

A Systematic Review of Field Studies on Criteria-Based Content Analysis

Siegfried L. Sporer & Jaume Masip (2024)

Supplemental Materials 1:

Description of Studies Excluded and Reasons for Exclusion

Selecting studies for inclusion in a systematic review or meta-analysis does not only entail finding studies but also making decisions about their inclusion/exclusion after scrutinizing the methodology employed. While authors of primary studies may feel uncomfortable with some of these decisions, we believe that such screening is necessary (Lipsey & Wilson, 2001; Valentine, 2009).

Specific Issues with Field Studies

Inclusion/exclusion criteria for experimental studies in general may be rather straightforward if one follows the general guidelines in the meta-analytic literature (e.g., Borenstein et al., 2019; Lipsey & Wilson, 2001). However, the problem is more complicated when studying the detection of deception, and it is particularly difficult in field studies on this topic. The primary area of application of SVA/CBCA used to be the assessment of the credibility of (child) sexual (and physical) abuse allegations. In many of these cases, only the purported victim's testimony is available as evidence, and the victim's statements are typically denied by the alleged perpetrator. Ground truth is particularly difficult to establish in CBCA field studies.

Over the years, several criteria have been proposed to determine ground truth, ranging from strong DNA evidence to more questionable guilt or innocence indicia (e.g., recantations, passing or failing a polygraph test...). For all CBCA (archival) field studies or quasi-experiments conducted in field settings that we were able to retrieve, two coders critically examined the ground truth criteria considered by the original authors and classified studies as “0” (not acceptable for inclusion), “1” (acceptable but with some reservation), and “2” (methodologically sound). In addition to the ground truth criteria, other methodological

aspects were considered to classify studies as “0”, such as inadequate statistical analyses, or errors or omissions in the reporting of the data needed to calculate the effect sizes.

The studies classified as “1” and “2” are included in the systematic review. We also performed some meta-analytical calculations, but only with the studies classified as “2”. In this appendix, we briefly describe the studies coded as “0” and explain the reasons why they were excluded. For some studies that other authors have referred to in narrative or meta-analytic reviews, we justify our decision in more detail; for others, we just list key critical points.

Yuille and Cutshall (1989)

Yuille and Cutshall (1989) summarized several studies testing a method they designed to tally the number of details in free accounts provided by victims, eyewitnesses, and suspects. At the end of their book chapter, they reported a field case study where not only did they use their method but also assessed a purported offender’s credibility with several CBCA criteria.

The case involved the 1972 murder of the wife and two children of Dr. Jeffery MacDonald. One day, Dr. MacDonald, who was a physician at an army base in North Carolina, phoned the military police reporting that he and his family had been attacked at their home by several intruders, including a White female. When the police arrived at MacDonald’s house, they found his wife and their two daughters stabbed to death. MacDonald had been injured. He reported that the intruders broke in while he was asleep on the living room couch, had asked him for drugs, and then had attacked him and his family.

The investigation was unsuccessful, and initially no one was charged. However, MacDonald’s father-in-law became concerned about some evidence that came out during the investigation, including indicia that MacDonald may have had sexual activity outside his marriage. Therefore, he persisted until MacDonald was first charged with the murders and eventually convicted. However, a White woman named Stoeckley had claimed to be one of the intruders. She was willing to testify during the proceedings but only if the prosecutor granted her immunity, which he did not. However, after the trial, she provided four recorded interviews to a police officer and an investigator who believed in MacDonald’s innocence.

Furthermore, a video tape was made to be shown on a TV show (though it was never shown). Yuille and Cutshall (1989) were granted access to all these materials.

The authors used their method to count the number of details in Interview #1, the details that were repeated across interviews, the consistency across interviews, and the total number of unique details. They also checked verifiable details against the available evidence, and compared some details provided by Stoeckley with MacDonald's statements. Finally, Yuille and Cutshall (1989) note that Stoeckley's statements had logical consistency (CBCA01), unstructured production (CBCA02), and many details (CBCA03), including unusual (CBCA08) and superfluous details (CBCA09). Stoeckley also described an unexpected complication (CBCA07). Yuille and Cutshall also considered some of the SVA validity checklist aspects, in particular those related to Stoeckley's possible motives to report a true event vs. to fabricate. Based on their assessment, the authors conclude Stoeckley's accounts were probably true.

There are at least two compelling reasons why we were not able to include this study in our systematic review. First, as Yuille and Cutshall (1989) themselves acknowledge, "In this case we don't know what actually happened, so the success or failure of the analyses remains indetermined" (p. 182). Second, even if ground truth had been firmly established, this is a case study with $N = 1$ and no comparison group.

Raskin and Esplin (1991a)

This study was first presented at a NATO conference on credibility assessment in Italy (Esplin, Houed, & Raskin, 1988) and subsequently published in a conference volume on children's suggestibility (Raskin & Esplin, 1991a). The study has often been described as a milestone in the establishment of SVA and CBCA.

Statements from 20 "confirmed" and 20 "doubtful" child sexual abuse cases were analyzed with CBCA. Confirmed statements were supported by the confession of the alleged perpetrator (four cases), physical evidence (two cases), or both confessions and physical evidence (14 cases). Cases were classified as doubtful if the perpetrator persisted in his denial (20 cases), the child recanted her accusation or there was a lack of prosecution or a judicial

dismissal, there was no corroborating evidence (19 cases), or the alleged perpetrator had passed a polygraph test (14 cases).

At face value, this study would provide the strongest support ever seen for any content-based credibility assessment study. In the same volume, Wells and Loftus (1991) raised a series of alternative explanations for the results obtained. For example, the age of the children in unconfirmed cases ($M = 6.9$ years) may have been lower than in confirmed cases ($M = 9.1$ years), a difference which Raskin and Esplin (1991b) point out not to be significant, $t(38) = 1.95$. For this difference, we calculate a $d = 0.617$, which is not negligible.

Wells and Loftus (1991) raise other plausible rival hypotheses which may make children in the unconfirmed group less convincing, and hence receive lower CBCA scores: Doubtful cases may have been cases of real abuse, but the victims were much less convincing than those in the confirmed group, perhaps because of deficiencies in logical reasoning, poor verbal skills, or similar factors. These deficiencies may have resulted in an impoverished statement with fewer CBCA criteria. At the same time, if those children were unconvincing, then prosecutors may not have pressed charges, judges may have dismissed the cases, and defense attorneys may have been unlikely to advise their clients to make an admission. Note that these were three ground truth criteria used by Raskin and Esplin (1991a). Also, although Raskin and Esplin said that the one expert coding all statements had no knowledge of the cases, Wells and Loftus provided several arguments suggesting that she may have had the opportunity of becoming knowledgeable. Wells and Loftus list a series of additional criticisms we will not reiterate here but urge readers to consult if they doubt our exclusion decision.

Raskin and Esplin's study also exemplifies a special problem regarding ground truth, viz. the non-independence of ground truth criteria. Because this issue pertains to many validity studies, we explain it in more detail in the Discussion section of the main article. But even if we were to accept Raskin and Esplin's rebuttal to defend their study, there are statistical reasons why we decided to exclude it from our systematic review.

There was absolutely no overlap between the distributions of the summary scores, which could range from 0 to 38 (0/1/2 codings * 19 criteria): All summary scores for

confirmed statements were between 16 and 34, while all scores for unconfirmed cases were from 0 to 10. The authors report a mean of 24.8 for confirmed cases and of 3.6 for unconfirmed cases, $t(38) = 16.53$. We calculate a d value of 5.227, the highest we have ever seen in a deception study.

Thus, both for the summary score and for many individual criteria, this study yields extreme outlier values. If effect sizes for binary data (criterion absent vs. present [0/1]) are calculated for this study, which contained zero values in one or more of the cells, the odds ratios (OR) for some criteria were as high as $OR > 500$, which would transform in d values > 6 , using conversion formulae by Lipsey and Wilson (2001) or Borenstein (2009; for a review of conversion formulae, see Sanchez-Meca et al., 2003). Even if we did not exclude this study because of ground truth reasons, it would have to be excluded based on outlier analyses.

Krahé and Kundrotas (1992)

In the study by Krahé and Kundrotas (1992; see also Krahé et al., 1995), 30 protocols of police interrogations of women alleging to have been sexually abused were examined. Half of them were classified as "credible" on the basis of suspect confessions and/or physical evidence and forwarded to the state attorney. The other half were considered false allegations because the women themselves admitted their falsity at the end of the interviews. All protocols were anonymized and evaluated by 53 police officers (34 females; mean number of years in office = 16.8 years, with all but four having had experience in the interrogation in rape allegations). Thirty-one officers assessed the protocols with the 19 CBCA criteria and 22 without them. Because the statements were rather long (on average about 3 typed pages), each officer evaluated only four accounts.

We decided to exclude this study for the following reasons: (1) the ground truth criteria were not sufficiently established nor independently corroborated; (2) police protocols are *not verbatim* transcripts of free reports by witness-victims but subjective summaries of what interviewers thought they ought to write down in the protocols (see Lamb et al., 2000; Milne et al., 2022). Hence, these protocols may reflect as much what the officers believed to be important and what they may have thought corroborated or discredited a witness's

allegation. In other words, these protocols may not only describe what really happened but also reflect rape myths held by these officers (see Greuel, 1993).

Lamers-Winkelmann, Buffing, and van der Zanden (1992)

Out of an original sample of 227 alleged victims of child sexual abuse in the Netherlands, 185 were interviewed and the interview transcripts were made available to Lamers-Winkelmann et al. (1992). The 185 cases were categorized as either “substantiated” (if the alleged perpetrator made a confession and/or was convicted), “(highly) probable” (based on the child’s statement, the statements of other children, eyewitnesses, corroborating evidence, or behavioral and emotional signs of abuse), or “unfounded” (no statement provided by the child, or the statement had been fabricated or highly influenced by an adult, or there was an alternative explanation for the child’s behavioral and emotional signs). Allocation of cases to the “(highly) probable” and the “unfounded” categories required the agreement of at least two independent raters.

Out of the 185 children who were interviewed, 73 made no statement describing the sexual abuse. Also, nine additional interviews could not be transcribed because of equipment failure. This left the authors with 103 usable interviews (75 of females, 25 of males; age range = 2 to 11 years). Out of these interviews, 36 were of substantiated cases, 61 of (highly) probable cases, and (only) six were of unfounded cases.

Two independent raters, who were blind to the truth status of the interviews and had been trained by Undeutsch and Steller on SVA (including CBCA), rated all transcripts. Each CBCA criterion was rated as present or absent.¹ Because the CBCA criteria had high internal consistency (Cronbach’s *alpha* = .879) and alpha would not increase by excluding any criterion, the authors mainly focused on the mean number of CBCA criteria present, though they also reported data for the individual criteria in the tables.

¹ The authors reported an inter-rater reliability of .85 for CBCA. However, it is unclear what specific reliability coefficient they calculated, as well as which codings they considered for the calculation--separate reliabilities for each of the 19 CBCA criteria should have been reported.

The main finding of this study is that the presence of CBCA criteria increased with age. This was so even when considering substantiated, probable, and unfounded cases separately. However, the means of CBCA summary scores did not differ significantly across substantiated ($M = 9.05$), (highly) probable ($M = 7.61$), and unfounded ($M = 8.33$) cases.²

Interestingly, the raters also made an overall credibility judgment based not only on the CBCA criteria but also considering other SVA aspects. This judgment was made on a five-point scale (very unlikely/unlikely/inconclusive/likely/very likely that the sexual abuse happened). The authors reported frequencies under each scale point, but based on these frequencies (and assuming that the scale ranged from $1 = \text{very unlikely}$ to $5 = \text{very likely}$), we calculated the means and standard deviations for the substantiated ($M = 4.36$, $SD = 0.68$), the probable ($M = 3.93$, $SD = 0.83$), and the unfounded group ($M = 3.17$, $SD = 0.75$). The effect was significant, $F(2, 100) = 7.37$, $p = .001$, partial $\eta^2 = .128$. Bonferroni-corrected post-hoc comparisons showed that the substantiated group differed significantly from both the (highly) probable ($p = .032$; $g = 0.54$, 95% CI [0.13, 0.96]) and the unfounded group ($p = .002$; $d = 1.69$, 95% CI [0.77, 2.62]), but the probable and the unfounded group did not differ significantly, $p = .071$, despite the effect size being very large, $d = 0.92$, 95% CI [-0.07, 1.76]. Note, however, that the number of cases in the unfounded group was extremely small ($n = 6$).

Though these latter findings seem to suggest that criteria other than the mere CBCA ratings are important to determine statement credibility, we have several concerns regarding this study. First, while some of the criteria considered to establish ground truth appear valid (e.g., physical evidence), others are certainly questionable (e.g., the child's statement, behavioral or emotional signs...). Also, the authors did not specify how many criteria were needed to allocate a case to the probable category, or whether certain criteria should be weighted more heavily than others, or how to proceed if criteria of the probable group were

² Based on the data provided by Lamers-Wilkelman et al. (1992) in their Table 6, we re-calculated these means. Surprisingly, we found different values: 8.19, 6.92, and 7.67, respectively.

present along with the “fabrication or adult-influence” criterion of the unfounded group. Second, given the small number of cases in the unfounded group ($n = 6$), any comparison involving this group is meaningless. Third, as we noted in footnote 3, there seem to be some errors in the way the mean number of criteria for each group was calculated. Based on these concerns, we decided to exclude this study from our systematic review.

Subsequently, Lamers-Winkelmann and Buffing (1996) and Lamers-Winkelmann (1999) published two reports based on the same data that focused on the covariation of age on the CBCA criteria. We cite these reports at the end of the systematic review.

Brodie (1993)

Brodie (1993) conducted a study to examine the influence of the personal and professional characteristics of child protective service workers on their accuracy in determining whether alleged child sexual abuse cases had occurred (i.e., the child had been abused) or not. They also tested, with a repeated-measure design, the impact of training workshop on accuracy.

The participants were 53 social workers drawn from a child protective service department in California. They first filled in a 175-item self-report (partly based on prior scales) to measure their personal demographics, personal abuse history, education and work experience, knowledge about abuse, and beliefs about abuse. The author would subsequently examine whether several of these variables were related to accuracy.

Brodie used 12 transcribed child sexual abuse statements. These statements were selected from those previously examined in the field study by Raskin and Esplin (1991a; see our summary of this study above). Six transcripts were from Raskin and Esplin’s “confirmed” group, and six from their “doubtful” group. Three of the statements from each group were of children with a modal age of 6 years, and three were of children with a modal age of 9 years. All children were female.

Each participant (rater) was assigned two transcripts (out of the pool of 12 selected accounts) to assess before the training. After the training, each participant rated the same two transcripts and two additional transcripts. The author counter-balanced the transcripts to ensure each account was rated the same number of times both (a) before and then after the

training, and (b) after the training as a new transcript (see Brodie, 1993, p. 15, for the design). The participants rated each assigned transcript on a 1-to-5 statement-validity scale, and scores were later transformed such that scores of 4 or 5 were coded as “high validity” (i.e., abuse occurred), scores of 1 or 2 were coded as “low validity” (i.e., abuse did not occur), and scores of 3 were coded as “undecided”. Statements coded as having either high or low validity were then re-coded as correctly or incorrectly judged based on Raskin and Esplin’s (1991a) ground truth determination for those statements. Note that although this was in essence an experimental study designed to test the effectiveness of the training workshop, after the training instructed coders rated accounts from actual cases, which is what is typically done in field studies.

The scope of the training workshop (whose duration and specific activities are not reported) was rather broad, covering “a review of the literature concerning what is currently known about professional judgments of child sexual abuse, interview techniques, an overview of child development, memory and testimonial capacities, an introduction to SVA/CBCA, and actual examples of interviews of child cases.” (Brodie, 1993, p. 15).

Most of the participants’ personal and professional characteristics had no significant relation with accuracy. More relevant for our purposes are the participants’ accuracy rates before and after the training. The authors found no significant effect of training on accuracy, either when comparing pre- vs. post-training accuracy in judging the same transcripts or in judging different transcripts. Specifically, 61% of all 106 judgements (53 participants times 2 statements) made before the training were accurate, 22% were incorrect, and 17% were undecided. After the training, the corresponding percentages for the same accounts were 60%, 26%, and 13%, respectively; and for different accounts the percentages were 54%, 33%, and 13%, respectively. In our view, the broad nature of the workshop, which was not specifically focused on credibility criteria, may have contributed to these null results.

Several aspects prevented us from including this study in our systematic review. Here, we mention three: First and foremost, while our goal was to examine the validity of the individual CBCA criteria, no scores were reported by Brodie (1993) for individual CBCA criteria (nor for the summary score). It is even unclear whether raters coded such criteria.

Second, the focus of the training workshop was extremely broad, and no information is reported about the duration of the workshop nor about specific teaching activities. If the coverage of credibility criteria in the workshop was meager (as suggested by Brodie's statement that it contained only *an introduction* to SVA/CBCA) and the workshop involved no practice and/or homework (with participants coding transcripts) followed by feedback, the raters may have been insufficiently trained. Third, the accounts were drawn from the pool of statements previously employed by Raskin and Esplin (1991a). As explained above in summarizing Raskin and Esplin's (1991a) study, their research suffers from several major limitations. Therefore, using a subset of their statements is problematic.

Dana-Kirby (1997, Study 3)

Dana-Kirby (1997) conducted three CBCA studies for her dissertation. However, only Study 3 is a field study. She examined 22 suspects' statements (seven truthful and 15 deceptive) collected from either a US police department (four statements) or from real-crime TV programs (18 statements). Many additional accounts were evaluated, but they were dismissed because ground truth could not be established. All statements in the study were transcribed.

Ground truth was determined on the basis of eyewitness accounts (nine statements), physical evidence (e.g., fingerprints, blood alcohol level; 12 statements), information provided by "reliable" sources (e.g., confidential informants; four statements), police data such as police reports or fingerprint databases (three statements), and eventual confession by the suspect (four statements). For 13 statements, ground truth was determined based on one criterion only (though for seven of these 13 cases, the one criterion was physical evidence), for eight statements it was established based on two criteria, and for one statement it was determined based on three criteria (see Appendix F in Dana-Kirby, 1997).

Two trained research assistants independently examined whether CBCA01 to CBCA14 were present or absent. They subsequently discussed disagreements to reach a consensus for each criterion across all accounts (hereafter "consensual ratings"). The author reported three separate reliability coefficients: proportion agreement, Cohen's *kappa*, and Maxwell's *RE* coefficient. Although proportion agreement was above 0.50 for most criteria,

reliability as measured by *kappa* or Maxwell's *RE* was quite poor. Cohen's *kappa* was negative (-.16) for CBCA08, it was 0 for four criteria, higher than 0 but lower than .25 for five criteria, and higher than .50 only for CBCA12 (*kappa* = .64). Maxwell's *RE* coefficient was 0 for CBCA11, higher than 0 but lower than .50 for six criteria, in the .50 to .70 range for four criteria, and higher than .70 for only three criteria.

Because the majority of statements came from TV programs, nothing is known about interviewers, interviewing protocol (e.g., rapport, open vs. closed questions, interviewer training). Also, statements on TV can hardly be compared with witness or suspect statements taken at the police station or in other legal settings. Furthermore, we question some of the ground truth criteria like the blood alcohol or confidential informants. Last but not least, the poor reliability of CBCA coding made us conclude that this study cannot be used for testing the validity of CBCA criteria.

Hershkowitz (1999)

In this study, 12 plausible and 12 implausible allegations of child sexual abuse were examined. The goal of the study was to assess the impact of the interviewers' utterance types on (a) the number of details provided by the children, (b) the total number of CBCA criteria present at least once in each statement (only CBCA01 to CBCA14 were considered), and (c) the strength of presence of criteria CBCA04 to CBCA14 in each statement (measured as the frequencies of occurrence). The CBCA ratings were made by two individuals who had been extensively trained to code CBCA criteria. They agreed in 90% of their decisions, and disagreements were resolved by consensus in a conference with a third coder.

Ground truth was established satisfactorily, as the accounts were a subset of those included in Lamb et al.'s (1997) study, which we selected for inclusion. However, precisely to avoid including the same data twice, we had to exclude Hershkowitz's (1999) study. Besides, Hershkowitz did not report enough information to calculate the effect sizes for the original criteria, which would have prevented us from including her study even if the data had been original.

Orbach and Lamb (1999)

The authors examined the presence of the first 14 memory-related CBCA criteria in a statement of sexual abuse provided by a 13-year-old girl. Because the incident had been surreptitiously audiotaped by the girl, its occurrence was established with absolute certainty. Ten of the 14 examined criteria were present. Unfortunately, we could not calculate effect sizes because this is a case study with $N = 1$.

Parker and Brown (2000)

Parker and Brown (2000) collected 55 rape allegations. However, 12 were screened out as unsuitable for analysis. The retained 43 allegations had been collected with the Cognitive Interview (Geiselman & Fisher, 1989) from 38 females and five males (age range: 13 to 80 years, $Mdn = 21$). The authors examined these allegations to test the discriminative value of 13 SVA validity checklist criteria and 18 CBCA criteria.

"False (unfounded) allegations" were drawn from metropolitan police databases, and "founded (genuine) rape cases" were randomly selected from the Metropolitan Police Service's Crime Recording Information System (CRIS). It is unclear based on the reading of the paper whether ground truth was established by the police or by the researchers.

Sixteen (37.2%) allegations were regarded as "true," 15 (34.9%) as "unsubstantiated," and 12 (27.9%) as "false." These were the criteria to classify allegations as true: "From the found statements there was one conviction (2.4%), 11 suspects were charged (25.7%) and four cases were substantiated but the suspect remains unidentified (9.4%)" (Parker & Brown, 2000, p. 243). Regarding the 15 unsubstantiated cases, "in seven of these a suspect was identified but the allegation was subsequently withdrawn by the alleged victim (16.4%); six cases involved 'acquaintance rapes', which did not proceed (13.9%); and in two cases there was no likelihood of a suspect being identified (4.7%)" (Parker & Brown, 2000, p. 243). Finally, for the 12 false allegations, "five of these involved serial allegeders found to be suffering from psychiatric disturbances (11.7%); four cases involved an admission of malicious intent (9.4%); two involved delusional alcoholics who later withdrew their allegations; and the remaining case involves an epileptic serial allegeder suffering post-seizure rape delusions (2.4%)" (Parker & Brown, 2000, pp. 243-244).

Some decisions seem questionable, such as classifying cases as unsubstantiated because they were "acquaintance rapes" or because witnesses later withdrew their allegations. Additional problems involve the small sample size, paired with such a wide age range (including both children and adults), and the inclusion of several alleged with presumed psychological/psychiatric disorders.

In addition, there are several statistical ambiguities and problems in the analyses reported. "Multidimensional Scalogram Analyses" are presented, selectively reporting only 10 CBCA criteria (of 18 criteria rated according to the appendix) that distinguished significantly between "true", "unsubstantiated" and "false" statements (CBCA01 to CBCA07, CBCA12, CBCA13 and CBCA15). Furthermore, in Parker and Brown's (2000) Appendix A only single mean values are presented, and only for those criteria that were significant, along with F and p values with $df = "42, 2"$ (which should be $df = "2, 40"$ for ANOVAs with three groups and $N = 43$). When only a single mean for three groups is presented, one cannot know the direction of effect of the group differences, let alone calculate effect sizes for the difference between "true" and "false" accounts. If only the overall "accuracy of classifications" could be extracted from this article, it would be an overestimate due to retaining only "significant" CBCA criteria, and to not using proper cross-validation procedures (Kleinberg et al., 2019; Sporer et al., 2021).

In short, this study does not fulfill our criteria for establishing ground truth nor do the methodological and statistical issues described allow conclusions about the validity data of CBCA criteria or SVA.

Hershkowitz (2001)

The author reported an interesting case of a 10-year-old girl who took a shortcut to get home through the woods after school. A man appeared in the woods and exposed his penis to the child, who turned around and ran away. Because her mother had forbidden her to walk through the woods, the child initially did not disclose the event, but after one week she decided to tell her mother. To avoid being questioned about the delay, she decided to describe the event as if it had just happened (rather than telling it had happened the week before). She arrived home crying and, because she could not articulate a word, her mother asked

affirmative questions so she could nod or shake her head. The girl nodded affirmatively to the questions whether somebody had touched her body and whether somebody had touched her privates. The mother examined the child's privates and saw a man's pubic hair. At this point, the mother called the police.

The child was interviewed by the police a few hours later, and she reported that someone assaulted her in the woods and forcefully penetrated her. After giving her statement, she was examined by a physician (6 hours after the assault had allegedly occurred), who found no evidence at all of any forceful penetration attempt, and who reported that the pubic hair the mother had seen belonged to the child. No evidence was found at the scene of the crime either.

Two days later, the child was interviewed again. She recanted her prior allegation and described the above chain of events. She specified that the time of the event differed, and that the man did not rape her, but that other details were true. She explained that she described a rape event to not contradict her mother, in the hope that this way the investigation would end immediately.

Two experienced coders established ground truth using an independent-case-facts scale that considers a variety of sources of information (Horowitz et al., 1995). Based mostly on the results of the medical examination (no signs of penetration; no bruises, scratches, or other injuries; pubic hair pertaining to the own child) and the detailed recantation of the child, as well as on the absence of any material evidence at the crime scene and the mother's statement concerning her suggestive interaction with the child, the coders agreed that the event appeared "very unlikely" to have occurred.

Two separate trained CBCA raters examined the child's first allegation to the police with 14 CBCA criteria. The criteria were coded as absent or present, and inter-coder reliability was 90%. The coders determined that seven CBCA criteria were present in the account. This may look like a relatively large number of criteria, but note that many details of the child's account were true--only the specific day and the penetration attempt were fabricated. This may explain the high CBCA score (Hershkowitz, 2001).

Unfortunately, we were unable to include Hershkowitz's (2001) study because, just like the report by Orbach and Lamb (1999) described above, it is based on $N = 1$, without a comparison between confirmed vs. unconfirmed cases. The case is however interesting on many grounds, and we encourage readers to consult Hershkowitz's original report for a more detailed description of the case, including the kinds of questions asked to the child and the number of details she provided.

Suiter (2001)

The goal of this dissertation was to examine Investigative Discourse Analysis (IDA; Rabon, 1996) and CBCA from the perspective of linguistic theory. IDA has been promoted in law-enforcement contexts (e.g., Adams, 1996).

Suiter (2001) compared IDA and CBCA predictions with linguistic theory. She also empirically tested some of these predictions. However, her approach was fundamentally exploratory and qualitative; she reported only frequencies and did not use inferential tests. Suiter employed three separate sets of data: (a) Nine true and seven false accounts written by college students (experimental data); (b) 30 narratives taken from the Cambridge University Press/Cornell University Corpus of natural language data (CUP-CORNELL), which contains natural speech of ordinary citizens covering all kinds of speech genres (e.g., causal conversations, action language, service encounters, etc.; natural language data); and (c) statements taken by two police forces from 12 victims, 18 witnesses, and 22 suspects (field data). Only the field data were of interest for our project.

A fundamental limitation of Suiter's (2001) field study is the uncertainty regarding ground truth. She is fully aware of this issue:

Every effort has been made to ensure that the true narratives are true, and the false narratives false. It is not known definitely, however, that the true narratives are true. For example, one narrative which was initially included in the data sample turned out to be a repetition of a story from the movie *Good Will Hunting*. In other words, I have no guarantee that the true narratives are true. Similarly, in the suspect narratives (most of which are disclosure narratives), I have no guarantee that the suspects are fully allocutating to the

elements of the crime. Even if, for example, a suspect admits to killing someone, there is no independent verification that the killing happened as the suspect stated. Finally, in the three suspect narratives which I classify as false, only one narrative is definitely deceptive, since the police charged the suspect with obstruction. The other two narratives I classified as false because in the first narrative the suspect's narrative varied significantly from the victim's, and the victim would have no reason to lie, and in the second narrative the suspect's version again differs from the victim's version. Even though in this case both subjects have a motivation to deceive, from other evidence the police has classified his narratives as probably deceptive, and I will keep that designation for this thesis (Suiter, 2001, pp. 16-17).

It also seems from this paragraph (and is clear from other passages in Suiter's dissertation) that only three out of the 52 field statements were coded as deceptive. Furthermore, apparently only the author (who was not blind to the purposes of the study) coded the statements. Last but not least, regarding CBCA, the aspects that Suiter (2001) examined were not specifically CBCA criteria. Instead, she measured linguistic features (operationalized as the frequencies of certain words) that she considered to be somehow related to some CBCA items. For instance, the frequencies of the words "think" (not used as a hedge), "feel" and "notice" were measured as indicative of CBCA12 (*Accounts of subjective mental state*) and CBCA13 (*Attribution of the interaction partner's mental state*). Indeed, this is not the way CBCA criteria ought to be coded.

Suiter's (2001) dissertation contains interesting insights from linguistic theory that researchers making predictions about verbal content cues to deception would be well advised to consider. However, for the reasons we enumerated above, we could not include her field study in our systematic review.

Adams (2002/2004) and Adams and Jarvis (2006)

The authors examined written accounts collected in the US during actual criminal investigations. From an initial pool of 347 statements, 60 were selected that met the following criteria: First, investigators had been able to determine veracity based on (a)

conviction by a judge or jury, and/or (b) overwhelming physical case evidence, and/or (c) a corroborated confession by the offender. If none of these pieces of information was available to determine veracity, the statement was not included in the study. According to Adams (2002), for many of the included statements ground truth could be established based on all three criteria. Deception could occur either by commission or by omission (i.e., the account was truthful except that the incriminating facts were omitted), which makes it rather unique in the CBCA literature. Second, to be included, accounts had to be obtained via open-ended instructions (e.g., "Write down what happened") rather than following specific questions. Third, statements had to be fully legible; those containing illegible words were excluded. Fourth, the accounts had to be written originally in English--some statements in the pool had been written in Spanish and subsequently translated into English; these statements were excluded. Finally, only one statement from each case and from each individual was included.

Out of the selected 60 accounts (30 truthful, 30 deceptive), 47 were from violent crime cases (abduction, assault, homicide, rape, and robbery), while the remaining 13 accounts referred to property crimes (arson, drugs, and theft). The participants ($N = 60$; 28 females and 32 males; M age = 27 years, range: 15 to 59, only five participants were under 20) were either suspects ($n = 37$) or victims ($n = 23$).

Raw frequencies and density measures (i.e., frequency divided by the number of words) were calculated for three deception and three truth criteria. Only one of the latter criteria, *Quoted Discourse*, corresponds to a CBCA criterion (CBCA06, *Reproduction of Conversations*; see Adams, 2002/2004, pp. 64-65). One half (i.e., 30) of the statements contained no *Quoted Discourse* at all, while the other half contained between one and 60 instances of it (across all 60 statements, $M = 4.7$, $SD = 8.6$).³ Several analyses were conducted to examine the relationship between the criteria (entered as either frequency counts or density) and truth status, including correlations and logistic regression analyses. Although the direction of effect was in the expected direction, in no case were the results

³ This pattern of data suggests that data for this variable were extremely skewed, making the following analyses problematic.

significant for *Quoted Discourse* (frequency count: $g = 0.170$, 95% CI [-0.331, 0.670]; density: $g = 0.219$, 95% CI [-0.282, 0.720]).⁴ Confirmed accounts tended to be shorter than deceptive statements, but the difference was not significant: $g = -0.396$, 95% CI [-0.901, 0.108].

This study suffers from a number of problems. First, the most potentially valid ground truth criterion is vague, as the authors do not describe what they mean exactly by "overwhelming evidence", nor do they provide any examples. Second, no information is given concerning coder training, the number of coders, whether the coder(s) was/were blind, or inter-coder reliability. Third, combining suspect and victim statements is problematic. Finally, CBCA is typically used with oral statements rather than written accounts. Because of these reasons, we decided not to include this study.

Rassin and van der Sleen (2004)

In a field study by Rassin and Van der Sleen (2004), the authors selected "true" and "false" accounts by mailing "approximately 220 Dutch police officers (known to the second author)" and asking them "to go through two files of recent sexual offences, one of which rested upon an almost certainly true allegation, and one of which had turned out to be based on a false allegation" (p. 591). The officers completed a checklist pertaining to the characteristics of the allegation. The authors received a total of

"... 88 checklists, 37 pertaining to true and 51 pertaining to false allegations. One of the items in the checklist referred to the outcome of the case at hand. Analysis of this item indicated that 27 of the presumably accurate allegations had actually resulted in a conviction of the perpetrator. Of the 51 allegedly false allegations, 14 had resulted in a conviction of the "victim" for making a false allegation. We limited our analyses to these 27 true and 14 false allegations because, in these cases, the allegations can—at least from a judicial sense—indeed be considered to be true and false, respectively."

⁴ The authors conducted similar analyses after dividing the statements into three partitions: prologue, criminal incident, and conclusion. Because this is not common practice in the CBCA/SVA research literature, these analyses are not reported here.

The 41 checklists included in the study were completed by 33 other police officers (20 women). The mean age of these officers was 45.8 yr. ($SD = 5.9$).” (p. 591; emphasis added).

The checklist contained 43 credibility criteria drawn from different sources, including 29 items from Burgess and Hazelwood (2001), 11 CBCA criteria, and eight RM items. Each item was scored for its absence/presence (0/1), allowing also a “Don’t know/Not applicable” option, for which no value was assigned. Some of the items were expected to be truth criteria, and others lie criteria, which were re-scaled before analysis (marked with “R” in the appendix of items in Rassin & van der Sleen, 2004).

Two summary scores were reported, one for a combination of CBCA and RM items (Cronbach’s $\alpha = .68$), and one on the Burgess and Hazelwood (2001) items (Cronbach’s $\alpha = .68$). Although there were highly significant differences between “true” and “false” allegations for both scales (Burgess-Hazelwood scales: $t(39) = 6.4, p < .001, d = 2.11$; CBCA/RM: $t(39) = 4.3, p < .001, d = 1.42$), there was considerable overlap between the distributions.

Although the authors raise a series of interesting points and issues (for example, the overlap of CBCA and RM criteria and the existence of potentially valid lie criteria), we decided to exclude this study for the following reasons:

(1) Asking police officers to subjectively select “an almost certainly true allegation” and “one of which had turned out to be based on a false allegation” was likely to have led to a highly subjective pre-selection of particularly convincing cases.

(2) Even though only a subset of 27 cases that resulted in a conviction of a perpetrator (retained as true cases) and 14 cases that resulted in a conviction of the “victim” for making a false accusation were retained for further analysis, no additional ground truth criteria from independent case evidence were presented.

(3) Considering that CBCA and SVA are also used in the Netherlands (Knoerschild & van Koppen, 2003; for a critique, see van Koppen & Saks, 2003), the police officers who selected the cases may have had received some training in credibility assessment, and

credibility experts may have been involved in these cases, resulting in circularity of validation.

(4) The 41 checklists were completed by 33 different police officers (20 female), but nothing is said about their training or knowledge about SVA, their professional experience, etc. (only their age). No data on the reliability of coding is presented.

(5) In addition, neither detailed descriptions of the criteria nor whether they were indicative of truth or deception were provided to raters. Thus, it appears that the police officers completing the checklists used their own personal interpretations of the items, which may reflect more their own *beliefs about truth/lie criteria* than the direction of effect postulated by CBCA and RM experts.

(6) The authors reported a total of 43 comparisons (one for each item), but no attempt was made to adjust for *alpha* inflation.

Chang (2008)

The author examined written statements voluntarily provided by 88 suspects, 26 victims, and seven witnesses of felony crimes (total $N = 121$) before formal police interrogations in the US. The cases were sorted into two major categories. They were allocated to the *felony conviction/other corroborative data supporting culpability* category if the individual had been found guilty based on a combination of different kinds of evidence--namely witness statements, physical evidence, forensic evidence (e.g., DNA tests), suspect confessions, co-defendant confessions, and "evidence reports that support a subject's culpability" (Chang, 2008, p. 51). Conversely, cases were allocated to the *subject exonerated/cleared/corroborative data indicating other subject(s)* category if someone claimed a crime had been committed but later either recanted or was refuted by others, or if someone was falsely accused but was eventually acquitted based on the types of evidence mentioned above. Cases that could not be allocated to either of these categories were considered *inconclusive*.

The written accounts were coded by four trained graduate students. Nineteen verbal content criteria were considered, including some that are related to CBCA. In particular,

Chang's (2008) *Quoted Discourse* is analogue to CBCA06, *Writer's Emotions* is similar to the emotional dimension of CBCA13, *Spontaneous Corrections* is similar (though not identical) to CBCA14, and *Lack of Memory* resembles CBCA15. Other criteria were also coded that were taken from the RM or the SCAN literature but were not separated clearly in the analyses.

However, this study has several major problems. First, presumably there is a confound between role (suspects vs. victims and witnesses) and truth/deception allocation, as it is very unlikely that victims and witnesses were allocated to the *felony conviction/other corroborative data supporting culpability* category. Second, the raters were not blind to condition because, in addition to coding the 19 credibility criteria, they also coded the characteristics employed to determine ground truth. Third, no inter-rater reliability information was provided, and it is unclear which rater's data were used for analyses, or whether (and, if so, how) a consensus was reached by the raters. Fourth, crucially, the number of truthful and deceptive accounts varies across analyses. According to Table 1 (in Chang, 2008), 20 accounts were deceptive and 49 were truthful. But in the "primary analyses" section, the number of deceptive accounts was 16 and the number of truthful ones was 53. In subsequent analyses, the author compared truthful accounts with a combination of deceptive and inconclusive statements; sample sizes were, respectively, 36 (rather than either 49 or 53) and 52 (rather than 85, that is, total sample size [$N = 121$] minus 36 truthful accounts). The reasons for these inconsistencies are unclear.

For all the above reasons, we decided to exclude this study from our systematic review.

Critchlow (2011)

This report describes an undergraduate project report testing CBCA with adult rape allegations. The author had access to a pool of 125 such allegations collected by British police officers with specific training to deal with sexual assault crimes. All statements were collected using Geiselman and Fisher's (1989) cognitive interview. Out of this pool, 28 allegations were extracted. However, no information is provided how the original pool had

been created, nor how this subsample of 28 statements were selected. The 28 statements were allocated to one of four categories:

(1) If the allegation had been determined to be truthful based on medical examination or CCTV, the case had been forwarded to the prosecution service, and the trial outcome had resulted in a conviction, the allegation was considered to be *genuine* ($n = 7$).

(2) If the allegation had been proven false based on evidence such as CCTV, or if conflicting medical evidence was available, or if the purported victim later recanted in writing, the allegation was deemed to be *false* ($n = 7$).

(3) Some allegations had passed the "threshold test" for evidence within police procedure and, therefore, had been forwarded to the prosecution service. However, either the prosecution service or the victim decided to take no further action. These cases were independently rated by two detective inspectors and a researcher, who considered the statements to be truthful. These allegations were tagged as *credible* ($n = 7$).

(4) Finally, there were some cases for which (a) the corresponding police force review officer had established there was not enough evidence to support the allegation, or (b) the victim had asked that no further action be taken. Also, these cases had been independently rated by two detective inspectors and a researcher, who considered the statements to be false or suspicious. These allegations were tagged as *non-credible* ($n = 7$).

Nothing is known about the qualification of these two inspectors or the researcher to make such classifications and on what basis.

The author rated each of the 19 CBCA criteria as 0 (absent), 1 (present) or 2 (strongly present) in the statements. However, prior to data analysis these scores were reduced to binary codings (0 = criterion absent; 1 = criterion present) by transforming scores of 2 into scores of 1. As the author had collected and selected the cases to be included, it is highly questionable that he was blind with respect to ground truth. A random sample of eight accounts was also rated by a forensic psychologist who was also a university lecturer. The Pearson correlation between the total CBCA scores of the two raters was $r = .98$. Inter-rater discrepancies were resolved by discussion. However, the reliability of summary scores does not imply that individual criteria were rated reliably (see Hauch et al., 2017; Sporer, 2012).

The author reported the CBCA summary scores (*Ms* and *SDs*) for each of the four allegation categories. In the report, there is also a reference and discussion of "Table 2", which supposedly contained the values for individual criteria. However, this table is missing from the report. We contacted the author but were unable to obtain these data, which were necessary for our analysis.

In summary, we believe that ground truth was not unequivocally established for the reasons noted above, and that neither the blindness of the rater nor the training and reliability of scoring was sufficiently established. Although one could argue that Groups 1 and 2 might be compared and hence included, this would be tantamount to using an extreme group comparison which would lead to an overestimate of effect sizes. Even if one were willing to accept ground truth as established, the effect size for the comparison of summary scores between Groups 1 and 2 was $g = 4.93$, which would have to be excluded from any review or meta-analysis as an outlier.

Johnston, Candelier, Powers-Green, and Rahmani (2014)

The authors created an 11-item *Deception Detection Checklist* that included some CBCA criteria and asked a large sample of college students to use it to rate a "truthful" and a "deceptive" statement, each made by a different individual accused of child molestation. Ground truth was based only on jurors' decision after trial, which is problematic.

Another questionable aspect concerns the kind of statements used. These were collected by a clinical psychologist to whom the alleged perpetrators had been referred to for psychological assessment. Instead of describing an event, these statements contained the alleged molesters' explanations of the charges against them and the reasons why they believed they had been falsely accused. Last but not least, only one "truthful" and one "deceptive" statement were used in this study.

Wojciechowski (2014)

Apparently based on the author's dissertation (in Polish), the author reported an archival analysis of a random selection of "over 130 criminal cases" (Wojciechowski, 2014, p. 122) that went through two court levels ("instances"). Of these, 79 transcripts were used (47 true and 32 false statements) but the selection of cases is not explained.

Ground truth was not established to our full satisfaction: "Testimonies were considered true if the courts of both first and second instance recognised them as reliable and there was additional evidence supporting the witnesses' accounts (such as physical traits of expert evidence). The testimonies were considered deceptive if the courts of the first and second instance found them unconvincing, the witness admitted to giving false evidence and was sentenced for perjury." (p. 122). In the absence of details about the type of evidence considered (e.g., type of physical evidence or expert testimony on eyewitness credibility) the courts determination as "reliable" or "unconvincing" does not appear sufficient to establish ground truth.

The author conducted statistical analyses via a "feature selection and variable screening tool" (p. 123), but this tool is not explained. He indicated that six CBCA criteria (and one SVA Validity Checklist criterion) discriminated between truthful and deceptive accounts. He reported chi-square values and *p* values for these six criteria, but not for the remaining CBCA items, and no *means* and standard deviations were reported for any criterion, making it impossible to determine the direction of the effects, nor to calculate effect sizes. Finally, CBCA and Validity Checklist data were analyzed together instead of separating them, which made it impossible to determine their respective discriminative value.

Wojciechowski (2014) also reported raters' assessment decisions (i.e., the dichotomous decisions whether each statement was truthful or deceptive), which were in stark contrast to their ratings of the CBCA criteria, some of which would appear promising if statistical details had been reported. The author also reported a "recursive partitioning analysis", which was based on a data mining algorithm for generating decision trees. This analysis resulted in 100% classification accuracy of true accounts and 84% of false accounts. However, the algorithm is not explained and there appears to have been no cross-validation.

Leal, Vrij, Warmelink, Vernham, and Fisher (2015, Study 1)

Leal et al. (2015, Study 1) recruited 40 participants who either had experienced ($n = 19$) or falsely pretended to have experienced ($n = 21$) theft, loss, or accidental damage of a valuable item. All participants claimed to have experienced the incident during a phone call of a caller of an insurance company (10 different callers were involved). All callers believed

they were talking to a genuine client of the company. Truthful mock claimants were instructed to answer all questions truthfully, and deceptive claimants were “instructed to answer the questions in such a way that the interviewer would believe that indeed something had been accidentally damaged, stolen, or lost” (Leal et al., 2015, p. 134). The participants were forewarned that they would receive a phone call from the insurance company at least three days ahead of time; therefore, they had some time to prepare their claims. All participants were asked an open question to describe the event. “True” responses were about equally as long ($M = 144$ words; $SD = 83$) as lies ($M = 140$; $SD = 84$), $t(38) = 0.139$, $p = .890$, $g = 0.043$, 95% CI [-0.565, 0.651].⁵

These responses were transcribed and coded “by two experienced CBCA raters, who were blind to the veracity status of the transcripts” (p. 135), but no further information is provided in the report about the extent of the raters’ experience or training. The raters coded all CBCA criteria but CBCA10, CBCA13, CBCA18, and CBCA19. All criteria were rated as absent or present, except “visual details” (which we do not know whether they are *Unusual details*, *Superfluous details*, or both) and *Contextual embedding*, for which their frequency of occurrence was coded and the corresponding scores were dichotomized later.

The authors focused on the CBCA sum score, which was 4.00 ($SD = 2.00$) for truth tellers, and 3.36 ($SD = 1.70$) for liars. The difference was not significant, $t(38) = 1.09$, $p = .281$, $g = 0.339$, 95% CI [-0.273, 0.952]. Despite focusing on the summary score, the authors also reported the scores for five CBCA criteria that separated between truth tellers and liars with $p < .10$, four with an effect size in the expected direction, one in the opposite direction (CBCA05, which had a $d = -0.60$, not 0.60 as reported). Since no data were reported on the other 10 criteria, such selective reporting prevented us from including this study (Matt & Cook, 2009, 2019).

We also concluded that the way ground truth was established is not satisfactory. Truth tellers were all known personally to the experimenter, and the experimenter either personally

⁵ We calculated the unbiased effect sizes g and confidence intervals from the M s and SD s reported.

witnessed the event or the future participant spontaneously mentioned the event to the experimenter. While we find the first of these ground truth criteria to be acceptable, the second is highly questionable. Liars were either known to the experimenter or to other participants, and were asked to describe a false incident involving theft, loss, or accidental damage. However, there was no independent corroboration whether liars had experienced such an incident or not--i.e., for liars, the only ground truth was the person's word. The types of daily life events investigated can easily be constructed based on script knowledge, which may account for the nonsignificant differences for most criteria and the summary score.

The selective reporting of a subset of significant findings only leaves open plausible rival hypotheses to be considered in a quasi-experiment that may threaten internal, construct and external validity (see the extensive discussion in Shadish et al., 2002), which also led us to exclude this study.

Johnston, Candelier, Powers-Green, and Johnston (2016)

This study is a follow-up of Johnston et al. (2014), and shares some of its problems. The authors used a new set of criteria that contained eight items that were discriminative in their prior study. However, considering the way the items are described (see Appendix A in Johnston et al., 2016), it is questionable they still reflect the original CBCA criteria. Unfortunately, only two "true" and two "false" statements were included, which is problematic (e.g., Levine et al., 2022; Wells & Widnischtl, 1999). The statements were rated by a large sample of university students. The four statements had been collected by a forensic clinical psychologist to whom the defendants had been referred to for psychological assessment. Ground truth was again established based solely on conviction or acquittal in a court of law, which cannot be considered sufficient to establish ground truth. Of note, the forensic psychologist's report may have had an impact on the court's decision (possible circularity issue).

Wojciechowski, Gräns, and Lidén (2018, Experiment 2)

Three groups of raters ($N = 34$) were either untrained, trained in using SVA and CBCA (including all 19 CBCA criteria), or trained in employing an alternative credibility-assessment procedure. They had to assess the credibility of large numbers of "true" and

"false" denials, as well as of "true" confessions (no "false" confessions were included). All statements were transcripts of suspect interrogations conducted in criminal proceedings in Poland.

This study has several problematic design and statistical issues, but two aspects are critical. First, the way ground truth was established is not described, and the authors themselves acknowledged that "ground truth was not objectively established" (Wojciechowski et al., 2018, p. 17). Second, it seems that the suspects' verbatim statements were not available because "in the Polish protocols only a summary of a suspect's answers are saved" (p. 13). This is a major limitation because CBCA can only be used if the interviewee's own words are completely and accurately recorded.

Manzanero, Scott, Vallet, Aróztegui, and Bull (2019), and Sporer, Manzanero, and Masip (2021)

This is a quasi-experiment conducted in the field. Fifteen individuals (truth tellers) with mild to moderate, non-specific intellectual disability went on a day trip by bus. During the trip, the bus got on fire. Two years after the event, the bus passengers were interviewed individually about the trip. An additional group of 16 individuals (liars) who had not gone on the day trip but who were informed about it were also interviewed. They also had an intellectual disability. Truth tellers were instructed to tell the truth, while liars were asked to pretend they had gone on the trip and had experienced the fire incident. All interviews were conducted by two experienced forensic psychologists. Both interviewers were blind to the truth status of the participants. The interviews began with an open prompt, followed by five focused questions.

All statements were transcribed, and the transcripts were subsequently analyzed with 17 CBCA criteria (CBCA17 and CBCA18 were excluded).⁶ *Logical structure* and *Unstructured production* were measured on a 0 to 10 scale. To assess *Amount of detail*, a chart was made of "micro-propositions" about what had actually happened during the event,

⁶ The transcripts were also analyzed in terms of additional verbal content aspects, including several Reality Monitoring criteria (see Manzanero et al., 2014; Sporer et al., 2021).

and then the authors tallied the number of micro-propositions mentioned by each participant. All other 14 CBCA criteria were first coded as present or absent, and then the strength of presence was measured by counting the frequencies of occurrence for each criterion. Inter-rater agreement was good.

Additionally, 33 college students with no specific knowledge about credibility assessment or intellectual disability judged whether each statement was truthful or deceptive (though not all statements were judged by all students). However, note that these observers watched the statements on videotape (rather than reading transcripts). Therefore, their accuracy rates may not be comparable to CBCA scores based on analyses of transcripts.

The outcomes of this study were first reported by Manzanero et al. (2019). Subsequently, the data were re-analyzed by Sporer et al. (2021), who excluded from all analyses three participants whose truthfulness could not be assessed by the 33 naïve observers (because of problems to understand the participants' speech). Sample size in Sporer et al. was, therefore, $N = 29$ (13 females and 16 males; $M_{age} = 32.2$, $SD = 5.84$, range = 23 to 51). The participants' mean WAIS score was 60.3 ($SD = 9.53$, $Mdn = 55$, range = 48 to 80). Truth tellers ($n = 13$) and liars ($n = 16$) did not differ in age or WAIS scores.

Only three CBCA criteria differed significantly between truth tellers and liars. However, this may be due to the low statistical power of such a small sample. *Quantity of details*, *Contextual embedding*, *Reproductions of conversations* and *Superfluous details* showed medium to large effect sizes in the expected direction. Nevertheless, small negative effect sizes were found for *Unstructured production*, *Unexpected complications*, *Accounts of other people's mental state* and *Spontaneous corrections*, and a medium negative effect size ($d = -0.45$) was obtained for *Details misunderstood*. Sporer et al. (2021) conducted multiple discriminant analyses with the 11 CBCA criteria that had both positive associations in the expected direction and a positive corrected item-total correlation. After cross-validation using the leave-one-out method, these analyses yielded a correct classification rate of 62% for truthful accounts and 81% for deceptive accounts.

Manzanero et al. (2019) tested the potential of High-Dimensional Visualization (HDV) graphs to classify this pool of truthful and deceptive statements on the basis of the

CBCA criteria. An accuracy rate of 81% was reached. Finally, the 33 lay observers who made dichotomous judgments about the truthfulness of the videorecorded statements reached an accuracy rate of 59% for truthful statements and 65% for deceptive statements. Several additional analyses were reported by Manzanero et al. (2019) and Sporer et al. (2021).

The main reason why we could not include this small but interesting quasi-experiment in our systematic review is that the participants suffered from intellectual disability. The question whether CBCA can discriminate between self-experienced and invented accounts of people with intellectual disability is relevant and worth exploring. However, CBCA was not designed to be used with people suffering from cognitive impairments, and combining the outcomes of this study with those of other studies would be inappropriate, as the cognitive functioning of these individuals may influence the presence of CBCA criteria. Individuals with intellectual disabilities may have a hard time in creating a detailed story; as argued by Sporer et al. (2021), this may either (a) hinder the generation of rich accounts by truth tellers, which would decrease the differences between truthful and deceptive accounts, or (b) limit the liars' ability to invent a detailed and convincing account without a memory base, which would increase the differences between truthful and deceptive accounts.

Niveau (2020)

In this study, 96 child sexual abuse interviews collected by Swiss specialized police officers were examined. Two SVA experts (a psychologist and a psychiatrist) independently assessed the credibility of the statements. They agreed in 78% of the cases. Subsequently, they discussed the disagreements to reach a consensus. For all 96 cases included in the study, the experts' credibility decision was confirmed by at least one independent piece of evidence such as "a confession, retraction, material element, medical element or confirmed testimony" (Niveau, 2020, p. 98).⁷ In all, 62 cases were considered credible by the experts, while the remaining 34 were classified as not credible.

⁷ The original pool contained 127 cases. Thirty-one cases were excluded due to several reasons, including the absence of independent case evidence confirming the experts' judgment.

Niveau (2020) reported several analyses. Here we will refer to only those involving CBCA. The author reported the number (and percentage) of credible and non-credible accounts containing each of the 19 CBCA criteria. Unsurprisingly, all criteria were more frequently present in credible than in non-credible accounts, with the difference being significant for six criteria.

The main problem with this study is circularity, as CBCA scores were first used to classify the accounts as credible or not credible, and, subsequently, credible and not credible accounts were compared in terms of the presence of CBCA criteria. While these analyses are valuable to inform about the relative impact of each criterion on the raters' classifications of the accounts as credible or not credible, they do not speak about the validity of the criteria. Therefore, we were unable to include this study in our review. However, we did include a previous study by Welle et al. (2016), which contained about two-thirds of the cases in this study but without the circularity encountered here. Readers should consider these two studies together to understand the difficulties in testing the validity of lie and truth criteria in field settings.

Some Additional Studies

The studies summarized above were either field studies or quasi-experiments conducted in the field that we had selected based on our eligibility criteria. They were therefore assessed in terms of the ground truth criteria and methodological quality, and they received a score of 0. Above, in addition to summarizing these studies, we specify the reasons why we excluded them.

But there are a few additional studies that we would like to summarize here for different reasons. The first is a quasi-experiment authored by Pezdek et al. (2004). Some readers may be surprised that we did not include this study. However, in reality we excluded it early on in the selection process. The reason for this exclusion is that it contains no deceptive group. This may not be immediately apparent, and this is why we describe the study below.

We also summarize a couple of recent dissertations that we retrieved in our updated literature search. Though they include only laboratory experiments, these dissertations are innovative and open up new avenues of inquiry.

Because the three reports summarized below were excluded before coding for ground truth (Pezdek et al.'s because of containing no lie condition, and the two dissertations because of reporting only laboratory experiments), they were not included in the inter-coder reliability calculations reported in the systematic review.

Pezdek, Morrow, Blandón-Gitlin, Goodman, Quas, Saywitz, et al. (2004)

The main goal of this quasi-experiment was to examine whether “accounts of familiar events, regardless of truth value, receive higher CBCA scores than do accounts of unfamiliar events” (Pezdek et al., 2004, p. 121). Because the authors were interested in child sexual abuse, the participants were children ($N = 114$; 86 girls and 28 boys) and the target event was a painful, very invasive medical procedure called “voiding cystourethrogram fluoroscopy” (VCUG) that involves intrusive, forced genital contact. The authors empirically established that VCUG shares several important features with child sexual abuse (see Pezdek et al., 2004, pp. 121-122).

Two groups of children having undergone VCUG and having been videotaped during the procedure participated. Children in the relatively unfamiliar condition ($n = 49$) received a VCUG once, and those in the relatively familiar condition ($n = 27$) had received it multiple times. All children in these two conditions were interviewed (with free-recall questions) about what happened during the VCUG procedure. The authors also included a control condition of children who had not undergone VCUG. Instead, they were submitted to an anal-genital exam that was part of a routine medical check ($n = 38$). This latter group of children was interviewed about the anal-genital routine examination.

All interviews were transcribed and examined by two coders who had been trained in CBCA and had experience in using CBCA. They rated the first 16 CBCA criteria (as either 0 (criterion absent), 1 (criterion present) or 2 (criterion strongly present)). Inter-rater reliability was very high, with Cohen's *kappas* for the individual criteria ranging between .81 and 1.00, with mean *kappa* = .92.

The authors calculated the sum score across the 16 CBCA criteria that they considered. As predicted by the authors, the mean sum score was significantly higher for the relatively familiar group ($M = 10.67$, $SD = 5.73$) than for the relatively unfamiliar group ($M = 7.68$, $SD = 5.97$), $g = 0.503$, 95% CI [0.031, 0.975].⁸ The control group had a score of 9.92 ($SD = 5.07$), which did not differ significantly from either of the other two conditions; $g = 0.138$, 95% CI [-0.350, 0.626], and $g = -0.397$, 95% CI [-0.821, 0.027], for the relatively familiar and the relatively unfamiliar conditions respectively. But note that children in the control group also described a *truthful* event.

Pezdek et al. (2004) did not report descriptive data for individual criteria, but they showed that the relatively familiar and the relatively unfamiliar groups differed significantly for *General characteristics* (CBCA01 to CBCA 03) and *Specific contents* (CBCA04 to CBCA 13) criteria, but not for *Motivation-related criteria* (CBCA14 to CBCA16). Additional analyses revealed that children who had discussed the VCGU procedure with their mothers (which suggests they were more familiar with the procedure) had higher scores than those who had not. CBCA scores also correlated significantly with age, $r = .20$, $p < .05$.

We could not include this creative study in our systematic review because all three groups of children told the truth. As indicated above, the goal of Pezdek et al.'s study was to test whether familiarity with the target event increased CBCA ratings, not whether CBCA can discriminate between self-experienced and fabricated accounts. Besides, we focus on individual criteria, which were not reported in the article.

Maier (2021)

In his doctoral dissertation, Benjamin Maier (2021) sought to expand the theoretical basis of the CBCA approach by focusing on the creative and strategic demands on storytellers, both in true and false reports, as well as by examining the utility of a weighting system for integrating individual criteria. Three laboratory experiments focused on these goals by investigating both memory-based criteria and script-related criteria (based on

⁸ We calculated the unbiased effect sizes g and confidence intervals from the values reported by Pezdek et al. (2004) in their Table 2.

Volbert and Steller's, 2014, revised model of CBCA criteria). In Study 1, 30 non-student participants reported on autobiographical events (two truthful, two invented) selected from a list of topics presented to them (similar to the study by Steller et al., 1992). Study 2 was a questionnaire study investigating 135 storytellers' content-related strategies. In Study 3, 90 participants (63 non-students, the rest students) told two truthful and two false events as in Study 1. Taken together, all three studies can be considered laboratory analogue experiments and thus were not eligible for our systematic review.

Roenspies-Heitman (2022)

In recent years, several authors have stressed the importance of witnesses' strategic considerations that may affect the presence of motivational criteria in their accounts (Niehaus, 2001; Volbert & Steller, 2008). Roenspies-Heitman (2022) used Volbert and Steller's (2014) re-grouped CBCA criteria-catalogue. She conducted two studies in which 248 pupils between 12 and 19 years either participated in a dyadic interaction game in a school classroom (an unpleasant activity that was considered a laboratory analogue for a negative emotional experience) or fabricated an account based on instructions parallel to the experience group. Hence, the study can be characterized as a laboratory analogue, not a field study, and could not be included in our systematic review.

References

- Adams, S. H. (1996). Statement analysis: What do suspects' words really reveal? *FBI Law Enforcement Bulletin*, 65, 12-20.
- Adams, S. H. (2002/2004). *Communication under stress: Indicators of veracity and deception in written narratives* (Publication No: 3110253) [Doctoral dissertation, Virginia Polytechnic Institute and State University]. ProQuest Information and Learning Company.
- Adams, S., & Jarvis, J. P. (2006). Indicators of veracity and deception: an analysis of written statements made to police. *International Journal of Speech Language and the Law*, 13, 1-22. <https://doi.org/10.1558/sll.2006.13.1.1>
- Borenstein, M. (2009). Effect sizes for continuous data. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2nd ed., pp. 221-235). Russell Sage Foundation.
- Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2019). *Introduction to meta-analysis*. Wiley.
- Brodie, L. A. (1993). *Making judgments regarding child sexual abuse: The impact of professional background, experience and knowledge base* (Publication No: 9329404) [Doctoral dissertation, Biola University]. ProQuest Information and Learning Company.
- Burgess, A. W., & Hazelwood, R. R. (2001). False rape allegations. In R. R. Hazelwood & A. W. Burgess (Eds.), *Practical aspects of rape investigation: A multidisciplinary approach* (pp. 177-197). CRC Press.
- Chang, G. H.-Y. (2008). *Effectiveness of content analysis in assessing suspect credibility: Counterterrorism implications* (Publication No: 3315323) [Doctoral dissertation, The University of Nebraska, Lincoln]. ProQuest Dissertations & Theses Full Text: The Humanities and Social Sciences Collection.
- Critchlow, N. (2011). *Applying criteria based content analysis to assessing the veracity of rape statements* [Undergraduate project report]. Manchester Metropolitan University. <https://e-space.mmu.ac.uk/576508/>

- Dana-Kirby, L. (1997). *Discerning truth from deception: Is Criteria-based Content Analysis effective with adult statements?* (Publication No: 9723291) [Doctoral dissertation, University of Oregon]. ProQuest Dissertations & Theses Full Text: The Humanities and Social Sciences Collection.
- Esplin, P. W., Houed, T., & Raskin, D. C. (1988, June). *Application of statement validity assessment* [Paper presentation]. NATO Advanced Study Institute on Credibility Assessment, Maratea, Italy.
- Geiselman, R. E., & Fisher, R. P. (1989). The Cognitive Interview technique for victims and witnesses of crimes. In D. C. Raskin (Ed.), *Psychological methods in criminal investigation and evidence* (pp. 191–215). Springer.
- Greuel, L. (1993). *Polizeiliche Vernehmung vergewaltigter Frauen* [Police interrogations of raped women]. Weinheim, Germany: Psychologie Verlags Union.
- Hauch, V., Sporer, S. L., Masip, J., & Blandón-Gitlin, I. (2017). Can credibility criteria be assessed reliably? A meta-analysis of Criteria-based Content Analysis. *Psychological Assessment*, 29, 819-834. <https://doi.org/10.1037/pas0000426>
- Hershkowitz, I. (1999). The dynamics of interviews involving plausible and implausible allegations of child sexual abuse. *Applied Developmental Science*, 3, 86-91. https://doi.org/10.1207/s1532480xads0302_3
- Hershkowitz, I. (2001). A case study of child sexual false allegation. *Child Abuse and Neglect*, 25, 1397-1411. [https://doi.org/10.1016/S0145-2134\(01\)00274-5](https://doi.org/10.1016/S0145-2134(01)00274-5)
- Horowitz, S. W., Lamb, M. E., Esplin, P. W., Boychuck, T. D., Reiter-Lavery, L., & Krispin, O. (1995). Establishing ground truth in studies of child sexual abuse. *Expert Evidence*, 4 (2), 42–51.
- Johnston, S., Candelier, A., Powers-Green, D., & Johnston, G. (2016). Attributes of true and deceptive statements made in evaluations of criminal defendants. *Journal of Forensic Psychology Practice*, 16, 347-373. <https://doi.org/10.1080/15228932.2016.1219218>
- Johnston, S., Candelier, A., Powers-Green, D., & Rahmani, S. (2014). Attributes of truthful versus deceitful statements in the evaluation of accused child molesters. *SAGE Open*, 4(3). <https://doi.org/10.1177/2158244014548849>

- Kleinberg, B., Arntz, A., & Verschuere, B. (2019). Being accurate about accuracy in verbal deception detection. *PLoS ONE* 14(8), e0220228.
<https://doi.org/10.1371/journal.pone.0220228>
- Knoernschild, C., & van Koppen, P. (2003). Psychological expert witnesses in Germany and the Netherlands. In P. J. van Koppen & S. D. Penrod (Eds.), *Adversarial vs. inquisitorial justice: Psychological perspectives on criminal justice systems* (pp. 255-282). Plenum.
- Krahé, B., & Kundrotas, S. (1992). Glaubwürdigkeitsbeurteilung bei Vergewaltigungsanzeigen: Ein aussagenanalytisches Feldexperiment [Assessing the credibility of rape reports: A field study of statement analysis]. *Zeitschrift für experimentelle und angewandte Psychologie*, 39, 598-620.
- Krahé, B., Reimer, T., & Scheinherger, R. (1995, October). *Kriterienorientierte Aussageanalyse als Instrument zur Glaubhaftigkeitsbeurteilung: Eine methodenkritische Untersuchung* [Criteria-based statement analysis as an instrument for assessing credibility: A method-critical study] [Conference presentation]. 6th Meeting of the German Psychology and Law Association in Bremen, Germany.
- Lamb, M. E., Orbach, Y., Sternberg, K. J., Hershkowitz, I., & Horowitz, D. (2000). Accuracy of investigators' verbatim notes of their forensic interviews with alleged child abuse victims. *Law and Human Behavior*, 24, 699-707.
<https://doi.org/10.1023/A:1005556404636>
- Lamers-Winkelmann, F. (1999). Statement Validity Analysis: Its application to a sample of Dutch children who may have been sexually abused. In K. Coulborn Faller & R. VanderLaan (Eds.), *Maltreatment in early childhood. Tools for research-based intervention* (pp. 59-81). The Haworth Press.
- Lamers-Winkelmann, F., & Buffing, F. (1996). Children's testimony in The Netherlands: A study of Statement Validity Analysis. In B. L. Bottoms & G. S. Goodman (Eds.), *International perspectives on child abuse and children's testimony* (pp. 45-61). Sage.
- Lamers-Winkelmann, F., Buffing, F., & van der Zanden, A. P. (1992, June 29 – July 3). *Interviewing child victims of sexual abuse: What children can or will tell about sexual*

- abuse: Preliminary results* [Lecture]. VIII postgraduate course “The victim and the criminal justice system”, Vrije Universiteit, Amsterdam, The Netherlands.
- Leal, S., Vrij, A., Warmelink, L., Vernham, Z., & Fisher, R. P. (2015). You cannot hide your telephone lies: Providing a model statement as an aid to detect deception in insurance telephone calls. *Legal and Criminological Psychology*, 20, 129-146.
<https://doi.org/10.1111/lcrp.12017>
- Levine, T. R., Daiku, Y., & Masip, J. (2022). The number of senders and total judgments matter more than sample size in deception detection experiments. *Perspectives on Psychological Science*, 17, 191-204. <http://doi.org/10.1177/1745691621990369>
- Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis*. Thousand Oaks, CA: Sage.
- Maier, B. G. (2021). *Which processes govern the emergence of CBCA criteria in witness statements? An integrated and theory-driven approach for analysis* [Doctoral Dissertation]. Freie Universität Berlin, Germany.
- Manzanero, A. L., Alemany, A., Recio, M., Vallet, R., & Aróztegui, J. (2015). Evaluating the credibility of statements given by persons with intellectual disability. *Anales de Psicología*, 31, 338–344. <https://doi.org/10.6018/analesps.31.1.166571>
- Manzanero, A. L., Scott, M. T., Vallet, R., Aróztegui, J., & Bull, R. (2019). Criteria-Based content analysis in true and simulated victims with intellectual disability. *Anuario de Psicología Jurídica*, 29, 55–60. <https://doi.org/10.5093/apj2019a>
- Matt, G. E., & Cook, T. D. (2009). Threats to the validity of generalized inferences. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis* (2nd ed., pp. 537-560). Russell Sage Foundation.
- Matt, G. E., & Cook, T. D. (2019). Threats to the validity of generalized inferences from research syntheses. In H. Cooper, L. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (3rd ed., pp. 489-516). Russell Sage Foundation.
- Milne, R., Nunan, J., Hope, L., Hodgkins, J., & Clarke, C. (2022). From verbal account to written evidence: Do written statements generated by officers accurately represent

- what witnesses say? *Frontiers in Psychology*, 6519.
<https://doi.org/10.3389/fpsyg.2021.774322>
- Niehaus, S. (2001). *Zur Anwendbarkeit inhaltlicher Glaubhaftigkeitsmerkmale bei Zeugenaussagen unterschiedlichen Wahrheitsgehaltes* [Applicability of content criteria to testimonies with different truth status]. Europäische Hochschulschriften.
- Niveau, G. (2020). Sensory information in children's statements of sexual abuse. *Forensic Sciences Research*, 6, 97-102. <https://doi.org/10.1080/20961790.2020.1814000>
- Orbach, Y., & Lamb, M. E. (1999). Assessing the accuracy of a child's account of sexual abuse: A case study. *Child Abuse & Neglect*, 23, 91-98.
[https://doi.org/10.1016/S0145-2134\(98\)00114-8](https://doi.org/10.1016/S0145-2134(98)00114-8)
- 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. <https://doi.org/10.1348/135532500168119>
- Pezdek, K., Morrow, A., Blandón-Gitlin, I., Goodman, G. S., Quas, J. A., Saywitz, K. J., Bidrose, S., Pipe, M.-E., Rogers, M., & Brodie, L. (2004). Detecting deception in children: Event familiarity affects Criterion-based Content Analysis ratings. *Journal of Applied Psychology*, 89, 119–126. <https://doi.org/10.1037/0021-9010.89.1.119>
- Rabon, D. (1996) *Investigative discourse analysis*. Carolina Academic Press.
- 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. Implications for eyewitness testimony* (pp. 153-164). American Psychological Association.
- Raskin, D. C., & Esplin, P. W. (1991b). Commentary: Response to Wells, Loftus, and McGough. In J. Doris (Ed.), *The suggestibility of children's recollections. Implications for eyewitness testimony* (pp. 172-176). American Psychological Association.
- Rassin, E., & van der Sleen, J. (2005). Characteristics of true versus false allegations of sexual offences. *Psychological Reports*, 97, 589-598.
<https://doi.org/10.2466/pr0.97.2.589-598>

- Roenspies-Heitman, J. (2022). *Kriterienorientierte Inhaltsanalyse von Zeugenaussagen: Eine empirische Untersuchung zur Validität ausgewählter Glaubhaftigkeitsmerkmale* [Criteria-based Content Analysis of witness statements: An empirical investigation of selected credibility criteria]. Springer Nature. <https://doi.org/10.1007/978-3-658-36655-1>
- Sánchez-Meca, J., Chacón-Moscoso, S., & Marín-Martínez, F. (2003). Effect-size indices for dichotomized outcomes in meta-analysis. *Psychological Methods*, 8, 448-467.
<https://doi.org/10.1037/1082-989X.8.4.448>
- Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Houghton Mifflin Company.
- Sporer, S. L. (2012). Making the subjective objective? Computer-assisted quantification of qualitative content cues to deception. In E. Fitzpatrick, J. Bachenko, & T. Fornaciari (Eds.), *Proceedings of the Workshop on Computational Approaches to Deception Detection* (pp. 78–85). Association for Computational Linguistics.
<https://aclanthology.org/W12-0412.pdf>
- Sporer, S. L., Manzanero, A. L., & Masip, J. (2021). Optimizing CBCA and RM research: Recommendations for analyzing and reporting data on content cues to deception. *Psychology, Crime, & Law*, 27, 1-39.
<https://doi.org/10.1080/1068316X.2020.1757097>
- Steck, P., Hermanutz, M., Lafrenz, B., Schwind, D., Hettler, S., Maier, B., & Geiger, S. (2010). *Die psychometrische Qualität von Realkennzeichen* [The psychometric quality of reality criteria]. <https://doi.org/10.25968/opus-263>
- Steller, M., Wellershaus, P., & Wolf, P. (1992). Realkennzeichen in Kinderaussagen: Empirische Grundlagen der kriterienorientierten Aussagenanalyse [Reality criteria in children's testimonies: Empirical principles of Criteria-based Content Analysis]. *Zeitschrift für Experimentelle und Angewandte Psychologie*, 39, 151–170.
- Suiter, T. L. (2001). *Linguistic foundations of Investigative Discourse Analysis and Content Based Criteria Analysis*. (Publication No: 9988194) [Doctoral dissertation, Cornell University]. ProQuest Information and Learning Company.

- Valentine, J. (2009). Judging the quality of primary research. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2nd ed., pp. 129-146). Russell Sage Foundation.
- van Koppen, P. J., & Saks, M. J. (2003). Preventing bad psychological scientific evidence in The Netherlands and The United States. In P. J. van Koppen & S. D. Penrod (Eds.), *Adversarial vs. inquisitorial justice: Psychological perspectives on criminal justice systems* (pp. 283-307). Plenum.
- Volbert, R., & Steller, M. (Eds.). (2008). *Handbuch der Rechtspsychologie* [Handbook of psychology and law]. Hogrefe Verlag.
- Volbert, R., & Steller, M. (2014). Is this testimony truthful, fabricated, or based on false memory? *European Psychologist*, 19, 207–220. <http://doi.org/10.1027/1016-9040/a000200>
- Welle, I., Berclaz, M., Lacasa, M.-J., & Niveau, G. (2016). A call to improve the validity of criterion-based content analysis (CBCA): Results from a field-based study including 60 children's statements of sexual abuse. *Journal of Forensic and Legal Medicine*, 43, 111-119. <https://doi.org/10.1016/j.jflm.2016.08.001>
- Wells, G. L., & Loftus, E. (1991). Commentary: Is this child fabricating? Reactions to a new assessment technique. In J. Doris (Ed.), *The suggestibility of children's recollections. Implications for eyewitness testimony* (pp. 168-171). American Psychological Association.
- Wells, G. L., & Windschitl, P. D. (1999). Stimulus sampling and social psychological experimentation. *Personality and Social Psychology Bulletin*, 25, 1115-1125. <https://doi.org/10.1177/01461672992512005>
- Wojciechowski, B. W. (2014). Content analysis algorithms: An innovative and accurate approach to statement veracity assessment. *European Polygraph*, 8 (3), 119-128.
- Wojciechowski, B. W., Gräns, M., & Lidén, M. (2018). A true denial or a false confession? Assessing veracity of suspects' statements using MASAM and SVA. *PLoS ONE*, 13(6), e0198211. <https://doi.org/10.1371/journal.pone.0198211>

Yuille, J. C., & Cutshall, J. (1989). Analysis of the statements of victims, witnesses and suspects. In J. C. Yuille (Ed.), *Credibility assessment* (pp. 175-191). Kluwer.