Validity of Content-Based Techniques to Distinguish True and Fabricated Statements: A Meta-Analysis.

Verena A. Oberlader[1*], Christoph Naefgen[2*], Judith Koppehele-Gossel[1], Laura Quinten[1], Rainer Banse[1], and Alexander F. Schmidt[3]

[1]University of Bonn, Bonn, Germany.

[2] University of Tübingen, Tübingen, Germany.

[3]University of Luxembourg, Walferdange, Luxembourg.

*The first two authors contributed equally to this article.

This pre-print has been accepted for publication in LAW AND HUMAN BEHAVIOR (March 2016)

Corresponding author:

Verena A. Oberlader, University of Bonn, Social and Legal Psychology, Department of Psychology, Kaiser-Karl-Ring 9, 53111 Bonn, Germany, oberlader@uni-bonn.de

Abstract

Within the scope of judicial decisions, approaches to distinguish between true and fabricated statements have been of particular importance since ancient times. Although methods focusing on "prototypical" deceptive behavior (e.g., psychophysiological phenomena, nonverbal cues) have largely been rejected with regard to validity, content-based techniques constitute a promising approach and are well established within the applied forensic context. The basic idea of this approach is that experience-based and nonexperience-based statements differ in their content-related quality. In order to test the validity of the most prominent content-based techniques, Criteria-Based Content Analysis (CBCA) and Reality Monitoring (RM), we conducted a comprehensive meta-analysis on English- and German-language studies. Based on a variety of decision criteria, 56 studies were included revealing an overall effect size of $g = 1.03$ (95% confidence interval [0.78, 1.27], $Q = 420.06$, $p < .001$, $I^2 = 92.48\%$, $N = 3429$). There was no significant difference in the effectiveness of CBCA and RM. Additionally, we investigated a number of moderator variables such as characteristics of participants, statements, and judgment procedures, as well as general study characteristics. Results showed that the application of all CBCA criteria outperformed any incomplete CBCA criteria set. Furthermore, statement classification based on discriminant functions revealed higher discrimination rates than decisions based on sum scores. Finally, unpublished studies showed higher effect sizes than studies published in peer-reviewed journals. All results are discussed in terms of their significance for future research (e.g., developing standardized decision rules) and practical application (e.g., user training, applying complete criteria set).

Keywords: meta-analysis; deception detection; content-based techniques; criteria-based content analysis (CBCA); reality monitoring (RM)

Deception is a topic that has concerned humanity since ancient times. The law “If a man came forward as a witness (in a lawsuit), and was shown to be a perjurer, he shall pay fifteen shekels of silver” (Finkelstein, 1968/69, p. 70) was found in the codex of Ur-Nammu 2100-2050 B.C., the oldest code of law known today. Accordingly, a particular interest in deception detection by judicial systems goes back just as far. The codex also includes the law “If a man had accused the wife of a(nother) man of fornication but the river-ordeal had proved her innocent, the one who had accused her must pay one-third mina of silver” (Finkelstein, 1968/69, p. 68). This suggests that the “river-ordeal” was a way of divining whether or not an allegation was true. To date, various techniques to discriminate between truth and lies have been developed for diverse applications (Volbert & Banse, 2014), ranging from non-scientific theories about eye-movements to technically elaborate brain-scanning techniques (see Gamer & Ambach, 2014, for a critical review of different methods).

Regarding cases of child sexual abuse, it is a frequent phenomenon that statements of alleged victims and alleged perpetrators are the only evidence at hand, often resulting in a statement-against-statement constellation. In such cases, the court has to assess the credibility of the statements to decide the verdict. Given that humans perform barely better than chance in judging the truth status of a statement (e.g., Bond & DePaulo, 2006; Ekman & O’Sullivan, 1991), there is a substantial need to establish objective, reliable, and valid techniques. As there is no simple single behavior that solely occurs when people are lying or telling the truth (DePaulo et al., 2003), research has focused on cues more likely to occur in one of the two cases. Köhnken (1990) classified different research approaches into four categories: Psychophysiological phenomena (e.g., skin-conductance level, heart rate), extra-linguistic cues (e.g., speech rate, pitch of voice), nonverbal cues (e.g., facial expressions, pupil dilation), and content of the statement (e.g., quantity of details, contextual embedding).

The psychophysiological phenomena approach uses bodily reactions to specific stimuli (offense-relevant vs. offense-irrelevant stimuli) to come to a diagnostic conclusion. However,

this approach lacks sufficient psychometric quality and, thus, has been rejected as evidence in criminal proceedings in Germany (Federal Court of Justice of Germany, 1998). The extra-linguistic and the nonverbal cues approach assume different behavioral displays for experience-based and nonexperience-based statements (Niehaus, 2001). Several meta-analyses have been conducted to assess the reliability and validity of extra-linguistic and nonverbal cues (e.g., DePaulo et al., 2003; Zuckerman, DePaulo, & Rosenthal, 1981; Zuckerman & Driver, 1985). Most studies revealed only few valid indicators and concluded that there is no prototypical deceptive behavior. Reviewing the literature, Vrij (2000) reported a complex behavioral pattern discriminating between truth tellers and liars, the latter displaying a higher-pitched voice; slower speech; more speech hesitations, errors and pauses; as well as fewer gestures and hand movements. However, study results are inconsistent (Vrij, Edward, & Bull, 2001) and at present it seems unwarranted to use these approaches as a basis of credibility assessment.

In comparison with behavioral lie-detection tools, the content-based assessment of statement veracity has a different diagnostic focus. The basic idea is that statement content varies as a function of cognitive processes involved to produce a statement (Undeutsch, 1954). In psychological research, the most prominent of these techniques are the criteria-based content analysis (CBCA; Steller & Köhnken, 1989) and the reality monitoring (RM; Johnson & Raye, 1981). The aim of the present meta-analysis is to synthesize research on the effectiveness of these content-based techniques by investigating empirical effect sizes.

CBCA

CBCA assessment is admissible as evidence in legal proceedings in the United States (Ruby & Brigham, 1997) and in Western European countries such as Germany and the Netherlands (Vrij, 2000). It is probably the most widely used veracity assessment technique in the world (Vrij, Akehurst, Soukara, & Bull, 2002). CBCA is based on the Undeutsch (1967)

hypothesis, which states that experience-based statements show higher quality than nonexperience-based statements (i.e., they are richer in detail and show more elaborate links to other external events) as they can be simply recalled from memory instead of being actively made up. Steller and Köhnken (1989) compiled a list of 19 content criteria to assess the quality of a statement, subdivided into five different categories: General characteristics, specific characteristics, content particularities, motivational components, and an offence-related component (Table 1). Their systematization represented an integration of previously worked out criteria listed by Undeutsch (1967), Arntzen (1970), Trankell (1971), Szewczyk (1973), and Dettenborn, Froehlich, and Szewczyk (1984).

(Insert Table 1)

CBCA criteria relate to narrative dimensions and are to be assessed in the transcript of a oral statement. They can be rated in terms of their presence and/or strength. This procedure has a positive bias: The presence of criteria enhances the probability that a statement is experience-based. The absence of criteria, on the other hand, does not imply that a statement is fabricated. For example, if a statement refers to a very simple event in question, its quality might be capped, even though it is experience-based. Low quality could also be caused by a person's lack of motivation to make a detailed statement. For this reason, among others, the CBCA should not be considered as a diagnostic checklist, as there is no standardized system to weigh criteria or to set a cutoff score for experience-based statements. In order to assess the credibility of a statement, quality analysis based on CBCA has to be complemented by the assessment of personal (e.g., age of interviewee, abilities to deceive, willingness to testify) and situational (e.g., complexity of event in question, time interval between event in question and interview, interview technique) factors (for a detailed description see Volbert, Steller, & Galow, 2010). The basic idea of this diagnostic procedure, known as Statement Validity Assessment (Köhnken & Steller, 1988), is that the quality of a statement cannot be assessed in an absolute manner, but only in relation to the aforementioned variables.

CBCA can be regarded as a formalization of a set of working hypotheses on credibility assessment, developed by practitioners in this context (e.g., Arntzen, 1970; Szewczyk, 1973; Trankell, 1971; Undeutsch, 1967). A precise description of underlying psychological processes is yet lacking (Lamb et al., 1997). However, Köhnken (1990) described two processes, primary and secondary deception, that are responsible for quality differences.

Primary deception refers to the act of communicating incorrect information and is based on the assumption that a statement the product of a cognitive performance. Constructing a lie requires cognitive resources as the person must be able to invent a story, remember the invented story, reproduce it (often several times and after long periods), conceivably modify and extend the story spontaneously when answering unexpected questions, and then remember these modifications. Köhnken (1990) hypothesized that cognitive schemata about true and fabricated statements are used to comply with these requirements. On the other hand, a true statement is a reconstruction of an experienced event that can simply be retrieved from an autobiographic episodic representation and therefore requires fewer cognitive resources. Secondary deception refers to a lying person's motivation for self-presentation relying on idiosyncratic cognitive schemata on the presumed behavior of liars and truth tellers. Of course, not only liars, but also truth tellers are motivated to appear credible. However, liars are typically strongly concerned about their impression and, hence, concentrate on resource consuming strategic self-presentation. In contrast, self-presentation of truth tellers is usually less strategic and resource-consuming as they are more likely to take their credibility as granted (Volbert & Steller, 2014). Thus, experience-based statements will be less impaired by the consumption of resources and, hence, more elaborate than nonexperience-based statements (Köhnken, 1990).

Elaborating these assumptions, Niehaus (2008) presented a revised classification of content criteria assigning criteria to cognitive processes or strategic self-presentation. Volbert and Steller (2014) further edited this version by classifying content criteria as memory-related

content criteria (further subdivided into episodic autobiographical memory, script-deviant memory, details that were not comprehended) and content criteria of strategic self-presentation (further subdivided into memory-related shortcomings, criteria questioning credibility, other problematic content).

RM

The RM approach was developed on the basis of theoretical assumptions from Johnson and Raye (1981), who also hypothesized differences in the quality of memories of true and fabricated events. In contrast to the CBCA approach, it is important to note that this approach was not originally a method to assess a statement's veracity status, but to describe how people discriminate between memories from external and internal sources (e.g., from imagination). However, the basic idea is similar to CBCA. Based on perceptual processes, experience-based memories should contain more sensory, contextual, and affective information than nonexperience-based memories. In contrast to CBCA, the RM approach contains not only "truth criteria", but also a "lie criterion": Nonexperience-based memories should contain more indicators of cognitive operations such as thoughts and reasoning (Vrij, Mann, Kristen, & Fisher, 2007). At the processing level, the corresponding recall of imagined events should be far more elaborate as compared with the more automatic recall of actually experienced events that includes external information such as spatial, temporal, sensory, and semantic details. Therefore, statements referring to internal information should differ from statements referring to external information (Short & Bodner, 2011).

The RM approach provides a list of criteria that are indicative either of true statements based on external information, or of fabricated statements based on internal information (Sporer, 1997, 2004). As RM has not yet been standardized, there are several versions with different criteria. For example, Sporer (1997) presents eight criteria (Table 2), but Vrij, Akehurst, Soukara, and Bull (2004b) use only four criteria (visual, auditory, temporal, and

spatial details). Depending on the RM version applied, discrimination rates vary (Masip, Sporer, Garrido, & Herrero, 2005). Therefore, RM research results should be carefully scrutinized.

(Insert Table 2)

How Valid Are Content-Based Techniques?

Two basic approaches are used to investigate the validity of content-based techniques: Field studies that examine already existing statements from crime victims in forensic contexts (e.g., Roma, San Martini, Sabatello, Tatarelle, & Ferracuti, 2011) and laboratory studies examing true and fabricated statements that are generated by participants[1] (e.g., Vrij, Akehurst, Soukara, & Bull 2004a).

The field study approach has the advantage of high ecological validity, as statements are derived from real cases. (Akehurst, Manton, & Quandte, 2011). However, the truth status of statements is often difficult to establish with certainty (Vrij, 2005). Their classification as true or fabricated is based on available case facts with varying validity. Confessions to the police, for example, as used by Krahé and Kundrotas (1992), can be assumed to be less valid than video recordings of an event filmed by the offender, as used by Akehurst et al. (2011). Notably, these case facts are not always independent from the quality of statements. For instance, a perpetrator's confession might depend on the body of evidence, i.e., the quality of a victim's statement. If the quality of a statement of an alleged victim is very poor, then the evidence is weak, and the confession of an actual guilty perpetrator less likely compared with cases in which a victim provides a high-quality statement (Craig, Scheibe, Raskin, Kircher, & Dodd, 1999). Moreover, it cannot be ensured that statements based on suggested memories of events are not falsely classified as intentionally fabricated or truthful (Volbert & Steller, 2014).

The laboratory study approach is more standardized and has the advantage that the

objective truth value of a statement is under experimental control, however, at the cost of decreased ecological validity. Laboratory studies can further be divided into two categories. In one category, participants are asked to make a statement about an event from their past, either made up or truthful. Santtila, Roppola, Runtti, and Niemi (2000) additionally included external confirmation that events reported as truthful were actually experience-based, asking the parents of participants. If no such independent validation is used, the truth status of a story is essentially based on self-report. In the other category, objective knowledge about the events is acquired by assigning participants to conditions of an experimentally manipulated event. For instance, participants took part in an event (truthful statement group) or received a description of that event (fabricated statement group) and were questioned about the event later on (e.g., Vrij et al., 2004b).

The validity of CBCA has been investigated in several laboratory (e.g., Akehurst, Köhnken, & Höfer, 2001; Vrij et al., 2002) and field (e.g., Parker & Brown, 2000) studies. Generally, results of validation studies support the basic CBCA assumption that experience-based statements have superior quality as compared with statements not based on experience (e.g., Blandon-Gitlin, Pezdek, Lindsay, & Hagen, 2009; Gödert, Gamer, Rill, & Vossel, 2005; Vrij & Mann, 2006). In a qualitative review, Vrij (2005) examined studies on CBCA with regard to characteristics such as accuracy, interrater agreement, or the frequency of occurrence of different criteria. Independent of the study design, the total CBCA score was higher for true statements than for lies in 11 out of 12 studies. The overall error rate was estimated at about 30% in laboratory studies. A recent meta-analysis (Amado, Arce, & Farina, 2015) of 20 quantitative CBCA studies with children samples found a significant positive effect size for the total CBCA score ($d = 0.78$, $\delta = 0.79$; see Lipsey & Wilson, 2000, for extended information on effect size measures). Within laboratory studies, 65% of truthful statements met more criteria than fabricated statements. Results of field studies showed that 97% of truthful statements met more criteria than fabricated statements and, thus, supported

the validity of CBCA in practical contexts.

As Statement Validity Assessment also includes a diagnostic evaluation of personal and situational factors, many studies took into account external factors that influence statement quality beyond its veracity status (e.g., age, verbal ability, fantasy proneness of the witness, interview style). Some of these studies identified constraints of CBCA. Familiarity with the event, for example, had a stronger influence on CBCA scores than veracity status did (e.g., Blandon-Gitlin, Pezdek, Rogers, & Brodie, 2005; Vrij et al., 2002). With respect to the cognitive load assumption, experiences based on a familiar event are not isolated in memory and can serve as generic information facilitating schematic generation of a fabricated statement (Blandon-Gitlin et al., 2005). Other studies show that inter-individual differences in performance levels and existing cognitive schemata as well as situational factors must be taken into account. For example, Vrij et al. (2004a) showed that performance in lying varied as a function of social skills.

Overall, Vrij (2005) considered CBCA as a robust tool to be applied not only in cases of child sexual abuse, as recommended by some authors (e.g., Honts, 1994; Horowitz et al., 1997). According to study results it can be used for adults as well as for children (Vrij, 2005). Amado et al. (2015) concluded that CBCA largely complies with US-American legal standards for scientific evidence to be admitted in a court (*Daubert standards*), but should not be applied as a checklist – as the criteria lack internal consistency and there is no objective decision rule.

In general, validation studies show that RM discriminates between truthful and fabricated statements as least as well as CBCA (Sporer, 2004). Similar to CBCA, findings are particularly ambiguous on the level of individual criteria (e.g., Memon, Fraser, Colwell, Odinot, & Mastroberardino, 2010; Vrij et al., 2004a).

Comparing the two diagnostic approaches there are some advantages of RM over CBCA. RM has often been shown to have a higher interrater reliability as compared with CBCA (e.g.,

Sporer, 1997; Strömwall, Bengtsson, Leander, & Granhag, 2004; Vrij, Edward, Roberts, & Bull, 2000). This might be because of the straightforward application of the method, which is less time-consuming and does not include as many subjective decisions as CBCA. Sporer (1997) also pointed out that RM criteria are more precise and, therefore, easier to operationalize (e.g., sensory details) as compared with the rather global CBCA criteria (e.g., accounts of subjective mental state). For example, it seems much easier to code sensory details than to code accounts of subjective mental state. Nearly any reported perception could be categorized as sensory detail. But what should be counted as an account of subjective mental state? Greuel et al. (1998) define accounts of subjective mental state as reports of emotional processes, thoughts, and physical sensations within the context of the crime scene. Using this example, coding RM criteria might be easier and require less subjective decisions. But, on the other hand, their diagnostic value could be rather unspecific compared with CBCA criteria.

The Present Study

The goal of this meta-analysis is to estimate the effectiveness of CBCA and RM in discriminating between truthful and fabricated statements. Besides the overall effectiveness of these techniques, differences in the effectiveness between CBCA and RM were of interest. To check whether modalities of application (e.g., age or motivation of participants) are influential, further potential moderators that might affect statement quality, and, thus, the discrimination between truthful and fabricated accounts were taken into consideration.

Moderators

Four groups of moderating variables were coded: Information about participants (age, gender, motivation, interview role and experience status, training status), information about the statement (event characteristics, production mode, type of lie), information about the judgment procedure (type of rater, type of dependent variables [DVs], number and scoring of

criteria), and general study characteristics (study design, publication status, type and year of publication).

Information about participants.

Age.

If statement quality varies as a function of cognitive processes involved to produce a statement, it can be assumed that it is related to the developmental status of participants. More precisely, the (re)production of an actual experienced or nonexperienced event requires cognitive, linguistic, and meta-cognitive abilities that gradually develop in children (e.g., cognitive processing resources; Vrij, 2005). Several studies have shown that CBCA scores are correlated with age (e.g., Blandon-Gitlin et al., 2005; Roma et al., 2011; Vrij et al., 2004a). For instance, Vrij et al. (2004a) found the highest CBCA scores for the oldest participants (14-15 years old), and the lowest for the youngest ones (five to six years old). In interviews on sexual abuse with children between two and 14 years, the total CBCA score and 13 out of 19 criteria were correlated with age (Buck, Warren, Betman, & Brigham, 2002; cited from Vrij, 2005).

Gender.

To the best of our knowledge, previous studies on content-based techniques revealed no gender effects, i.e. statement quality of female and male participants did not differ in general (e.g., Roma et al., 2011; Sporer, 1997). Nevertheless, we tested the moderating effect of participants' gender.

Motivation.

Statement quality should also be related to the motivation of participants to produce a convincing statement, either truthful or fabricated (Köhnken, 1990). Highly motivated persons can be assumed to provide more details (Hauch, Sporer, Michael, & Meissner, 2014), but, for example, might be less likely to express self-criticism. Strategic self-presentation should be a more or less automatic component in practical contexts, especially in cases that

are identity-relevant, such as sexual abuse cases. In laboratory studies, however, we can not assume with certainty that participants are motivated to perform the task and to seem truhful. For this reason, numerous studies provided incentives to motivate both truth tellers and liars to make up compelling stories (e.g., Gödert et al., 2005; Nahari, Vrij, & Fisher, 2012; Vrij et al., 2007).

Interview role and experience status.

With regard to motivational influences, the role of participants in the interview must be taken into account in laboratory studies. As a result of motivational differences statements of accused participants (e.g., Nahari et al., 2012) might be of higher quality than statements of participants that are interviewed as witnesses (e.g., Vrij, Kneller, & Mann, 2000). At the same time, the experience status of participants should be relevant. In terms of statement quality (e.g., accounts of subjective mental state, descriptions of interactions, reproduction of conversation), it seems crucial whether a person actually experienced an event (e.g., being part of a mock crime scene; Vrij et al., 2007) or just observed it (e.g., watching a video of a crime; Vrij et al., 2001).

Training of participants.

Training of participants is likely to constitute another influential factor in laboratory studies. To this effect, children and young adults who were taught CBCA criteria subsequently obtained higher scores than an untrained control group (Vrij et al., 2002). Training effects were found especially for CBCA criteria that represent characteristics of episodic memories, such as patterns of space and time, conversations, or intra-psychic processes (Rutta, 2001). However, participants were not able to produce unusual details, complications, or details they did not seem to understand, despite explicit instructions to do so (Wrege, 2004).

Information about the statement.

Event characteristics.

Regarding the experienced event itself, its relevance should be inherent in the practical forensic context of sexual abuse in field studies, but the situation is different for artificial events in laboratory studies. Therefore, Steller (1989) argued for tailoring laboratory events after cases of sexual abuse on the basis of three criteria: Personal involvement, negative emotional tone, and extensive loss of control. Some studies meet these requirements, for example by asking participants to remember or fabricate an event that caused financial, emotional, and/or physical harm (Merckelbach, 2004).

Production mode.

CBCA criteria are related to narrative dimensions. Thus, the type of statement, reported orally (e.g., Akehurst et al., 2001) or in writing (e.g., Nahari et al., 2012), can be expected to produce differences in quality. Writing is generally less practiced than speaking, requires more deliberateness and commitment, and, thus, it should be less productive and elaborated. As a consequence, differences between liars and truth tellers should be more pronounced in written accounts (Hauch, Blandón-Gitlin, Masip, & Sporer, 2015). In their meta-analysis on computer-based techniques, lies contained fewer sensory perceptual words than true stories only when written down by hand ($g_u = 0.34$), and contained fewer spatial details only when typed on a keyboard ($g_u = 0.13$). Motivational CBCA criteria such as “spontaneous corrections”, in contrast, should be more likely to occur in oral statements.

Type of lie.

Another potential moderator refers to the type of lie that people tell in fabricated statements (outright or concealment lies). Outright lies are complete fabrications, and concealment lies are based on modifications of actually experienced events but include the addition of fabricated or omission of truthful information (Nahari et al., 2012). As concealment lies include activities, events, or conversations participants has actually experienced in some way, they should be more difficult than outright lies to distinguish from truthful statements.

Information about the judgment procedure.

Type of rater.

Eventually, the functioning of a content-based technique depends on the assessment of a person applying the respective criteria. In an ideal situation, an expert rater familiar with the technique is involved. Some studies involved such experienced practitioners (e.g., Vrij et al., 2007), some involved trained laypersons (e.g., Merckelbach, 2004), and others asked amateurs to evaluate the statements (e.g., Nahari et al., 2012). Although a truth bias in judgment has been revealed for laypersons (e.g., Bond & DePaulo, 2006), an "investigator bias" or lie bias seems prominent in professionals' ratings (Meissner & Kassin, 2002; Hauch et al., 2015).

Type of DVs.

Several basic approaches have been used to test the effectiveness of content-based techniques in discriminating truthful statements from lies. They are based on statements that can be classified objectively as truth or lies. Thus, the independent variable always is the actual truth status of a statement, either ascertained by experimental manipulation or external information. DVs, on the other hand, vary more widely. In the studies included, three different types occurred and were examined for possibly moderating effects. The first was the form of DV used in the judicial process: A binary classification of a statement as either truthful or fabricated by a person applying a content-based technique implicating hit rates, false alarms, misses, and correct rejections (e.g., Landry & Brigham, 1992). The second was the DV as the difference between the average sum scores of true and fabricated statements (e.g., Nahari et al. 2012). Here, the presence of individual criteria has been evaluated in two different ways: Adding 1 to a sum score for each criterion present or adding the rated strength of presence of each criterion measured on a Likert scale to a sum score. Third, discriminant analyses can be used; developing a function of individual criteria that optimally classifies statements from this sample as truthful or fabricated (e.g., Memon et al., 2010). With regard

to discriminant analysis it must be noted that this method tends to overestimate classification rates; a correction by cross-validation or bootstrapping methods would be necessary to obtain realistic results (Babyak, 2004). As not all included studies considered this problem, we tested the moderating effect of correcting or not correcting for overestimated classification rates in discriminant analyses.

Number and scoring of criteria.

It has to be noted that selection and number of the content-based technique criteria vary across studies. For RM criteria, different versions exist, and not all CBCA studies apply the full set of 19 criteria from the original catalogue (Steller & Köhnken, 1989). Some experimental designs do not provide the narrative complexity necessary for some of the criteria to occur, and some events do not show the characteristic features of events in the forensic context – as it is not possible to pardon a non-existent perpetrator, for example. If fewer criteria are taken as a basis, the way of excluding criteria might be relevant as well. Practically, this leads to the question of whether criteria were excluded nonsystematically or whether the 14-item version of the CBCA by Raskin, Esplin, and Horowitz (1991) was used.. Raskin et al. (1991) modified the content criteria list by Steller and Köhnken (1989), eliminating the five motivational criteria. Additionally, in some studies raters scored the strength of criteria on a Likert scale, but in others just rated criteria dichotomously as present or not.

General study characteristics.

Study design.

Compared with field studies, fewer differences between liars and truth tellers, and generally less credibility criteria, were observed in laboratory studies (Vrij, 2005). Likewise, Amado et al. (2015) found larger effect sizes for field studies. The quality of statements – and, hence, quality differences between true and fabricated statements – crucially depend on the testimony setting, such as the personal significance of the situation

(Steller, 1989). Obviously, with regard to statement situations, laboratory and field studies differ. Larger effects in field studies could be due to higher ecological validity (e.g., motivation of a participant to make a statement). Moreover, case facts that are used to validate the truth or falsity status of a statement are not always independent of the statement's quality and therefore, may lead to larger effects (Vrij, 2005). Based on these considerations, the study design (field or laboratory study) was tested as a moderator. As participants could be part of both groups within the same study (i.e., make a truthful and a fabricated statement), we further tested the moderating effect of within-subjects and between-subjects designs.

Publication.

The tendency that significant findings are more likely to be published in peer-reviewed journals, referred to as publication bias (Sutton, 2009), might restrict the validity of results. Therefore, publication status was tested as a moderator. Furthermore, a moderator analysis was conducted using the year of publication.

Research Questions

Based on the existing empirical literature on content-based techniques, we investigated the following research questions: (a) Are content-based techniques valid, and if so, how well do they work? (b) What are favorable boundary conditions for the validity of content-based techniques? (c) What should be considered for future research? (d) What practical implications can be derived from the results? Compared with the recent meta-analysis by Amado and colleagues (2015) that focused on the effectiveness of CBCA within Anglo-American samples of children, we extended our analysis to both CBCA and RM within Anglo-American and German samples of children and adults. Additionally, we included further moderators above and beyond the ones tested in Amado et al. (2015) by also taking information about participants, the statement, the judgement procedure, and general study

characteristics into account.

Method

Study Acquisition

Exclusion and inclusion criteria.

For inclusion, a study had to contain at least one group of truth tellers and at least one group of liars who provided statements. Participants could be part of both groups within the same study (i.e., make a truthful and a fabricated statement). Furthermore, statements had to be analyzed by raters using at least one content-based technique and had to be classified as truthful or fabricated by a personal decision, a statistical decision, or by the sum score of criteria. Thus, results needed to be presented as categorization frequencies or as average sum scores and standard deviations for truthful and fabricated accounts. Authors were contacted if results were not presented in the required form or if any information was missing. If scores were reported per criterion or as a total score, average sum scores and standard deviations were calculated. As this meta-analysis focuses on fabricated statements, study results on suggested memories were excluded. Similar to lies, suggested memories are not experience-based and are internally generated, but – and this is a critical difference – not intentionally. Suggested memories are based on the subjective belief that the remembered event has actually happened (Volbert & Steller, 2014; see Loftus & Pickrell, 1995, on the creation of suggested memories). Therefore, CBCA and RM cannot be applied to differentiate between true and suggested memories, as the respective core features do not differ (e.g., subjective conviction of telling the truth eliminating the need for strategic self-presentation). In the case of field studies, suggestive influences on the persons making statements could not be ruled out completely. Nevertheless, these studies were included as their high ecological validity is of special importance.

Dependent variables.

The included studies feature three forms of dependent variables. Statements were judged dichotomously as truthful or fabricated by a personal decision, by statistical decision (i.e., discriminant analysis), or by a scoring of criteria (i.e., sum score).

Database.

To acquire eligible studies, the databases Journals@Ovid, PsycINFO, PSYNDEXplus Literature and Audiovisual Media, and the database of psychological diploma theses (ZPID-Datenbank Diplomarbeiten Psychologie) were searched for the following terms in English: “Criteria-based content analysis”, “CBCA”, “reality monitoring”, “RM”, “scientific content analysis”, ”SCAN”, “statement validity assessment”, “SVA”, “validity checklist” (connected by OR) AND “psychology of evidence”, “statement analysis”, “credibility”, “credibility judgment”, “deception” (connected by OR). The same databases were also searched for the following terms in German: “Kriterienbasierte Inhaltsanalyse”, “CBCA”, “reality monitoring”, “RM”, “scientific content analysis”, “SCAN”, “statement validity assessment”, “SVA”, “validity checklist” (connected by OR) AND “Aussagepsychologie”, “Aussagebeurteilung”, “Glaubhaftigkeit”, “Glaubwürdigkeit” (connected by OR). Several terms of the combined German search terms were English as these are established terms. Additionally, researchers known to work in this field were contacted and asked to provide unpublished literature and any relevant theses. Unpublished and published studies in English or German were included. The database search included studies published or retrieved prior to March 18, 2015.

Final data sample.

From a total of 74 identified studies, 56 matched the inclusion criteria (Figure 1). Some studies included multiple effect sizes of different results that were based on the same sample of statements. To counter the problem of dependent data, the following decision rules were

applied: If studies explored different content-based techniques, only effect sizes for CBCA results were used (35, 37, 40, 42, 43, 47, 48, 52, 56).[2] If studies investigated training effects, only effect sizes of the contrast between trained liars and truth tellers were included, as this comparison is more conservative with regard to statistical analyses (41, 44, 45, 50, 56). If studies contrasted concealment and outright liars with a control group of truth tellers, only effect sizes of the contrast between concealment liars and truth tellers were included (39, 49). Again, this comparison is more conservative with regard to statistical analyses. Study 2 used two different RM procedures. In this case, the effect size of the averaged result was included. Moreover, three studies compared different intervals between event and interview. The following decisions were made: In study 36, the effect size of results with a one-year interval was favored over results with a 10-year interval; in study 55, the effect size of a second interview about an event was favored over the first interview; and in study 37, the effect size of the interview results regarding an event that had been experienced or imagined four times was favored over the interview results regarding an event that had been experienced or imagined one time. Again, these comparisons represent more conservative effect estimations. Furthermore, some samples formed the basis for or were reported on in multiple studies. In these cases, effect sizes of studies of a peer-reviewed journal (25, 30) were favored over book chapters, the effect size of a study with a bigger sample size (16) was favored over a smaller sample size, and effect sizes in studies published in journals with higher impact factors (16, 32) were favored over those in journals with lower impact factors. Finally, we exluded studies on scientific content analysis (SCAN) as we found only five studies using this content-based technique.

Moderator Variables

The moderators were coded either continuously or categorically. Continuous moderators comprised the (a) gender ratio of the sample, in terms of a proportion from zero (no women in

the sample) to one (no men in the sample), and (b) the year of publication. Categorical moderators comprised (a) the content-based technique used (CBCA or RM); (b) information about participants including age (< 18 years or ≥ 18 years), motivation (absence or presence of incentives; only applied to laboratory studies), experience status in the truth condition (event not experienced or event experienced), role in the interview (not accused or accused), and training of participants (not trained, trained); (c) information about the statement including event characteristics (absence or presence of personal involvement, negative emotional tone, and extensive loss of control; presence was only coded if all three criteria were fulfilled), type of lie (concealment lie or outright lie), and production mode (oral or written form); (d) information about the judgment procedure including type of rater (laypersons, trained laypersons, or experts), set of RM criteria (not complete or complete), set of CBCA criteria (not complete, complete, or 14-item version), scoring of the criteria (absence/presence or scoring on a Likert scale), type of dependent variable (personal decision, discriminant analysis, or decision based on sum scores), correction for potential overestimation of results based on discriminant analysis (correction or no correction); and (e) general study characteristics including the study design (field study or laboratory study; within-subject design or between-subject design) and publication status (published or not published). Absent or ambiguous information was coded as missing values, except for motivation, where missing information was coded as “no incentive”.

Coding Procedure and Intercoder Reliability

Two independent coders (first and third author) manually calculated effect sizes, corrected effect sizes, standard errors, and inverse variance weights and coded the moderator variables with the help of a detailed coding manual. Intraclass correlation coefficients (two-way mixed, single measure) for continuous variables ranged from .80 to 1.00 and Cohen’s kappa for categorical variables ranged from .74 to 1.00. Any inconsistencies were cleared by

subsequent mutual discussions.

Statistical Methods

Effect size measure.

We used the standardized mean difference to measure effect sizes because it is applicable to all different forms of results in content-based technique research. The standardized mean difference for between-subjects comparisons is the difference between two groups divided by their pooled standard deviation, Cohen's $d_s = (M_{true} - M_{fabricated})/ SD_{pooled})$. For within-subjects comparisons it is the difference between measures devided by their averaged standard deviation, Cohen's $d_{av} = (M_{true} - M_{fabricated}) / ([SD_{true} + SD_{fabricated}] / 2)$.[3] Effect sizes for studies reporting mean sum scores and standard deviations of one content-based technique for truthful and fabricated statements were calculated on this basis.

However, study results that are based on binary classification rates, either by a personal or a statistical decision (i.e., discriminant analyses), report hit rates, false alarms, misses, and correct rejections instead of mean sum scores and standard deviations. To compare group differences measured on dichotomous scales equivalent to the standardized mean difference, Lipsey and Wilson (2000) propose a probit-transformation of results. Therefore, differences between the probit-transformed hit rate of correctly classified truthful accounts and the probit-transformed false alarms rate of incorrectly classified fabricated accounts were used as a measure of Cohen's *d*. If all statements of one type were correctly or incorrectly classified (i.e., hit rate of 100% or miss rate of 0%), 0.5 was added to the absolute values of hits, misses, false alarms, and correct rejections, as it is not possible to use the probit-transformation on a rate of 0 or 1 (Lipsey & Wilson, 2000).

Given that these calculation methods can result in a slight overestimate, effect sizes were corrected as proposed by Hedges (1981), $g = (1 - 3 / [4 \times N - 9]) \times d$ (Lipsey & Wilson, 2000). Moreover, all effect sizes were weighted by the accuracy of their estimate as

represented by their inverse variance weight, $w = 2 \text{ x } n_{true} \text{ x } n_{fabricated} \text{ x } (n_{true} + n_{fabricated})/2 \text{ x } (n_{true} + n_{fabricated}) + n_{true} \text{ x } n_{fabricated} \text{ x } g$ (Lipsey & Wilson, 2000).

To illustrate the practical implication of the overall effect size, detection rates were calculated. More precisely, exemplary hit rates (HR) and corresponding false alarm rates (FA) were computed based on the formulas HR = Φ * (g + probit[FA]), FA = Φ * (probit[HR] – g), with Φ as cumulative distribution function of x that returns x if used on probit(x), g = 1.03, and different fixed values for HR and FA.

Meta-analytical models.

Following Viechtbauer (2005), we used random-effects models to estimate the overall effect size of content-based techniques as well as the effect sizes of subsets of categorical moderators, $\theta_i = \mu + u_i$. Random-effects models assume that true effect values are normally distributed around μ with a variance of τ^2 and u_i being normally distributed around zero with a variance of τ^2.

However, the influence of moderators was estimated by mixed-effects models, $\theta = ß_0 + ß_1 x_{i1} + u_i$. The mixed-effects model introduces a moderator variable x, where x_i is the value for the variable in study i, which explains part of the variance in true effects. The residual variance u_i that cannot be explained by the rest of the model is assumed to be normally distributed with a variance of τ^2. To estimate τ^2, the restricted maximum likelihood estimator was used, as it is approximately unbiased (Viechtbauer, 2005). To test for significant differences, Q-tests were performed with mixed-effects models in which the moderators were used as grouping variables.

In order to examine the study sample for outliers, two random-effects models on different data sets were conducted, on one data set excluding the lowest effect size and on one excluding the highest effect size. If the recalculated Q statistic of the reduced study sample was significant and changed over 50%, the respective study was excluded as outlier (Babchishin, Nunes, & Hermann, 2013).

Testing publication bias.

To check whether our results are influenced by publication bias, we used a nonparametric trim and fill method developed by Duval and Tweedie (2000a). The basic idea of this procedure is that an unknown number of studies are suppressed (e.g., because of publication policy) and, hence, not included in a meta-analysis. Based on the distribution of included studies in a funnel plot, a simple rank-based data augmentation technique is used to estimate and adjust the effect size for the number of missing studies. Two different non-parametric algorithms, the R_0 and the L_0 algorithm, that differ with regard to their robustness against certain data configurations were used to estimate the number of missing studies (see Duval & Tweedie, 2000b, for extended information). If these algorithms produce different results, a possibility mentioned by Duval and Tweedie (2000a), both have to be reported. Subsequently, the augmented data set was then used in a random-effects model for the overall effectiveness of content-based techniques. Furthermore, as specifically peer-reviewed published studies should be affected by a potential publication bias, the trim and fill method was also applied to this subset.

Software used.

Effect sizes, standard errors, and inverse variance weights were calculated manually. All other calculations were conducted in R (R Department Core Team, 2012) using the packages RStudio (RStudio, 2012) and metafor (Viechtbauer, 2010).

Results

Overall Effect Size

Figure 1 gives an overview of all studies analyzed. Using the final sample of 56 effect sizes, a random-effects model resulted in an estimated effect size of $d = 1.05$ (95% CI [0.79, 1.30], $Q = 435.84$, $p < .001$, $I^2 = 92.95$) and $g = 1.03$ (95% CI [0.78, 1.27], $Q = 420.06$, $p < .001$, $I^2 = 92.48\%$), $k = 56$, $N = 3{,}429$.[4] These results indicate that, irrespective of specific

boundary conditions, content-based techniques discriminate between true and fabricated statements with an overall corrected effect size of g = 1.03. No study was excluded as a statistical outlier.

(Insert Figure 1)

Table 3 illustrates exemplary hit rates and corresponding false alarm rates of the overall effect size g = 1.03. With regard to an equally high sensitivity and specificity, the overall effect size resulted in a hit rate of 69.70% and a false alarm rate of 30.30%. Further calculations showed that higher sensitivity or specificity rates demand rapidly growing costs (e.g., hit rate of 99% demanded false alarm rate of 90.30%).

(Insert Table 3)

Moderator Variables

Results of random-effects models for subsets of the studies based on the categorical moderators variables are displayed in Table 4. For all subsets, effect sizes were positive. Furthermore, none of the 95% CIs included an effect size of zero with the exception of the subset of trained participants. Q-tests with mixed-effects models for categorical moderators revealed significant differences between studies using different sets of CBCA criteria (Q = 6.09, p = .047, k = 46), different dependent variables (Q = 23.58, p < .001, k = 56), and different types of publication (Q = 5.49, p = .002, k = 56).

(Insert Table 4)

Post-hoc tests revealed that studies using the whole set of 19 CBCA criteria showed significantly higher effect sizes than studies using a limited set of CBCA criteria (p = .017). With regard to the 14-item version, there were no significant differences compared with the complete or any incomplete CBCA criteria set. Furthermore, post-hoc tests showed that studies using discriminant analyses to classify statements as truthful or fabricated produced significantly higher effect sizes than studies referring to scores (p < .001). Regarding personal

decisions, there were no significant differences compared with decisions by discriminant analysis or by scores. Finally, post-hoc tests showed that unpublished studies showed higher effect sizes than studies published in peer-reviewed journals (p = .019).

In order to examine differences between published and unpublished studies, χ^2-tests for each moderator variable were run. Results showed that only the distribution of the type of rater (χ^2 (2) = 5.98, p = .05) differed significantly. With regard to published studies, the majority of raters were trained laypersons (k = 34), only k = 8 studies included experts, and k = 2 studies untrained laypersons. Within unpublished studies, frequencies were more comparable (laypersons: k = 2, trained laypersons: k = 4, and experts: k = 4).

All other catergorcial moderator variables revealed no significant effect. Q-tests with mixed-effects models for continuous moderator variables revealed that year of publication (Q = 0.57, p = .451, k = 56) as well as gender ratio in the sample (Q = 2.46, p = .117, k = 44) had no significant influence on effect sizes.

Publication Bias

Results of the trim and fill method entering the final data set yielded no missing studies when using the L_0 estimator and 14 missing studies when using the R_0 estimator. A random-effects model of the augmented data set of k = 70 effect sizes based on the estimates of the R_0 estimator (Figure 2) revealed an estimated effect size of 0.58 (95% CI [0.26, 0.91], Q = 765.16, p < .001, I^2 = 96.01%).

(Insert Figure 2)

Results of the trim and fill method entering the data set of published studies again resulted in no missing studies with the L_0 estimator and eight missing studies with the R_0 estimator. A meta-analysis on this augmented data set of k = 54 effect sizes based on the estimates of the R_0 estimator (Figure 3) revealed an estimated effect size of 0.60 (95% CI [0.29, 0.90], Q = 505.99, p < .001, I^2 = 95.18%, k = 54).

(Insert Figure 3)

Discussion

Are Content-Based Techniques Valid?

The present meta-analysis found strong evidence that content-based techniques discriminate between truthful and fabricated statements above chance level. According to the classification of Cohen (1988), overall effect size of the final data set indicated a large effect of g = 1.03 (95% CI [0.78, 1.27], k = 56). In light of the fact that content-based techniques are frequently used in the practical context and often provide a foundation of far-reaching single-case judgments, this result first of all implicates a necessary condition for a practical application in forensic settings. However, the 95% CI of the overall effect size included medium to large effects, and after controlling for publication bias with the trim and fill method by using the R_0 algorithm, the augmented data set of k = 70 only resulted in a medium effect size of g = 0.58 with a 95% CI [0.26, 0.91] including small to large effect sizes. In addition, further analyses revealed that published studies (g = 0.89, 95% CI [0.66, 1.12], k = 46) showed smaller effect sizes than unpublished studies (g = 1.71, 95% CI [0.90, 2.52], k = 10). As published studies may be particularly affected by a potential publication bias, the trim and fill method was applied to this subset. Again, the augmented data set of k = 54 resulted in a medium effect size of g = 0.60 with a 95% CI [0.29, 0.90] including small to large effect sizes by using the R_0 algorithm. With regard to these results, the estimated overall effect size of g = 1.03 may be overoptimistic. However, applying the trim and fill method by using the L_0 algorithm revealed no missing studies, neither for the total set nor for the subset of published studies. Obviously, the two algorithms generated quite different results. Nevertheless, as Duval and Tweedie (2000a) recommend using both, the differentiating results should be taken as a basis for sensitivity analysis. In general, it must be noted here that the trim and fill method provides information of the effect of potential missing studies. That

means the effect size could change only if unpublished studies are not included. However, the present meta-analysis already included unpublished studies. Therefore, the effect sizes for the total set as well as for the subset of published studies estimated on the augmented data sets should be interpreted as a lower threshold. Moreover, Q and I^2 statistics indicated high heterogeneity between single study effects (Borenstein, Hedges, Higgins, & Rothstein, 2009). The wide range of the effect sizes ($-0.25 < g < 3.67$) might be caused by the great variability in study designs that largely differed with regard to methodological aspects (e.g., participants, study design, motivation of participants). This dovetails with the fact that we have found several moderators that at least partly elucidate the sources underlying this heterogeneity (number of criteria, publication status, type of dependent variables).

In view of the separate effect sizes for the two content-based techniques, it is obvious that both CBCA and RM can discriminate between truthful and fabricated statements above chance level. Moreover, post-hoc analyses revealed no significant differences between content-based techniques, even though RM (g = 1.26, 95% CI [0.78, 1.75], k = 10) descriptively performed slightly better than CBCA (g = 0.97, 95% CI [0.70, 1.25], k = 46). However, as RM has originally not been developed for the context of credibility assessment, its criteria (e.g., affective information, contextual information) might be easier to operationalize in an experimental setup than CBCA criteria that stemmed from a forensic context (e.g., unexpected complications during the incident; Sporer, 1997).

What Are Favorable Boundary Conditions for the Validity of Content-Based Techniques?

Concerning our second research question, moderator analyses showed that truthful and fabricated statements could be significantly better differentiated by using the complete set of CBCA criteria than any incomplete set. This result underpins the validity of the full set of 19 CBCA criteria. According to classical test theory, the superiority of the complete criteria set

was actually to be expected, as test validity increases with the number of converging but conceptually different construct facets measured. However, using any incomplete set of CBCA criteria was often a result of specific experimental conditions in which some CBCA criteria could not even occur (e.g., pardoning the perpetrator, unexpected complications during the incident). In this context, the internal validity of the study design has to be evaluated. Finding an ecologically valid operationalization that is also ethically acceptable is challenging, of course, but this objective must be pursued to be able to investigate the incremental validity of each criterion on this basis. However, the effect size of studies using the 14-item version did not significantly differ from the effect sizes of studies using the complete or any incomplete set. Descriptively, its effect size is just between the effect sizes of the complete and any incomplete CBCA version – in line with expectations from test theory. The absence of significant differences could be explained by the fact that only $k = 5$ studies used the 14-item version. Amado et al. (2015) also found comparable effect sizes for both versions including considerably fewer studies using the short version. Further research on the 14-item version is necessary to get a clearer picture and particular attention should be paid to the fact that motivational items are specifically neglected.

Furthermore, moderator analyses revealed significant differences concerning the dependent data used. More specifically, statistical decisions based on discriminant analyses outperformed statement classifications based on mean sums. This result could partly be explained by the tendency of discriminant analyses to produce overoptimistic estimates of classification rates if no correction for this bias, such as cross-validation, bootstrapping, or leave-one-out, is applied (Babyak, 2004). Only three of 22 studies reported a correction for overoptimistic classification rates based on discriminant analysis. A meta-analytic random-effects model including the three studies still estimated a large effect size of $g = 1.19$ with a 95% CI [0.83, 1.56], descriptively smaller than the effect size of studies not correcting for bias, $g = 1.78$ (95% CI [1.28, 2.29]). However, post-hoc tests revealed no significant

differences between studies correcting or not correcting for a potentional overestimate. Either way, as the subset of studies correcting for a potential overestimate was very small (reducing statistical power to detect any difference), the results must be taken with caution, but we can still assume that studies using discriminant analyses to classify statements outperform the use of sum scores. With regard to personal classification, there were no significant differences. Again, descriptively the effect size of studies referring to personal decisions just fell between the two other effect sizes. Once more it must be noted that only $k = 6$ studies belonged to this category. Future research should investigate the effect of personal decisions (and possible boundary conditions impacting on these), particularly in comparison with the performance of discriminant functions.

Somewhat surprisingly, moderator analyses showed that unpublished studies ($g = 1.71$, 95% CI [0.90, 2.52], $k = 10$) revealed effect sizes approximately twice as large as those in published studies ($g = 0.89$, 95% CI [0.66, 1.12], $k = 46$). In order to examine potential sources of this difference, distributions of moderator characteristics within the two subsets of studies were compared. Results showed that published and unpublished studies significantly differed only with regard to the type of rater. Within unpublished studies ($k = 10$), raters were just as often experts as trained laypersons (each $k = 4$). Only two unpublished studies used untrained laypersons. However, within published studies ($k = 44$), raters were mostly trained laypersons ($k = 34$), eight experts, and, again, only two published studies used untrained laypersons. Therefore, differences between unpublished and published studies might be due to a proportionally higher presence of experts within unpublished studies than within published ones. It is highly likely that experts differ from (trained) laypersons in their interpretation of the criteria. While experienced users might have a precise understanding of criteria such as accounts of subjective mental state, (trained) laypersons might only have a rough idea of the specific conditions necessary to fulfill this criterion. This raises the question of whether the high variability of study results might be explained – at least in parts – by the

idiosyncratic interpretation of the criteria (that could even differ among experts). However, the type of rater revealed no independent effect apart from being confounded with publication status.

What Should Be Considered for Future Research?

Existing content-based techniques require an expert to make a judgment based on detected criteria. However, there is no established cutoff score that can be used for the subjective decision of whether a statement is experience-based or not. This fact introduces a methodological black box into the assessment process, making it difficult to disentangle individual and measure-related sources of classification errors. Moreover, it hampers the standardization of content-based techniques. To counter this problem, standardized decision rules would be necessary. Based on the present findings, such guidance could ideally be developed by running discriminant analyses on large samples of truthful and fabricated statements. As the quality of a statement differs with regard to several characteristics assessed within the Statement Validity Assessment (e.g., development of the statement, verbal competence of participants), no single cutoff value can feasibly be defined. However, developing standardized decision rules based on the results of discriminant analyses could provide information about the level of sensitivity and specificity that must be taken into account for higher or lower cutoff scores and moreover, about the optimal integration of criteria (e.g., scoring). In addition, with more studies reporting classification tables from discriminant analyses it would become testable whether content-based techniques suffer from diagnostic biases (e.g., positivity bias; Rassin, 2000).

Unlike the comparison of sum scores that is independent of the criteria compilation, discriminant analyses further allow to test the predictive power of the particular constellation. The ratio of criteria variance between and within statement groups is optimized while multivariately controlling for possible criteria overlap. This way, theoretical assumptions –

for example on the different effectiveness of characteristics indicating episodic memory and script-deviant details – could be examined irrespective of the number of criteria that tap into overlapping diagnostic constructs. At least for laypersons, the former content characteristics (e.g., temporal information, emotions, thoughts) might be an obvious truth indicator, whereas the latter (e.g., effort to remember, admitting lack of memory, spontanous corrections) might not. Thus, the probability of including script-deviant information in order to make a convincing statement should be lower than the probability of embedding details characteristic of episodic memories (Volbert & Steller, 2014). Accordingly, discriminant analytic results should reveal lower predictive power for characteristics indicating episodic memories compared with script-deviant details.

As discussed before, coding the criteria of content-based techniques requires a large amount of interpretation. There are attempts to specify and clarify coding decisions (e.g., Arnzten, 2011; Greuel et al., 1998), but criteria ratings are still highly subjective and amenable to idiosyncratic operationalization. For example, Arnzten (2011) defines the criterion “descriptions of interactions” as chains of reciprocal actions and reactions. However, this definition leads to further questions: How many actions and reactions does it take to have a chain and code the criterion as present? Are all types of action-reaction chains equally meaningful, independent of content and context? With regard to the latter question it must be stressed that the case-related interpretation of CBCA and RM criteria is an inherent problem of Statement Validity Assessment. The evaluation of criteria is based on the statement-relevant competence of participants and not on normative rules. If a person is eloquent and, hence, able to invent a high quality statement, coders should expect a more complex description of interactions compared with that given by a participant with lower competence. However, there are no norms that specify what exact statement quality has to be expected from an individual with a given degree of competence. Therefore, future studies should empirically address this problem, as the individual quality-competence comparison is a core

tenet of the Statement Validity Assessment.

Additionally, content-based techniques could be improved by further including promising criteria from other approaches. For example, Fuller, Biros, and Delen (2011) developed an automated, text-based deception detection method that includes cues from different methodological backgrounds, for example, RM, linguistic inquiry, and word count (Newman, Pennebaker, Berry, & Richards, 2003), with an overall discrimination rate of 74%. Such an eclectic approach could increase the predictive power of established content-based techniques if these approaches tapped into conceptually distinct shares of criterion variance and if potential overoptimistic biases were statistically controlled for (Babyak, 2004).

In future research, the ecological validity of studies must be a central aim. Even though moderator analyses did not reveal any significant differences between laboratory and field studies, nor for different characteristics of laboratory studies (e.g., motivation, event characteristics, interview role, type of lie), a realistic operationalization is essential – specifically if study findings are to be generalized to forensic settings. Although these lacking context effects corroborate the general robustness of content-based assessment approaches it must be noted that the absence of significant moderator effects could, at least in part, be explained by the high heterogeneity between studies – even within the specific moderator subsets – and, hence, by a lack of statistical power. Addtionally, for some moderator tests statistical power was further affected by small subset groups (e.g., $k = 6$ for field studies). In consequence, suggestions for future research and practical application are in some way limited by methodical restrictions. However, based on strong theoretical reasoning it must be stressed that the quality of statements – and, hence, quality differences between truthful and fabricated statements – should crucially depend on the cognitive demand of the situation and the motivation of a participant. In turn, these variables are likely to be influenced by the setting, e.g. by the complexity and personal significance of the situation. Thus, in order to create optimal conditions, study scenarios should provide a lifelike set-up, although

ecologically valid manipulations of testifying motivation will be difficult to achieve. Of course, for ethical reasons it is not possible to replicate or even approximate extremely stressful real-world scenarios such as child sexual abuse in the laboratory. Moreover, the consequences of a guilty verdict in a criminal trial are obviously not comparable with an absent gratification for a non-convincing statement. Nonetheless, content-based technique research conditions that are as realistic as possible are still mandatory as invalid results have serious consequences in real-life settings. Although event characteristics (personal involvement, negative emotional tone, and extensive loss of control) did not moderate the present findings, we still endorse Steller's (1989) recommendation to tailor study events towards cases of sexual abuse. Despite our results corroborating the general diagnostic capacity of content-based approaches, it can be assumed that such a set up allows more ecologically valid conclusions, particularly since we were only able to test whether the presence of all three or fewer criteria (not specifically which and how many) made a difference. Future research needs to add an increased database to test this empirical notion on more thorough statistical grounds.

What Practical Implications Can Be Derived from the Results?

With an overall effect size of $g = 1.03$, content-based techniques seem to provide substantial potential to discriminate between truthful and fabricated statements, especially as they are only one component of statement analysis. To illustrate the practical consequences of this result, corresponding detection rates were calculated (Table 3). With regard to an equally high sensitivity and specificity, the overall effect size resulted in a hit rate or correct rejection rate of roughly 70%. Although these rates are quite satisfying, content-based technique applications still produce misclassifications, and in practice, every single mistake has severe and far-reaching consequences. For example, this could mean that in 30% of the cases a truthful victim statement is not classified as such and in turn, a guilty offender will not be

convicted. Or, conversely, in 30% of the cases a fabricated victim statement is classified as truthful and an alleged perpetrator gets sentenced although he is not guilty. On the other hand, higher hit or lower false alarm rates demand rapidly growing costs, e.g., a hit rate of 99.90% produces a false alarm rate of 98.00%, or a false alarm rate of 0.10% produces a hit rate of 2%.

The question of whether content-based techniques are good enough to be used as the centerpiece of an expert witness testimony is only partly empirical. Whether a cue is considered valid enough or how it will be weighted is ultimately a judicial issue. For example, eyewitnesses' ability to discriminate and identify faces is a widely accepted type of evidence in court. In a meta-analysis, Meissner and Brigham (2001) estimated an overall effect size of $g = 0.82$ ($k = 56$), which is roughly comparable to the present results. Moreover, meta-analytic results on the validity of CBCA within statements of children ($\delta = 0.79$, $k = 20$) by Amado et al. (2015) are similar to our findings. Additionally, Vrij, Fisher, and Blank (2015) found a comparable overall detection rate of 71% with regard to a cognitive approach (i.e., imposing cognitive load, encouraging interviewees to say more, and asking unexpected questions) of lie detection. From this comparison one cannot derive a clear authorization for the use of content-based techniques, but the latter findings from other deception detection approaches might serve as a benchmark to interpret the CBCA effect size with regard to the practical applicability. Diagnostic precision is not an either-or question in applied contexts, but rather falls onto a dimension of more or less accuracy that needs to be gauged against available alternatives. Importantly, it is an empirical fact that CBCA and RM outperform any unstandardized decision purely based on judges' subjective evaluation, which hardly exceeds chance level (Bond & DePaulo, 2006).

Hence, particularly regarding cases of child sexual abuse where it is a frequent phenomenon that the statements of the alleged victim and/or perpetrator are the only evidence at hand, legal certainty is difficult to establish. In such statement-against-statement

constellations, decision aids based on content-based techniques can form an important psychological part of the judicial investigation (as it is the standard procedure in German courts; Niehaus, 2000). Notably, CBCA or RM scores have never been designed to form the sole basis for the veracity assessment of statements, but should rather be integrated into a comprehensive evaluation of personal (e.g., cognitive capacities) and situational factors (e.g., statement development) gathered by file studies and diagnostic interviews. As this procedure represents a sophisticated task, users should be familiar with the theoretical background and well trained in the practical use of CBCA or RM. However, to the best of our knowledge, the incremental validity of the other parts of Statement Validity Assessment have not yet been tested empirically. Additionally, we can state that – at least for the use of CBCA– coding the complete set of criteria increases the reliability and validity of the assessment. Moreover, both methods are equally applicable, as moderator analysis did not show a significant difference in the effectiveness of CBCA and RM. In summary, it is safe to conclude that content-based techniques are at present among the best available empirically validated methods for veracity assessement of statements.

References

References marked with an asterisk indicate 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., Köhnken, G., & Höfer, E. (2001). Content credibility of accounts derived from live and video presentations. *Legal and Criminological Psychology, 6,* 65-83. doi: 10.1348/135532501168208

*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

Amado, B. G., Arce, R., & Farina, F. (2015). Undeutsch hypothesis and Criteria-Based Content Analysis: A meta-analytic review. *The European Journal of Psychology Applied to Legal Context, 7,* 3-12. doi: 10.1016/j.ejpal.2014.11.002

Arntzen, F. (1970). *Psychologie der Zeugenaussage: Einführung in die forensische Aussagepsychologie [Psychology of testimony: Introduction to forensic psychology of statement analysis]*. Göttingen, Germany: Verlag für Psychologie.

Arntzen, F. (2011). *Psychologie der Zeugenaussage: System der Glaubhaftigkeitsmerkmale* (5. Auflage) *[Psychology of testimony: System of credibility criteria* (5. edition)*]*. München, Germany: Beck.

Babchishin, K. M., Nunes, K. L., & Hermann, C. A. (2013). The validity of Implicit Association Test (IAT) measures of sexual attraction to children: A meta-analysis. *Archives of Sexual Behavior, 42,* 487-499. doi: 10.1007/s10508-012-0022-8

Babyak, M. A. (2004). What you see may not be what you get: A brief, nontechnical introduction to overfitting in regression-type models. *Psychosomatic Medicine, 66,*

411-421. doi: 10.1097/01.psy.0000127692.23278.a9

* Berger, O. (2005). *Aspekte der Zeugenkompetenz und Validierung der kriterienorientierten Aussageanalyse von Jugendlichen mit Intelligenzminderung. [Aspects of the competence of witnesses and validation of the criteria-based content analysis for adolescents with mental retardation]* (unpublished doctoral dissertation). Universität Regensburg, Regensburg, Germany. http://epub.uni-regensburg.de/10348/

* 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

* Bogaard, G., Meijer, E. H., & Vrij, A. (2014). Using an example statement increases information but does not increase accuracy of CBCA, RM, and SCAN. *Journal of Investigative Psychology and Offender Profiling*, *11*, 151-163. doi: 10.1002/jip.1409

Bond, C. F., & DePaulo, B. M. (2006). Accuracy of deception judgments. *Personality and Social Psychology Review*, *10*, 214-234. doi: 10.1207/s15327957pspr1003_2

* Bond, G. D., & Lee, A. Y. (2005). Language of lies in prison: linguistic classification of prisoners' truthful and deceptive natural language. *Applied Cognitive Psychology*, *19*, 313-329. doi: 10.1002/acp.1087

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction into meta-analysis.* Chichester, UK: John Wiley & Sons, Ltd. doi: 10.1002/9780470743386.refs

Buck, J. A., Warren, A. R., Betman, S., & Brigham, J. C. (2002). Age differences in criteria-based content analysis scores in typical child sexual abuse interviews. *Applied Developmental Psychology*, *23*, 267-283. doi: 10.1016/S0193-3973(02)00107-7

* Colwell, K., Hiscock-Anisman, C. K., Memon, A., Taylor, L., & Prewett, J. (2007). Assessment criteria indicative of deception (ACID): An integrated system of investigative interviewing and detecting deception. *Journal of Investigative*

Psychology and Offender Profiling, *4*, 167-180. doi: 10.1002/jip.73

Cohen, J. (1988). *Statistical power analysis for the behavioral science* (second edition). New Jersey, USA: Erlbaum.

* 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

DePaulo, B. M., Lindsay, J. J., Malone, B. E., Muhlenbruck, L., Charlton, K., & Cooper, H. (2003). Cues to deception. *Psychological Bulletin*, *129*, 74-118. doi: 10.1037//0033-2909.129.1.74

Dettenborn, H., Froehlich, H.-H., & Szewczyk, H. (1984). *Forensische Psychologie [Forensic psychology]*. Berlin, Germany: VEB Deutscher Verlag der Wissenschaften.

* Dolezych, N. (2006). *Die Umsetzung von intuitiven Täuschungsstrategien in nicht erlebnisbasierten Aussagen [The implementation of intuitive deceptive strategies within nonexperience-based statements]* (unpublished diploma thesis). Universität Potsdam, Potsdam, Germany.

Duval, S. J., & Tweedie, R. L (2000a). Trim and Fill: A Simple Funnel-Plot–Based Method of Testing and Adjusting for Publication Bias in Meta-Analysis. *Biometrics*, *56*, 455–463. doi: 10.1111/j.0006-341X.2000.00455.x

Duval, S. J., & Tweedie, R. L. (2000b). A non-parametric 'trim and fill' method of assessing publication bias in meta-analysis. *Journal of the American Statistical Association*, *95*, 89-98. doi: 10.2307/2669529

Ekman, P., & O'Sullivan, M. (1991). Who can catch a liar? *American Psychologist*, *46*, 913-920. doi:10.1037/0003-066X.46.9.913

Federal Court of Justice of Germany (1998). BGH 1 StR 156/98, verdict of 17.12.1998 (German regional court Mannheim). http://www.hrr-strafrecht.de/hrr/1/98/1-156-98.php3?view=print

* Fiegler, S. (2009). *Zur Gültigkeit der Undeutsch-Hypothese unter Berücksichtigung der Schwierigkeit aus einer untrainierten Stichprobe [On the validity of the Undeutsch-Hypothesis in consideration of difficulty within an untrained sample]* (unpublished diploma thesis). Julius-Maximilians-Universität Würzburg, Würzburg, Germany.

Finkelstein, J. J. (1968/69). The Laws of Ur-Nammu. *Journal of Cuneiform Studies*, *22*(3/4), 66-82.

Fuller, C. M., Biros, D. P., & Delen, D. (2011). An investigation of data and text mining methods for real world deception detection. *Expert Systems with Applications, 38*, 8392-8398. doi:10.1016/j.eswa.2011.01.032

* Gödert, H. W., Gamer, M., Rill, H. G., & Vossel, G. (2005). Statement validity assessment: Inter-rater reliability of criteria-based content analysis in the mock-crime paradigm. *Legal and Criminological Psychology, 10,* 225-245. doi: 10.1348/135532505X52680

Gamer, M., & Ambach, W. (2014). Deception research today. *Frontiers in psychology, 5*, Article 256. doi: 10.3389/fpsyg.

Greuel, L., Offe, S., Fabian, A., Wetzels, P., Fabian, T., Offe, H. & Stadler, M. (1998). *Glaubhaftigkeit der Zeugenaussage: Theorie und Praxis der forensisch-psychologischen Begutachtung [Credibility of witness statement: Theory and practice of psychological statement assessment]*. Weinheim, Germany: Beltz.

Hauch, V., Blandón-Gitlin, I., Masip, J., & Sporer, S. L. (2015). Are Computers Effective Lie Detectors? A Meta-Analysis of Linguistic Cues to Deception. *Personality and Social Psychology Review, 19,* 307-342. doi: 10.1177/1088868314556539.

Hauch, V., Sporer, S. L., Michael, S. W., & Meissner, C. A. (2016). Does Training Improve the Detection of Deception? A Meta-Analysis. *Communication Research, 43,* 283-343. doi: 10.1177/0093650214534974

Hedges, L. V. (1981). Distribution theory for glass's estimator of effect size and related estimators. *Journal of Educational Statistics, 6,* 107-120.

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., 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

* Jang, K.-W., Kim, D.-Y., Cho, S., & Lee, J.-H. (2013). Effects of the combination of P3-based GKT and reality monitoring on deceptive classification. *Frontiers in Human Neuroscience*, *7*, Article 18. doi: 10.3389/fnhum.2013.00018

*Janka, C. (2003). *Der Einfluß des Zeitintervalls zwischen Ereignis und Aussage auf die inhaltliche Qualität wahrer und intentional falscher Aussagen [The influence of the time interval between event and statement on the content quality of true and deceptive statements]* (unpublished diploma thesis). Technische Universität Berlin, Berlin, Germany.

Johnson, M. K., & Raye, C. L. (1981). Reality monitoring. *Psychological Review*, *88*, 67-85. doi: 10.1037/0033-295X.88.1.67

Köhnken, G. (1990). *Glaubwürdigkeit: Untersuchungen zu einem psychologischen Konstrukt [Credibility: Investigations of a psychological construct]*. München, Germany: Psychologie Verlags Union.

* Köhnken, G., Schimossek, E., Aschermann, E., & Höfer, E. (1995). The cognitive interview and the assessment of the credibility of adults' statements. *Journal of Applied Psychology*, *80*, 671-684. doi: 10.1037/0021-9010.80.6.671

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

* Krahé, B., & Kundrotas, S. (1992). Glaubwürdigkeitsbeurteilung bei Vergewaltigungsanzeigen: Ein aussagenanalytisches Feldexperiment [Credibility

assessment of rape reports: A field study on statement assessment]. *Zeitschrift für experimentelle und angewandte Psychologie, 39,* 598-620.

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for *t*-tests and ANOVAs. *Frontiers in Psychology, 4*, Article 863, doi: 10.3389/fpsyg.2013.00863

* Lamb, M. E., Sternberg, K. J., Esplin, P. W., Hershkowitz, I. Orbach, Y., & Hovav, M. (1997). Criterion-based content analysis: A field validation study. *Child Abuse and Neglect, 21*, 255-264. doi: 10.1016/S0145-2134(96)00170-6

* Landry, K. L., & Brigham, J. C. (1992). The effect of training in criteria-based content analysis on the ability to detect deception in adults. *Law and Human Behavior, 16,* 663-676. doi: 10.1007/BF01884022

* 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. doi: 10.1111/lcrp.12017

Lipsey, M. W., & Wilson, D. B. (2000). *Practical meta-analysis (Applied Social Research Methods)*. Thousand Oaks, CA: Sage Publication.

Loftus, E. F., & Pickrell, J. (1995). The formation of false memories. *Psychiatric Annals, 25*, 720- 724.

* Lüdke, S. (2008). *Der Einfluss des Unspezifitätseffekts auf die Aussagequalität: Werden erlebnisbasierte Aussagen depressiver Frauen für unwahr gehalten? [The influence of the unspecificity-effect the on statement quality: Are experience-based statements of depressive women rated as decpetive?]* (unpublished diploma thesis) Freie Universität Berlin, Berlin.

Masip, J., Sporer, S. L., Garrido, E., & Herrero, C. (2005). The detection of deception with the reality monitoring approach: A review of the empirical evidence. *Psychology,*

Crime & Law, 11, 99-122. doi: 10.1080/10683160410001726356

Meissner, C. A., & Brigham, J. C. (2001). Thirty years of investigating the own-race in memory for faces. A meta-analytic review. *Psychology, Public Policy, and Law, 7*, 3-35. doi: 10.1037/1076-8971.7.1.3

Meissner, C. A., & Kassin, S. M. (2002). "He`s guilty!": Investigator bias in judgments of truth and deception. *Law and Human Behavior, 26*, 469-480. doi: 10.1023/A:1020278620751

* Memon, A., Fraser, J., Colwell, K., Odinot, G., & Mastroberardino, S. (2010). Distinguishing truthful from invented accounts using reality monitoring criteria. *Legal and Criminological Psychology, 15*, 177-194. doi: 10.1348/135532508X401382

* Merckelbach, H. (2004). Telling a good story: Fantasy proneness and the quality of fabricated memories. *Personality and Individual Differences*, *37*, 1371-1382. doi: 10.1016/j.paid.2004.01.007

* Nahari, G., Vrij, A., & Fisher, R. P. (2012). Does the truth come out in the writing? SCAN as a lie detection tool. *Law and Human Behavior, 36*, 68-76. doi: 10.1037/h0093965

* Naumann, T. (2005). *Zur Anwendbarkeit der Kriterienorientierten Inhaltsanalyse bei nicht-erlebnisbegründeten Aussagen nach Vorabinformation unterschiedlichen Ausmaßes [On the applicability of the criteria-based content analysis on nonexperience-based statements with regard to different forms of pre-information]* (unpublished diploma thesis). Technische Universität Braunschweig, Braunschweig, Germany.

Newman, M. L., Pennebaker, J. W., Berry, D. S., & Richards, J. M. (2003). Lying words: Predicting deception from linguistic styles. *Personality and Social Psychology Bulletin, 29*, 665-675. doi: 10.1177/0146167203029005010

* Niehaus, S. (2000). *Zur Anwendbarkeit inhaltlicher Glaubhaftigkeitsmerkmale bei Zeugenaussagen unterschiedlichen Wahrheitsgehaltes [The applicability of content-based credibility criteria within statements differing in reality content]* (unpublished

doctoral dissertation). Universität Osnabrück, Osnabrück, Germany.

Niehaus, S. (2001). *Zur Anwendbarkeit inhaltlicher Glaubhaftigkeitsmerkmale bei Zeugenaussagen unterschiedlichen Wahrheitsgehaltes [The applicability of content-based credibility criteria within statements differing in reality content]*. Frankfurt am Main, Germany: Peter Lang.

Niehaus, S. (2008). Merkmalsorientierte Inhaltsanalyse [Criteria-based content analysis]. In R. Volbert, & M. Steller (Eds.), *Handbuch der Rechtspsychologie* (pp. 311-321). Göttingen, Germany: Hogrefe.

* 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. doi: 10.1348/135532500168119

* Porter, S., & Yuille, J. C. (1996). The language of deceit: An investigation of the verbal clues to deception in the interrogation context. *Law and Human Behavior, 20,* 443-458. doi: 10.1007/BF01498980

*Porter, S., Yuille, J. C., & Lehman, D. R. (1999). The nature of real, implanted, and fabricated memories for emotional childhood events: Implications for the recovered memory debate. *Law and Human Behavior, 23,* 517-537.

R Development Core Team (2012). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. http://www.R-project.org.

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, Salt Lake City, UT.

Rassin, E. (2000). 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

Ruby, C. L., & Brigham, J. C. (1997). The usefulness of the criteria-based content analysis technique in distinguishing between truthful and fabricated allegations: A critical review. *Psychology, Public Policy, and Law*, *3*, 705-737. doi:10.1037/1076-8971.3.4.705

* Ruby, C. L., & Brigham, J. C. (1998). Can criteria-based content analysis distinguish between true and false statements of african-american speakers? *Law and Human Behavior*, *22*, 369-388. doi: 10.1023/A:1025766825429

* Rutta, Y. (2001). *Der Effekt von Hintergrundwissen über aussagepsychologische Methodik auf die inhaltliche Qualität von intentionalen Falschaussagen [The effect of background knowledge on credibility assessment on the content quality of decpetive statements]* (unpublished diploma thesis). Freie Universität Berlin, Berlin, Germany.

* 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

* Scheinberger, R. (1993). *Inhaltliche Realkennzeichen in Aussagen von Erwachsenen – Eine Simulationsstudie zur wissenschaftlichen Evaluation der kriterienorientierten Aussageanalyse. [Content-based reality criteria in adults` statements – a simulation study on the scientific evaluation of criteria-based content analysis.]* (unpublished diploma thesis). Freie Universität Berlin, Berlin, Germany.

* Schelleman-Offermans, K., & Merckelbach, H. (2010). Fantasy proneness as a confounder of verbal lie detection tools. *Journal of Investigative Psychology and Offender Profiling*, *7*, 247-260. doi: 10.1002/jip.121

* Short, J. L., & Bodner, G. E. (2011). Differentiating accounts of actual, suggested and fabricated childhood events using the judgment of memory characteristics

questionnaire. *Applied Cognitive Psychology, 25*, 775-781. doi: 10.1002/acp.1756

* Sporer, S. L. (1997). The less travelled road to truth: verbal cues in deception detection in accounts of fabricated and self-experienced events. *Applied Cognitive Psychology, 11*, 373-397. doi: 10.1002/(SICI)1099-0720(199710)11:5<373::AID-ACP461>3.0.CO;2-0

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

* Sporer, S. L., & Küpper, B. (1995). Realitätsüberwachung und die Beurteilung des Wahrheitsgehalts von Erzählungen: Eine experimentelle Studie. [Reality monitoring and the assessment of reality content in narratives: an experimental study.] *Zeitschrift für Sozialpsychologie, 26*, 173-193.

* Sporer, S. L., & Sharman, S. J. (2006). Should I believe this? Reality monitoring of invented and self-experienced events from early and late teenage years. *Applied Cognitive Psychology, 20*, 837-854.

* 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.]* (unpublished research paper). Universität Konstanz, Konstanz. http://opus.bsz-bw.de/fhhv/frontdoor.php?source_opus=321

Steller, M. (1989). Recent developments in statement analysis. In J. C. Yuille (Ed.), *Credibility assessment* (pp. 135-154). New York, USA: Kluwer/ Plenum.

Steller, M., & Köhnken, G. (1989). Criteria-based statement analysis. In D. C. Raskin (Ed.). *Psychological methods in criminal investigation and evidence* (pp. 217-245). New York, USA: Springer.

* Steller, M., Wellershaus, P., & Wolf, T. (1992). Realkennzeichen in Kinderaussagen. [Reality criteria in children`s statements.] *Zeitschrift für experimentelle und angewandte Psychologie, 39*, 151-170.

* 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

* Strömwall, L. A., & Granhag, P.-A. (2005). Children`s repeated lies and truths: Effects on adults judgments and reality monitoring scores. *Psychiatry, Psychology & Law*, *12*, 345-356. doi: 10.1375/pplt.12.2.345

* Suckle-Nelson, J. A., Colwell, K., Hiscock-Anisman, C., Florence, S., Youschak, K. E. & Duarte, A. (2010). Assessment criteria indicative of deception (ACID): Replication and gender differences. *The Open Criminology Journal*, *3*, 23-30. doi: 10.2174/1874917801003010023

Sutton, A. J. (2009). Publication Bias. In H. Coopeer, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook* of *reserach synthesis and meta-analysis* (pp. 435-452). New York, USA: Russell Sage Foundation.

Szewczyk, H. (1973). Kriterien der Beurteilung kindlicher Zeugenaussagen [Criteria for the assessment of childrens' statements]. *Probleme und Ergebnisse der Psychologie, 46*, 47-66.

Trankell, A. (1971). *Der Realitätsgehalt von Zeugenaussagen: Methoden der Aussagepsychologie. [Reality content in testimonies: Methods of psychology of statement analysis.]* Göttingen, Germany: Vandenhoeck & Ruprecht.

* 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. (1954). Die Entwicklung der gerichtspsychologischen Gutachtertätigkeit. [The development of expert activity in forensic psychology.] In A. Wellek (Ed.), *Bericht über den 19. Kongress der Deutschen Gesellschaft für Psychologie in Köln 1953*

[Report of the 19th congress of the German Association of Psychology, Cologne, 1953] (pp. 132-154). Göttingen, Germany: Hogrefe.

Undeutsch, U. (1967). Beurteilung der Glaubhaftigkeit von Zeugenaussagen. [Assessment of statement credibility.] In U. Undeutsch (Ed.), *Forensische Psychologie* (pp. 26-181). Göttingen, Germany: Hogrefe.

Viechtbauer, W. (2005). Bias and efficiency of meta-analytic variance estimators in the random-effects model. *Journal of Educational and Behavioral Statistics, 30*, 261–293. doi: 10.3102/10769986030003261

Viechtbauer, W. (2010). Conducting Meta-Analyses in R with the metafor Package. *Journal of Statistical Software, 36*(3), 1-48.

Volbert, R., & Banse, R. (2014). Deception detection. *European Psychologist, 19*, 159-161. doi: 10.1027/1016-9040/a000209

Volbert, R., & Steller, M. (2014). Is this testimony truthful, fabricated, or based on false memory? Credibility assessment 25 years after Steller and Köhnken (1989). *European Psychologist, 19*, 207-220. doi:10.1027/1016-9040/a000200

Volbert, R., Steller, M. & Galow, A. (2010). Das Glaubhaftigkeitsgutachten. [Credibility assessment.] In H.-L. Kröber, D. Dölling, N. Leygraf, & H. Saß (Eds.), *Handbuch der Forensischen Psychiatrie* (pp. 623-689). Heidelberg, Germany: Springer.

Vrij, A. (2000). *Detecting lies and deceit: The psychology of lying and its implications for professional practice*. Chichester, UK: Wiley.

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

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, S., & Bull, R. (2004a). Let me inform you how to tell a

convincing story: CBCA and reality monitoring scores as a function of age, coaching, and deception. *Canadian Journal of Behavioural Science/Revue canadienne des sciences du comportement*, *36*, 113-126. doi:10.1037/h0087222

* Vrij, A., Akehurst, L., Soukara, S., & Bull, R. (2004b). 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

* Vrij, A., Edward, K., & Bull, R. (2001). Stereotypical verbal and nonverbal responses while deceiving others. *Personality and Social Psychology Bulletin*, *27*, 899-909. doi: 10.1177/0146167201277012

* Vrij, A., Edward, K., Roberts, K. P., & Bull, R. (2000). Detecting deceit via analysis of verbal and nonverbal behavior. *Journal of Nonverbal Behavior*, *24*, 239-263. doi: 10.1023/A:1006610329284

Vrij, A., Fisher, R. P., & Blank, H. (2015). A cognitive approach to lie detection: A meta‐analysis. *Legal and Criminological Psychology*. Advance online publication. doi: 10.1111/lcrp.12088

* Vrij, A., Kneller, W., & Mann, S. (2000). The effect of informing liars about criteria-based content analysis on their ability to deceive CBCA-raters. *Legal and Criminological Psychology*, *5*, 57-70. doi: 10.1348/135532500167976

*Vrij, A., & Mann, S. (2006). Criteria-based content analysis: An empirical test of its underlying processes. *Psychology, Crime & Law*, *12*, 337-349. doi: 10.1080/10683160500129007

* Vrij, A., Mann, S., Kristen, S., & Fisher, R. P. (2007). Cues to deception and ability to detect lies as a function of police interview styles. *Law and Human Behavior*, *31*, 499-518. doi: 10.1007/s10979-006-9066-4

* Willén, R. M., & Strömwall, L. (2012). Offenders' uncoerced false confessions: A new

application of statement analysis? *Legal and Criminological Psychology*, *17*, 346-359. doi: 10.1111/j.2044-8333.2011.02018.x

* Wolf, P., & Steller, M. (1997). Realkennzeichen in Aussagen von Frauen. Zur Validierung der Kriterienorientierten Aussageanalyse für Zeugenaussagen von Vergewaltigungsopfern [Reality criteria in statements of women: Validation of the CBCA for statements of rape victims]. In L. Greuel, T. Fabian, & M. Stadler (Eds.), *Psychologie der Zeugenaussage* (pp. 121-130). Weinheim, Germany: Psychologie Verlags Union.

* Wrege, J. (2004). *Der Einfluss von Hintergrundinformationen auf spezielle Glaubwürdigkeitsmerkmale. [The influence of background information on certain credibility criteria.]* (unpublished diploma thesis). Freie Universität Berlin, Berlin, Germany.

* 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

Zuckerman, M., DePaulo, B. M., & Rosenthal, R. (1981). Verbal and nonverbal communication of deception. *Advances in Experimental Social Psychology*, *14*, 1-59.

Zuckerman, M., & Driver, R. E. (1985). Telling lies: Verbal and nonverbal correlates of deception. In W. A. Siegman, & S. Feldstein (Eds.), *Multichannel integration of non-verbal behavior* (pp. 129-147). Hillsdale: Erlbaum.

Footnotes

[1] In the following the term “participants” refers to participants making a truhthful or fabricated statement.

[2] The numbers in brackets indicate the study IDs (see Figure 1).

[3] For within-subjects comparisons, effect sizes could be also calculated as Cohen’s $d_{rm} = ([M_{true} - M_{fabricated}] / \sqrt{[SD_{true}^2 + SD_{fabricated}^2 - 2 \times r \times SD_{true} \times SD_{fabricated}]}) \times \sqrt{(2\,[1 - r])}$ (Lakens, 2013). The formula takes the correlation between measures of depedent groups into account. As this was not regularly provided in the included studies, we ran simulation analyses for all within-subjects comparisons with varying correlation coefficients ($r = 0.1$, $r = 0.25$, $r = 0.5$, $r = 0.75$, $r = 0.9$). Results showed that effect sizes of within-subjects comparisons calculated as Cohen’s d_{rm} with varying correlation coefficients are virtually identical to each other and to effect sizes calculated as Cohen’s d_{av} differing only at the third decimal. Therefore, effect sizes of within-study comparisons were calculated as Cohen’s d_{av}.

[4] Overall effect sizes are reported for uncorrected (d) and corrected (g) effect sizes. All further results are reported for corrected effect sizes.

Table 1

CBCA Criteria (Adapted From Steller & Köhnken, 1989)

General characteristics
1. Logical consistency
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 state
13. Attribution of perpetrator's mental state
Motivation-related contents
14. Spontaneous corrections
15. Admitting lack of memory
16. Raising doubts about one's own testimony
17. Self-deprecation
18. Pardoning the perpetrator
Offense-specific element
19. Details characteristic of the offense

Table 2

RM Criteria (Adapted From Sporer, 1997)

1. Clarity
2. Realism
3. Reconstructability of the story
4. Emotions
5. Perceptual information
6. Spatial information
7. Temporal information
8. Cognitive operations

Table 3

Example Calculations for Targeted Hit Rates at the Price of Corresponding False Alarm Rates for the Overall Effect Size of g = 1.03

Hit rate	False alarm rate
69.70	30.30
95.00	73.00
99.00	90.30
99.90	98.00
30.00	**5.00**
9.70	**1.00**
2.00	**0.10**

Notes: All values in % and bolded values were fixed.
First line shows values for equal hits and correct rejections.

Table 4

Meta-analytical Results for Categorical Moderator Variables

Moderator	Moderator category	*g*	95% CI	*Q*	I^2 in %	*k*	*n*	Study IDs
Content-based technique	$Q_{between}$ (*df* = 1)			0.98				
	CBCA	0.97	[0.70, 1.25]	333.17***	92.51	46	2669	1, 3 – 7, 9, 10, 12 – 24, 27, 28, 30 – 33, 35 – 37, 39 – 50, 52 – 54, 56
	RM	1.26	[0.78, 1.75]	75.99***	90.75	10	760	2, 8, 11, 25, 26, 29, 34, 38, 51, 55
Age	$Q_{between}$ (*df* = 1)			0.37				
	< 18	0.90	[0.39, 1.41]	96.30***	91.76	11	767	7, 15, 16, 19, 21, 30, 32, 37, 46, 49, 55
	≥ 18	1.12	[0.79, 1.45]	300.45***	92.97	39	2027	1 – 6, 8 – 12, 14, 17, 20, 22 – 24, 26 – 29, 31, 33 – 36, 38, 43 – 45, 47, 48, 50 – 54, 56
Incentives	$Q_{between}$ (*df* = 1)			0.16				
	Yes	0.93	[0.39, 1.46]	75.45***	94.14	13	976	1, 11, 17, 23, 30, 36, 39 – 42, 45, 47, 52
	No	1.04	[0.76, 1.32]	321.77***	91.97	42	2403	2 – 10, 12 – 16, 18 – 21, 24 – 29, 31 – 35, 37, 38, 43, 44, 46, 48 – 51, 53 – 56
Experience status in truthful condition	$Q_{between}$ (*df* = 1)			0.06				
	Event experienced	1.02	[0.75, 1.30]	382.10***	93.46	49	3041	1 – 4, 6 – 8, 10 – 17, 19 – 33, 36 – 49, 51, 53 – 56
	Event not experienced	1.11	[0.72, 1.50]	21.16**	72.80	7	388	5, 9, 18, 34, 35, 50, 52
Interview role in lying condition	$Q_{between}$ (*df* = 1)			0.17				
	Accused	1.21	[0.30, 2.11]	35.38***	92.07	6	369	1, 11, 23, 38, 47, 51
	Not accused	1.01	[0.75, 1.26]	384.41***	92.57	50	3060	2 – 10, 12 – 22, 24 – 37, 39 – 46, 48 – 50, 52 – 56
Participant training	$Q_{between}$ (*df* = 1)			0.47				
	Yes	0.79	[-0.06, 1.65]	28.18***	92.71	6	380	41, 42, 44, 45, 50, 56

	No	1.07	[0.80, 1.33]	387.17***	92.57	49	2966	1 – 40, 43, 46 – 49, 51 – 53, 55
Event characteristics	$Q_{between}$ (df = 1)			0.74				
	All 3 present	1.24	[0.63, 1.84]	166.80***	93.37	16	663	3, 4, 6, 12, 14, 16 – 19, 21, 30, 33, 36, 41, 45, 49
	At least 1 missing	0.94	[0.70, 1.19]	250.80***	90.83	40	2766	1, 2, 5, 7 – 11, 13, 15, 20, 22 – 29, 31, 32, 34, 35, 37 – 40, 42 – 44, 46 – 48, 50 – 56
Lie	$Q_{between}$ (df = 1)			1.41				
	Concealment	1.29	[0.71, 1.87]	37.38***	88.90	8	471	2, 11, 34, 35, 38, 39, 51, 52
	Outright	0.89	[0.61, 1.17]	270.38***	92.90	40	2523	1, 4 – 10, 12, 13, 15, 17, 18, 20, 22, 24 – 26, 28 – 32, 36, 37, 40 – 50, 53 – 56
Statement	$Q_{between}$ (df = 1)			0.20				
	Oral	1.05	[0.79, 1.31]	344.53***	91.44	48	2756	1 – 5, 7 – 9, 11 – 23, 27 – 37, 39 – 50, 52 - 55
	Written	0.92	[0.16, 1.68]	68.21***	96.31	8	673	6, 10, 24 – 26, 38, 51, 56
Rater	$Q_{between}$ (df = 2)			0.18				
	Laypersons	1.12	[0.22, 2.01]	37.48***	90.66	4	367	22, 24, 26, 38
	Trained laypersons	0.99	[0.71, 1.26]	213.55***	90.90	38	2253	2, 4 – 6, 8 – 16, 20, 21, 23, 25, 27, 28, 30, 32– 37, 39, 40, 42 – 44, 46, 48 – 50, 52, 55, 56
	Experts	0.92	[0.33, 1.50]	134.20***	93.54	12	715	1, 3, 7, 17 – 19, 31, 41, 45, 47, 53, 54
CBCA criteria	$Q_{between}$ (df = 2)			6.09*				
	Complete	1.54	[0.82, 2.26]	108.32***	92.16	12	422	3, 18, 22, 24, 33, 36, 37, 41, 44, 45, 48, 49
	Not complete	0.70	[0.43, 0.96]	130.38***	88.21	29	1835	1, 4 – 7, 9, 10, 12 – 15, 17, 20, 21, 23, 27, 28, 30, 31, 39, 40, 42, 43, 46, 47, 50, 53, 54, 56
	14-item version	1.20	[0.47, 1.93]	45.48***	92.80	5	412	16, 19, 32, 35, 52
RM criteria	$Q_{between}$ (df = 1)			1.67				
	Complete	1.65	[0.43, 2.87]	22.68***	93.78	4	215	25, 38, 51, 55
	Not complete	0.80	[0.30, 1.28]	25.34***	83.76	4	419	2, 8, 26, 29
Scoring	$Q_{between}$ (df = 1)			0.79				
	Absence/ presence	1.22	[0.70, 1.73]	181.34***	92.99	17	971	7, 9, 11, 12, 16, 19, 21, 22, 31, 33 – 35, 50, 51, 53 – 55
	Scoring on a	0.95	[0.66, 1.23]	221.46***	92.15	37	2354	1 – 6, 8, 10, 13 – 15, 17, 18, 20, 23 – 26, 28, 30,

	scale							32, 36 – 49, 52, 56
Dependent variables	$Q_{between}$ (*df* = 2)			23.58***				
	M and SD	0.51	[0.29, 0.73]	139.62***	85.40	27	2153	1, 6, 7, 10, 16, 17, 19, 21, 23, 24, 26, 27, 30, 31, 35, 39 – 42, 46 – 48, 50, 52 – 54, 56
	Personal decision	1.09	[0.60, 1.59]	15.09**	66.40	6	232	8, 12, 18, 22, 44, 49
	Discriminant function (DF)	1.70	[1.26, 2.14]	144.87***	91.99	23	1044	2 – 5, 9, 11, 13 – 15, 20, 25, 28, 29, 32 – 34, 36 – 38, 43, 45, 51, 55
	$Q_{between}$ (*df* = 1)			0.18				
	DF with correction bias	1.19	[0.83, 1.56]	2.53	13.32	3	157	2, 5, 11
	DF without correction bias	1.78	[1.28, 2.29]	142.33***	93.09	20	887	3, 4, 9, 13–15, 20, 25, 28, 29, 32-34, 36-38, 43, 45, 51, 55
Study designs	$Q_{between}$ (*df* = 1)			2.91				
	Field study	1.66	[0.73, 2.59]	49.66***	90.28	6	335	3, 14, 16, 19, 21, 33
	Lab study	0.95	[0.71, 1.19]	330.32***	92.00	50	3094	1, 2, 4 – 13, 15, 17, 18, 20, 22 – 32, 34 – 56
	$Q_{between}$ (*df* = 1)			1.75				
	Within-subjects	0.80	[0.40, 1.19]	125.15***	93.37	18	929	4, 6, 8, 10, 17, 18, 25, 29 – 32, 35, 36, 41, 43, 45, 48, 56
	Between-subjects	1.16	[0.85, 1.47]	283.80***	91.64	37	2488	1 – 3, 5, 7, 9, 11 – 16, 19 – 24, 26 – 28, 33, 34, 37 – 40, 44, 46, 47, 49 – 55
Publication	$Q_{between}$ (*df* = 1)			5.49*				
	Published	0.89	[0.66, 1.12]	314.66***	90.91	46	3165	1 – 16, 19, 21, 23, 25, 26, 28 - 35, 37 – 40, 42, 43, 46 – 56
	Not published	1.71	[0.90, 2.52]	79.09***	90.59	10	264	17, 18, 20, 22, 24, 27, 36, 41, 44, 45

Note: CI = confidence interval; ID = identification; *df* = degrees of freedom; CBCA = criteria-based content analysis; RM = reality monitoring; DF = discriminant function; * $p < .05$; ** $p < .01$; *** $p < .001$

Figure 1. Forest plot of the corrected effect sizes for each study. Square location indicates the magnitude and square size the weighting of the single effect size. The length of the lines corresponds to the 95% CI.

Figure 1

Authors (year)	*d*	*g*	*n*	Study ID
Akehurst, Bull, Vrij, & Köhnken (2004)	0.16	0.15	32	46
Akehurst, Köhnken, & Höfer (2001)	1.04	1.02	66	13
Akehurst, Manton, & Quandte (2011)	0.96	0.94	31	14
Berger (2005)	1.52	1.5	25	18
Blandon-Gitlin, Pezdek, Rogers, & Brodie (2005)	0.27	0.27	94	7
Bogaard, Meijer, & Vrij (2014)	0.53	0.53	128	56
Bond & Lee (2005)	1.11	1.11	64	29
Colwell, Hiscock-Anisman, Memon, Taylor, & Prewett (2007)	1.69	1.66	38	11
Craig, Scheibe, Raskin, Kircher, & Dodd (1999)	0.6	0.59	48	21
Dolezych (2006)	0.11	0.11	27	17
Fiegler (2009)	1.99	1.96	50	22
Gödert, Gamer, Rill, & Vossel (2005)	0.59	0.58	68	39
Jang, Kim, Cho, & Lee (2013)	3.76	3.66	30	51
Janka (2003)	3.76	3.65	14	36
Köhnken, Schimossek, Aschermann, & Höfer (1995)	1.23	1.21	59	5
Krahé & Kundrotas (1992)	1.95	1.89	30	3
Lamb, Sternberg, Esplin, Hershkowitz, Orbach, & Hovav (1997)	0.64	0.64	89	16
Landry & Brigham (1992)	0.3	0.28	12	12
Leal, Vrij, Warmelink, Vernham, & Fisher (2015)	0.35	0.34	40	53
Leal, Vrij, Warmelink, Vernham, & Fisher (2015)	0.65	0.65	83	54
Lüdke (2008)	1.9	1.86	36	24
Memon, Fraser, Colwell, Odinot, & Mastroberardino (2009)	0.93	0.92	60	2
Merckelbach (2004)	-0.25	-0.25	38	6
Nahari, Vrij, & Fisher (2012)	0.68	0.67	41	38
Naumann (2005)	0.85	0.77	10	44
Niehaus (2000)	0.59	0.59	80	49
Parker & Brown (2000)	3.63	3.53	28	33
Porter & Yuille (1996)	0.94	0.93	60	23
Porter, Yuille, & Lehman (1999)	-0.15	-0.15	75	31
Roma, Martini, Sabatello, Tatarelli, & Ferracuti (2011)	2.69	2.67	109	19
Ruby & Brigham (1998)	0.91	0.88	12	4
Rutta (2001)	3.22	3.14	16	45
Santtila, Roppola, Runtti, & Niemi (2000)	0.85	0.85	68	32
Scheinberger (1993)	3.76	3.66	30	20
Schelleman-Offermans & Merckelbach (2010)	0.43	0.43	60	10
Short & Bodner (2011)	1.11	1.1	55	8
Sporer (1997)	0.78	0.77	40	43
Sporer & Küpper (1995)	0.99	0.99	100	25
Sporer & Sharman (2006)	0.15	0.15	240	26
Steck, Hermanutz, Lafrenz, Schwind, Hettler, Maier, & Geiger (2010)	0.74	0.73	40	27
Steller, Wellershaus, & Wolf (1992)	0.29	0.29	88	30
Strömwall, Bengtsson, Leander, & Granhag (2004)	0.21	0.21	87	37
Strömwall & Granhag (2005)	1.65	1.62	44	55
Suckle-Nelson, Colwell, Hiscock-Anisman, Florence, Youschak, & Duarte (2010)	1.75	1.74	88	34
Tye, Amato, Honts, Devitt, & Peters (1999)	2.49	2.42	28	15
Vrij, Akehurst, Soukara, & Bull (2004a)	0.34	0.33	180	42
Vrij, Akehurst, Soukara, & Bull (2004b)	0.37	0.37	196	40
Vrij, Edward, & Bull (2001)	0.74	0.73	73	35
Vrij, Edward, Roberts, & Bull (2000)	1.15	1.14	73	52
Vrij, Kneller, & Mann (2000)	0.05	0.04	30	50
Vrij & Mann (2006)	0.47	0.46	80	1
Vrij, Mann, Kristen, & Fisher (2007)	0.36	0.35	120	47
Willén & Strömwall (2012)	-0.08	-0.08	30	48
Wolf & Steller (1997)	3.76	3.66	30	28
Wrege (2004)	0.29	0.28	16	41
Zaparniuk, Yuille, & Taylor (1995)	1.38	1.35	40	9

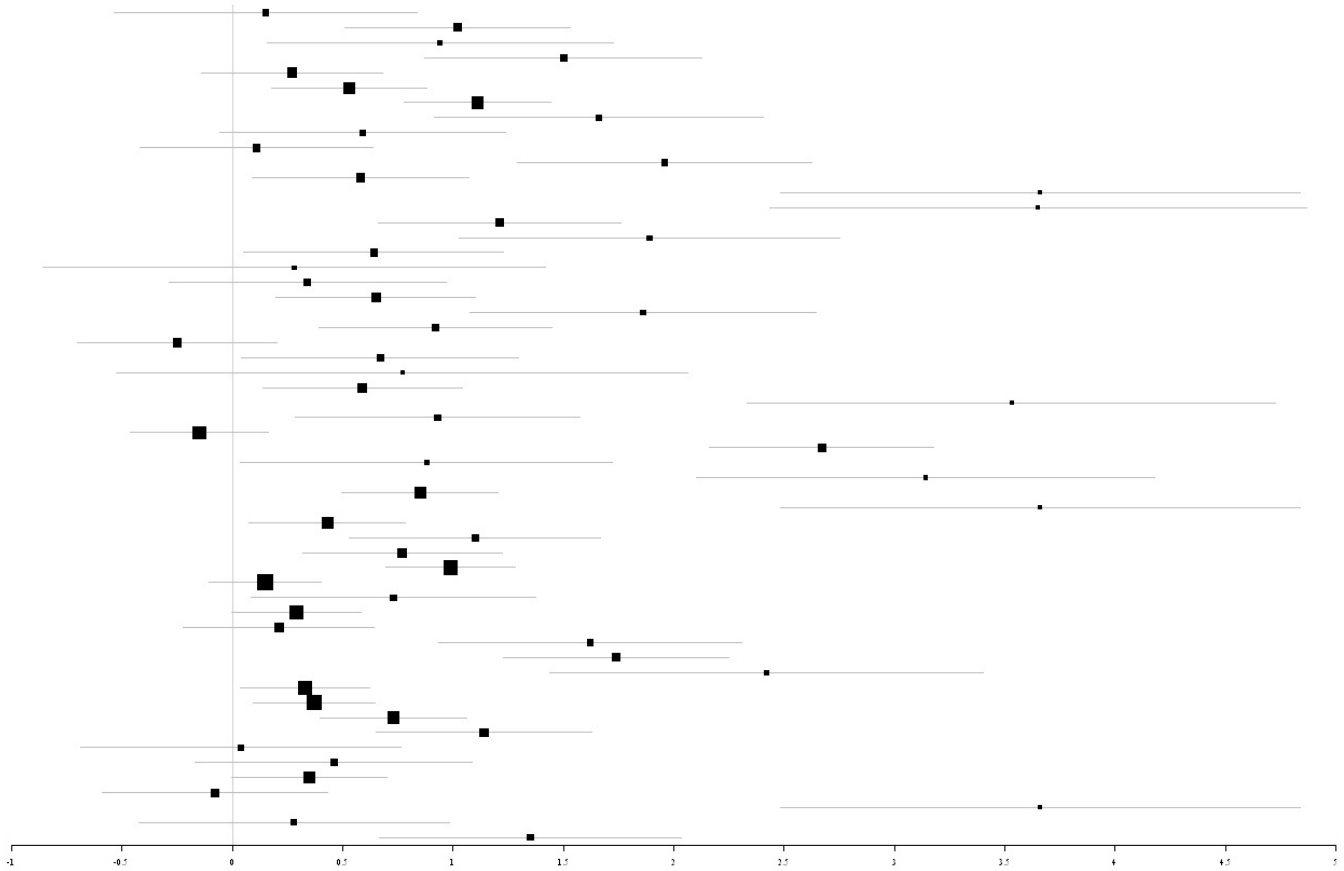


Figure 2. Funnel plot of the R_0 algorithm on the augmented data set of $k = 70$ studies for the total set of studies (white points indicate missing studies; black points indicate included studies).

Figure 2

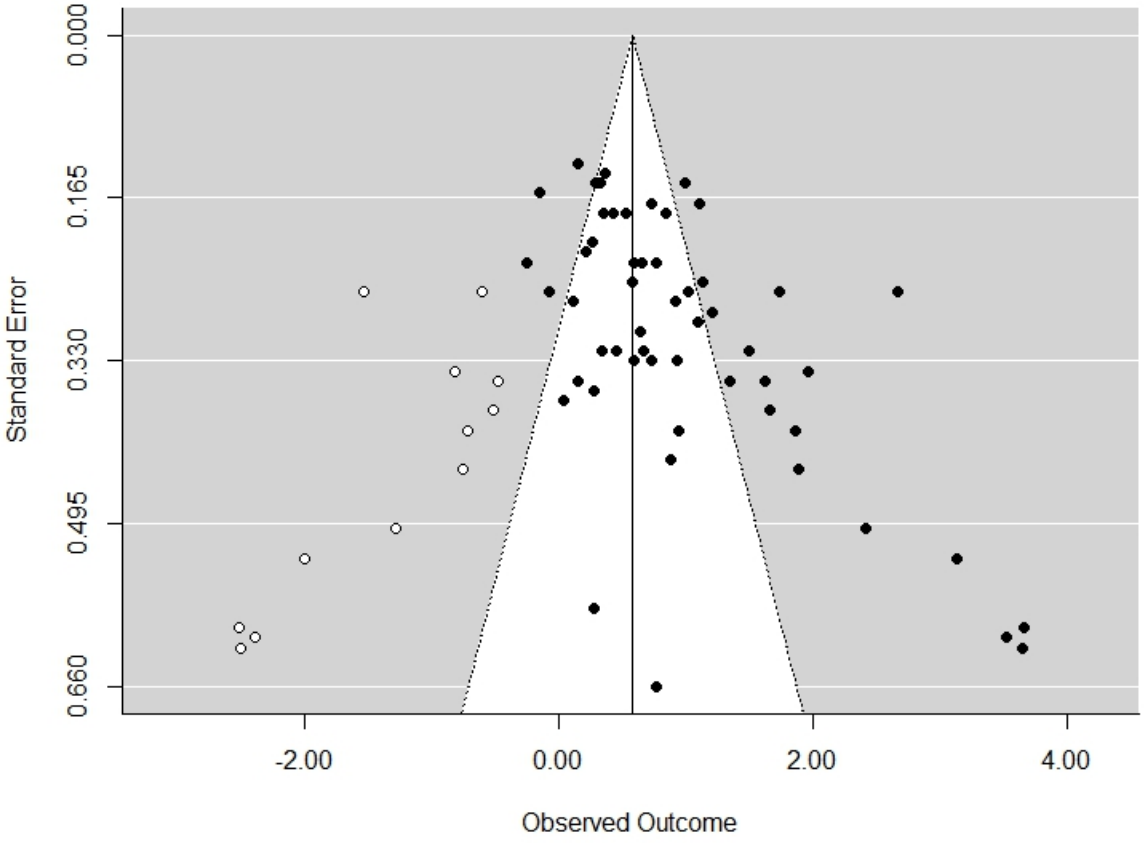
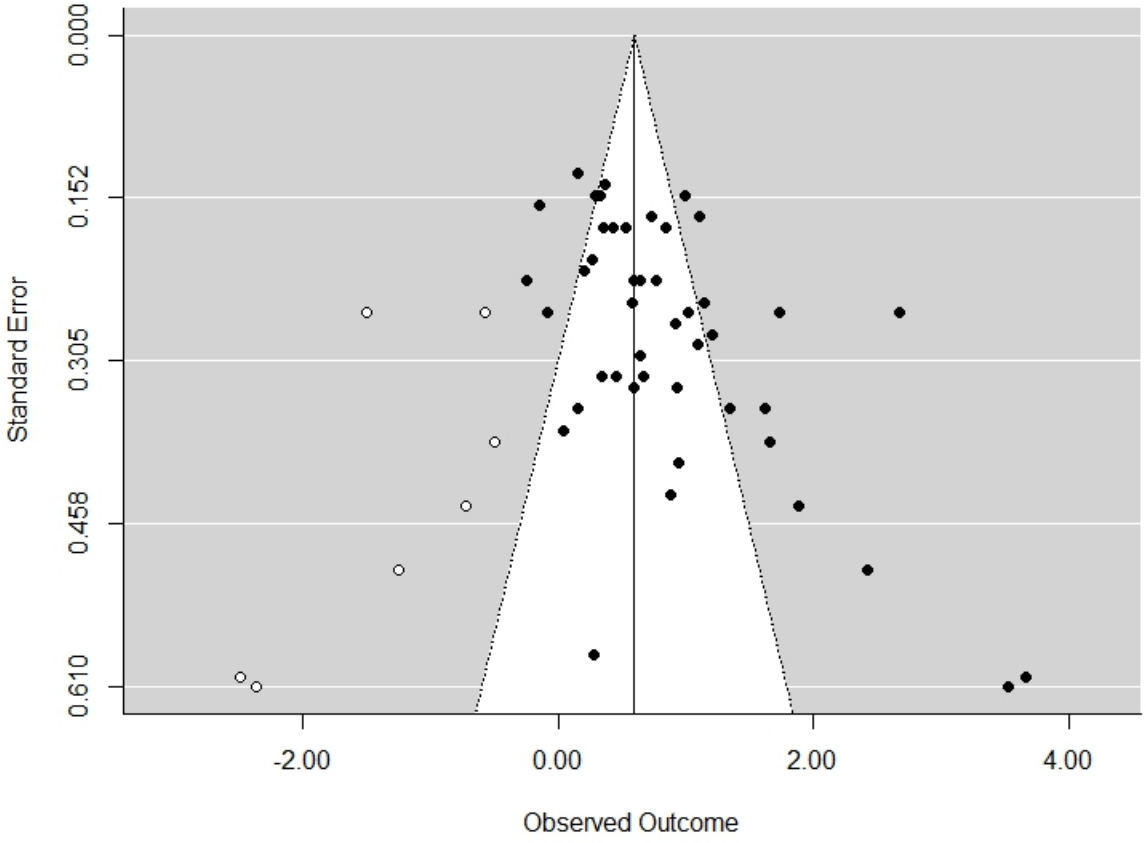

Standard Error
0.000
0.165
0.330
0.495
0.660
-2.00
0.00
2.00
4.00
Observed Outcome

Figure 3. Funnel plot of the R_0 algorithm on the augmented data set of $k = 54$ studies for the subset of published studies (white points indicate missing studies; black points indicate included studies).

Figure 3


Standard Error
0.000
0.152
0.305
0.458
0.610
-2.00
0.00
2.00
4.00
Observed Outcome