Careful calculation or a leap of faith?:

A field study of the translation of CBCA ratings to final credibility judgements

Running head: CBCA scores and credibility judgements

Key words: deception, child witnesses, CBCA, field investigation

Abstract

This field experiment investigated the influence of CBCA criteria ratings on ultimate decision accuracy regarding the credibility of children's statements of sexual abuse. Following a selection procedure, based on case facts independent of statement quality, 21 truthful accounts and 10 fabricated accounts of 6 to 17 year olds were analysed. Two experts rated the presence of the CBCA criteria and made overall credibility judgements for each statement. Rater one achieved an overall hit rate of 84% (95% for truthful statements and 60% for fabricated statements) and rater two a hit rate of 81% (81% for both truthful and fabricated statements) but the raters did not always agree. The CBCA criteria appeared more often in the truthful statements compared to the fabricated statements. Additional factors that influenced raters' credibility judgements, besides CBCA scores, are discussed.

Careful calculation or a leap of faith?: The translation of CBCA ratings to final credibility judgements

The most commonly used procedure for assessing the credibility of accounts, based on speech content, is Statement Validity Assessment (SVA). This technique evolved in order to evaluate, systematically, the veracity of children's allegations of sexual abuse (Undeutsch, 1984; Arntzen, 1983). SVA assessments are accepted as evidence (concerning children's credibility) in some US courts (Ruby & Brigham, 1997) and in criminal courts in several European countries such as Germany (Köhnken, 2002) and the Netherlands (Lamers-Winkelman & Buffing, 1996). A review of the first 37 studies into the efficacy of the technique (Vrij, 2005) has highlighted mixed findings. In general, methodologically sound experiments have supported the technique as an improvement not only on simple 'gut reaction' credibility judgements but also on judgements based on unreliable nonverbal cues. SVA comprises four stages, a collection of relevant case information, a semi-structured interview, Criteria-Based Content Analysis (CBCA) of a transcript of the interview and referral to a Validity Checklist.

The major component of the SVA technique, CBCA, focuses on specific content characteristics which, if present in a statement, support the Undeutsch hypothesis that the account is based on genuine personal experience (i.e. that it is truthful). This investigation will only focus on the CBCA part of the overall SVA technique.

The present field investigation was designed not only to assess the feasibility of using CBCA (using materials for which it was originally designed) but further to look at the relationship between CBCA coders' ratings of the criteria and their final

judgements of truthfulness. Are CBCA scores clearly influencing overall credibility judgements or are there more subjective decision making processes at play?

As Vrij's (2005) review highlighted, the vast majority of CBCA studies concerning children's statements that have calculated differences in CBCA scores between truthful and fabricated statements, have been laboratory based (e.g. Akehurst, Köhnken & Höfer, 2001; Ruby & Brigham, 1997; Sporer, 1997; Vrij, Akehurst, Soukara & Bull, 2002). Laboratory studies have found evidence that CBCA is a 'successful' technique for judging the veracity of statements (Akehurst, Köhnken & Höfer, 1995; Köhnken, Schimossek, Aschermann & Höfer, 1995; Sporer, 1997; Vrij, Akehurst, Soukara & Bull, 2002), nonetheless, their inability to mimic sexually abusive experiences means that they lack ecological validity. Trankell (1972) and Undeutsch (1984) have argued that such lab studies cannot experimentally manufacture statements such that they imitate *real* truths and lies.

Establishing ground truth: Field investigations into the CBCA technique, using real statements from alleged child victims, offer a more ecologically valid alternative. However, the difficulty in determining whether a child's account of a sexually abusive experience is unequivocally true or false is crucial. Esplin, Boychuk and Raskin (1988), in the first published field CBCA evaluation, found that CBCA criteria strongly differentiated between the statements they labelled 'confirmed' and 'not confirmed'. However, Wells and Loftus (1991) criticized Esplin et al (1988) for a number of reasons. They made reference to their use of only one evaluator, the fact that independent case facts were not utilized in the establishment of ground truth[1] and that the findings could be as a result of an age effect, as the children in the

[1] Similarly more recent studies by Craig, Scheibe, Raskin, Kircher and Dodd (1999) and Parker and Brown (2000) failed to adequately establish the ground truth of their target statements in field evaluations of CBCA.

'confirmed' group were older than those in the 'not confirmed' group and were therefore more likely to give fuller and richer accounts regardless of credibility status.

In the past researchers have neglected to use independent case facts to categorise their cases into those where sexual abuse has occurred and those where it has not[2]. Instead they have used grouping criteria that include persistent denial by the accused and judicial dismissal for 'fabricated' statements and confessions by the suspect and guilty verdicts in court for the 'truthful' statements. These statement dependent classification tools are not sufficiently robust for establishing ground truth. For instance, the fact that a judge dismisses a case may be as a result of the poor quality of the child's statement and the fact that s/he was unconvincing which would in turn lead to a low CBCA score. Moreover, once a suspect learns that a child has given a poor statement s/he is less likely to plead guilty. Wells and Loftus (1991) labelled this problem that of 'classification circularity' (p.170). The most important factor to bear in mind here is that when ground truth is established, criteria for classification must be independent of statement quality and based on independent case facts (e.g. supporting physical evidence, videorecorded evidence and so on). The current study fulfilled this requirement by developing a conservative criteria set which was used to classify the statements.

Notwithstanding the value of this investigation with regard to the methodological issue outlined above, other aims, yet to be addressed in the extant literature, were to investigate the effect of CBCA criteria ratings on the final decisions made by raters and to explore the nature of disagreements (if any) between two raters.

From rating criteria to overall judgements of credibility: In the past, investigators have reported differences in CBCA ratings between objectively

[2] A notable exception being that of Lamb, Sternberg, Esplin, Hershkowitz, Orbach and Hovav (1997) however this study did not utilise the full set of CBCA criteria

'truthful' and 'fabricated' accounts (Akehurst, Köhnken & Höfer, 2001; Sporer, 1997; Vrij, Akehurst, Soukara & Bull, 2002). In Vrij's (2005) summary of the occurrence of CBCA criteria in field studies, he highlights the well replicated finding that total CBCA scores are higher for truthful compared to fabricated accounts. These criteria-related findings have generally been reported in conjunction with hit rates (i.e. how often a CBCA rater has made overall accurate decisions with regard to the credibility of a sample of statements). However, whilst researchers often make the link between CBCA scores and objective truth status (e.g. using discriminant analyses) they have never made the link between CBCA scores and judgements. Is it the case that a high total CBCA score *always* translates to a judgment of truthfulness - probably not, and nor should it according to pioneers of the technique who suggest that raters should not use specific thresholds to make their decisions i.e. they should not use objective *decision rules* (e.g. "if there are more than five criteria present then I must rate this statements as truthful"). They suggest instead that raters should consider for example the distribution of criteria, the impact of particular criteria (e.g. accurately reported details misunderstood, see the discussion section), characteristics of the interviewee and interviewer and motives for false reporting (the latter being factors in the Validity Checklist) (Steller, & Köhnken, 1989). In line with this guidance from the originators of the technique, we did not recommend any decision rules relating to the criteria to our raters. However, as a consequence of this, the 'leap' from CBCA score to final credibility judgement becomes, to a certain extent, subjective and opens up the possibility that raters will disagree on their final decision for the same statement as they are influenced to differing extents by different factors after they have completed their CBCA analysis.

When raters disagree: When used in real life investigations, usually only one expert in the CBCA technique will be asked to rate the credibility of a child's account. To assess reliability issues, this experiment allowed for the comparison of two independent raters' evaluations of the same statements. Our focus was not only on inter-rater reliability for individual criteria (as for previously reported research in this area e.g. Akehurst, Köhnken & Höfer, 1995; Gödert, Gamer, Rill & Vossel, 2005; Köhnken, Schimossek, Aschermann & Höfer, 1995; Sporer, 1997; Vrij, Akehurst, Soukara & Bull, 2002) but also on the overall true/lie judgement of our raters and whether they were in agreement. We recorded all instances of disagreement between them and compared total CBCA scores for these cases. Two scenarios must be considered; the first where raters disagree over a truthful statement and the second where they disagree over a fabricated statement. In the first case, in order to support the use of the CBCA technique used on it's own (as it has been in all previous experimental studies) i.e. without supporting case information or the consideration of the Validity Checklist it should be that the erroneous rater, that has failed to accurately label a truthful child, has failed to spot key CBCA criteria that have successfully persuaded the accurate rater to come to the correct conclusion and therefore the inaccurate evaluator should record fewer CBCA criteria than the accurate evaluator. In the second case, in order to support the use of the CBCA technique used on it's own, it should be that the erroneous rater, that has incorrectly labelled a lying child as a truth-teller, has over-inflated his/her CBCA ratings and should therefore record more presence of CBCA criteria than the accurate evaluator. However, as mentioned above, it is unlikely to be as straightforward as this and we predict that additional influences will come to bear on any differences of opinion

between evaluators. These findings have repercussions for the real world application of this technique and have not been reported in previous CBCA studies.

Method

Classification of statements. Initial selection was based on computer searches of police records of allegations of sexual offences against children, to one police force, in central England between 2003 and 2005, using the key words 'detected' (for potentially truthful reports) and 'no crime' (for potentially unfounded reports). In total 175 cases were targeted and the computer and paper files on these cases were scrutinised. The investigative material examined for each case included verbatim copies of medical reports, videorecordings of the alleged incidents filmed by the alleged offenders, CCTV footage of alleged perpetrators, summaries/transcripts of suspect interviews, sworn witness/police statements and child investigation logs. These cases were then classified as truthful, fabricated or unclassifiable based on a selection of criteria.

The criteria evolved with the selection process. Initially we looked for criteria highlighted by Lamb et al (1997) and Horowitz, Lamb, Esplin, Boychuk, Reiter-Lavery and Krispin (1996) (e.g. medical evidence, suspect statements, confessions and verdicts at court) however we added to these as we scrutinised cases that included very persuasive evidence, independent of statement quality e.g. a videorecording, made by the alleged offender, of the event for truthful cases and documentary evidence to prove the suspect was elsewhere for the fabricated statements (e.g. the perpetrator attending a meeting with his social worker at the time of the alleged incident).

All truthful statements satisfied at least three of the following criteria and therefore had to include at least one of those marked with *. Those marked with *

were considered strong indicators of truthfulness and were, crucially, independent of the quality of the child's statement.

(i) Medical evidence e.g. DNA evidence*

(ii) Videorecording of the event filmed by the offender*

(iii) Corroboration from another independent victim or witness (not known to the alleged victim)*

(iv) Guilty verdict at court

(v) Confession from suspect (prior to plea negotiation)

Questions have been raised about the use of confessions to establish 'ground truth' as authors such as Steller and Köhnken (1989) argue that there are problems associated with this. They comment that in Germany a confession may or may not be obtained depending on the judgement of the psychological expert as to the veracity of the child's statement thus creating a circular classification process. Further, authors such as Gudjonsson (1992), Leo (1996) and Kassin (1997) have stated that as a result of a growing number of documented cases, there are justifiable concerns about the risk of false confessions. In view of this in the current study, confessions, whilst noted, were not considered independent case facts.

All fabricated statements satisfied at least three of the following criteria and therefore had to include at least one of those marked with *. Those marked with * were considered strong indicators of fabrication and were, crucially, independent of the quality of the child's statement.

(i) Contradictory evidence (e.g. evidence that indicated that an offence could not have been committed in the way that it was alleged i.e. descriptions from the children that were contrary to the laws of nature)*

(ii) Proof that the incident could not have occurred when it was alleged to have occurred* (e.g. CCTV or validated documentary evidence or photographic evidence to prove the suspect was elsewhere at the time of the alleged incident)

(iii) Retraction by witness at a later date (comprehensive and plausible)

(iv) Persistent not guilty plea from alleged offender.

With reference to criterion (ii) above, it is noted that a particular allegation might have been based on an event that did happen but at another time or did happen but with a different perpetrator. However, in a credibility assessment for a court case, these allegations would need to be assessed for whether they had happened at the alleged time and with the alleged perpetrator. With the proof that the alleged offender could not have committed the offence at the alleged time there must have been some elements of fabrication in the statements for which this criterion was met.

Sample: Following the vigorous selection procedure from the original 176 cases; 68 were removed as the recording of the forensic interview was not available or recordings were too poor to allow for transcription, 43 were removed as they were unclassifiable and 21 were removed as, although they fulfilled some criteria for truthful or fabricated statements there were no independent case facts to help with classification. Furthermore, it was ensured that none of the truthful cases included any of the criteria that were used to classify fabrication (e.g. a case was dismissed if, even though there was corroborating evidence and a guilty verdict in court, it included a retraction by the witness) and none of the fabricated cases contained any of the truthful criteria (e.g. a case was dismissed if, even though there was contradictory evidence and a persistent guilty plea from the alleged offender, it included corroboration from an independent witness). This resulted in the dismissal of a

further 10 cases. Three cases were removed for miscellaneous reasons (e.g. an interpreter was present). This resulted in 31 statements being used.

All allegations were of a sexual nature. Of the 31 cases, 14 involved intra familial alleged perpetrators (step-fathers, fathers, siblings, cousins), 14 extra familial (persons known to the children but not family members) and 3 cases involved persons not known to the children.

The statements comprised 10 fabricated accounts from 2 males and 8 females whose ages ranged from 8 years to 15 years (M = 12.60 years, SD = 1.95 years) and 21 truthful statements, 3 were given by males and 18 by females. Ages for the truthful group ranged from 6 years to 16 years (M = 10.05 years, SD = 3.07 years). The children who gave the fabricated statements were significantly older than those who gave the truthful statements ($t(29) = 2.39$, $p < 0.05$). However there was no significant correlation between age and total CBCA score ($r(29) = 0.03$, n.s).

Interviews: Each of the selected interviews had been conducted by specialist UK police officers, male and female, who had been specifically trained to interview in accordance with the guidelines of Achieving Best Evidence in Criminal Proceedings for Vulnerable Witnesses including Children (Home Office, 2007). All interviews comprised the phased or step-wise interview protocol (i.e. rapport building, free narrative, questioning and closure) as advocated for SVA (see Yuille, 1988). All statements were transcripts of the first police/social services interviews with the children following disclosure. In the UK, it is standard practice to video record interviews with children making allegations of sexual abuse and from each video recording a word for word statement was produced including all child witness and police officer utterances. There was no significant difference between the word

lengths of the truthful statements (M = 3,464 words, SD = 1, 557 words) and those of the fabricated statements (M = 5, 160 words, SD = 2, 357 words),(t (29)= 1.63, n.s).

In completing the transcripts all identifying material (e.g. names/locations) was removed. Statements were passed to two CBCA coders who were blind with regard to the veracity of the statements. Both raters evaluated all the statements independent of one another. When conducting CBCA, evaluators should account for the age of the child witness, assuming that the child has a 'normal' level of social, cognitive and linguistic development. That is, they should adjust their expectation with regard to the presence of criteria expecting fewer criteria in the truthful statement of a younger child compared to the truthful statement of an older child (see Steller & Köhnken, 1989). As such, 15 transcripts of the 6-10 year old children were handed to CBCA raters first and were all coded before 16 transcripts of the 11-16 year old children were coded. Evaluators were aware of the age of the child for each statement. The raters were told that the sample of statements comprised both truthful and fabricated accounts but that there was not a 50/50 split. Thus, they were blind to the base rate.

CBCA rater training**:** In order to ensure a standardized and consistent approach when rating each criterion, extensive CBCA rater training was undertaken. Both raters initially received a booklet providing detailed descriptions and examples of the 18 criteria and were instructed to read these thoroughly. It should be noted that criterion number 19 (Details characteristic of the offence) from the original 19 criteria described by Steller and Köhnken (1989) was not used in this study as raters did not have the relevant knowledge. This criterion can only be accurately coded by professionals with lengthy professional experience of investigations of child sexual abuse.

The first training session was conducted a few days later. Each criterion was discussed with an expert rater and the detection of such criteria in interview transcripts (of an unrelated event) were rehearsed. The second training session focused on the use of 5-point Likert scales to rate the presence of each criterion using example transcripts of different incidents. Köhnken (2004) pointed out that 5-point Likert scales were desirable for investigations such as these as they afford more sensitivity, to differences between truthful and fabricated statements, than the 3-point scales often used in research (e.g. Blandón-Gitlin, Pezdek, Lindsay & Hagen, 2009; Strömwall, Bengtsson, Leander & Granhag, 2004; Tye, Amato, Honts, Devitt & Peters, 1999). Both the trainee raters and the expert rater then rated a number of transcripts in respect of the content criteria and results were compared. When no disagreements of more than one point on the scale occurred, the training was concluded. The rating scales and training procedure had been adapted from another study where inter-rater agreement for criteria had been established (Akehurst, Köhnken & Höfer, 2001).

CBCA rating measures: Each CBCA rater filled in a form relating to each of the statements. Each CBCA criterion was rated on a 5 – point rating scale, ranging from 1 = 'absent' to 5 = 'strongly present'. Following the evaluation of each of the 18 criteria, each rater was required to make an assessment regarding whether the statement was true or false.

Results

Inter-rater reliability: Pearson product-moment correlations were calculated between rater one's scores and rater two's scores for each of the eighteen CBCA criteria used as well as between the total CBCA score for each rater (see Table 1).

Anson, Golding and Gully (1993) suggested that r values concerning inter-rater reliability of .5 and higher were adequate, r values of .6 and higher were good and values of .75 and higher were excellent. It can be seen that 8 out of 18 criteria showed excellent reliability, 6 others achieved good reliability, 2 adequate reliability and 2 did not reach adequate reliability (related external associations and self deprecation). Inter-rater reliability for the total CBCA scores was excellent (r = .91). When testing the effectiveness of techniques it is valuable to test both the reliability and validity of those measures. However, in real life, it would generally be the case that only one rater assessed the presence of the CBCA criteria. With this in mind, all criteria (regardless of inter-rater reliability) were maintained for subsequent analyses and rather than calculating a mean score across raters each evaluator was considered separately for the remainder of the paper.

Overall judgement accuracy: So that results could be compared to those of similar studies, judgement accuracy was computed for each rater.

Rater one achieved an overall hit rate of 84% with 95% correct classification of truthful statements and 60% correct classification of fabricated statements. A series of two-tailed binomial tests indicated that she accurately classified truthful statements at a level significantly better than chance ($p < .001$) however she did not perform better than chance level when classifying fabricated statements. A 2 (accuracy) x 2 (truth status) X^2 analysis revealed a significant relationship between truth status and accuracy (X^2 (1, df = 1) = 10.22, $p < 0.005$) showing that rater one was significantly more accurate at classifying truthful statements compared to fabricated statements[3].

[3] Moreover, signal detection analysis for response bias (where the signal was deemed to be 'truth') revealed that rater one held a truth bias ($B''_D = -.22$)

Rater two achieved an overall hit rate of 81% with 81% correct classification of truthful statements and 81% correct classification of fabricated statements. A series of two-tailed binomial tests indicated that she performed significantly better than chance level for truthful statements ($p < .01$) and for fabricated statements ($p < 0.01$). A 2 (accuracy) x 2 (truth status) X^2 analysis revealed no significant relationship between truth status and accuracy (X^2 (1, df = 1) = 2.32, n.s.) showing that for rater two there was no significant difference in accuracy for truthful and fabricated statements[4].

Discriminating objective truths from lies: To determine whether total CBCA scores were predictive of *objective* truth status, discriminant analyses were conducted for each rater with actual truth status of statements as the dependent grouping variable and the total CBCA scores as the predictor variable (individual criteria ratings were not used as predictor variables due to the relatively small sample size). For rater one, this yielded a significant discriminant function (Wilk's $\lambda = .84$, $X^2 = 5.03$, df = 1, $p < .05$) and an overall classification rate of 68% (67% for truthful statements and 70% for fabricated statements). For rater two, the same analysis yielded a significant discriminant function (Wilk's $\lambda = .86$, $X^2 = 4.20$, df = 1, $p < .05$) and an overall classification rate of 68% (67% for truthful statements and 70% for fabricated statements). These findings indicate that, for both raters, total CBCA score was able to discriminate between truthful and fabricated accounts 68% of the time.

Discriminating judgements: Further discriminant analyses were conducted this time with *judgements* of credibility as the grouping variable. For rater one, this yielded a significant discriminant function (Wilk's $\lambda = .73$, $X^2 = 8.90$, df = 1, $p < 0.01$) with a classification accuracy of 81% (79% for truthful statements and 86% for

[4] Moreover, signal detection analysis for response bias (where the signal was deemed to be 'truth') revealed that rater two held a small lie bias ($B''_D = .17$)

fabricated statements). For rater two, the discriminant function was significant (Wilk's $\lambda = .46$, $X^2 = 20.65$, df = 1, $p < .001$) with a classification accuracy of 81% (79% for truthful statements and 83% for fabricated statements) In sum, the ratings made by raters one and two predicted their credibility judgements 81% of the time.

Effectiveness of individual criteria: Two ANOVAs (one for each rater) were conducted on the data with the truth status of statements as the independent variable (truthful and fabricated). The dependent variables were the scores for each criterion and the total CBCA scores.

For raters one and two, the criteria admitting lack of memory, unstructured production and contextual embedding appeared significantly more often in truthful compared to fabricated accounts. The total CBCA scores for both raters were also significantly higher for truthful compared to fabricated accounts. Thus these three criteria were the most *effective* at discriminating between truthful and fabricated statements. Means, F values and Cohen's d values for these criteria can be seen in Table 2.

Influence of individual criteria: Two further ANOVAs (one for each rater) were conducted this time with the rating of a statement (truthful or fabricated regardless of accuracy) as the independent variable and criteria ratings as the dependent variables. For rater one, contextual embedding, reproduction of conversation, description of interactions, unstructured production and logical structure were present significantly more often in those statements judged as truthful compared to those judged as fabricated. For rater two, unstructured production, contextual embedding, description of interactions, logical structure, quantity of detail, reproduction of conversation, unexpected complications and reports of own mental state appeared significantly more often in the statements judged as truthful compared

to those that were judged fabricated. Thus these criteria were the most *influential* during CBCA analysis for this investigation. Means, F values and Cohen's d values for these criteria can be seen in Tables 3 and 4.

When raters disagreed: The cases where rater one and rater two disagreed with regard to their final judgement of credibility were scrutinised. The raters disagreed about the credibility of 7 of the 31 children (22% of the statements evaluated for this study). In six cases, rater one thought the children were telling the truth and rater two thought they were lying. Of these cases rater one was correct 50% of the time. The reverse was true for one case (i.e. rater one thought a child was lying and rater two thought she was telling the truth). In this case rater one was accurate. A 2 (rater) x 2 (decision) X^2 analysis revealed a significant difference between the raters' decisions (X^2 (1, df = 1) = 5.9, $p < 0.05$. The total CBCA scores for each rater for the statements where they disagreed showed that for six of the seven cases in question CBCA scores did not differ significantly (for three cases there was no difference at all in CBCA total score and for three cases only a one point difference separated them). For the final case, a fabricated statement, rater one made an incorrect judgement of 'truthful' and her total CBCA score was 41 and rater two made an accurate judgement of 'fabricated' and her total CBCA score was 35.

Discussion

The two expert raters who evaluated statements for this investigation both distinguished between truthtellers and liars at a level significantly above chance. The results provided support for the Undeutsch hypothesis that accounts of actually experienced events (truthful) would contain more CBCA criteria than those that were fabricated. These findings support those of previous research both in the field and in

the laboratory (for a review see Vrij, 2005). However, when results were more thoroughly scrutinised a common theme emerged.

Discrimination: Objective truths and subjective judgements: Discriminant analyses were performed to investigate the predictive nature of the total CBCA scores for firstly, the objective truth status of the statements and secondly, the decisions of the raters. For both raters, total CBCA scores could accurately predict the objective truth status of the accounts only 68% of the time (67% for truthful accounts and 70% for fabricated accounts). Total CBCA scores accurately predicted raters' *decisions* 81% of the time (79% for truthful accounts and 83% for fabricated accounts).

As the total CBCA scores were only predictive of objective truthfulness for 68% of cases, raters must have utilised other factors to classify statements. In some cases where CBCA totals were relatively small, raters correctly chose to label a statement truthful and conversely in other cases, where CBCA totals were relatively high, raters chose to label a statement fabricated. What led them to override, on some occasions but not others, their CBCA scores in favour of an accurate (or inaccurate) decision? As evaluators were not provided with any more information than the statements themselves, and the ages of the interviewees, the answer must lie to some extent in the weight that evaluators attached to the presence of certain criteria. This is best illustrated with an example from one of our statements. The criterion, 'accurately reported details misunderstood' never occurred in our fabricated statements and only appeared twice in two of our truthful statements. This endorses a previous finding that this criterion seldom occurs in statements (Vrij, 2000). However, it can be a very powerful criterion. The raters in the current study provided anecdotal evidence that when this criterion did occur it added considerable weight to the rating of a

statement as truthful. Here is an example, taken from a truthful statement, of a 6 year old's description of a penis;

Child It was a whitey, creamy colour….

P.C. Yeah….

Child …..I seen the bones and everything as well

P.C. Saw the bones? Oh!

Child These black things

P.C. Black things

Child …yeah, the ones like that (points to veins on hand)

P.C. Oh, you mean the veins

In the authors' view herein lies the advantage of verbal cues over nonverbal cues when judging the credibility of children. Where a child is observed at interview displaying a great deal of nonverbal movement this might be due to nervousness, excitement, copying the movements of the interviewer or simply needing to go to the toilet, rather than lying! Thus nonverbal cues can be very misleading. However, when we consider the verbal content of a statement, a child who is talking about an event which s/he has not witnessed cannot include details that are not available from his or her knowledge base. Yet, if a description is given by a child who clearly does not understand what it is s/he is describing this must unequivocally lead to the assumption that the event was actually witnessed.

A particularly powerful section of text is likely to be only one of many factors that influence CBCA evaluators. Our findings indicate that they are basing their decisions on more than simply the frequency of the content criteria. For example, in the current investigation, rater one exhibited a truth bias. This has been documented in previous work surrounding the effectiveness of CBCA (e.g. Akehurst, Köhnken, &

Höfer, 2001) and may be as a result of striving to seek out cues to truthfulness alone (i.e. not searching for lie signs as well). Rater two did not hold such a truth bias and this may have accounted for the differences between some of the raters' judgements. An important applied issue concerns the possibility that two raters, rating the same statement independently, could be influenced so strongly, for various and differing reasons, that they come up with differing overall credibility ratings.

Rater disagreement**:**. For the current study the raters disagreed 22% of the time. Those differences would have huge consequences in real life, so why might they have occurred? One answer might be that raters assessed the frequency of criteria in opposing ways. However, results show a striking similarity in total CBCA scores between raters for the statements where there were disagreements. Scores were within one point of one another for six of the seven statements where there were differences in final credibility assessment. This again highlights that subjectivity and additional influences must account for the differing final decisions of our raters who were in strong agreement on total CBCA scores.

Although not utilized in the current study, the application of the Validity Checklist, may have allowed the raters to contemplate and examine any additional factors of the investigation in a more systematic manner and may have helped them to agree more often on their final assessments. This Checklist (which includes factors relating to characteristics of the interviewee and interviewer and motives for false reporting), whilst likely to reduce some subjectivity, is not however exhaustive. Whether decisions are more accurate when based on simply the application of CBCA or when external factors are more systematically accounted for, remains to be seen. A preliminary study by Gumpert and Lindblad (1999) suggested that using the Validity Checklist did not make a significant difference to final ratings of credibility.

Notwithstanding this research evidence, we recognize that the CBCA technique was not designed for use in isolation and guidelines pertaining to the entire SVA technique (Raskin & Esplin, 1991; Steller & Köhnken, 1989) make clear that judgements should be made in the context of all known case facts and all other evidence which was not possible for this study.

The effectiveness and influence of individual criteria: For both raters, admitting lack of memory, unstructured production and contextual embedding were the only criteria to appear significantly more often in the *actually* truthful compared to *actually* fabricated accounts. However, when we studied the criteria most often appearing in the statements *judged* as truthful (compared to those *judged* as fabricated) a different picture emerged. Both raters were influenced by contextual embedding and unstructured production which stood them in good stead for making accurate decisions[5]. However, they were also influenced by reproduction of conversation, description of interactions and logical structure (for raters one and two) as well as quantity of detail, unexpected complications and own mental state[6] (for rater two). So again, the same picture emerges. Not only were raters recording very similar total CBCA scores, but they were also influenced by the same criteria (including helpful and unhelpful criteria) and still they came to differing overall decisions for 22% of the cases.

In sum, decisions that are made in the real world by legal professionals and psychologists are not done so in a vacuum. Personal and professional experience will play a role in decision making with regard to children's allegations of sexual abuse. Whilst techniques for lie detection should aim to minimise these influences and biases

[5] These criteria *were* present significantly more often in truthful compared to fabricated statements in this sample

[6] These criteria *were not* present significantly more often in truthful compared to fabricated statements in this sample

it is unrealistic to believe that they can ever be negated. The current seemingly high overall accuracy score for raters is somewhat overshadowed, in our view, by the discrepancy in overall credibility decision for a significant number of the statements and follow up research will aim to investigate further CBCA evaluators reasoning behind their final judgements of credibility. What other mediating factors are considered by evaluators, over and above their CBCA ratings? Until this issue is addressed via further methodologically sound and ecologically valid research, the subjective nature of the leap between CBCA criteria coding and final credibility judgement calls into question the reliability of the technique.

References

Akehurst, L., Köhnken, G., & Höfer, E. (2001). Content credibility of accounts derived from live and video presentations. *Legal and Criminological Psychology, 6*, 65-83.

Arntzen, F. (1983). *Psychologie der Zeugenaussage: Systematik der Glaubwürdigkeitsmerkmale*. Munich: Beck.

Anson, D.A., Golding, S.L., & Gully, K.J. (1993). Child sexual abuse allegations: Reliability of criteria-based content analysis. *Law and Human Behavior, 17,* 331 – 341.

Blandón-Gitlin, I., Pezdek, K., Lindsay, D.S., & Hagen, L. (2009). Criteria-Based Content Analysis of true and suggested accounts of events. *Applied Cognitive Psychology, 23*, 901-917.

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., Boychuk, T., & Raskin, D.C. (1988, June). *A field validity study of Criteria-Based Content Analysis of children's statements in sexual abuse cases.* Paper presented at the NATO Advanced Study Institute on Credibility Assessment. Maratea, Italy.

Gödert, H.W., Gamer, M., Hill, H.G., & Vossel, G. (2005). Statement Validity Assessment: Inter-rater reliability of Criteria-Based Content Analysis in the mock-crime paradigm. *Legal and Criminological Psychology, 10 (2),* 225 – 242.

Gudjonsson, G.H. (1992). *The psychology of interrogations, confessions and testimony*. Chichester: Wiley.

Gumpert, C.H., & Lindblad, F. (1999). Expert testimony on child sexual abuse: A qualitative study of the Swedish approach to statement analysis. *Expert Evidence, 7,* 279-314.

Home Office (2007). *Achieving best evidence in criminal proceedings: Guidance for vulnerable or intimidated witnesses including children.* London: Home Office.

Horowitz, S.W., Lamb, M.E., Esplin, P.W., Boychuk, T.D., Reiter-Lavery, L, & Krispin, O. (1996). Establishing ground truth in studies of child sexual abuse. *Expert Evidence, 4,* 42-52.

Kassin, S.M. (1997). The psychology of confession evidence. *American Psychologist*, *52*, 221-233.

Köhnken, G. (2004). Statement Validity Analysis and the 'detection of truth'. In P.A. Granhag & L.A. Strömwall (Eds.), *The Detection of Deception in Forensic Contexts* (pp. 41-63). Cambridge, UK: Cambridge University Press.

Köhnken, G. (2002). A German perspective on children's testimony. In H.L. Westcott, G.M. Davies, & R.H.C. Bull (Eds.), *Children's testimony: A handbook of psychological research and forensic practice* (pp.233-244). Chichester: Wiley.

Köhnken, G., Schimossek, E., Aschermann, E., & Höfer, E. (1995). The cognitive interview and the assessment of the credibility of adult's statements. *Journal of Applied Psychology*, *80*, 671-684.

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 & Neglect*, *21*, 255-264.

Lamers-Winkelman, F., & Buffing, F. (1996). Children's testimony in the Netherlands: A study of Statement Validity Analysis. In B.L. Bottoms & G.S. Goodman (1996), *International perspectives on child abuse and children's testimony* (pp. 45-62). Thousand Oaks, CA: Sage.

Leo, R.A. (1996). Inside the interrogation room. *Journal of Criminal Law and Criminology, 86*, 266-303.

Parker, A.D., & Brown, J. (2000). Detection of deception: Statement Validity Analysis as a means of determining truthfulness or falsity of rape allegations. *Legal and Criminological Psychology, 5,* 237-259.

Raskin, D.C., & Esplin, P.W. (1991). Assessment of children's statements of sexual abuse. In J. Doris (Ed.), *The suggestibility of children's recollections* (pp. 153-165). Washington, DC: American Psychological Association.

Reinhard, M., Burghardt, K., Sporer, S., & Bursch, S. (2002). Alltagsvorstellungen über inhaltliche Kennzeichen von Lügen. *Zeitschrift für Sozialpsychologie, 33 (3),* 169 – 180.

Ruby, C.L., & Brigham, J.C. (1997). The usefulness of the Criteria-Based Content Analysis technique in distinguishing between truthful and fabricated allegations. *Psychology, Public Policy and Law, 3*, 705-737.

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

Steller, M. (1989). Recent developments in statement analysis. In J.C. Yuille (Ed.), *Credibility assessment* (pp. 135-154). Dordrecht: Kluwer Academic Publishers.

Steller, M., & Köhnken, G. (1989). Criteria-Based Statement Analysis. In D. Raskin (Ed.), *Psychological methods for criminal investigation and evidence* (pp. 217-245). New York: Springer-Verlag.

Strömwall, L.A., Bengtsson, L., Leander, L., & Granhag, P.A. (2004). Assessing children's statements: The impact of a repeated experience on CBCA and RM ratings. *Applied Cognitive Psychology, 18*, 653-668.

Trankell, A. (1972). *Reliability of evidence*. Stockholm: Beckmans.

Tye, M.C., Amato, S.L., Honts, C.R., Devitt, M.K., & Peters, D. (1999). The willigness of children to lie and the assessment of credibility in an ecologically relevant laboratory setting. *Applied Developmental Science, 3(2)*, 92-109.

Undeutsch, U. (1984). Courtroom evaluation of eyewitness testimony. *International Review of Applied Psychology*, *33*, 51-67.

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

Vrij, A. (2000). *Detecting lies and deceit*. Chichester: Wiley.

Vrij, A., Akehurst, L., Soukara, S., & Bull, R. (2002). Will the truth come out? The effect of deception, age, status, coaching and social skills on CBCA scores. *Law and Human Behavior, 26*, 261-283.

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.

Table 1

Inter-rater reliability for individual criteria and total CBCA scores

Criterion	r
Logical structure	.85**
Unstructured production	.60**
Quantity of detail	.75**
Contextual embedding	.96**
Description of interactions	.68**
Reproduction of conversation	.88**
Unexpected complications	.87**
Unusual details	.82**
Superfluous details	.57**
Accurately reported details misunderstood	.71**
Related external associations	.48**

Subjective mental state	.72**
Perpetrator's mental state	.73**
Spontaneous corrections	.87**
Admitting lack of memory	.89**
Raising doubts about one's own testimony	.54**
Self-deprecation	.25
Pardoning the perpetrator	.70**
TOTAL CBCA SCORE	.91**

** Significant at the 0.01 level

Table 2

Means, F-values and Cohen's d scores for the criteria that significantly discriminated between truthful and fabricated statements (i.e. the most effective criteria)

	Rater one				Rater two			
	M (true)	M (false)	F	d	M (true)	M (false)	F	d
Admitting lack of memory	2.24	1.70	7.19**	0.75	3.33	2.30	10.10**	1.21
Unstructured production	3.47	2.60	6.97**	0.97	2.24	1.70	7.19**	0.75
Contextual embedding	3.71	3.10	6.63**	0.79	3.71	3.20	4.37*	0.64
TOTAL CBCA SCORE	43.57	38.90	5.60*	2.09	42.52	38.10	4.68*	1.96

* Significant at the 0.05 level ** Significant at the 0.01 level

Table 3

Rater one: Means, F-values and Cohen's d scores for the criteria that significantly discriminated between 'true' and 'false' judgements (i.e. the most influential criteria)

	M (truthful)	M (fabricated)	F	d
Contextual embedding	3.71	2.86	11.60**	1.20
Reproduction of conversation	2.71	1.71	8.45**	1.12
Description of interactions	3.33	2.71	6.87**	0.85
Unstructured production	3.38	2.57	4.35*	0.93
Logical structure	3.75	3.14	4.34*	0.73
TOTAL CBCA SCORE	43.58	36.86	10.63**	3.16

* Significant at the 0.05 level ** Significant at the 0.01 level

Table 4

Rater two: Means, F-values and Cohen's d scores for the criteria that significantly discriminated between 'true' and 'false' judgements (i.e. the most influential criteria)

	M (truthful)	M (fabricated)	F	d
Unstructured production	3.56	2.23	26.04**	1.59
Contextual embedding	3.89	3.08	16.59**	1.11
Description of interactions	3.61	2.77	12.33**	1.10
Logical structure	3.83	3.08	10.96**	0.95
Quantity of detail	4.67	4.00	9.73**	0.86
Reproduction of conversation	3.17	2.15	7.97**	1.04
Unexpected complications	2.11	1.62	7.85**	0.71
Own mental state	2.61	2.00	6.66*	0.75
TOTAL CBCA SCORE	44.89	35.92	46.22**	4.68

* Significant at the 0.05 level ** Significant at the 0.01 level