (in press). Emotional memory and memory for emotions. Chapter to appear in O. Luminet, A. Curci and M. Conway (Eds.), *Flashbulb memories: New issues and new perspectives,* Psychology Press.

Fivush, R. (1998). Children's recollections of traumatic and nontraumatic events. *Development and Psychopathology, 10,* 699-716.

Foley, M. A., & Johnson, M. K. (1985). Confusions between memories for performed and imagined actions: A developmental comparison. *Child Development, 56,* 1145-1155.

Fremouw, W., Miller, C., & Nangle, D. (1995). Real versus imagined memories of children and adults: Implications for assessment of child abuse. *American Journal of Forensic Psychology, 13*, 21-29.

Joffe, R. D. (1994). *Content-based criteria analysis: An experimental investigation with children*. Unpublished doctoral dissertation, Department of Psychology, University of British Columbia, British Columbia, Canada.

Johnson, M. K., Foley, M. A., Suengas, A. G., & Raye, C. L. (1988). Phenomenal characteristics of memories for perceived and imagined autobiographical events. *Journal of Experimental Psychology: General, 117*, 371-376.

Johnson, M. K., Hashtroudi, S., & Lindsay, D. S. (1993). Source monitoring. *Psychological Bulletin, 114*, 3-28.

Johnson, M. K., & Raye, C. L. (1981). Reality monitoring. *Psychological Review, 88*, 67-85.

Johnson, M. K., & Suengas, A. G. (1989). Reality monitoring judgements of other people's memories. *Bulletin of the Psychonomic Society, 27*, 107-110.

Lamb, M. E., Hershkowitz, I., Orbach, Y., & Esplin, P. W. (2008). *Tell me what happened: Investigative interviews of child victims and witnesses*. Chichester: Wiley.

Lamb, M.E., Orbach, Y., Hershkowitz, I., Esplin, P.W., & Horowitz, D. (2007). Structured forensic interview protocols improve the quality and informativeness of investigative interviews with children: A review of research using the NICHD Investigative Interview Protocol. *Child Abuse and Neglect, 31, 1201-1231*

Lamb, M. E., Sternberg, K. J., Esplin, P. W. (2000). Effects of age and delay on the amount of information provided by alleged sex abuse victims in investigative interviews. *Child Development, 71,* 1586-1596.

Lamb, M. E., Sternberg, K. J., Esplin, P. W., Hershkowitz, I., Orbach, Y., & Hovav, M. (1997). Criterion-based content analysis: A field validation study. *Child Abuse and Neglect, 21*, 255-264.

Masip, J., Sporer, S.L., Garrido, E., Herrero, C. (2005). The detection of deception with the reality monitoring approach: A review of the empirical evidence. *Psychology, Crime, and Law, 11,* 99-102.

McGinnis, D., & Roberts, P. (1996). Qualitative characteristics of vivid memories attributed to real and imagined experiences. *American Journal of Psychology, 109*, 59-77.

O'Neill, D. K., & Gopnik, A. (1991). Young children's ability to identify the sources of their beliefs. *Developmental Psychology, 27,* 390-397.

Orbach, Y., & Lamb, M.E. (2007). Young children's references to temporal attributes of allegedly experienced events in the course of forensic interviews. *Child Development, 78,* 1100-1120.

Pezdek, K., & Taylor, J. (2000). Discriminating between accounts of true and false events. In D. Bjorklund (Ed.), (pp. 69-91). *Research and theory in false-memory creation in children and adults*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

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

Porter, S., & Yuille, J. C. (1996). The language of deceit: An investigation of the verbal clues to deception in the interrogation context. *Law and Human Behavior, 20*, 443-458.

Raskin, D. C., & Esplin, P. W. (1991a). Assessment of children's statements of sexual abuse. In J. Doris (Ed.), *The suggestibility of children's recollections* (pp. 153-176). Washington, DC: American Psychological Association.

Raskin, D. C., & Esplin, P. W. (1991b). Commentary: Response to Wells, Loftus, and McGough. In J. Doris (Ed.), *The suggestibilty of children's recollections* (pp. 172-176). Washington, DC: American Psychological Association.

Roberts, K. P., & Blades, M. (1995). Children's discriminations of memories for actual and pretend actions in a hiding task. *British Journal of Developmental Psychology, 13,* 321-334.

Roberts, K. P., Lamb, M. E., & Randall, D. W. (1997, July). *Assessing the plausibility of allegations of sexual abuse from children's accounts*. Paper presented at the biennial meeting of the Society for Applied Research in Memory and Cognition, Toronto, Canada.

Santtila, P., Roppola, H., & Niemi, P. (1999). Assessing the truthfulness of witness statements made by children (aged 7-8, 10-11, and 13-14) employing scales derived from Johnson and Raye's model of reality monitoring. *Expert Evidence, 6*, 273-289.

Schooler, J. W., Gerhard, D., & Loftus, E. F. (1986). Qualities of the unreal. *Journal of Experimental Psychology: Learning, memory, and Cognition, 12*, 171-181.

Sporer, S. L. (1997). The less traveled road to truth: Verbal cues in deception detection in accounts of fabricated and self-experienced events. *Applied Cognitive Psychology, 11*, 373-397.

Sporer, S. L. (2004). Reality monitoring and the detection of deception. In P. A. Granhag & L. A. Strömwall (Eds.), The detection of deception in forensic contexts (pp. 64-102). Cambridge: Cambridge University Press.

Sporer, S. L., & Küpper, B. (1995). Realitätsüberwachung und die Beurteilung des Wahrheitsgehaltes von Erzählungen: eine experimentelle studie. *Zeitschrift fur Sozialpsychologie, 26*, 173-193.

Strömwall, L. A., Granhag, P.-A. (2005). Children's repeated lies and truths: effects on adults' judgments and reality monitoring scores. *Psychiatry, Psychology, and the Law, 12,* 345-356.

Suengas, A. G., & Johnson, M. K. (1988). Qualitative effects of rehearsal on memories for perceived and imagined complex events. *Journal of Experimental Psychology: General, 117*, 377-389.

Undeutsch, U. (1982). Statement reality analysis. In A. Trankell (Ed.), *Reconstructing the past: The role of psychologists in criminal trials* (pp. 27-56). Stockholm, Sweden: Norstedt & Sons.

Vrij, A., Edward, K., Roberts, K.P., & Bull, R. (2000). Detecting deceit via analysis of verbal and nonverbal behavior. *Journal of Nonverbal Behavior, 24,* 239-263.

Vrij, A. (2005). Criteria-Based Content Analysis: A Qualitative review of the first 37 studies. *Psychology, Public Policy, & Law, 11,* 3-41.

Vrij, A. (2008). *Detecting lies and deceit: Pitfalls and opportunities* (2nd ed.). New York, NY: John Wiley & Sons Ltd.

Wells, G. L., & Loftus, E. F. (1991). Commentary: Is this child fabricating? Reactions to a new assessment technique. In J. Doris (Ed.), *The suggestibility of children's recollections* (pp. 168-171). Washington, DC: American Psychological Association.

Yuille, J. C. (1988). The systematic assessment of children's testimony. *Canadian Psychology, 29*, 247-262.

Appendix 1

Abbreviated version of the Report Characteristics Questionnaire.

Criterion	Rating system	Examples
	General Characteristics	
Clarity	0 = dim and vague 1 = somewhat clear 2 = sharp, vivid, clear	0 = no specific information, e.g., "he did it" 2 = vivid details, easy to imagine
Complexity	0 = simple 1 = somewhat complex 2 = complex	0 = few details such as sequence of events or actions 2 = complex storyline, e.g., child's description includes clear sequence of events
Realism	0 = bizarre 1 = somewhat plausible 2 = plausible	0 = implausible or unlikely, e.g. "he took me in his spacecraft" 2 = could reasonably have happened
Order of Event	0 = confusing 1 = somewhat comprehensible	0 = order of events does not make sense or is not provided

	2 = comprehensible	2 = order of events makes sense and is clear
Event Detail	0 = vague 1 = somewhat detailed 2 = very detailed	0 = very few details, e.g., “it was gross” 2 = rich in detail, e.g., includes emotional and perceptual information
Event Duration[a]	1 = short 2 = long 5 = no judgment possible	1 = event was likely short in duration 2 = event was likely long, e.g., involved extended interaction between child and perpetrator
Tone[a]	1 = negative 2 = neutral 3 = mixed 4 = positive 5 = no judgment possible	1 = overall tone of the memory is negative, e.g. “I got scared” 4 = overall tone of the memory is positive
Setting[a]	1 = unfamiliar 2 = familiar 5 = no judgment possible	1 = event occurred in a place the child had never been before 2 = event occurred in a place the child is familiar with, e.g., at home

	Specific Characteristics[b]	
Perceptual – People (Visual Detail, Sound, Smell, Physical Sensation, Taste)	0 = absent 1 = weak presence 2 = present 3 = strongly present	Perceptual details about people, e.g., child describes aspects of individual's appearance, such as "he unbuttoned his *pants*"
Perceptual – Objects (Visual Detail, Sound, Smell, Physical Sensation, Taste)	As above	Details about objects, such as shape, sound, e.g. "the TV was making loud scratchy noises"
Actions	As above	Details about actions experienced by the child, e.g. "I was watching TV"
Spatial (Location, Arrangement of People, Arrangement of Objects, Environment)	As above	Specific spatial details, e.g. "one time he took me in his bedroom"
Temporal (Year, Season, Month, Day, Hour, Time)	As above	Specific details about when the event occurred, e.g., "they went out on New Years Eve"

Supporting Memories (Events Before, Events After, Events Between)	As above	Details about other events that anchor the main event, e.g., “I went on vacation the next day”
Affective information	As above	Child describes feelings experienced at the time, e.g., “I started crying”
Rehearsal (Covert, Overt)	As above	Evidence that the child has thought (covert) or talked (overt) about the event, e.g., “I thought about it for a long time after”
Cognitive Operations (Remembered Thoughts, Cognitive Operations)	As above	Descriptions of how the child thought at the time of the event, e.g., “I trusted him”

Notes.

[a] These criteria were coded and analyzed as categorical variables.

[b] 0 = absent; 1 = 1-5 lines; 2 = 6-30 lines; 3 = more than 31 lines.

[c] 0 = absent; 1 = 1-5 instances; 2 = 6-30 instances; 3 = more than 31 instances.

[d] These criteria were excluded from analyses.

Table 1 *Means (and standard errors) for individual criteria in Study 1*

	Age Group			
	Younger		Older	
Status	Confirmed	Doubtful	Confirmed	Doubtful
	General Characteristics [a,b]			
Clarity	1.88 (.15)	.78 (.14)	2.00 (.16)	2.00 (.19)
Complexity	1.63 (.20)	.33 (.19)	2.00 (.21)	2.00 (.25)
Realism	1.88 (.35)	2.78 (.33)	2.00 (.37)	2.00 (.44)
Order of Event	1.75 (.36)	1.11 (.34)	2.00 (.39)	2.00 (.46)
Event Detail	1.75 (.19)	.33 (.18)	2.00 (.21)	2.00 (.25)
	Specific Characteristics[c]			
Perceptual – People	1.75 (.17)	1.22 (.16)	2.29 (.18)	2.60 (.21)
Perceptual – Objects	1.13 (.19)	.78 (.18)	1.57 (.20)	1.80 (.24)
Actions	2.00 (.19)	1.33 (.18)	2.43 (.20)	2.60 (.24)
Spatial	1.50 (.19)	.56 (.17)	2.29 (.20)	2.40 (.23)
Temporal	1.38 (.25)	.89 (.23)	2.00 (.26)	2.40 (.31)
Supporting Memories	1.50 (.30)	.78 (.29)	2.43 (.33)	2.40 (.38)
Affective Information	.38 (.23)	.22 (.21)	1.00 (.24)	.80 (.29)
Rehearsal	.25 (.21)	.44 (.20)	1.25 (.22)	.80 (.26)
Thoughts/Cognitive Operations	1.25 (.22)	.56 (.21)	1.43 (.23)	2.00 (.28)

Notes.
[a] Based on a scale of 0 (low) – 2 (high).
[b] Duration, Setting, and Tone are not included here as these variables were categorical.
[c] Based on a scale of 0 = absent, 1 = weak presence, 2 = present, 3 = strongly present

Table 2 *Means (and standard errors) for individual criteria in Study 2*

	Age Group			
	Younger		Older	
Status	Confirmed	Doubtful	Confirmed	Doubtful
	General Characteristics [a, b]			
Clarity	1.42 (.17)	.50 (.20)	1.87 (.12)	1.80 (.26)
Complexity	1.25 (.18)	.50 (.22)	1.83 (.13)	2.00 (.28)
Realism	1.75 (.13)	1.75 (.16)	2.00 (.09)	1.80 (.20)
Order of Event	1.08 (.21)	.75 (.26)	1.87 (.15)	2.00 (.32)
Event Detail	1.08 (.19)	.50 (.23)	1.74 (.14)	2.00 (.30)
	Specific Characteristics[c]			
Perceptual – People	1.67 (.14)	1.57 (.18)	2.13 (.10)	2.40 (.22)
Perceptual – Objects	1.25 (.19)	.86 (.25)	1.74 (.14)	2.20 (.29)
Actions	1.58 (.16)	1.29 (.20)	1.96 (.11)	2.60 (.24)
Spatial	1.42 (.18)	1.29 (.23)	2.04 (.13)	2.60 (.28)
Temporal	1.25 (.20)	1.00 (.26)	1.96 (.14)	2.20 (.31)
Supporting Memories	.92 (.21)	.57 (.27)	1.91 (.15)	2.40 (.32)
Affective Information	.08 (.14)	<.01 (.18)	.52 (.10)	.40 (.21)
Rehearsal	.17 (.16)	.14 (.21)	.61 (.12)	.40 (.25)
Thoughts/Cognitive Operations	.67 (.16)	.29 (.22)	1.09 (.12)	1.00 (.25)

Notes.

[a] Based on a scale of 0 (low) – 2 (high).

[b] Duration, Setting, and Tone are not included here as these variables were categorical.

[c] Based on a scale of 0 = absent, 1 = weak presence, 2 = present, 3 = strongly present

Table 3

Means (and standard errors) for individual criteria in Study 3

	Age Group			
	Younger		Older	
Status	Confirmed	Doubtful	Confirmed	Doubtful
	General Characteristics [a, b]			
Clarity	1.00 (.20)	.58 (.19)	1.78 (.22)	1.70 (.21)
Complexity	1.55 (.19)	1.17 (.18)	1.22 (.21)	1.50 (.20)
Realism	1.36 (.19)	.92 (.18)	2.00 (.21)	1.80 (.20)
Order of Event	1.00 (.20)	.50 (.19)	1.89 (.22)	1.60 (.21)
Event Detail	1.00 (.22)	.58 (.21)	1.44 (.24)	1.30 (.23)
	Specific Characteristics[c]			
Perceptual – People	1.73 (.14)	1.42 (.15)	1.78 (.16)	1.70 (.15)
Perceptual – Objects	.82 (.12)	1.08 (.13)	1.00 (.14)	1.00 (.13)
Actions	1.73 (.11)	1.67 (.10)	2.00 (.12)	2.00 (.11)
Spatial	1.46 (.18)	1.33 (.17)	1.89 (.19)	1.50 (.18)
Temporal	1.46 (.17)	1.08 (.16)	1.89 (.19)	1.80 (.18)
Supporting Memories	1.64 (.30)	1.42 (.28)	1.78 (.33)	1.80 (.31)
Affective Information	.73 (.19)	.25 (.18)	.44 (.20)	.80 (.19)
Rehearsal	1.00 (.17)	.67 (.16)	.56 (.19)	.90 (.18)
Thoughts/Cognitive Operations	.55 (.19)	.42 (.18)	1.11 (.21)	.90 (.20)

Notes.

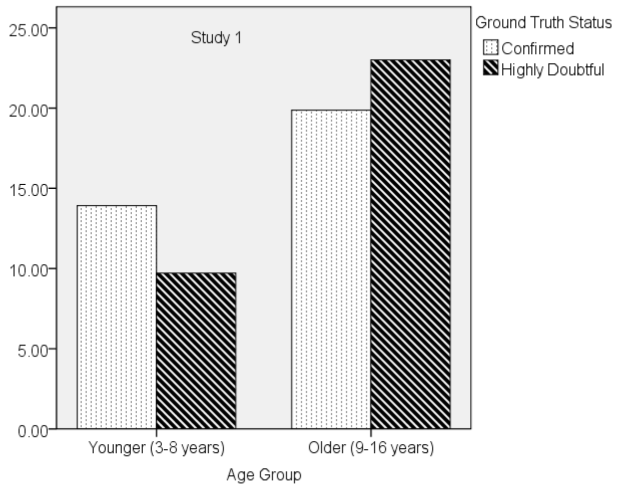
[a] Based on a scale of 0 (low) – 2 (high).

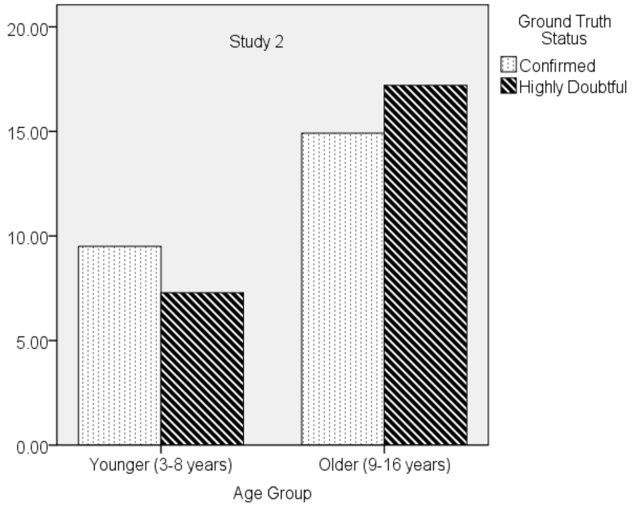
[b] Duration, Setting, and Tone are not included here as these variables were categorical.

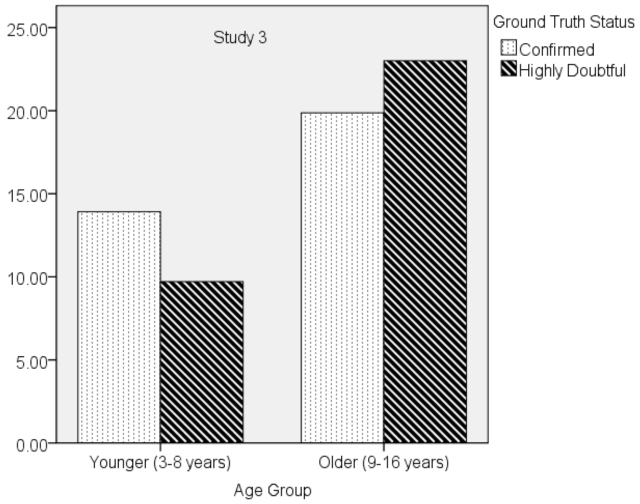
[c] Based on a scale of 0 = absent, 1 = weak presence, 2 = present, 3 = strongly present

Figure caption

Total RCQ score by age group and ground truth status


Study 1
Ground Truth Status
Confirmed
Highly Doubtful
25.00
20.00
15.00
10.00
5.00
0.00
Younger (3-8 years)
Older (9-16 years)
Age Group


Study 2
Ground Truth Status
Confirmed
Highly Doubtful
20.00
15.00
10.00
5.00
0.00
Younger (3-8 years)
Older (9-16 years)
Age Group


Study 3
Ground Truth Status
Confirmed
Highly Doubtful
25.00
20.00
15.00
10.00
5.00
0.00
Younger (3-8 years)
Older (9-16 years)
Age Group