The European Journal of Psychology Applied to Legal Context 7 (2015) 3-12

The European Journal of Psychology Applied to Legal Context

www.elsevier.es/ejpal

Undeutsch hypothesis and Criteria Based Content Analysis: A meta-analytic review

Bárbara G. Amado[a], Ramón Arce[a], and Francisca Fariña[b]

[a]Department of Organizational and Legal Forensic Psychology, and Behavioral Sciences Methodology, University of Santiago de Compostela, Spain
[b]AIPSE Department, University of Vigo, Spain

ARTICLE INFORMATION

Manuscript received: 03/01/2014
Revision received: 30/07/2014
Accepted: 04/08/2014

Keywords:
Meta-analysis
CBCA
Credibility
Testimony
Sexual abuse
Child

ABSTRACT

The credibility of a testimony is a crucial component of judicial decision-making. Checklists of testimony credibility criteria are extensively used by forensic psychologists to assess the credibility of a testimony, and in many countries they are admitted as valid scientific evidence in a court of law. These checklists are based on the Undeutsch hypothesis asserting that statements derived from the memory of real-life experiences differ significantly in content and quality from fabricated or fictitious accounts. Notwithstanding, there is considerable controversy regarding the degree to which these checklists comply with the legal standards for scientific evidence to be admitted in a court of law (e.g., *Daubert standards*). In several countries, these checklists are not admitted as valid evidence in court, particularly in view of the inconsistent results reported in the scientific literature. Bearing in mind these issues, a meta-analysis was designed to test the Undeutsch hypothesis using the CBCA Checklist of criteria to discern between memories of self-experienced real-life events and fabricated or fictitious accounts. As the original hypothesis was formulated for populations of children, only quantitative studies with samples of children were considered for this study. In line with the Undeutsch hypothesis, the results showed a significant positive effect size that is generalizable to the total CBCA score, $\delta = 0.79$. Moreover, a significant positive effect size was observed in each and all of the credibility criteria. In conclusion, the results corroborated the validity of the Undeutsch hypothesis and the CBCA criteria for discriminating between the memory of real self-experienced events and false or invented accounts. The results are discussed in terms of the implications for forensic practice.

Palabras clave:
Meta-análisis
CBCA
Credibilidad
Testimonio
Abusos sexuales
Menores

La hipótesis Undeutsch y el "Criteria Based Content Analysis": una revisión meta-analítica

RESUMEN

Con frecuencia, la evaluación de la fiabilidad de un testimonio se lleva a cabo mediante el uso de sistemas categoriales de análisis de contenido. Concretamente, el instrumento más utilizado para determinar la credibilidad del testimonio es el Criteria Based Content Analysis (CBCA), el cual se sustenta en la hipótesis Undeutsch, que establece que las memorias de un hecho auto-experimentado difieren en contenido y calidad de las memorias fabricadas o imaginadas. Las opiniones y resultados contradictorios encontrados en la literatura científica respecto al cumplimiento de los criterios judiciales (*Daubert standards*) así como el abundante número de trabajos existentes sobre la materia, nos llevó a diseñar un meta-análisis para someter a prueba la hipótesis Undeutsch, a través de la validez de los criterios de realidad del CBCA para discriminar entre la memoria de lo auto-experimentado y lo fabricado. Se tomaron aquellos estudios cuantitativos que incluían muestras de menores, esto es, con edades comprendidas entre los 2 y 18 años. En línea con la hipótesis Undeutsch, los resultados mostraron un tamaño del efecto positivo, significativo y generalizable para la puntuación total del CBCA, $\delta = 0.79$. Asimismo, en todos los criterios de realidad se encontró un tamaño del efecto positivo y significativo. En conclusión, los resultados avalan la validez de la hipótesis Undeutsch y de los criterios del CBCA para discriminar entre memorias de hechos auto-experimentados y fabricados. Se discuten las implicaciones de los resultados para la práctica forense.

*Correspondence concerning this article should be sent to Ramón Arce. Departamento de Psicología Organizacional, Jurídico-Forense y Metodología de las Ciencias de Comportamiento. Universidad de Santiago de Compostela. Facultad de Psicología. Campus Vida, s/n. E-15782 Santiago de Compostela, A Coruña, Spain.
E-mail: ramon.arce@usc.es

http://dx.doi.org/10.1016/j.ejpal.2014.11.002
1889-1861/

Hans and Vidmar (1986) estimated that in around 85% of judicial cases, the evidence bearing most weight is the testimony, which underscores that the evaluation of a testimony is crucial for judicial judgement making. In terms of the application of Information Integration Models to legal judgements (Kaplan, 1982), the reliability and validity of the testimony are the mechanisms underlying the evaluation of a testimony. The validity of a testimony, i.e., the value of a testimony for judgement making is easily estimated and is to be determined by the rulings of judges and the courts. As for the reliability of a testimony, the courts and scientific studies have tended to estimate it in terms of the credibility of a testimony (Arce, Fariña, & Fraga, 2000), which entails the design of methods for its estimation. Traditionally, judges and the courts have performed this function on the basis of legal criteria, jurisprudence, and their own value judgements. Alternatively, numerous scientific techniques (and pseudoscientific) have been proposed such as non-verbal indicators of deception, paraverbal indicators of deception, physiological indicators (e.g., polygraph tests or functional magnetic resonance imaging), and categorical systems of content analysis. Of these, categorical systems of content analysis are currently the most systemically used technique by the courts. Thus, the courts in countries such as Germany, Sweden, Holland, and several states in the USA admit these categorical systems as scientific evidence (Steller & Böhm, 2006; Vrij, 2008). In Spain, where they are also admitted as legally admissible evidence and extensively used by the courts, an analysis of legal judgements showed that when a forensic psychological report based on a categorical system of content analysis (i.e., Statement Validity Analysis, SVA) confirmed the credibility of a testimony, the conviction rate was 93.3%, but when it failed to do so, the acquittal rate was 100%. In contrast, in other countries such as the UK, the US, and Canada these checklists are not admitted as legally valid evidence (Novo & Seijo, 2010).

Underlying categorical content systems is what is commonly referred to as the *Undeutsch hypothesis* that asserts that the memory of a real-life self-experienced event differs in content and quality from a fabricated or imagined event (Undeutsch, 1967, 1989). On the basis of this hypothesis, Steller and Köhnken (1989) have integrated all the categorical systems (e.g., Arntzen, 1970; Dettenborn, Froehlich, & Szewczyk, 1984; Szewczyk, 1973; Undeutsch, 1967) into what is known as Criteria Based Content Analysis (CBCA), which has become the leading categorical system for evaluating the credibility of a testimony (Griesel, Ternes, Schraml, Cooper, & Yuille, 2013; Vrij, 2008).

CBCA, which is part of SVA, consists of three elements: 1) semi-structured interview, i.e., the free narrative interview; 2) content analysis on CBCA criteria; and 3) evaluation of CBCA outcomes using the Validity Checklist. The semi-structured interview involves a narrative format that, unlike other types of interview such as standard, interrogative or structured interviews, facilitates the emergence of criteria (Vrij, 2005). Moreover, this type of interview generates more information (Memon, Meissner, & Fraser, 2010), which meets the requirement that CBCA criteria content analysis be performed on sufficient material (Köhnken, 2004; Steller, 1989).

The Checklist of CBCA criteria (Steller & Köhnken, 1989) consists of 19 criteria structured around 5 major categories: general characteristics, specific contents, peculiarities of the content, contents related to motivation, and specific elements of aggression (see Table 1). These criteria of reality do not constitute a methodic categorical system (Bardin, 1977; Weick, 1985), but rather stem from the authors' personal experiences of cases (Steller & Köhnken, 1989). Though this checklist was originally developed as a comprehensive system of credibility criteria grounded on the Undeutsch hypothesis, Raskin, Esplin, and Horowitz (1991) highlighted that only the first 14 criteria are related to the Undeutsch hypothesis, and the remaining 5 criteria are not associated to the aforementioned hypothesis as they are not linked to the concept of memory of actual events. This reclassification overlaps, though not entirely, with the theoretical model proposed by Köhnken (1996), who regroups these major categories into two main factors: cognitive (criteria 1 to 13) and motivational (criteria 14 to 18). The cognitive factor encompasses cognitive and verbal skills, and implies that a self-experienced statement contains CBCA criteria from 1 to 13. The motivational factor, however, relies on the individual's ability to avoid appearing deceitful and ways of managing a positive self-impression of oneself as an honest witness. Thus, the motivational factor covers criteria 14 to 18, which are contrary-to-truthfulness-stereotype criteria though they really appear in true statements. Thus, these criteria have been suggested to be useful for assessing the hypothesis of the (partial) fabrication of statements (Köhnken, 1996, 2004).

Initially, the CBCA criteria were intended for populations of child alleged victims of sexual abuse. However, CBCA criteria have been applied to other types of events and age ranges. This generalization has been extended to professional practice too. Thus, the guidelines of the Institute of Forensic Medicine in Spain, which is the official public institution responsible for forensic evidence, recommends SVA as part of the protocol for women alleging intimate partner violence (Arce & Fariña, 2012). Moreover, there is no consensus regarding the term *minor*, particularly since studies use the term range from 2 to 18-year-olds and the concept of minor is generally associated to the legal age of criminal responsibility. In relation to the context of application, only field studies involve real cases of sexual abuse, since it would be unethical to subject children to conditions or instructions of victims of sexual abuse. Hence, most research is experimental and certain authors have expressed their reservations regarding validity (Konecni & Ebbesen, 1992). Moreover, real eyewitnesses and subjects under high fidelity laboratory conditions have been found to perform different tasks (Fariña, Arce, & Real, 1994). In order to overcome this limitation, some experimental studies have recreated high fidelity simulated conditions in order to mimic the context of recall of child alleged victims of sexual abuse. These conditions have been defined as personal involvement, negative emotional tone of an event, and extensive loss of control over the situation (Steller, 1989). Accordingly, this achieves face validity, with

Table 1
CBCA-Criteria (adapted from Steller & Köhnken, 1989)

General characteristics
1. Logical structure
2. Unstructured production
3. Quantity of details
Specific contents
4. Contextual embedding
5. Descriptions of interactions
6. Reproduction of conversation
7. Unexpected complications during the incident
Peculiarities of content
8. Unusual details
9. Superfluous details
10. Accurately reported details misunderstood
11. Related external associations
12. Accounts of subjective mental states
13. Attribution of perpetrator's mental state
Offence-specific elements
14. Spontaneous corrections
15. Admitting lack of memory
16. Raising doubts about one's own testimony
17. Self-deprecation
18. Pardoning the perpetrator
Offence-specific elements
19. Details characteristic of the offence

external validity remaining entirely untested (Konecni & Ebbesen, 1992). Nevertheless, in spite of the weaknesses of this experimental paradigm in generalizing CBCA outcomes to the forensic context, experimental laboratory studies are useful for assessing certain variables that may lead to further internal validity research (Griesel et al., 2013). In contrast, the limitation of field studies resides in their difficulty in sustaining the ground truth in accurate objective criteria. These differences in research paradigms imply more credibility criteria are observed in field studies than in experimental ones (Vrij, 2005).

The CBCA criteria are measured on two response scales, presence vs. absence and the degree of presence. The unit of analysis is the full statement for the first major category, *general characteristics*, and for the remainder frequency counts. The presence of reality criteria is assumed to be indicative of memory based on real-life events, but the absence of criteria does not imply recall is based on fabricated accounts. Additionally, fictitious memory may contain reality criteria. Thus, the evaluation rests on a clinical judgement (Köhnken, 2004), which is semi-objective. Nonetheless, a replicable objective evaluation system, i.e. stringent, is a fundamental standard for forensic practice. Alternatively a variety of decision rules have been proposed such as the presence of 3 criteria to judge a statement as true (Arntzen, 1983); of 7 criteria, criteria 1 to 5, plus 2 others; or the first three ones, plus 4 others (Zaparniuk, Yuille, & Taylor, 1995). Unfortunately, these decision rules lack empirical rigour.

Furthermore, the application of criteria derived from the Undeutsch hypothesis to the forensic assessment of the credibility of a testimony (to be more precise, the truthfulness of a testimony given that credibility is a legal concept defined by the ruling of judges and the courts) must fulfil the legal standards for scientific evidence to be admitted as such in court. The standards governing the admission of expert testimony in court have been laid down by the Supreme Court of the United States in Daubert v. Merrell Dow Pharmaceuticals (1993) and are as follows: 1) is the scientific hypothesis testable?; 2) has the proposition been tested?; 3) is there a known error rate?; 4) has the hypothesis and/or technique been subjected to peer review and publication; and 5) is the theory upon which the hypothesis and/or technique is based on generally accepted among the appropriate scientific community? Several authors have raised their doubts on whether SVA/CBCA fulfil the above criteria, and this has led to seemingly contradictory viewpoints (Honts, 1994; Vrij, 2008). Bearing in mind these contradictory results and interpretations, and the wealth of scientific evidence, the aim of this meta-analysis was to review the literature on both experimental and field studies in order to assess the degree to which the Undeutsch hypothesis meets the judicial standards of evidence by estimating the effect size of the CBCA criteria. As initially the hypothesis was formulated for a population of children, though it has also been applied to adults, this review was restricted to studies on samples of children.

Method

Literature Search

The aim of the scientific literature search was to identify all of the empirical studies assessing the efficacy of CBCA criteria in discriminating between true statements of actual experiences and the invented, imagined, fictitious, fabricated, or false accounts of children. An exhaustive multi-method search was undertaken in the following international psychology databases of reference: PsycInfo and all of the databases of Web of Science; the Spanish language databases of reference Psicodoc (database of the Official Spanish College of Psychology); the Italian databases (ACPN, Archivio Collettivo Nazionale dei Periodici); the German Psychlinker databases; and the French human, social sciences, and economics Francis databases; the Google Scholar meta-search engine of scientific articles; a manual search in books; crosschecking all the references included in published reviews of articles and manuals; and directly contacting authors to request copies of unavailable studies. The keywords entered in the search engines were: Criteria Based Content Analysis or CBCA (kriteriumbasierte inhaltsanalyse, análisis de contenido basado en criterios, analisi del contenuto basata su criteri), credibility (glaubwürdigkeit, credibilidad, credibilità), content analysis (inhaltsanalyse, análisis de contenido, analisi del contenuto), child sexual abuse (kindesmissbrauch, abusos sexuales a menores, abuso sessuale perpetrato su minori), child testimony (kindliche zeugenaussage, testimonio del menor, bambini testimoni). The searches in the databases of non-English speaking countries were undertaken in the corresponding language. In line with the method of successive approximations, all of the keywords in the selected articles were revised in search of other potential descriptors. However, successive searches with these new descriptors failed to produce any further studies for the meta-analysis.

Inclusion and Exclusion Criteria

Though the Undeutsch hypothesis was initially formulated for children, the exact age group it encompasses has never been clearly specified. Nevertheless, as it was intended for judicial contexts and victims of sexual abuse, it is understood that it refers to children under the age of consent. First, the literature has taken the legal concept of minor as below the age of criminal responsibility (< 18 years) (Raskin & Esplin, 1991). Likewise, the lower age group was not related to the model, but the hypothesis is supported in that the memory of genuine life experiences differs in content and quality to memory of fictitious accounts, with the criteria for discriminating between both types of memory being derived from the witness' verbal account. Thus, the child is expected to have the sufficient narrative capability to express these criteria (Köhnken, 2004). Once again, the lower age limit was crosschecked in the studies reviewed, observing the lowest was a 2-year-old child (Buck, Warren, Betman, & Brigham, 2002; Lamers-Winkelman & Buffing, 1996), whose narrative skills, memory, gaps in memory, and recovery may be insufficient. Notwithstanding, given that the scientific literature has set a minimum age of 2 years and a maximum of 18 years, all of the studies with witnesses between these ages were included. Thus, studies with samples of children or that calculated the effect size for the subsamples of children were included.

Second, delimiting the testimony to sexual abuse would compel studies to focus on this type of victim, as it would be unethical to subject children to memories of feigned victims. Consequently, the studies reviewed can be subdivided into low fidelity experimental studies (i.e., the scenarios neither involved sexual abuse nor was the implication or motivation of the participants controlled), high fidelity experiments (i.e., the scenarios do not involve sexual abuse, but they create an emotionally charged contexts close to the victimization of sexual abuse, and the implication of the participants is controlled), and field studies (i.e., real cases of sexual abuse where the ground truth is based on judicial judgements, the confession of the accused, medical evidence, and polygraph tests). The effects of the context of the research (i.e., field vs. laboratory high fidelity studies) on the results of the quality of an eyewitness' identification have been found to be significant, and even contradictory (Fariña et al., 1994). Thus, it would be plausible to believe that this same bias may also affect the testimony of children alleged victims of sexual abuse. Hence, according to the circumstances, the context of the research was considered as a moderator.

Third, for the criteria to discriminate between memories of real events and fabricated accounts, SVA proposes statements should be obtained using a free narrative interview, e.g., step-wise interview,

cognitive interview, Memorandum of Good Practices (Köhnken, 2004; Steller, 1989; Undeutsch, 1989), as they facilitate the emergence of criteria (Vrij, 2005). Likewise, compliance with other SVA criteria was strictly observed, i.e., studies noncompliant with any of the basic characteristics of the interview, such as inappropriate prompts or suggestions, were excluded.

Fourth, the studies admitted as legal evidence should be published in scientific peer-reviewed journals (Daubert v. Merrell Dow Pharmaceuticals, 1993). Notwithstanding, the literature has identified as key references studies that have not been published in these journals (i.e., Boychuk, 1991; Esplin, Houed, & Raskin, 1988), and these have been included in the meta-analysis. Thus, according to the circumstances, compliance with the peer-review publication Daubert standard was taken as a moderator.

Fifth, the effect size was calculated from the data obtained and, if required, the authors were contacted to request the effect size or data for computing it as well as to clarify errors or queries regarding the data.

Sixth, studies with samples shared with other studies were excluded (i.e., Hershkowitz, 1999), to avoid empirical redundancy (duplicity in publishing data) – only the original study and the outlier values [IQR ± 1.5] were included. An independent analysis and control of outliers was carried out for each meta-analysis.

A total of 20 publications fulfilled the selection criteria, ranging from 5 effect sizes for *self-deprecation* and *details characteristic of the offence* to 17 effect sizes for the criterion quantity of details (see Table 2), and 22 for the total CBCA score.

Procedure

Having scanned the literature and selected the articles for this meta-analysis, they were coded according to the variables that could function as a moderator. The literature cites the age of the child, the research paradigm (field studies vs. experimental laboratory studies), and the type of design (within- or between-subject), which may mediate the results. Numerous studies have found a correlation between age and total CBCA score, i.e., the older the child, the greater probability of scoring high on this instrument (Anson, Golding, & Gully, 1993; Buck et al., 2002; Craig, Scheibe, Raskin, Kircher, & Dodd, 1999; Horowitz, Lamb, Esplin, Boychuk, Krispin, & Reiter-Lavery, 1997; Lamb, Sternberg, Esplin, Hershkowitz, & Orbach, 1997; Roma, San Martini, Sabatello, Tatarelli, & Ferracuti, 2011). Moreover, if we take into account the child's development, it is probable that to some extent the presence of certain criteria is influenced by the child's age. In relation to the research paradigm, Vrij (2005) found the differences between statements of genuine experiences and invented ones were greater in field studies than in experimental studies. Nonetheless, the results obtained from both field studies and laboratory experiments corroborated the Undeutsch hypothesis. As for the type of design (within- vs. between-subject), criteria have been found to be sensitive to the type of design, whereas total CBCA scores were not (Bensi, Gambetti, Nori, & Giusberti, 2009). Moreover, within-subject designs enhance the value of CBCA criteria for discriminating between memory of truthful and false statements.

These variables were complemented with others obtained from the content analysis of the studies themselves. Through a method of successive approximations, two raters examined the studies to identify potential moderator variables described in them, that later underwent a *Thurstone style* evaluation by 10 judges on the degree of independence and pertinence for the study aim. Having identified the moderators in the coded studies, the non-productive ones were eliminated, $p \leq .05$ (for a more detailed description of the method see Arce, Velasco, Novo, & Fariña, 2014). Thus, the following moderators were identified as productive by the raters coding the studies: sex of

Table 2
Results of the Meta-Analyses for Each CBCA Criterion and the Total CBCA Score

Criterion	*k*	*N*	d_w	SD_d	δ	SD_δ	%VE	99% CI_d	90% CI_δ
1	16	1381	0.47	0.334	0.52	0.272	46	[0.42, 0.52]	[0.17, 0.87]
2	15	1217	0.40	0.465	0.53	0.588	25	[0.35, 0.45]	[-0.15, 1.22]
3	17	1477	0.77	0.575	0.87	0.594	16	[0.64, 0.89]	[0.10, 1.63]
4	15	1341	0.69	0.307	0.78	0.571	17	[0.64, 0.74]	[0.05, 1.51]
5	16	1407	0.44	0.443	0.50	0.434	25	[0.39, 0.49]	[-0.06, 1.06]
6	16	1407	0.53	0.430	0.59	0.415	27	[0.48, 0.58]	[0.06, 1.12]
7	11	1111	0.29	0.196	0.33	0.000	100	[0.24, 0.34]	[.33]
8	16	1437	0.27	0.366	0.31	0.335	34	[0.23, 0.31]	[−0.12, 0.74]
9	13	1199	0.42	0.323	0.47	0.274	44	[0.37, 0.47]	[0.12, 0.82]
10	13	1062	0.31	0.409	0.35	0.385	30	[0.26, 0.36]	[−0.14, 0.84]
11	10	916	0.28	0.360	0.32	0.328	35	[0.22, 0.34]	[−0.10, 0.74]
12	15	1194	0.46	0.437	0.52	0.419	28	[0.41, 0.51]	[−0.01, 1.06]
13	10	1052	0.18	0.278	0.21	0.222	50	[0.13, 0.23]	[−0.08, 0.49]
14	15	1367	0.20	0.400	0.23	0.383	28	[0.15, 0.25]	[−0.27, 0.72]
15	13	1076	0.15	0.358	0.17	0.318	38	[0.10, 0.20]	[−0.24, 0.58]
16	10	809	0.19	0.308	0.22	0.238	53	[0.02, 0.35]	[−0.09, 0.52]
17	5	447	0.16	0.476	0.18	0.480	20	[0.08, 0.24]	[−0.43, 0.80]
18	6	517	0.23	0.374	0.25	0.343	34	[0.16, 0.30]	[−0.18, 0.69]
19	5	318	1.25	0.773	1.40	0.807	14	[1.10, 1.40]	[0.38, 2.44]
Average			0.40		0.46		35		
Total score	18	1122	0.78	0.587	0.79	0.275	20	[0.72, 0.84]	[0.11, 1.47]

Note. *k* = number of studies; *N* = total sample size; d_w = effect size weighted for sample size; SD_d = observed standard deviation of *d*; δ = effect size corrected for criterion unreliability; SD_δ = standard deviation of δ; %VE = variance accounted for by artifactual errors; 99% CI_d = 99% confidence interval for *d*; 90% CI_δ = 90% credibility interval for δ.

the participants, the type of scale used for rating the presence of each criterion (i.e., present vs. absent; the weighted presence of a criterion), number of raters, training of the raters, type of interview, coding criteria, and the type of material for content analysis (transcript, video or both).

The coding of these variables (see Appendix 1) was undertaken by two independent raters. Ratings were crosschecked and found to agree totally as was expected, given the clarity and mutual exclusion of the variables.

Data Analysis

The calculation of the effect size was homogenized in Cohen's *d*, which was computed, when this statistic was not provided by the author of a study, according to the following statistics and availability: the means and standard deviations/standard mean error and, in absence of these, the *t*-test value or the associated probability.

When the results were proportioned in Fisher's *F* values, the effect size on eta^2 was transformed to *d*, and only if *F* was available without the exact probability (*d* being obtained from the exact probability), *F* was transformed in *t*-scores and, from this, *d* was computed.

When the results are provided in proportions an effect size δ is obtained (Hedges & Olkin, 1985) using the procedure of Kraemer and Andrews (1982), which is equivalent to Cohen's *d*, whereas when the results are expressed in 2x2 contingency tables, phi is obtained, and in turn the effect size of *d* (Cohen, 1988; Rosenthal, 1994).

The unit of analysis (*n*) was the number of statements, weighting the estimates of the effect sizes by the number of statements instead of the conventional number of participants.

The meta-analysis was performed according to the procedure of Hunter and Schmidt (2004), with a total of 20 Bare-Bones type meta-analysis: one for each of the CBCA criteria and one for all of the criteria of a statement.

The relationship between both distributions for which the effect sizes were calculated was crucial for interpreting the effect sizes. As for studies with practical utility, Fritz, Morris, and Richler (2012) recommended three statistics: U1, the Binomial Effect Size Display (BESD), and the Probability of Superiority (PS).

Reliability Criterion

The reliability of each criterion was assessed in the primary studies using between-rater reliability or between-rater agreement (e.g., intra-class correlation, Maxwell RE, Pearson's correlation, Cohen's kappa, Fleiss' kappa, concordance index, Spearman's rho). Nevertheless, some studies failed to report estimates of reliability or reported several estimates. In the latter case, the reliability index (between-rater agreement indexes were excluded as they do not measure reliability) best fits the data distribution according to the conclusions of Anson et al. (1993) and Horowitz et al. (1997). Moreover, reliability was not systematically and homogeneously informed (i.e., the statistic provided was either on each criteria and the total criteria, only the total, only the criteria, or only the range for each criterion). In short, not all of the studies informed of reliability, and the reliability for each criterion, nor were they estimated using the same statistic. Thus, for criteria 17 to 18 no estimates on the reliability of these measures were found. This prompted us to calculate one estimate of reliability for the criteria and another for the total CBCA score since the reliability for the criteria is different to that for the whole instrument, i.e., the total CBCA score (Horowitz et al., 1997). The estimate of reliability for the individual criteria was calculated using the reliability coefficients of each study to obtain an average *r* of .79 (*SEM* = 0.02). To estimate the reliability of the total CBCA score, the Spearman-Brown prediction formula was used with the *r* extracted from .98 of the total score.

Results

The results of the effect sizes calculated for each criterion and for the total CBCA score, the total number of statements, the weighted sample effect size (*d*), the standard deviation, the effect size corrected for criterion unreliability (δ), the percent of variance accounted for by artifacts, and confidence and credibility intervals (when neither of the intervals contained zero, the estimated effect size is deemed to be significant and generalizable, respectively), are shown in Table 2.

Moreover, the results revealed a large positive effect size for the total CBCA score, for both the weighted sample effect size of *d*, 0.78, and the effect size corrected for the criterion unreliability (δ), 0.79; it was significant, 99% CI_d [0.72, 0.84], and generalizable, 90% CI_δ [0.11, 1.47]; that is, with an expected minimum value of 0.11, and a maximum of 1.47 standard deviation. In practical terms, the results reveal that the rate of correct classifications of the total CBCA score was 68.5% for statements based on genuine experiences (true positives), with a failed detection rate (false negatives) of 31.5% (BESD) and 47% (U1 = 47) of the areas covered by both populations did not overlap, i.e., they were totally independent; with a .712 probability (PS) of obtaining more CBCA criteria in a statement based on true self-experienced accounts.

Though the results are statistically generalizable, the data from experimental and field studies have, as previously mentioned, practical implications in terms of meeting the Daubert standards. Moreover, concerns have been raised regarding the limitations on generalizing the results of experimental studies, and more criteria have been observed in field studies. Thus, a new meta-analysis was performed on the total CBCA score of studies according to the research paradigm (experimental studies vs. field studies). The results show (see Table 3) the effect size corrected for criterion unreliability (δ) was positive, significant, moderate (0.56), and generalizable to experimental studies; and positive, significant, more than large (2.71, $p < .01$), and generalizable to field studies. The comparison of the effect sizes of the experimental and field studies revealed the latter were significantly greater, $t(7.3) = 3.86$, $p < .01$, $d = 2.17$, than those obtained for the former. The subsequent meta-analysis grouping studies according to the degree of experimental fidelity was not undertaken due to the lack of low fidelity studies ($k = 1$). Thus, as expected, the results of high fidelity studies were practically the same as those for experimental studies (δ = 0.58). The results require the error rate, independence of distributions, and the probability of superiority to be estimated for each research paradigm. For experimental studies, the rate of correct classifications of truthful memories was 63.5%, with a failure rate of 36.5%; an independence of distributions of 36.1%; and the probability of obtaining more criteria in a statement based on memory of real-life events of .654. For field studies, the rate of correct classifications was 90.2%, with a margin of error of 9.8%; the distributions were completely independent in 90.4%; and the probability of superiority was .972.

In addition to the initial conception that all CBCA criteria are grounded on the Undeutsch hypothesis, an alternative contention sustains that only the first 14 criteria are derived from the abovementioned hypothesis. Thus, a new meta-analysis of the total CBCA score was performed in studies that were restricted to these criteria. The results (see Table 3) showed an effect size corrected for criterion unreliability (δ) positive, significant, large, 0.96, and generalizable.

As for the criteria independently, the results (see Table 2) show a significantly positive effect for all of them with values of effect size corrected for criterion unreliability (δ) ranging from 0.17 for the criterion of reality *admitting lack of memory,* to 1.40 for the criterion *details characteristic of the offence*. Moreover, the effect sizes for the criteria *logical structure*, *quantity of details*, *contextual embedding*, *reproduction of conversations*, *unexpected complications during the incident*, *superfluous details,* and *details characteristic of the offence* were generalizable. Of the literature reviewed, the results contrary to the Undeutsch hypothesis were only found in the criteria *unstructured*

Table 3
Results of the Meta-Analyses of Moderators

Moderators	*k*	*N*	d_w	SD_d	δ	SD_δ	%VE	99% CI_d	*90% CI_δ*
Field studies	8	413	2.40	1.37	2.71	1.48	9	[2.10, 2.70]	[0.82, 4.60]
Restricted Field studies[1]	6	325	2.33	1.01	2.63	1.05	16	[1.99, 2.66]	[1.28, 3.97]
Experimental studies	12	810	0.50	0.19	0.56	0.00	100	[0.33, 0.66]	[0.56]
High fidelity studies[2]	11	730	0.51	0.20	0.58	0.00	100	[0.34, 0.68]	[0.58]
CBCA 14 criteria version[3]	4	301	0.85	0.46	0.96	0.43	29	[0.56, 1.13]	[0.40, 1.52]
CBCA 19 criteria version	14	821	0.84	0.83	0.95	0.88	11	[0.67, 1.01]	[-0.18, 2.07]

Note. *k* = number of studies; *N* = total sample size; d_w = effect size weighted for sample size; SD_d = observed standard deviation of *d*; δ = effect size corrected for criterion unreliability; SD_δ = standard deviation of δ; %VE = variance accounted for by artifactual errors; 99% CI_d = 99% confidence interval for *d*; 90% XI_δ = 90% credibility interval for δ. [1]limited to those with estimations of the reliability of the evaluation, matching groups, and/or independence of measures in the grouping factor (i.e., confirmed vs. non-confirmed cases); [2]recreation of conditions facilitating personal involvement, negative emotional tone of the event, and/or extensive loss of control; [3]CBCA criteria related to Undeutsch hypothesis (Raskin et al., 1991).

production, *unusual details* (Granhag, Strömwall, & Landström, 2006), and *self-deprecation* (Steller, Wellershaus, & Wolf, 1988). For criteria without generalizable results, the moderators could not be assessed due to the lack of information regarding potential moderators in the primary studies with insufficient *N*s to proceed.

Discussion

The results of this study have the following implications in terms of the Undeutsch hypothesis and the Checklist of CBCA criteria. First, the results of the total CBCA score supported the validity of the Undeutsch hypothesis in discriminating between truthful statements based on self-experienced events and fictitious accounts. Hence, the presence of these criteria is associated to truthful statement, and a large and generalizable effect size. Furthermore, no study reported the opposite trend, i.e., significantly more criteria in fabricated statements. Thus, the hypothesis is valid and generalizable to other conditions (e.g., children of all ages, different contexts of sexual abuse, and research paradigms). Second, all of the CBCA criteria discriminated significantly between the real-life memories of children and fabricated accounts. Consequently, CBCA criteria were validated for 5 major categories, i.e., *general characteristics; specific contents;* peculiarities of content; *motivation-related contents;* and offence-specific elements. Third, in relation to the magnitude of the effect size, the criteria *quantity of details* and *details characteristic of the offence* discriminated significantly high with a large effect size. In comparison, the effect size was medium for the categories *logical structure*, *unstructured production*, *contextual embedding*, *description of interactions*, *reproduction of conversations*, and *accounts of subjective mental states*. Finally, the effect size was small for the remaining categories (i.e., *unexpected complications during the incident, unusual details, superfluous details, accurately reported details misunderstood, related external associations, attribution of perpetrator's mental state, spontaneous corrections, admitting lack of memory, raising doubts about one's own testimony, self-deprecation, pardoning the perpetrator*). Fourth, the effect size of the major categories ranged from medium to large (from 0.52 to 0.87) for *general characteristics*; from small to large (from 0.33 to 0.78) for *specific contents*; from small to moderate (from 0.21 to 0.52) for *peculiarities of content*; small for *motivation-related contents* (from 0.17 to 0.25); and more than large (> 1.20 *SD*) for *offence-specific elements*. In other words, the biggest effect size was for *general characteristics*, that bear considerable weight on judgement-making concerning the credibility of a testimony (Zaparniuk et al., 1995), and *offence-specific elements* major categories, that is, elements difficult to fabricate. Fifth, components (Köhnken, 1996), in line with the qualitative findings of Vrij (2005), cognitive criteria (criteria 1 to 13) had larger effect sizes than motivational criteria (criteria 14 to 18), that are not generalizable. If the motivational component is useful for assessing (partially) fabricated memories (alternative hypothesis to the truth), the system is of less value for such a task, and what is more, these criteria are subject to moderators so their use is not generalizable for this function. Sixth, the results shed some light on the degree of compliance with the Daubert standard. In relation to the first and second standards – *Is the scientific hypothesis testable?, 2) has the proposition been testable?* – this meta-analysis answers the question affirmatively, not only in meeting Daubert standard, but also in validating the hypothesis. Moreover, the error rate was quantified (third Daubert standard). As shown above, three error rates were obtained: a general one for all the studies, a specific one for experimental studies, and one for field studies. The results for all the studies (*k* = 18), and the experimental studies (*k* = 12) were similar to previous reports of more than 30%. In contrast, in field studies (*k* = 8) the error rate fell sharply to 10%. Moreover, 97% of honest statements contained more criteria than false statements. Notwithstanding, the data of field studies has been called into question for neither estimating the reliability of the evaluation of raters, nor matching groups, and the lack of safeguards in the independence of measures in the grouping factor. Though the meta-analytical technique is not concerned with these issues, and includes all of the studies excluding outliers, a new meta-analysis was undertaken restricted to studies with a grouping factor of confirmed cases vs. fabricated/un-confirmed cases, with reliability estimates for each criteria, and matched groups (*k* = 6), exhibiting a positive, significant, and generalizable effect size (δ) of 2.63 (see Table 3). The results of this meta-analysis corroborated robustness was similar to that obtained in all the field studies: a correct classification rate of truthful statements (BESD) of 89.8%, an independence between distributions (U1) of 89.6%, and a probability of superiority (PS) of .969 (i.e., 96.9% of truthful statements contained more criteria than fabricated statements). As for the fourth Daubert standard – *has the hypothesis and/or technique been subjected to peer review and publication?* – the answer is self-evident from the studies themselves. Concerning the fifth criterion – *is the theory upon which the hypothesis and/or technique is based on generally accepted in the appropriate scientific community?* –, this meta-analysis has no clear response. Honts (1994) is of the opinion that, save minor controversy, the hypothesis is widely accepted by the scientific community. However, Vrij (2008) has pointed out that this information is unknown since the scientific community was not consulted. Seventh, though the results of experimental studies were similar, they cannot be directly generalized to field studies, and at best exhibit face validity (Konecni & Ebbesen, 1992). Therefore, the findings of experimental studies require field studies to validate them, as they are insufficient on their own.

The results of this meta-analysis are subject to several limitations that should be borne in mind when generalizing the results. First, the reliability of the interview was not systematically contrasted and frequently the interviewers received poor or no train-

ing. It is well known that the abilities of the interviewer mediate the contents and quality of the interview (Bembibre & Higueras, 2010; Fisher, Geiselman, & Amador, 1989; Steller & Köhnken, 1989). Second, it is unknown (unreported) if the material for content analysis was sufficient for assessing the veracity of statements (Köhnken, 2004; Steller & Köhnken, 1989), an aspect which is crucial for children since their memory is less productive (Memon et al., 2010). Third, the measures of reliability of the coding rely on the raters themselves without contrasting if reliability is generalizable to other independent raters, but this procedure does not guarantee the data is actually reliable. Reliability in content analysis is derived from the measure of within-rater consistency (of the coder her/himself through time), and between-raters and between-contexts (with other raters independent to the study, and other materials) (Weick, 1985). Moreover, for the criteria in which the unit of analysis was not the entire statement, the estimate of reliability was generally calculated using frequency counts without verifying the exact correspondence between the counts (with some exceptions, Boychuk, Vrij). This practice overestimates reliability. Furthermore, reliability for each of the measures was not systematically reported (e.g., sometimes ranges, total reliability, or only the reliability of the major categories were documented), as well as the coder training that it is related to coding accuracy (Akehurst, Bull, Vrij, & Köhnken, 2004). Fourth, the Undeutsch hypothesis, and thus SVA/CBCA, assert the presence of criteria is indicative of true statements but, conversely, the absence of criteria does not imply a false statement (i.e., there are other alternative hypotheses to a false statement), and numerous studies have used the criteria for classifying false statements, when the categories are operative for classifying real statements but not fabricated ones. This bias arises from the design of experimental studies, in which the researcher has under control the memory of fictitious accounts, but in the forensic evaluation of real cases, hypotheses other than deception must be considered such as the evaluee's unwillingness to cooperate, insufficient memory of events for an analysis to be undertaken on the credibility of a statement, or impaired cognitive capacity (Köhnken, 2004). In terms of forensic applications, the analysis of the credibility of a testimony is admissible as incriminating evidence, being inoperative in classifying false statements. Thus, in terms of fidelity with the Undeutsch hypothesis and its application to forensic assessment, the results should be in the direction of the classification of true statements. Fifth, the results of some meta-analysis may be subject to a degree of variability given that *N*s < 400 do not guarantee the stability of sampling estimates (Hunter & Schmidt, 2004). Nevertheless, as the results of the inter-meta-analysis were consistent, these may affect the statistical data, with expected effects in terms of the Undeutsch hypothesis. Sixth, the reliability of each criterion was an estimate, given the aforementioned reporting problems in the primary studies. Seventh, the results were somewhat biased towards supporting the hypothesis since some studies failed to publish so called *conflicting data* i.e., data with criteria that failed to discriminate significantly between real-life self-experienced events and fabricated accounts were excluded (e.g., Akehurst, Manton, & Quandte, 2011). In any case, none of these studies reported results that contradicted the Undeutsch hypothesis, but rather criteria that failed to discriminate significantly between both types of memory. Nevertheless, most of these limitations are reflected in the increase in the error variance, reducing the estimated effect sizes, which means the true effect sizes would have been greater, lending even more support to the Undeutsch hypothesis.

As for the practical implications of this meta-analysis, the findings support the Undeutsch hypothesis and several Daubert standards, but this does not imply the use of Checklist of CBCA criteria can be directly generalized to the context of forensic evaluation. First, the categorical system proposed is not methodic, i.e., it fails to comply with stringent methodic conditions: mutual exclusion, homogeneity, pertinence, objectivity, fidelity and productivity (Bardin, 1997). For instance, as non-mutual exclusion between categories is guaranteed, the duplicity of measures may arise, the criteria are neither objective nor exhaustive – e.g., Roma et al. (2011) and Horowitz et al. (1997) have proposed the integration or redefinition of criteria due to rating difficulties –, the checklist may need additional criteria, or the checklist lacks internal consistency, in other words, it is not reliable. Second, the forensic application of a checklist of categories is driven by clinical judgements (Köhnken, 2004), or on quantitative decision rules that are not supported by empirical data (Arntzen, 1983; Zaparnuik et al., 1995). However, in the field of forensics an objective and strict decision criterion based on stringent standards of evidence should prevail over subjective clinical judgements, i.e., the rate of classification of false statements as true (false positives) should be 0 (i.e., the burden of proof is on the prosecution; it is entirely inadmissible to present incriminating expert forensic testimony on the basis of unsubstantiated evidences). It would be a decision rule based on data for controlling false positives and to ensure reliable coding (i.e., within-raters, between-raters, and between-context consistency), which would offset the potential effects of a truth bias or a response bias associated to the application of reality criteria (Griesel et al., 2013; Rassin, 1999; Sporer, 2004). The results of previous meta-analyses have shown it is possible, i.e., in field studies approximately 97% of truthful statements contained more reality criteria than fabricated accounts, with an approximately 90% total independence between the distributions of both groups of statements.

Conflict of Interest

The authors of this article declare no conflict of interest.

Financial Support

This research has been carried out within the framework of research project with the Reference Ref. GPC2014/022, funded by the Xunta de Galicia [Galician Autonomous Government], Spain.

Acknowledgements

The authors acknowledge and thank Professor Jesús F. Salgado for his methodological support and the revision of this manuscript.

References

[Asterisks refer to studies included in the meta-analysis]

*Akehurst, L., Bull, R., Vrij, A., & Köhnken, G. (2004). The effects of training professional groups and lay persons to use Criteria-Based Content Analysis to detect deception. *Applied Cognitive Psychology, 18*, 877-891. doi: 10.1002/acp.1057

*Akehurst, L., Manton, S., & Quandte, S. (2011). Careful calculation or a leap of faith? A field study of the translation of CBCA ratings to final credibility judgements. *Applied Cognitive Psychology, 25*, 236-243. doi: 10.1002/acp.1669

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. doi: 10.1007/BF01044512

Arce, R., & Fariña, F. (2012). Psicología social aplicada al ámbito jurídico [Applied social psychology to the legal context]. In A. V. Arias, J. F. Morales, E. Nouvilas, & J. L. Martínez (Eds.), *Psicología social aplicada* (pp. 157-182). Madrid, Spain: Panamericana.

Arce, R., Fariña, F., & Fraga, A. (2000). Género y formación de juicios en un caso de violación [Gender and juror judgment making in a case of rape]. *Psicothema, 12*, 623-628.

Arce, R., Velasco, J., Novo, M., & Fariña, F. (2014). Elaboración y validación de una escala para la evaluación del acoso escolar [Development and validation of a scale to assess bullying]. *Revista Iberoamericana de Psicología y Salud, 5*, 71-104.

Arntzen, F. (1970). *Psychologie der zeugenaussage. Einführung in die forensische aussagepsychologie* [Psychology of eyewitness testimony. Introduction to forensic psychology of statement analysis]. Göttingen, Germany: Hogrefe.

Arntzen, F. (1983). *Psychologie der zeugenaussage: Systematik der glaubwürdigkeitsmerkmale* [Psychology of witness statements: The system of reality criteria]. Munich, Germany: C. H. Beck.

Bardin, L. (1977). *L'Analyse de contenu* [Content analysis]. Paris, France: Presses Universitaires de France.

Bembibre, J., & Higueras, L. (2010). Eficacia diferencial de la entrevista cognitiva en función de la profesión del entrevistador: Policías frente a psicólogos [Differential efficacy of the cognitive interview as a function of the interviewer's job: Police officers vs. psychologists]. In F. Expósito, M. C. Herrera, G. Buela-Casal, M. Novo, & F. Fariña (Eds.), *Psicología jurídica. Áreas de intervención* (pp. 141-149). Santiago de Compostela, Spain: Consellería de Presidencia, Xustiza e Administracións Públicas.

Bensi, L., Gambetti, E., Nori, R., & Giusberti, F. (2009). Discerning truth from deception: The sincere witness profile. *European Journal of Psychology Applied to Legal Context, 1*, 101-121.

*Blandon-Gitlin, I., Pezdek, K., Rogers, M., & Brodie, L. (2005). Detecting deception in children: An experimental study of the effect of event familiarity on CBCA ratings. *Law and Human Behavior, 29*, 187-197. doi: 10.1007/s10979-005-2417-8

*Boychuk, T. D. (1991). *Criteria Based Content Analysis of children's statements about sexual abuse: A field-based validation study* (Unpublished doctoral dissertation). Arizona State University.

Buck, J. A., Warren, A. R., Betman, S. I., & Brigham, J. C. (2002). Age differences in Criteria-Based Content Analysis scores in typical child sexual abuse interviews. *Journal of Applied Developmental Psychology, 23*, 267-283. doi: 10.1016/S0193-3973(02)00107-7

*Casado del Pozo, A. M., Romera, R. M., Vázquez, B., Vecina, M., & De Paúl, P. (2004). Análisis estadístico de una muestra de 100 casos de abuso sexual infantil [Statistical analysis of a 100-case sample of child sexual abuse]. In B. Vázquez (Ed.), *Abuso sexual infantil. Evaluación de la credibilidad del testimonio* (pp. 73-105). Valencia, Spain: Centro Reina Sofía para el Estudio de la Violencia.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: LEA.

*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. doi: 10.1207/s1532480xads0302_2

Daubert v. Merrell Dow Pharmaceuticals, 509 U.S. 579 (1993).

Dettenborn, H., Froehlich, H., & Szewczyk, H. (1984). *Forensische psychologie* [Forensic Psychology]. Berlin, Germany: Deutscher Verlag der Wissenschaften.

*Di Blasio, P., & Conti, A. (2000). L'applicazione del "Criteria-Based Content Analysis" (C.B.C.A.) a racconti di storie vere e inventate [Application of Criteria-Based Content Analysis to accounts of real and fabricated stories]. *Maltrattamento e Abuso all'infanzia, 2*(3). 57-78. doi: 10.1400/62968

*Erdmann, K., Volbert, R., & Böhm, C. (2004). Children report suggested events even when interviewed in a non-suggestive manner: What are its implications for credibility assessment? *Applied Cognitive Psychology, 18*, 589-611. doi: 10.1002/acp.1012

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

Fariña, F., Arce, R., & Real, S. (1994). Ruedas de identificación: De la simulación y la realidad [Linepus: A comparison of high-fidelity research and research in a real context]. *Psicothema, 6*, 395-402.

Fisher, R. P., Geiselman, R. E., & Amador, M. (1989). Field test of the cognitive interview: Enhancing the recollection of actual victims and witness of crime. *Journal of Applied Psychology, 74*, 722-727. doi: 10.1037/0021-9010.74.5.722

Fritz, C. O., Morris, P. E., & Richler, J. J. (2012). Effect size estimates: Current use, calculations, and interpretation. *Journal of Experimental Psychology: General, 141*, 2-18. doi: 10.1037/a0024338

*Granhag, P. A., Strömwall, L. A., & Landström, S. (2006). Children recalling an event repeatedly: Effects on RM and CBCA scores. *Legal and Criminological Psychology, 11*, 81-98. doi: 10.1348/135532505X49620

Griesel, D., Ternes, M., Schraml, D., Cooper, B. S., & Yuille, J. C. (2013). The ABC's of CBCA: Verbal credibility assessment in practice. In B. S., Cooper, D. Griesel, & M. Ternes (Eds.), *Applied issues in investigative interviewing, eyewitness memory, and credibility assessment* (pp. 293-323). New York, NY: Springer. doi: 10.1007/978-1-4614-5547-9_12

Hans, V. P., & Vidmar, N. (1986). *Judging the jury*. New York, NY: Plenum Press.

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. San Diego, CA: Academic Press.

Hershkowitz, I. (1999). The dynamics of interviews involving plausible and implausible allegations of child sexual abuse. *Applied Developmental Science, 3*, 86-91. doi: 10.1207/s1532480xads0302_3

Honts, C. R. (1994). Assessing children's credibility: Scientific and legal issues in 1994. *North Dakota Law Review, 70*, 879-903.

Horowitz, S. W., Lamb, M. E., Esplin, P. W., Boychuk, T. D., Krispin, O., & Reiter-Lavery, L. (1997). Reliability of criteria-based content analysis of child witness statements. *Legal and Criminological Psychology, 2*, 11-21. doi: 10.1111/j.2044-8333.1997.tb00329.x

Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis: Correcting errors and bias in research findings*. Thousand Oaks, CA: Sage.

Kaplan, M. F. (1982). Cognitive processes in the individual juror. In N. L. Kerr & R. M. Bray (Eds.), *The psychology of the courtroom* (pp. 197-220). New York, NY: Academic Press.

Köhnken, G. (1996). Social psychology and the law. In G. R. Semin & K. Fiedler (Eds.), *Applied social psychology* (pp. 257-282). Thousand Oaks, CA: Sage.

Köhnken, G. (2004). Statement Validity Analysis and the detection of the truth. In P. A. Granhag & L. A. Strömwall (Eds.), *The detection of deception in forensic contexts* (pp. 41-63). Cambridge, UK: Cambridge University Press. doi: 10.1017/CBO9780511490071.003

Köhnken, G., & Steller, M. (1988). The evaluation of the credibility of child witness statements in the German procedural System. *Issues in Legal and Criminological Psychology, 13*, 37-45.

Konecni, V. J., & Ebbesen, E. B. (1992). Methodological issues on legal decision-making, with special reference to experimental simulations. In F. Lösel, D. Bender & T. Bliesener (Eds.), *Psychology and law. International perspectives* (pp. 413-423). Berlin, Germany: Walter de Gruyter.

Kraemer, H. C., & Andrews, G. (1982). A non-parametric technique for meta-analysis effect size calculation. *Psychological Bulletin, 91*, 404-412. doi: 10.1037/0033-2909.91.2.404

*Lamb, M. E., Sternberg, K. J., Esplin, P. W., Hershkowitz, I., & Orbach, Y. (1997). Assessing the credibility of children's allegations of sexual abuse: A survey of recent research. *Learning and Individual Differences, 9*, 175-194. doi: 10.1016/S1041-6080(97)90005-4

Lamers-Winkelman, F., & Buffing, F. (1996). Children's testimony in the Netherlands: A study of Statement Validity Analysis. *Criminal Justice and Behavior, 23*, 304-321. doi: 10.1177/0093854896023002004

*Mazzoni, G., & Ambrosio, K. (2002). L'analisi del resoconto testimoniale in bambini: Impiego del metodo di analisi del contenuto C.B.C.A. in bambini di 7 anni [Assessment of child witness statements: Application of CBCA method in a 7 year old children sample]. *Psicologia e Giustizia, 3*(2).

*Mazzoni, G., & Pezzati, S. (2002). Esame della validità del C.B.C.A. in racconti di bambini di 4-5 anni [An exam of validity of the CBCA in the account's four-five year old children]. *Età Evolutiva, 73*, 5-17.

Memon, A., Meissner, C. A., & Fraser, J. (2010). Cognitive interview. A meta-analytic review and study space analysis of the past 25 years. *Psychology, Public Policy, and Law, 16*, 340-372. doi: 10.1037/a0020518

Novo, M., & Seijo, D. (2010). Judicial judgement-making and legal criteria of testimonial credibility. *European Journal of Psychology Applied to Legal Context, 2*, 91-115.

Raskin, D. C., & Esplin, P. W. (1991). Statement Validity Assessment: Interview procedures and content analysis of children's statements of sexual abuse. *Behavioral Assessment, 13*, 265-291.

Raskin, D. C., Esplin, F. W., & Horowitz, S. (1991). *Investigative interviews and assessment of children in sexual abuse cases* (Unpublished manuscript). Department of Psychology, University of Utah.

Rassin, E. (1999). Criteria Based Content Analysis: The less scientific road to truth. *Expert Evidence, 7*, 265-278. doi: 10.1023/A:1016627527082

*Roma, P., San Martini, P., Sabatello, U., Tatarelli, R., & Ferracuti, S. (2011). Validity of Criteria-Based Content Analysis (CBCA) at trial in free-narrative interviews. *Child Abuse & Neglect, 35*, 613-620. doi: 10.1016/j.chiabu.2011.04.004

Rosenthal, R. (1994). Parametric measures of effect size. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 231-244). New York, NY: Russell Sage Foundation.

*Santtila, P., Roppola, H., Runtti, M., & Niemi, P. (2000). Assessment of child witness statements using Criteria-Based Content Analysis (CBCA): The effects of age, verbal ability, and interviewer's emotional style. *Psychology, Crime & Law, 6*, 159-179. doi: 10.1080/10683160008409802

Sporer, S. L. (2004). Reality monitoring and detection of deception. In P. A. Granhag & L. A. Strömwall (Eds.), *The detection of deception in forensic contexts* (pp. 64-102). Cambridge, UK: Cambridge University Press. doi: 10.1017/CBO9780511490071.004

Steller, M. (1989). Recent developments in statement analysis. In J. C. Yuille (Ed.), *Credibility assessment* (pp. 135-154). Deventer, Holland: Kluwer. doi: 10.1007/978-94-015-7856-1_8

Steller, M., & Böhm, C. (2006). Cincuenta años de jurisprudencia del Tribunal Federal Supremo alemán sobre la psicología del testimonio. Balance y perspectiva [Fifty years of the German Federal Court jurisprudence on forensic psychology]. In T. Fabian, C. Böhm, & J. Romero (Eds.), *Nuevos caminos y conceptos en la psicología jurídica* (pp. 53-67). Münster, Germany: LIT Verlag.

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

*Steller, M., Wellershaus, P., & Wolf, T. (1988). *Empirical validation of Criteria-Based Content Analysis*. Paper presented at the NATO Advanced Study Institute on Credibility Assessment, Maratea, Italy.

*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. doi: 10.1002/acp.1021

Szewczyk, H. (1973). Kriterien der Beurteilung kindlicher zeugenaussagen [Criteria for the evaluation of child witnesses]. *Probleme und Ergebnisse der Psychologie, 46*, 47-66.

*Tye, M. C., Amato, S. L., Honts, C. R., Devitt, M. K., & Peters, D. (1999). The willingness of children to lie and the assessment of credibility in an ecologically relevant laboratory setting. *Applied Developmental Science, 3*, 92-109. doi: 10.1207/s1532480xads0302_4

Undeutsch, U. (1967). Beurteilung der glaubhaftigkeit von aussagen [Evaluation of statement credibility/ Statement validity assessment]. In U. Undeutsch (Ed.), *Handbuch der Psychologie* (Vol. 11: Forensische Psychologie, pp. 26-181). Göttingen, Germany: Hogrefe.

Undeutsch, U. (1989). The development of statement reality analysis. In J. Yuille (Ed.), *Credibility assessment* (pp.101-119). Dordrech, Holland: Kluwer Academic Publishers.

Vrij, A. (2005). Criteria-Based Content Analysis: A qualitative review of the first 37 studies. *Psychology, Public Policy, and Law, 11*, 3-41. doi: 10.1037/1076-8971.11.1.3

Vrij, A. (2008). *Detecting lies and deceit: Pitfalls and opportunities* (2nd ed.). Chichester, England: John Wiley and Sons.

*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. doi: 10.1023/A:1015313120905

*Vrij, A., Akehurst, L., Soukara, R., & Bull, R. (2004). Detecting deceit via analyses of verbal and nonverbal behavior in children and adults. *Human Communication Research, 30*, 8-41. doi: 10.1111/j.1468-2958.2004.tb00723.x

Weick, K. E. (1985). Systematic observational methods. In G. Lindzey & E. Aronson (Eds.), *The handbook of social psychology* (Vol. 1, pp. 567-634). Hillsdale, NJ: LEA.

Zaparniuk, J., Yuille, J. C., & Taylor, S. (1995). Assessing the credibility of true and false statements. *International Journal of Law and Psychiatry, 18*, 343-352. doi: 10.1016/0160-2527(95)00016-B

Appendix 1
Moderator Variables

Primary studies	*N*	Age	Sex	Paradigm	Design	Raters	Coding Scale	Coders training	Type of interview	Criteria	Transcript/ video
Akehurst, Bull, Vrij, and Köhnken (2004)	151	7-11	M: 23 F: 26	Experimental: active	Within	58	0-4	Extensive training in CBCA coding	Step-wise interview	13	Transcript
Akehurst, Bull, Vrij, and Köhnken (2004)	132			Experimental: video	Within					13	Transcript
Akehurst, Manton, and Quandte (2011)[2,6]	31	6-17	M: 5 F: 26	Field	Between	2	1-5	Extensive training in CBCA coding	Step-wise interview	3	Transcript
Blandon-Gitlin, Pezdek, Rogers, and Brodie (2005)[5]	94	9-12	–	Experimental: active	Between	2	0-1	Extensive training in CBCA coding	Step-wise interview	16	Transcript
Boychuk (1991)	75	4-16	M: 15 F: 60	Field	Between	2	0-1	Expert raters	No standardized interview procedures	19	Transcript
Casado del Pozo, Romera, Vázquez Mezquita, Vecina, and de Paúl (2002)[6]	96	4-18	M: 28 F: 72	Field	Between	2	0-1	Expert raters	SVA guidelines	19	–
Craig, R. A., Scheibe, R., Raskin, D. C., Kircher, J. C., yDodd, D. H. (1999)[1]	48	3-16	M: 11 F: 37	Field	Between	4	0-1	8 hours training	SVA guidelines	14	Transcript
Di Blasio and Conti (2000)	88	9	M: 25 F: 19	Experimental: memory	Within	2	0-1	Expert raters	Step-wise interview	19	–
Erdmann, Volbert, and Böhm (2004)	70	6-8	M: 36 F: 31	Experimental: memory	Within	2	0-1/0-2	Experts raters	Step-wise interview	15	Transcript
Esplin, Houed, and Raskin (1988)	40	3-15	–	Field	Between	1	0-2	Intensive training in CBCA coding	–	19	Transcript
Granhag, Strömwall, and Landström (2006)	80	12-13	M: 42 F: 38	Experimental: staged	Between	2	0-2	Intensive training in CBCA coding	Cognitive Interview	10	–
Lamb, Sternberg, Esplin, Orbach, and Hovav (1997)[6]	89	4-13	M: 28 F: 70	Field	Between	2	0-1	Intensive training in CBCA coding	No standardized interview procedures.	14[1]	Transcript
Mazzoni and Pezzati (2002)[5]	60	4-5	M: 20 F: 21	Experimental: memory	Within	2	0-2	Training in CBCA	Step-wise interview	19	Transcript
Mazzoni and Ambrosio (2002)[5]	60	7	–	Experimental: memory	Within	2	0-2	–	Step-wise interview	19	Transcript
Roma, San Martini, Sabatello, Tatarelli, and Ferracuti (2011)[1,6]	109	4-14	M: 23 F: 86	Field	Between	2	0-1	Expert raters	Step-wise interview	14	Transcript
Santtila, Roppola, Runtti, and Niemi (2000)[1,5]	136	7-14	M: 34 F: 34	Experimental: memory	Within	2	0-1/0-2	Expert raters	Step-wise interview	14	Transcript
Steller, Wellershaus, and Wolf (1988)	176	10-13	–	Experimental: memory	Within	3	0-3	–	SVA guidelines	16	–
Strömwall, Bengtsson, Leander, and Granhag (2004)[3,5]	41	10-13	M: 45 F: 42	Experimental: active	Between	2	0-1/0-2	Extensive training in CBCA coding	Cognitive interview	15	Transcript
Strömwall, Bengtsson, Leander, and Granhag (2004)[4,5]	46										
Tye, Amato, Honts, Devitt, and Peters (1999)	28	6-10	M: 21 F: 27	Experimental: active	Between	3	0-2	Expert raters	SVA guidelines	12	Transcript
Vrij, Akehurst, Soukara, and Bull (2002)[5]	36	5-6	M: 16 F: 20	Experimental: active	Between	2	1-5	Extensive training in CBCA coding	Step-wise interview	9	Transcript
	56	10-11	M: 22 F: 34								
	57	14-15	M: 33 F: 24								

(Continued)

Appendix 1
Moderator Variables *(cont.)*

Primary studies	*N*	Age	Sex	Paradigm	Design	Raters	Coding Scale	Coders training	Type of interview	Criteria	Transcript/ video
Vrij, Akehurst, Soukara, and Bull (2004)[5]	35	5-6	M: 16 F: 19	Experimental: active	Between	2	1-5	Extensive training in CBCA coding	Step-wise interview	9	Transcript
	54	10-11	M: 22 F: 32								
	55	14-15	M: 32 F: 23								

Note.[1]CBCA14 criteria version, [2]limited to those criteria discriminating significantly between self-experienced and fabricated statements, [3]event experienced once, [4]event experienced four times, [5]high fidelity study, [6]restricted field study.