Meta-Analyses on the Validity of Verbal Tools for Credibility Assessment

Inaugural-Dissertation zur Erlangung der Doktorwürde

der

Philosophischen Fakultät

der

Rheinischen Friedrich-Wilhelms-Universität

zu Bonn

vorgelegt von

Verena Oberlader

aus

München

Bonn, 2019

Gedruckt mit der Genehmigung der Philosophischen Fakultät

der Rheinischen Friedrich-Wilhelms-Universität Bonn

Zusammensetzung der Prüfungskommission:

Prof. Dr. Ulrich Ettinger, Institut für Psychologie, Universität Bonn

(VORSITZENDER)

Prof. Dr. Rainer Banse, Institut für Psychologie, Universität Bonn

(BETREUER UND GUTACHTER)

Prof. Dr. Renate Volbert, Psychologische Hochschule Berlin

(GUTACHTERIN)

PD Ina Grau, Institut für Psychologie, Universität Bonn

(WEITERES PRÜFUNGSBERECHTIGTES MITGLIED)

Tag der mündlichen Prüfung: 13.09.2019

DANKSAGUNG

Allen voran bedanke ich mich bei dir, Rainer. Vor mittlerweile acht Jahren hast du mir die Chance eröffnet, die rechtspsychologische Forschung kennen und schätzen zu lernen und in diesem Fachbereich Fuß zu fassen. Zunächst als studentische Hilfskraft, später als Doktorandin hast du mir großes Vertrauen entgegengebracht, mich von Beginn an verantwortlich mitwirken lassen, mich gefordert und gefördert. Deine Offenheit und Leidenschaft für neuartige Fragestellungen waren dabei ein großer Antrieb.

Mein Dank gilt auch dir, Alex. Als Nachfolgerin auf deinem Bürostuhl hast du mir zunächst aus Luxemburg, später aus Hamburg und zuletzt aus Mainz immer mit Rat und Tat zur Seite gestanden. Deine Begeisterung für die Rechtspsychologie ist ansteckend und motivierend.

Ich bedanke mich auch bei meinen (ehemaligen) Mitstreiterinnen und Kolleginnen, Kathrin, Jelena, Lisa, Anja, Laura, Christine, Charis, Michaela und Ina. Auch wenn wir nicht immer Tür an Tür saßen, war unser (natürlich stets rein fachlicher) Austausch via Skype eine verlässliche Größe. Zum Forschungsalltag zwischen DKS, KKR und Innenministerium waren unsere Ausflüge auf diverse Konferenzen eine willkommene Abwechslung, die wir voll auszuschöpfen wussten und uns keinen Gesellschaftsabend oder Saunaausflug haben entgehen lassen.

Natürlich bedanke ich mich auch bei dir, Carolin. Du wusstest nicht nur auf jede organisatorische Frage eine verlässliche Antwort, du hast dir auch immer Zeit für einen netten Plausch zwischen Tür und Angel genommen.

Ich bedanke mich auch bei dir, liebe Judith. Als ehemaliges SRP-Mitglied habe ich dich beim Codieren von Studienergebnissen im SHK-Büro kennengelernt. Schnell haben wir festgestellt, dass wir noch anderes mit unserer gemeinsamen Zeit anzufangen wissen. Ich freue mich, dich seither als treue Lebensbegleiterin gewonnen zu haben.

Zum Schluss bedanke ich mich bei meiner Familie und meinen Freunden, die mich, kurz gesagt, in jeder Lebenslage glücklich machen. Allen voran danke ich dir, Mathias. Du sorgst für mich wie kein Zweiter, deine Worte bringen mich immer weiter, deine Unterstützung ist grenzenlos. Gerade in den letzten Monaten, seit unsere Kleine bei uns ist, hast du mich an allen Ecken und Enden entlastet, um mir freie Zeit zum Arbeiten einzuräumen. Ihr zwei seid mein Herz.

ABSTRACT

Since ancient times, approaches to distinguish between true and deceptive statements have been of particular importance in the context of court decisions. However, the applicability of most psychophysiological or behavioral measures of deception is critically discussed. Verbal tools for credibility assessment, nonetheless, are widely used. They rest on the assumption that the quality of statements that are experience-based differs from the quality of fabricated accounts. In order to test the validity of two prominent procedures, Criteria-Based Content Analysis (CBCA) and Reality Monitoring (RM), a random-effects meta-analysis (REMA) was conducted on 52 English- and German-language studies in Meta-Analysis 1. The REMA revealed a large point estimate with moderate to large effect sizes in the confidence interval. This finding applied for both CBCA and RM, despite the fact that (1) there was a high level of heterogeneity between studies that could not be resolved by moderator analyses and, (2) it cannot be ruled out that effect size estimates are biased and thus verbal tools for credibility assessment only work to a smaller extent. However, a recent simulation study cast doubt on these findings: It showed that the meta-analytic methods used in Meta-Analysis 1 lead to false-positive rates of up to 100% if data sets are biased. To test the robustness of previous findings, a reanalysis with different bias-correcting meta-analytic methods was conducted on an updated set of 71 studies in Meta-Analysis 2. The overall effect size estimates ranged from a null effect to conventionally large effect sizes. Taking into account specific strengths and limitations of each meta-analytic method, results indicated that CBCA and RM distinguish between experience-based and fabricated statements with moderate to large effect sizes. In contrast, the Scientific Content Analysis (SCAN) – a third verbal tool for credibility assessment that was also tested in the updated data set of Meta-Analysis 2 – did not discriminate between truth and lies and should thus not be used in practice.

"AT LEAST, LYING IS THINKING THE TRUTH."

Oliver Hassenkamp (translated from German)

INTRODUCTION

The assessment of the credibility of statements in criminal proceedings is an important and demanding task of any court and goes back a long way in history. The Ur-Nammu, the oldest code of law known today, determined in 2100-2050 B. C. that a "river ordeal" should decide whether a man's accusation that his wife had committed fornication was true or not (Finkelstein, 1968/69). Although today this task is no longer carried out by a "river ordeal", but by a judge, the question has remained the same: Is a statement based on real experience or is it invented? Credibility assessment of statements is particularly relevant when there is no other evidence (e.g., camera recordings, DNA evidence) at hand, as is often the case with child sexual abuse. In such statement-against-statement constellations the judge must decide who is telling the truth: the alleged victim or the alleged perpetrator. However, one criticism is that judges are not adequately trained in credibility assessment and hence often use invalid, everyday theories (e.g., Jahn, 2001). The application of lay theoretical approaches is alarming given the fact when using lay approaches humans distinguish between truth and lie hardly better than chance (e.g., Bond & DePaulo, 2006; Ekman & O'Sullivan, 1991; Hartwig & Bond, 2011). It is therefore paramount to establish objective, reliable, and valid procedures for the assessment of statement credibility in court and other contexts (e.g., border security, customs control).

There are various approaches to investigate differences between true and deceptive responding. These approaches use either psychophysiological or behavioral measures of deception, which can be further specified. Köhnken (1990) distinguished, for example, non-verbal, para-verbal, and verbal content cues. Others have used reaction time as a behavioral indicator of lying or telling the truth (e.g., Sartori, Agosta, Zogmaister, Ferrara, & Castiello, 2008). Regardless of their nature, measures of deception have been extensively studied for over a hundred years (for a historical overview on lie detection research see for example Lykken, 1998). The present meta-analyses are the first to synthesize the vast research on different tools for verbal credibility assessment.

Meta-Analysis 1 and 2 are presented below in chronological order. Meta-analysis 1 summarised the state of research on verbal tools of credibility assessment up to the year 2015 and, in comparison to previous research syntheses up to that point, enabled a comprehensive assessment of verbal tool's validity that is highly relevant for legal psychologists working in science and practice. However, a simulation study by Carter et al. (2019), which investigated the performance of different meta-analytical methods, cast doubt on the results of Meta-Analysis 1. These findings motivated me to reanalyse previous and newly added data using different meta-analytic methods. Meta-Analysis 2 is thus an update that synthesized research up to the year 2018 and was designed taking into account the latest state of research on meta-analytical methods.

MEASURES OF DECEPTION

Psychophysiological measures capture parameters of the autonomous (e.g., electrodermal activity, heart rate) or central nervous system (e.g., event-related potential) that are expected to differ depending on whether a person lies or responds truthfully. A recent meta-analysis by Leue and Beauducel (2019) demonstrated that the parietal P3 amplitude of event-related potential reliably reflected (a) the recognition of salient information that had to be concealed (larger P3 following more salient information compared to true responding to unknown stimuli; $\delta = 0.95$) and (b) increased mental effort of concealing knowledge compared to true responding to known stimuli (smaller P3 following more demanding information; $\delta = -0.52$). Suchotzki, Verschuere, Bockstaele, Ben-Shakhar, and Crombez (2017) investigated various behavioral measures of deception that used reaction time. A meta-analysis of 114 studies using the *autobiographical Implicit Association Test* (Sartori et al., 2008), the *Concealed Information Test* (Lykken, 1959), the *Sheffield Lie Test* (Spence et al., 2001), or the *Differentiation of Deception Paradigm* (Furedy, Davis, & Gurevich, 1988) revealed a large effect for standardized reaction time differences between true and deceptive responses ($d = 1.049$). Although these results sound promising, it must be noted that there is no omnipotent measure of deception, as the famous Pinnochio nose suggests (Volbert & Banse, 2014), neither a psychophysiological (e.g., Steller, 2008) nor a behavioral one (see for example results of the meta-analysis on 158 behavioral cues of deception by DePaulo et al., 2003). Whether a measure is a *measure of deception* depends on the paradigm applied, or vice versa, it is the paradigm that determines the psychological processes that a measure reflects (Meijer, Verschuere, Gamer, Merckelbach, & Ben-Shakar, 2016). For example, time-delayed responses to given words could signal the concealment of crime-related knowledge when combined with a Concealed Information Test (Lykken, 1959) or emotional valence when combined with an *emotional Stroop task* (Ben-Haim et al., 2016). Hence, research on deception detection seeks paradigms that capture psychological processes involved in lying.

There are plenty of theories about what psychological processes are related to deception (e.g., Buller & Burgoon, 1996; DePaulo et al., 2003; Walczyk, Harris, Duck, & Mulay, 2014; Zuckerman et al., 1981). Zuckerman et al. (1981), for example, supposed four factors: (1) an increased arousal; (2) negative emotions like guilt, shame, and fear; (3) cognitive aspects; and (4) attempts at behavioral control. However, the first two factors in particular are often critically discussed: It is repeatedly pointed out that although lying may be associated with an increased arousal or negative emotions, increased arousal or negative emotions does not necessarily indicate lying (so-called *fallacy of reverse inference*, Meijer et al., 2016). For example, a person who is telling the truth may have a heightened arousal level if he or she is suspected of having committed a

crime, or may feel ashamed if he or she reports a sexual offence. These examples illustrate that the psychological processes "arousal" and "negative emotions" are not uniquely associated with deception. Although individual studies have shown that liars have an increased arousal that is unaffected by cognitive load (Vincent & Furedy, 1992) or actually express more negative emotional words than truth tellers (e.g., Hauch, Blandón-Gitlin, Masip, & Sporer, 2015), the validity of these results should at least be critically questioned as long as the relationship between the psychological processes and deception is not established.

COGNITIVE APPROACH OF DECEPTION DETECTION

In contrast, the cognitive approach of deception detection looks more promising. It is based on the notion that lying is typically cognitively more demanding, or requires more cognitive resources in terms of executive control (i.e., response inhibition, working memory updating, and shifting; Miyake et al., 2000), than telling the truth. In comparison to experience-based statements, lies cannot simply be recalled from memory, but must be constructed while inhibiting the truth. In addition, liars have to be careful not to get caught up in contradictions regarding their own statements and/or the knowledge of the person being lied to. At the same time, liars have to make sure they appear credible and thus constantly monitor their own behavior as well as the reaction of the target person to see if their deception is believed or has to be adjusted. To accomplish these tasks, the truth needs to be activated in working memory and the mental sets of truth and deception must be constantly calibrated. However, these demands of lying do not apply under all conditions. There are situations where telling the truth imposes a high amount of cognitive load, too. If, for example, an experienced event has not been retrieved for a long time and must be actively searched in memory (for further examples see Walczyk et al., 2014), then truth telling can require extensive cognitive effort. Conversely, there are situations in which lying does not require much cognitive effort. If, for example, a target asks a simple question and cannot verify the answer, then telling a lie is not necessarily cognitively taxing.

Within their *Activation-Decision-Construction-Action Theory* (ADCAT), Walczyk and colleagues (2014) specified under which conditions "serious lying" (i.e., lying in situations where much is at stake) actually imposes cognitive load. The authors structure the process of lying in four components: The activation component (1) refers to aspects of the social environment that lead respondents to understand that true information is requested and then, if possible, retrieved from or encoded in working memory. The decision component (2) includes the social context that leads respondents to deceive in a certain way or reminds them of their decision to lie. The construction component (3) describes the manipulation of information. The action component (4) represents execution of the lie. Whether lying demands a high amount of cognitive resources depends,

roughly speaking, on the social context, on the evaluation of the consequences of a true or deceptive response, on the type of lie, on whether and how well a person is prepared to lie, and on the familiarity and complexity of the situation. The authors specified further boundary conditions under which each component can impose additional cognitive load and integrate empirical evidence of deception detection research in their theoretical framework (for detailed information see Walczyk et al., 2014). Although the ADCAT suggests that several boundary conditions have to be considered, empirical evidence stresses the assumption that lying is associated with a higher cognitive effort than telling the truth. For example, participants reported that they experienced lying as cognitively more demanding than telling the truth (e.g., Caso, Gnisci, Vrij, & Mann, 2005). Moreover, brain-imaging studies showed that lying is associated with activation in brain regions that are also activated in other cognitively demanding tasks (for example, the prefrontal cortex; Abe, 2009). In addition, results of a comprehensive meta-analysis on reaction time-based measures of deception confirmed the hypothesis that "lying takes time" as it is cognitively challenging (Suchotzki et al., 2017, p. 34).

The fallacy of reverse inference also applies to the cognitive approach of deception detection: "That is, even if deceptive responses are differentially associated with brain activation in areas associated with cognitive control, we cannot conclude that differential activation in these areas necessarily implies that the subject is deceptive (i.e., responses to questions may be associated with enhanced cognitive control even when they are truthful). Similarly, the fallacy of reverse inference applies to the absence of differential activation: a lack of activation in areas associated with inhibition does not necessarily imply that the subject is responding truthfully" (Meijer et al., 2016, p. 598). It is therefore a great challenge for research on deception detection to develop paradigms that tap into psychological processes associated with lying and to capture these processes through valid measures.

VERBAL CREDIBILITY ASSESSMENT

Verbal tools for credibility assessment are linked to the cognitive approach of deception detection. Most of these procedures serve to substantiate the truth and not to uncover lies, which is why they are labeled as tools for *credibility assessment*. In principle, verbal tools for credibility assessment are based on the notion that experience-based statements are of higher content quality than fabricated statements and that these differences are reflected in verbal cues. Verbal cues have a long tradition in the history of deception detection. As early as 900 B. C. a papyrus Veda stated that a poisoner is recognized, among other characteristics, by the fact that "he speaks nonsense" (Trovillo, 1939, p. 849). Since this rather unspecific description, research on verbal cues – either measured in isolation (i.e., verbal uncertainty, verbal immediacy; DePaulo et al.,

2003) or as part of a procedure – has developed. The two most prominent procedures, at least in science, are the *Criteria-Based Content Analysis* (CBCA; Steller & Köhnken, 1989) and *Reality Monitoring* (RM; Johnson & Raye, 1981).

CRITERIA-BASED CONTENT ANALYSIS

The CBCA is based on the assumption that experience-based statements are of higher content quality than fabricated statements, meaning they are richer in detail and show more elaborate links to external events (so-called *Undeutsch hypothesis*). Since a 1955 ruling of the German Supreme Court mandated that psychological experts be consulted in cases of child sexual abuse, a large amount of case material was available that led to this observation. In 1967, Udo Undeutsch was the first to list reality criteria to capture differences in quality between experience-based and fabricated statements. In the following years, Swedish and German experts compiled further lists of reality criteria (Arntzen, 1970; Dettenborn, Froehlich, & Szewczyk, 1984; Szewczyk, 1973; Trankell, 1971). On this basis, Steller and Köhnken formalized in 1989, more than 30 years after the ruling of the German Supreme Court, a set of 19 content[1] criteria for statement analysis, organized in five categories (see Table 1).

The first category comprises general characteristics of a statement as a whole, including *logical consistency*, *unstructured production*, and *quantity of details*. The second group refers to specific contents, such as *descriptions of interactions* or *reproduction of conversations*. The third category includes peculiarities of content, such as *unusual details* or *accurately reported details that were not comprehended*. The fourth group addresses motivation-related content, which is concerned with identifying statement details that a witness who makes a false statement and does not take his/her credibility for granted would not embed, e.g., *pardoning the perpetrator*. Finally, the last category includes an offence-specific element that is *detailed characteristic of the offence.*

The CBCA criteria are rated with reference to a verbatim transcript of the statement on an alleged event, either as absent/present or with regard to their strength (there are different scorings, for example 0 = absent, 1 = present, 2 = strongly present). The presence of criteria is interpreted as indicator of truth. Absences of criteria, however, do not indicate deception, because there are other reasons why criteria may not be present (e.g., lack of motivation to make a statement, simple and short event). Although Arntzen (1970) reported a rule of thumb according to which "at least three reality criteria must be given […] in order to classify a

[1] Steller and Köhnken (1989) criticised the undifferentiated use of the term "reality criteria", which refers to several aspects of credibility assessment and not only to content analysis. Therefore, the authors preferred the term "content criteria".

testimony as credible" (p. 46), Steller and Köhnken (1989) stressed that there are "no formalized decision rules […] for determining cutoff scores to differentiate between true and deceptive statements" (p. 231). Steller and Köhnken (1989) also pointed out that the occurrence of criteria does not depend exclusively on the truth status of a statement, but additionally on personal and situational factors. That is, CBCA criteria must not be misunderstood as a simple checklist to reveal experience-based or fabricated statements. Rather, the CBCA is one part of a whole diagnostic process, the *Statement Validity Assessment* (SVA) that examines various alternative hypotheses for the development a statement.

Table 1

CBCA Criteria (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. Detailed characteristic of the offense

In the first step of the SVA, the case file is examined in order to derive alternative hypotheses. The second step is a semi-structured interview on the event in question. It is important to obtain a comprehensive statement in a free narrative style that is not influenced by the interviewer. Using a verbatim transcript of the statement, the CBCA is performed in the third step. Finally, to answer the question of whether a person could or would have made a particular statement if it were not experience-based, the quality of the statement is considered in context of personal and

situational factors that have been summarized in slightly different versions of the so-called *Validity Checklist* (e.g., verbals skills, event characteristics, motives to report, interview style). Accordingly, for example, a high quality of a statement might be attributed to an interviewee's being extremely eloquent or verbally skilled. Conversely, a low-quality statement might given if the event in question was so simple and short that many criteria could just not occur.

Furthermore, Steller and Köhnken (1989) have pointed out that individual criteria are of different diagnostic value, which is another reason why the CBCA should not be used as a simple checklist. For example, *reporting a misunderstood detail* is more meaningful than a *description of contextual embedding*. In a modification of CBCA criteria 25 years later, Volbert and Steller (2014) took up this idea. On the basis of Niehaus (2008), the authors sorted CBCA criteria by new aspects that refer to different underlying processes: characteristics of episodic autobiographical memory, script-deviant details, and criteria of strategic self-presentation. In addition, they listed characteristics for the statement as a whole.

Characteristics of episodic autobiographical memory include spatiotemporal (e.g., *contextual embedding*, *spatial information*, *temporal information*) and self-related criteria (e.g., *emotions and feelings*, *own thoughts*, *sensory impressions*). This category refers to the fact that people who tell the truth can just fall back on episodic representations, whereas liars must use cognitive scripts and deliberate effort to construct a statement. Therefore, experience-based statements should generally be more detailed and elaborate than fabricated statements. However, cognitive scripts of liars may also include characteristics of episodic autobiographical memory (Volbert & Steller, 2014). In a study on the strategic meaning of individual CBCA criteria, Maier, Niehaus, Wachholz, and Volbert (2018) showed that deceivers would actually try to embed memory-related information in fabricated statements. The authors conclude that both truth tellers and liars are in principle motivated to use such criteria, but that embedding such details is more cognitively demanding if they are not based on an actual experience. For script-deviant criteria (e.g., *unusual details*, *unexpected complications during the incident*) the opposite is true. Volbert and Steller (2014) do not assume that liars will come up with the idea of incorporating these criteria into their invented statements, as they do not fit into the common script of a true statement. Study results of Maier et al. (2018) supported this assumption: Deceivers stated to avoid the use of script-deviant criteria. The criteria Volbert and Steller (2014) classified as strategic self-presentation should reflect the effort to present oneself as credible. Accordingly, liars should avoid, for instance, *spontaneous corrections*, *doubts about one's own testimony*, or *the expression of uncertainty*. Maier et al. (2018) confirmed this assumption for seven out of nine criteria of strategic self-presentation. Taken together, current research indicates that CBCA criteria differ in their strategic meaning, i.e., liars

are either motivated to include or avoid them. Maier and colleagues (2018) pointed out, however, that the mere motivation to use a criterion does not necessarily mean that it is actually incorporated into a fabricated statement. Whether or not a criterion is included in a statement also depends on how cognitively demanding it is. The study results illustrated that the diagnostic value of the CBCA criteria needs to be examined more closely.

The CBCA is admissible evidence in some courts of North America (Ruby & Brigham, 1997) and in several West European countries such as Austria, Germany, Sweden, Switzerland, and the Netherlands (Köhnken, 2004). Even though the CBCA has been developed on the basis of statements in alleged cases of child sexual abuse, many authors assume that it can also be used to assess other issues and adult testimonies (e.g., Köhnken, 2004; Köhnken, Schimossek, Aschermann, & Höfer, 1995; Porter & Yuille, 1996).

REALITY-MONITORING

While the atheoretical character of the CBCA reflects its historical development, namely its derivation from practice, the RM approach has a theoretical basis. Johnson and Raye (1981) described reality monitoring as a cognitive process by which a person distinguishes between experience-based and imagined memories. Based on the idea that memories of experienced events have stronger external links than memories of things that have only been imagined, the authors described criteria to differentiate the two memory types. Accordingly, externally generated memories that originate in perceptual experience should be characterized by contextual, sensory, and semantic information, whereas internally generated memories that originate from thought or in the imagination should be characterized by references to cognitive operations. Johnson, Foley, Suengas, and Raye (1988) tested this hypothesis: In Study 1, participants were expected to remember either an autobiographical or an imagined event (e.g., recent fantasy, recent dream, unfulfilled intention) and rated the memories with respect to 39 reality criteria using the *Memory Characteristics Questionnaire*. Results showed that memories of autobiographical events were characterized by more sensory (e.g., visual detail, sound, smell, taste), contextual (e.g., location, time, year, season), and semantic information (e.g., events before, events after), whereas memories of imagined events contained more references to thought processes. These study results provided empirical evidence for reality monitoring of one's own memories.

In addition to the research group of Alonso-Quecuty and Hernández-Fernaud (e.g., Alonso-Quecuty, 1992; Alonso-Quecuty, Hernández-Fernaud, & Campos, 1997; Hernández-Fernaud & Alonso-Quecuty, 1997), Sporer and Küpper (1995) explored whether the RM approach can also be used to assess the quality of someone else's memory and to distinguish between truth and lies.

As true statements are based on real experiences and fabricated statements are internally generated, the logic of RM should also apply for detecting deceit. At this point it should be noted that lies could of course also contain experience-based elements and could therefore be only partially deceptive. Nevertheless, Sporer and Küpper (1995) developed a procedure to rate the RM criteria as verbal cues of credibility: the *Judgement of Memory Characteristic Questionnaire*, wherein they summarized a total of 35 items on the basis of factor-analytical results into eight RM scales (see Table 2). The scales include seven cues for experience-based and one cue for fabricated statements, namely *cognitive operations* during the event, which must be recoded to calculate the total score. Study results indicated that individual RM criteria are not only suitable for distinguishing between one's own internally or externally generated memories, but are also effective in distinguishing experience-based and fabricated statements (for details see Sporer & Küpper, 1995).

Table 2

RM Criteria (Sporer & Küpper, 1995)

1. Clarity of memory
2. Sensory experiences
3. Spatial information
4. Time information
5. Emotions and feelings
6. Reconstructability of the story
7. Realism of the story
8. Cognitive operations

There is no standardized set of RM criteria and so researchers use different versions (i.e., different criteria, different operationalizations of criteria). Moreover, to the best of our knowledge, in contrast to the CBCA, the RM approach is not used in practice (Vrij, 2015). Vrij (2008) also doubted that RM is capable of assessing the truth status of children's statements or of statements that relate to events far back in time – two cases that are highly relevant in practice. Assessment of children's statements would be a challenge because children have rich imaginations and hence imagined things resemble real memories. Assessment of memories from events far prior would be a challenge because people use cognitive operations to facilitate the retrieval of external memories of events far back in time, so the memories resemble internally generated memories.

EMPIRICAL EVIDENCE OF VERBAL TOOLS FOR CREDIBILITY ASSESSMENT

STUDY DESIGNS

Basically, there are two approaches to investigate the validity of verbal tools for credibility assessment: Field studies using real-life statements (e.g., Roma, San Martini, Sabatello, Tatarelle, & Ferracuti, 2011) and laboratory studies using statements that are experimentally generated (e.g., Vrij, Akehurst, Soukara, & Bull, 2004a). Obviously, field studies on verbal tools for credibility assessment benefit from high ecological validity. The downside of real-life statements, however, is the difficulty to establish ground truth. Cases in which the credibility of a statement is at stake are generally characterized by a lack of objective evidence that can be used as validation criterion. Other criteria must therefore be used to determine the truth status of statements. For this purpose, studies use more or less hard criteria. Confessions to the police, for example, as used by Krahé and Kundrotas (1992), are certainly less objective than video recordings of an event filmed by the offender, as used by Akehurst, Manton, and Quandte (2011), but also harder to get. In addition, it should be noted that validation criteria might depend on the quality of statements. A perpetrator might only confess if the evidence against him or her is strong – if, for example, a victim provides a high-quality statement. Conversely, an innocent person could also make a false confession under the pressure of a false testimony. Due to the possible dependence of the statement quality and the validation criterion, Vrij (2005) assumed that the empirical evidence for field studies is inflated.

On the other hand, the experimental control of ground truth in laboratory studies is at the cost of decreased ecological validity. Laboratory studies investigate experience-based and fabricated statements that are produced under conditions that are more or less comparable to the field of practical application. For example, participants experience an event that they are later instructed to truthfully report, or receive a description of an event that they should claim to have experienced (e.g., Vrij, Akehurst, Soukara, & Bull, 2004b). In other studies, participants have been tasked with reporting autobiographical experiences that were either actually experienced or invented (e.g., Santtila, Roppola, Runtti, & Niemi, 2000). In these studies, ground truth is of course not guaranteed, since these statements cannot be verified. Santtila et al. (2000) addressed this problem by asking parents of study participants to validate the information.

EMPIRICAL EVIDENCE FOR CBCA

Thus far, the validity of the CBCA has mainly been investigated in the laboratory and only rarly in the field. Hence, empirical evidence is largely based on experimentally produced statements and lacks reference to contexts of practical application. In addition, the validity of external

validation criteria used in field studies is often criticized as not being independent of CBCA ratings (Vrij, 2008). These limitations need to be taken into account when evaluating the empirical evidence.

Results of individual studies, which differ with regard to their study design (e.g., type of lie, age of participants, role of participants), are summarized in two reviews (Vrij, 2005, 2008) and two recent meta-analyses (Amado, Arce, & Fariña, 2015; Amado, Arce, Fariña, & Vilariño, 2016). Vrij (2005, 2008) included all research on CBCA published in English. In summary, studies largely provided support for CBCA, showing the CBCA criteria occurred more often in experience-based than in fabricated statements. In cases where the Undeutsch hypothesis was not confirmed, studies typically showed no stastically significant difference between experience-based and fabricated statements. Only in very few cases the opposite was true, with CBCA criteria appearing more frequently in fabricated than in experience-based statements. Most of these contrasting findings occurred in two studies by Landry and Brigham (1992) and Ruby and Brigham (1998), which had methodical limitations that could be responsible for these results: First, the raters received only 45 minutes of training and, second, the collected statements were very short. With short statements, there is less opportunity for criteria to occur. With regard to the total score, Vrij (2008) outlined that 80% of the CBCA studies showed higher scores for experience-based than for fabricated statements. In only one of 20 studies, fabricated statements had higher total scores than experience-based statements. This study also had the same methodical limitations as described above (Ruby & Brigham, 1998). In addition, for 19 studies the classification rate was calculated. On average, in 71%, experience-based and fabricated statements were correctly classified (Vrij, 2008). The results of a meta-analysis on the validity of the CBCA in children samples by Amado and colleagues (2015) confirmed these findings and revealed significant positive effect sizes for each criterion (δ = 0.17-1.40) and the total score (δ = 0.79). Within laboratory studies, 65% of experience-based statements met more criteria than fabricated statements; within field studies it was even 97%, whereby, however, the limited validity of the external criteria must be considered. In a second meta-analysis, Amado et al. (2016) investigated the validity of the CBCA in adult samples and, again, found significant positive effect sizes for almost all criteria (δ = 0.11-0.71; exceptions: *self-deprecation*, *pardoning the perpetrator*) and the total score (δ = 0.56). Moreover, Vrij (2005) reported good interrater reliabilities for most criteria (exceptions: *unstructured production*, *spontaneous corrections*) and excellent interrater reliabilities for the total score.

In summary, previous research indicates that the empirical evidence for the validity of the CBCA is consistent across different study designs and populations. Especially, when compared to non-

verbal indicators of deception (e.g., gaze, smile), which often show erratic patterns, i.e., occur both in true and deceptive responding (Vrij, 2008). Based on these findings, Vrij (2008) and Amado et al. (2015) concluded that the CBCA is largely compatible with the *Daubert* standards, the guidelines of the United States Supreme Court for admitting scientific evidence in court. Accordingly, on the basis of existing CBCA research, the following five questions can be confirmed: Is the scientific hypothesis testable? Has the hypothesis been tested? Is there a known error rate? Has the hypothesis and/or technique been subjected to peer review and publication? Has research supported the hypothesis and/or technique? However, the final requirement of the Daubert standards that the theory on which the hypothesis and/or technique is based has to be generally accepted in the appropriate scientific community is not fully met. Thus, the CBCA is repeatedly criticized as atheoretical: For example, Sporer (1997) stated that it is unclear which psychological processes are responsible for quality differences in statements and under which conditions they occur.

EMPIRICAL EVIDENCE FOR RM

As for the CBCA, most studies investigating RM were conducted in the laboratory and only a few in the field. In addition, external validation criteria of field studies were, again, often not independent of the assessment of statement quality. These limitations must be considered when interpreting empirical evidence for RM.

Masip, Sporer, Garrido, and Herrero (2005) and Vrij (2008) summarized study results on RM. Both reviews found a mixed pattern at criteria level: Although some criteria were more pronounced in experience-based than in fabricated statements, they were not diagnostic in all studies (for example, *clarity*, *visual details*, *sound details*, *temporal information*, *realism*, and *reconstructability*; Vrij, 2008). For other criteria, there was an erratic pattern such that they were sometimes more strongly expressed in experience-based and sometimes more strongly expressed in fabricated statements (for example, *sensory information*, *contextual embedding*, *affect*, and *cognitive operations*; Vrij, 2008). These inconsistencies could be explained by different operationalizations of the criteria or different study procedures. For example, Vrij (2008) reported that one study found contradicting results for the criterion *spatial information*. This finding could be traced back to a methodical peculiarity of the study: Bond and Lee (2005) used an automatic computerized coding system and no human rater to assess RM criteria. Of course, these erratic findings could also indicate that some criteria do not work. The pattern looked clearer at the level of the total score: Multivariate analyses showed that the total set of RM criteria significantly discriminated between experience-based and fabricated statements. Only one study found no differences (Vrij et al., 2004a). This study, however, examined statements from 5- to 6-year-old children, for whom

differentiation using the RM approach should be more difficult. In both reviews, classification rates were comparable to the CBCA, the average accuracy scores ranging from 65% to 85% (Masip et al., 2005) and from 63% to 82% (Vrij, 2008). Interrater reliabilities were also comparable to the CBCA and in a satisfactory range (see Sporer, 2004). However, RM coding is often described as being easier because there are fewer criteria and less room for interpretation. For example, raters experience fewer difficulties in distinguishing *spatial* and *temporal details* (RM criteria) than *unusual*, *superfluous*, and *unexpected details* (CBCA criteria).

Although study results are partly contradictory at criteria level, previous research indicates that the RM total score discriminates between experience-based and fabricated statements across different study designs and populations. As for the CBCA, the Daubert standards are met with one exception. Thus, again, the following five questions can be affirmed: Is the scientific hypothesis testable? Has the hypothesis been tested? Is there a known error rate? Has the hypothesis and/or technique been subjected to peer review and publication? Has research supported the hypothesis and/or technique? However, Nahari (2018) pointed out that RM neglects an important characteristic of lying: the intention to deceive. As memory source monitoring approach, RM refers to internally generated false memories, but not to "self-manipulated memories". Thus, RM should not be able to fully explain differences between experience-based and fabricated statements.

META-ANALYSIS 1

The aim of Meta-Analysis 1[2] was to estimate the effectiveness of verbal tools for credibility assessment in distinguishing between experience-based and fabricated statements on a meta-analytic level. There is a large amount of data that examines the performance of CBCA and RM in different settings and indicates that both procedures work to a certain extent. Through synthesis, a meta-analysis enables both testing of whether these effects are robust across different populations and study designs, as well as the estimation of effect sizes more precisely than on the basis of individual studies alone (Borenstein et al., 2009). In comparison with a recent meta-analysis by Amado and colleagues (2015) that focused exclusively on the effectiveness of the CBCA within Anglo-American samples of children, we extended our meta-analysis to both CBCA and RM within Anglo-American and German samples of children and adults. In addition, we tested whether further boundary conditions influence the performance of verbal tools for credibility assessment and took several moderators into account. Hence, we sought to answer the following research question: How well do verbal tools for credibility assessment work? What are optimal boundary conditions? What should be considered for future research?

MODERATORS

CHARACTERISTICS OF PARTICIPANTS

AGE

Under the assumption that lying requires executive control, it must also be proposed that the ability to lie, like executive control, varies over the life span. The relationship of age and executive control is characterized by an inverted U-curve (Craik & Bialystok, 2006). Debey, Schryver, Logan, Suchotzki, and Verschuere (2015) showed that parameters of the Sheffield Lie Test were also associated with age (partly in a U-shape): Lying accuracy increased with age during childhood, was highest in young adulthood, and decreased in the elderly. Although lying speed did not significantly change from young childhood to young adulthood, it also declined during adulthood. In addition, research showed that verbal indicators of truth are also age-dependent. Studies have repeatedly demonstrated that CBCA scores increase with age because the production of criteria depends on verbal, cognitive, and meta-cognitive abilities (e.g., Blandon-Gitlin, Pezdek, Rogers, & Brodie, 2005; Buck, Warren, Betman, & Brigham, 2002; Roma et al.,

[2] Meta-Analysis 1 was published as Oberlader, V. A., Naefgen, C., Koppehele-Gossel, J., Quinten, L., Banse, R., & Schmidt, A. F. (2016). Validity of Content-Based Techniques to Distinguish True and Fabricated Statement: A Meta-Analysis. *Law and Human Behavior*, *40*, 440-457. doi: 10.1037/lhb0000193. For this reason, I refer to "we" when reporting on Meta-Analysis 1.

2011; Vrij et al., 2004a). With regard to RM, Vrij (2008) doubted that this approach is capable of assessing the truth status of children's statements because children have rich imaginations and hence their imagined memories resemble real ones. These data suggest studying participants' age as a moderator.

SEX

To the best of our knowledge, previous studies on verbal tools for credibility assessment revealed no statistically significant differences for statement quality of female and male participants (e.g., Roma et al., 2011; Sporer, 1997). However, since it is repeatedly discussed whether women and men differ in verbal abilities (e.g., meta-analysis by Hyde & Linn, 1988 revealed a small effect size of $d = 0.33$ indicating superior female performance in speech production), we investigated the influence of participants' sex.

MOTIVATION TO REPORT

In addition to demographic characteristics of participants, it can be supposed that the motivation to provide an experience-based or fabricated statement influences its quality. Within field studies, a high level of motivation can be expected to lead to a detailed and convincing statement, since there is usually a lot at stake (e.g., conviction, arrest). It is, however, difficult to establish such high motivation in the laboratory. To address this problem, numerous studies have offered incentives to motivate both truth tellers and liars to report compelling stories (e.g., Gödert et al., 2005; Nahari, Vrij, & Fisher, 2012; Vrij, Mann, Kristen, & Fisher, 2007).

EXPERIENCE STATUS

As described above, the CBCA was originally developed to assess the statement quality of alleged victims of child sexual abuse. However, many authors assume that the CBCA can also be used in other cases. Within laboratory studies researchers have investigated not only witnesses or victims (e.g., Vrij, Kneller, & Mann, 2000), but also suspects (e.g., Nahari et al., 2012). Moreover, studies differ in whether participants actively experienced an event (e.g., being part of a mock crime; Vrij et al., 2007) or only passively observed it (e.g., watching a video of a crime; Vrij, Edward, & Bull, 2001). According to theories on episodic memory, real behavior should lead to more intense memory than imagining an event (Schacter et al., 2007) and thus influence the performance of verbal tools for credibility assessment.

TRAINING OF PARTICIPANTS

It can be assumed that the performance of verbal tools for credibility assessment decreases when participants know the underlying rational, i.e., know which criteria indicate an experience-based or fabricated statement. Some studies have investigated the influence of training participants in the criteria used by credibility assessments. For example, a study by Vrij et al. (2000) showed that a CBCA expert could correctly classify only 27% of the statements if participants were trained. If participants were naïve, it was 69%.

CHARACTERISTICS OF THE STATEMENT

EVENT CHARACTERISTICS

As already mentioned, it is difficult to design ecologically valid laboratory studies. How can an ethically acceptable situation be created that is comparable, for example, to the experience of sexual abuse? To depict real situations in the best possible way, Steller (1989) recommended creating events that are characterized by personal involvement, a negative emotional tone, and a certain loss of control. Some studies meet these requirements by asking participants to remember or fabricate an event that caused financial, emotional, and/or physical harm (e.g., Merckelbach, 2004).

PRODUCTION MODE

Criteria rating of verbal tools for credibility assessment is usually based on oral statements (e.g., Akehurst, Köhnken, & Höfer, 2001). However, some studies have examined written statements (e.g., Nahari et al., 2012). Against this background arises the question of whether the production mode influences the quality of the statement. Horowitz and Newman (1964) showed that speaking is more productive and elaborate than writing, meaning that participants produced more words, phrases, and sentences in oral statements. Kellogg (2007) suggested that writing places higher demands on working memory than speaking, as it is less practiced, and demonstrated that participants' reports on a recalled story were more complete and accurate when made orally. Based on these findings, Hauch et al. (2015) assumed that differences between liars and truth tellers should be more pronounced in written than in verbal reports, since liars should use comparatively less *sensory* and *contextual details* (RM criteria). In fact, their meta-analysis revealed that lies contained fewer *sensory details* than true stories only when written down by hand (gu = 0.34) and contained fewer *spatial details* only when typed on a keyboard (g_u = 0.13). However, for other indicators of deception, results were less clear (for details see Hauch et al., 2015).

TYPE OF LIE

There are several types of lies that can be distinguished in different ways. An important distinction is whether a lie is completely fabricated (outright lie; e.g., Blandon-Gitlin, Pezdek, Rogers, & Brodie, 2005) or partly based on the truth (concealment lie; e.g., Bensi, Gambetti, Nori, & Giusberti, 2009). As concealment lies include experience-based aspects, it should be more difficult to distinguish them from true statements than outright lies.

CHARACTERISTICS OF THE ASSESSMENT PROCEDURE

TYPE OF RATER

The application of verbal tools for credibility assessment requires expertise in coding the criteria. In some studies statements have been rated by professionals (e.g., Vrij et al., 2007), and in others by trained laypersons (e.g., Merckelbach, 2004) or persons who are not familiar with the methods (e.g., Nahari et al., 2012).

NUMBER OF CBCA CRITERIA AND SCORING OF CRITERIA

In addition, studies differ in the selection of criteria. Regarding the CBCA, not all studies apply the full set of 19 criteria. Some study designs are simply not suited to produce certain CBCA criteria, thus they were omitted from the outset (e.g., Bogaard, Meijer, & Vrij, 2014). For example, it is not possible to pardon a perpetrator if there was none. Other studies have used the 14-item version of the CBCA by Raskin, Esplin, and Horowitz (1991), which excludes motivational criteria (e.g., Lamb et al., 1997). Furthermore, studies differ in the scoring of criteria: Either the criteria are measured using a Likert scale (e.g., Bradford, 2006) or only dichotomously in the form of absence/presence (e.g., Craig, Scheibe, Raskin, Kircher, & Dodd, 1999).

TYPE OF DEPENDENT VARIABLES

Studies have used different dependent variables to examine the effectiveness of verbal tools for credibility assessment. In some studies, raters have classified statements as true or deceptive (e.g., Berger, 2005). Other studies have used a statistical approach and determined classification rates using discriminant analysis (e.g., Bogaard, Meijer, Vrij, & Merckelbach, 2016). Since this approach optimizes the classification of statements by building and testing a model on the same sample, it must be cross-validated. Studies that have used discriminant analysis differ in whether they address this issue or not. Finally, some studies have compared means of verbal tools for credibility assessment in experience-based and fabricated statements (e.g., Bensi et al., 2009).

GENERAL STUDY CHARACTERISTICS

STUDY DESIGN

Previous reviews and meta-analyses on verbal tools for credibility assessment found larger effects for field than for laboratory studies (Amado et al., 2015; Vrij, 2005). This could be explained by the higher ecological validity of field studies, but also by the fact that validation criteria are not always independent of the quality of statements. Moreover, laboratory studies differ in whether participants were required to make both a experience-based and a fabricated statement (within-subjects design; e.g., Elntib, Wagstaff, & Wheatcroft, 2014) or were only part of one experimental group (between-subjects design; e.g., Flieger, 2009).

PUBLICATION STATUS

In order to investigate publication bias, it was also examined whether effect sizes of unpublished and published studies differ. Furthermore, the year of publication was taken into account.

METHOD

DATABASE

INCLUSION AND EXCLUSION CRITERIA

We included unpublished and published English- and German-language studies that compared the quality of experience-based and fabricated statements using CBCA or RM. Studies that compared true and suggestive statements were excluded. Just like lies, suggested statements are not experience-based and are internally generated, but – and that is a major difference – they are not created intentionally. Suggested memories, whether auto- or externally suggested, are based on the subjective belief that the remembered event has actually taken place (e.g., Loftus & Pickrell, 1995; Volbert & Steller, 2014). The rationale of verbal tools for credibility assessment, which is based on the fact that lying is cognitively demanding and involves motivational aspects like strategic self-presentation, is not appropriate in this case. This applies at least to the CBCA. Although minimal, there is empirical evidence that RM can distinguish between experience-based and suggested statements. A study by Schooler, Gerhard, and Loftus (1986) showed that RM-trained raters outperformed untrained raters in classifying suggested statements. Nevertheless, we excluded laboratory studies on this subject.

KEYWORD SEARCH

We ran the keyword search in the following databases: PsycARTICLES, PsycINFO, and PSYNDEXplus Literature and Audiovisual Media. For English-language studies we used the following terms: "Criteria-Based Content Analysis", "CBCA", "Reality Monitoring", "RM", "Scientific Content Analysis", "SCAN", "Statement Validity Assessment", "SVA", OR "Validity Checklist"; AND "psychology of evidence", "statement analysis", "credibility", "credibility assessment", OR "deception". For German-language studies we used following keywords: "Kriterienbasierte Inhaltsanalyse", "CBCA", "Reality Monitoring", "RM", "Scientific Content Analysis", "SCAN", "Statement Validity Assessment", "SVA", OR "Validity Checklist"; AND "Aussagepsychologie", "Aussagebeurteilung", "Glaubhaftigkeit", OR "Glaubwürdigkeit". We did not translate some of the keywords into German, as the respective English technical terms have been established in the German literature. The keyword search in the databases was completed on March 18, 2015. In addition, we have contacted researchers on verbal credibility assessment and asked for their unpublished studies.

FINAL DATA SAMPLE

From a total of 186 identified studies, 52 matched the inclusion criteria[3] (see Figure 1). In some studies, several comparisons were calculated based on one sample of experience-based and fabricated statements. To avoid the problem of dependent data in these cases, we applied the following decision rule: If studies investigated different verbal tools for credibility assessment, namely CBCA and RM, in one data set, we included only results for the CBCA to estimate the overall meta-analytic effect size and to run moderator analyses. In addition, we computed separate effect sizes for both procedures, each including all comparisons of one technique, to enhance statistical power. All other single-case decisions can be found in the data table (column: description of effect size basis; see Appendix A).

[3] Studies that examined SCAN were excluded due to the small number of studies (k = 3).

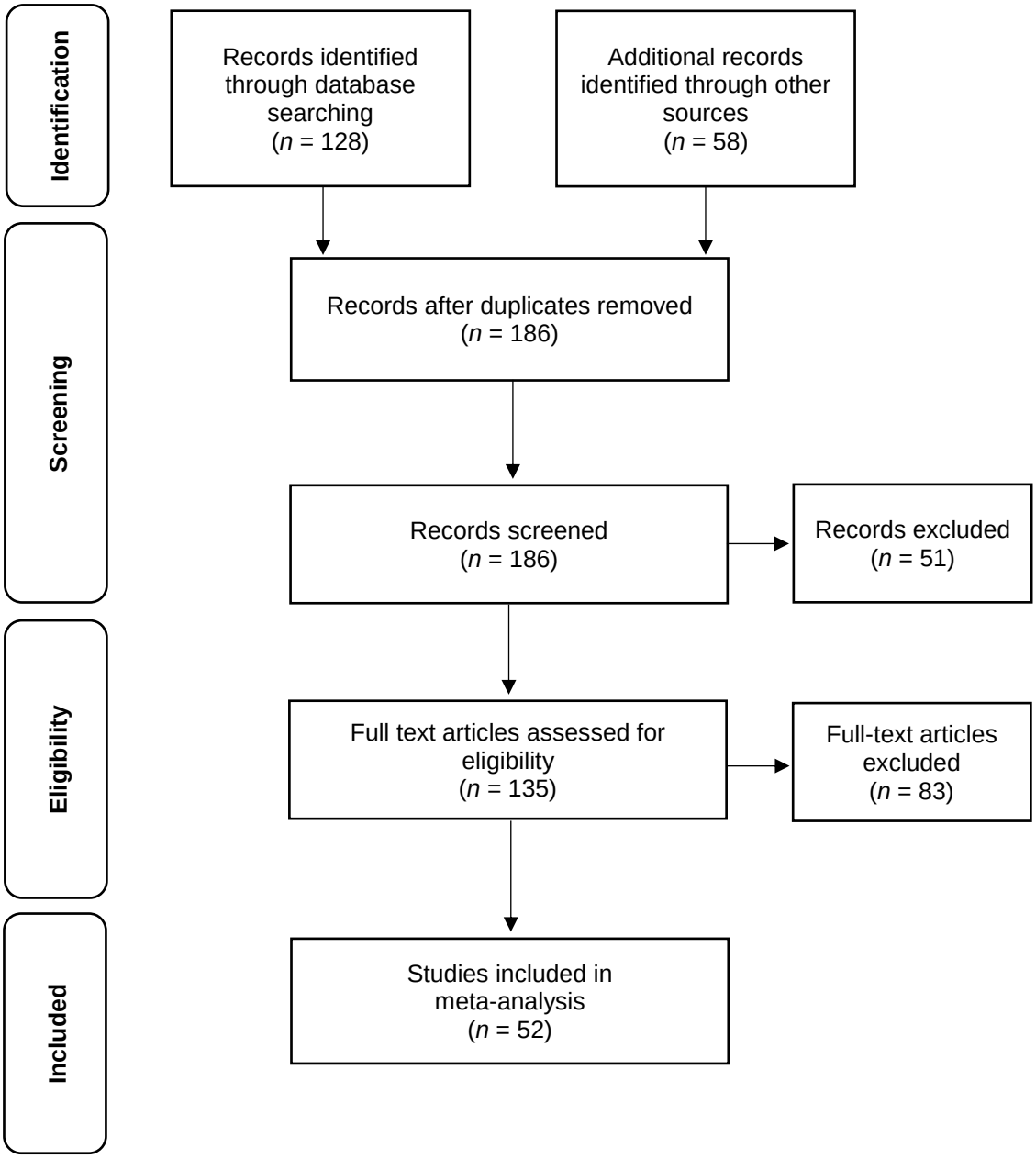

Figure 1. Full PRISMA diagram of the literature search of Meta-Analysis 1.

MODERATOR VARIABLES

Moderator variables were either continuous or categorical. The following continuous moderators were examined: sex ratio of participants in terms of a proportion from 0 (all men) to 1 (all women), and year of publication. Categorical moderators are displayed in Table 3.

Table 3

Categorical Moderators

Moderator	Coding
Age of participants	< 18 years or ≥ 18 years
Motivating incentive	Absence or presence of financial or other motivating incentives
Experience status	Event personally experienced or event not personally experienced; accused or not accused
Participant training	Trained or not trained
Event characteristics	Absence or presence of personal involvement, negative emotional tone, and extensive loss of control
Type of lie	Concealment lie or outright lie
Statement mode	Oral or written
Type of rater	Laypersons, trained raters, or professionals
Set of CBCA criteria	Not complete, complete set of 19 CBCA criteria by Steller & Köhnken (1989), or 14-item version by Raskin et al. (1991)
Scoring of criteria	Absence/presence scoring or scoring on a Likert scale
Decision basis	Rater decision, discriminant analysis, or mean comparison
Cross-validation in studies using discriminant analysis	Cross-validation or no cross-validation
Study design	Field study or laboratory study; within-subjects design or between-subjects design
Publication status	Published or not published

CODING PROCEDURE AND INTERCODER RELIABILITY

Two independent coders (first and third author of Oberlader et al., 2016) calculated effect sizes (Cohen's *d*, Hedges' *g*), standard errors, variances, and inverse variance weights. Based on a coding manual (see Appendix B), all moderator variables were rated. 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. Cases of disagreement were discussed after computing the interrater reliability and a consent decision was made.

STATISTICAL ANALYSES

EFFECT SIZE MEASURE

As a measure of effect size, we used Cohen's $d = (M_{true} - M_{fabricated}) / SD_{pooled}$)[4]. If studies provided means and standard deviations for CBCA or RM scores of experience-based and fabricated statements, Cohen's *d* could be calculated directly on this basis. For studies that classified statements using statistical or rater decisions, results first had to be probit-transformed (Lipsey & Wilson, 2000). Probit-transformed hit rates of correctly classified experience-based statements and probit-transformed false alarm rates of incorrectly classified fabricated statements were used to calculate Cohen's *d*. Since Cohen's *d* overestimates the effect size for small samples, we additionally calculated Hedges' $g = d * (1 - 3 / (4 * [n_{true} + n_{fabricated}] - 9)$. For the estimation of meta-analytical effects, effect sizes of the individual studies were weighted by their inverse variance weight (Lipsey & Wilson, 2000), $w = (2 * n_{true} * n_{fabricated} * [n_{true} + n_{fabricated}]) / (2 * [n_{true} + n_{fabricated}] + n_{true} * n_{fabricated} * g)$.

META-ANALYTICAL MODELS

We used random-effects meta-analysis (REMA) for effect size estimation. The REMA, modeled as $\theta_i = \mu + u_i$, addresses variation across studies and assumes that true effects have a mean of μ, and u_i is a random error that is normally distributed around zero with a variance of τ^2. For moderator analyses, we ran random-effects meta-regressions, $\theta = ß_0 + ß_1 x_{i1} + u_i$, with x_i as moderator variable in study i and the residual variance u_i that is normally distributed around zero

---

[4] 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 dependent 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 $d = (M_{true} - M_{fabricated}) / SD_{pooled}$, differing only at the third decimal. Therefore, effect sizes of within-study comparisons were calculated as Cohen's *d*.

with a variance of τ^2. To illustrate the practical significance of the effect size estimates, we additionally calculated the common language effect size indicator (CLES; McGraw & Wong, 1992) where applicable. The CLES indicates the probability of cases where experience-based statements have higher scores than fabricated statements. Since the CLES requires the *n*s of the experience-based and fabricated statements and these were not available for the effect size estimation based on trim-and-fill, the CLES could only be calculated for REMA.

In addition to effect size estimates, we provide information on the 95% CI, the z-statistic, the number of independent studies (k), and, where possible, the total number of statements (n). Moreover, we report the Q-statistic of moderation tests and I^2 as measure of heterogeneity, which indicates the percentage of observed variance that reflects real differences between studies. According to Higgins, Thompson, Deeks, and Altman (2003), I^2 values of 25% could be considered as low, 50% as medium, and 75% as high.

TEST FOR OUTLIERS

To test for outliers, we computed two additional REMAs for the overall effect, one without the lowest and one without the highest effect size. If the Q-statistic of heterogeneity of one or both of these data sets was statistically significant and changed over 50%, the study with the lowest and/or highest effect size would have been excluded as an outlier (Babchishin, Nunes, & Hermann, 2013).

TEST OF PUBLICATION BIAS

To test for publication bias, we used the trim-and-fill method by Duval and Tweedie (2000a) that is based on the graphical display of the effect sizes plotted against the standard error in a funnel plot. It is supposed that this funnel plot is asymmetrical if publication bias is present, i.e., more studies are on the side of positive effects. In this case, the trim-and-fill estimator iteratively removes individual study effects from one side of the funnel plot until the funnel plot is symmetrical (we used the estimators L_0 and R_0; Duval & Tweedie, 2000b). A corrected effect size is then calculated on the reduced data set. The previously removed studies are now refilled and further studies that are reflected at the recalculated mean are added.

SOFTWARE

To calculate the interrater reliabilities, we used IBM SPSS Statistics 24. For the meta-analytical calculations, we used the following R packages in R Statistical Software (version 3.4.1; R Core Team, 2017): compute.es (AC Del Re, 2013) and metafor (Viechtbauer, 2010).

RESULTS

OVERALL EFFECT SIZE ESTIMATION

A REMA on the total data set of 52 studies ($N_{statements}$ = 3,892) revealed a large point estimate with moderate to large effect sizes in the confidence interval and high heterogeneity between studies, d = 1.00 (95% CI [0.75, 1.25], z = 7.94, p < .001, I^2 = 92.12%) and g = 0.98 (95% CI [0.74, 1.22], z = 7.99, p < .001, I^2 = 91.72%). In 76%, experience-based statements had descriptively higher scores than fabricated statements. No study was excluded as statistical outlier.

PUBLICATION BIAS

The two trim-and-fill estimators yielded different results: The L_0 estimator indicated that no studies needed to be filled in. The R_0 estimator showed 12 missing studies. A REMA on the R_0-supplemented data set of 64 studies revealed a moderate point estimate for the bias-corrected effect size with small to large effects in the confidence interval and high heterogeneity, d = 0.58 (95% CI [0.27, 0.89], z = 3.63, p < .001, I^2 = 95.60%).

EFFECT SIZE ESTIMATION PER PROCEDURE

Figure 2 displays the forest plot of the effect sizes. Point estimates ranged from -0.25 to 3.66. Three point estimates were negative, i.e., in contrast to the hypothesis, but not statistically significantly different from zero; 17 confidence intervals included negative effect sizes.

CBCA

A REMA showed that the CBCA discriminated statistically significantly between experience-based and fabricated statements with a large point estimate and moderate to large effects in the confidence interval and high heterogeneity, d = 0.96 (95% CI [0.69, 1.24], z = 6.96, p < .001, I^2 = 91.76%) and g = 0.94 (95% CI [0.68, 1.21], z = 6.98, p < .001, I^2 = 91.36%), k = 44, $N_{statements}$ = 3,070. In approximately 75% of the cases, experience-based statements had descriptively higher scores than fabricated statements.

RM

A REMA showed that the RM discriminated statistically significantly between experience-based and fabricated statements with a large point estimate and moderate to large effects in the confidence interval and high heterogeneity, d = 0.87 (95% CI [0.61, 1.13], z = 6.67, p < .001, I^2 = 85.19%) and g = 0.86 (95% CI [0.61, 1.11], z = 6.73 p < .001, I^2 = 84.57%), k = 17, $N_{statements}$ = 1,892. In 73%, experience-based statements had descriptively higher scores than fabricated statements.

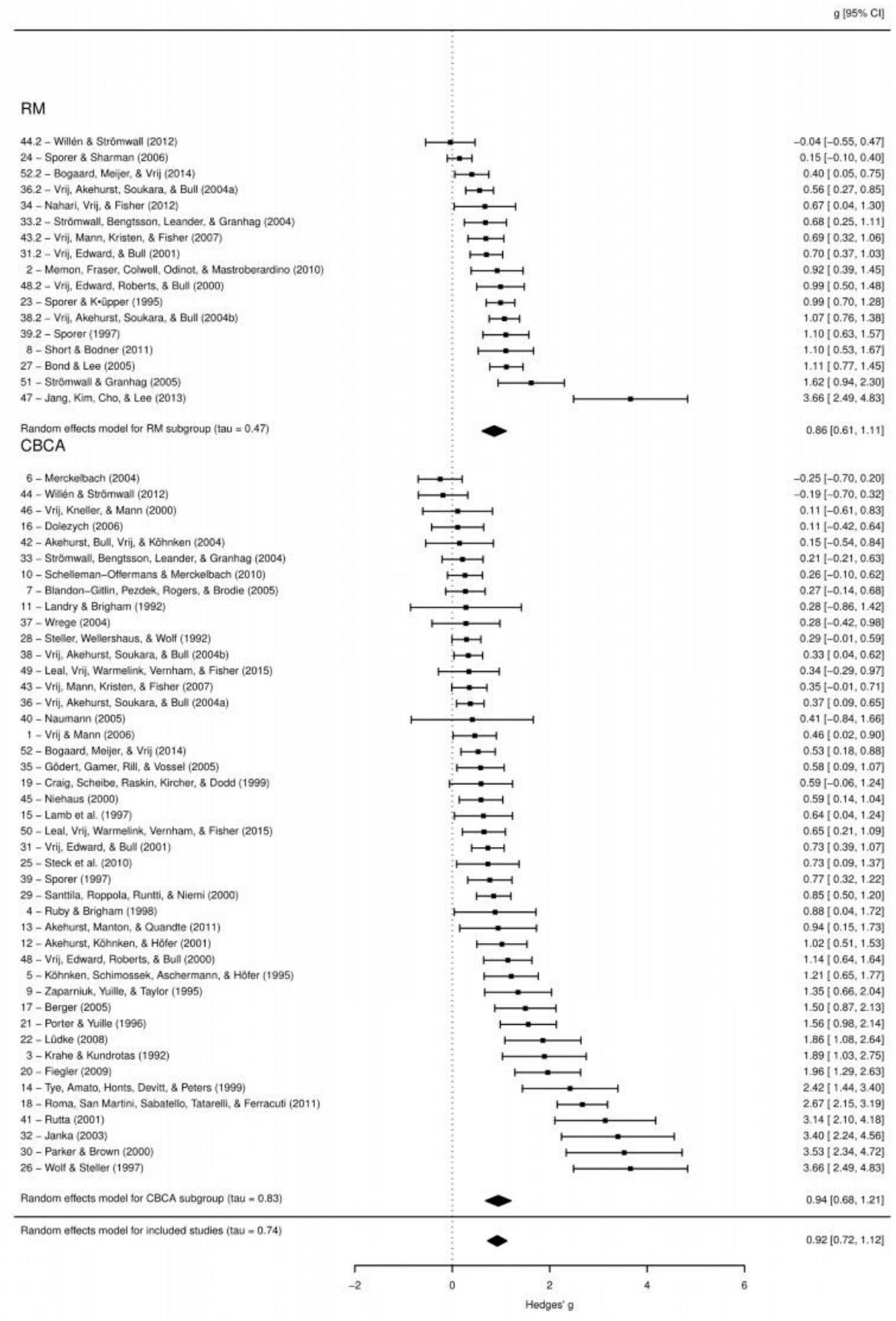

Figure 2. Forest plot separate for CBCA and RM including subset effect sizes and overall effect size estimation based on REMA.

MODERATOR ANALYSES

Table 4 displays the results of categorial moderator analyses. Despite the subsets *participants training: yes* and *type of rater: laypersons*, all moderator subsets showed statistically significant positive effect sizes. Although moderator analyses were intended to explain true variance between studies, the heterogeneity within the subsets was still high. With two exceptions (subset *discriminant analysis with bias correction*; subset *event not personally experienced*), the I^2 statistic revealed scores ≥ 81.85%.

The Q-tests of moderation revealed two statistically significant results. First, it showed that the complete version of CBCA criteria outperformed any incomplete set. There was a large point estimate for studies using 19 CBCA criteria compared to a moderate point estimate for studies using incomplete sets. Studies using the 14-item version of the CBCA by Raskin et al. (1991) did not differ statistically significantly from the other moderator subsets. Second, moderator analyses showed that studies classifying experience-based and fabricated statements by discriminant analysis outperformed studies comparing means of CBCA/RM scores in experience-based and fabricated statements. There was a large point estimate for studies using discriminant analysis compared to a moderate point estimate for studies using mean comparisons. Studies classifying experience-based and fabricated statements by rater decisions did not differ statistically significantly from the other moderator subsets.

All other categorical moderations were statistically non-significant. Only one of these results should be mentioned here, as it is a focal moderator closely related to the issue of publication bias: Effect sizes for published and unpublished studies yielded no statistically significant difference. Moreover, continuous moderators, the year of publication, $Q(1) = 0.59$, $p = .442$, $k = 52$, as well as the sex ratio in the sample, $Q(1) = 1.36$, $p = .244$, $k = 41$, were statistically non-significant.

Table 4

Results of categorical moderator analyses

Moderator	Moderator category	*g*	95% CI	*Q* *z*	I^2 in %	*k*	*n*	Study IDs
Age of participants				0.49				
	< 18	0.90	[0.39, 1.41]	3.47***	91.76	11	767	7, 14, 15, 18, 19, 28, 29, 33, 42, 45, 51
	≥ 18	1.01	[0.73, 1.39]	6.29***	92.23	35	1,796	1–6, 8–11, 13, 16, 20–22, 24–27, 30–32, 34, 37, 39–41, 43, 44, 46–50, 52
Motivating incentive				1.48				
	No	0.93	[0.65, 1.20]	6.69***	89.95	35	1,962	2, 4–12,14, 17, 22–27, 29, 31–34, 39, 40, 42, 44–52
	Yes	0.67	[0.21, 1.13]	2.83***	91.65	10	851	1, 16, 21, 28, 35–38, 41, 43
Experience status				2.22				
	Event not personally experienced	1.01	[0.65, 1.36]	5.50***	61.46	6	300	5, 9, 17, 31, 46, 48
	Event personally experienced	0.88	[0.61, 1.15]	6.41***	92.12	40	2,563	1, 2, 4, 6–8, 10–12, 14, 16, 20–29, 32–45, 47, 49, 50–52
				0.32				
	Not accused	0.85	[0.62, 1.09]	7.09***	89.76	41	2,532	2, 4–12, 14, 16, 17, 20, 22–29, 31–33, 35–42, 44–46, 48–52
	Accused	1.26	[0.15, 2.37]	2.22*	95.16	5	331	1, 21, 34, 43, 47
Participant training				0.48				
	No	0.92	[0.67, 1.17]	7.14***	90.68	39	2,400	1, 2, 4–12, 14, 16, 17, 20–29, 31–36, 39, 42–45, 47–49, 51
	Yes	0.75	[-0.10, 1.60]	1.74	92.71	6	380	37, 38, 40, 41, 46, 52

Table 4 (*continued*)

Moderator	Moderator category	*g*	95% CI	*Q* *z*	I^2 in %	*k*	*n*	Study IDs
Event characteristics (negative tone, personal involvement, loss of control)				0.06				
	At least one missing	0.87	[0.64, 1.10]	7.41***	88.58	36	2,535	1, 2, 5, 7–10, 12, 14, 20–29, 31, 33–36, 38–40, 42–44, 46–52
	All three present	0.96	[0.22, 1.70]	2.53*	93.70	10	328	4, 6, 11, 16, 17, 28, 32, 37, 41, 45
Type of lie				3.61				
	Concealment	1.19	[0.40, 1.97]	2.95**	92.47	6	345	2, 31, 34, 35, 47, 48
	Outright	0.84	[0.58, 1.10]	6.31***	91.12	38	2,418	1, 4–12, 14, 16, 17, 20, 22–24, 26–29, 32, 33, 36–46, 49–52
Statement mode				0.18				
	Oral	0.99	[0.75, 1.25]	7.75***	90.19	44	2,525	1–5, 7–9, 11–21, 25–33, 35–46, 48–51
	Written	0.90	[0.13, 1.67]	2.29*	96.42	8	673	6, 10, 22–24, 34, 47, 52
Type of rater				2.07				
	Laypersons	0.72	[-0.02, 1.47]	1.91	91.84	6	437	6, 20, 22, 24, 34, 42
	Trained participants	0.80	[0.59, 1.02]	7.29***	83.14	31	2,023	2, 4, 5, 8–13, 15, 19, 21, 23, 25, 26, 28, 31–33, 35, 36, 38–40, 43–46, 48, 51, 52
	Professionals	1.25	[0.59, 1.90]	3.74***	93.99	12	616	1, 3, 7, 16–18, 29, 30, 37, 41, 49, 50
Set of CBCA criteria				6.51*				
	Incomplete sets	0.66	[0.43, 0.88]	5.73***	81.85	27	1,730	1, 4–7, 9, 10, 11–14, 16, 19, 21, 25, 26, 28, 35, 36, 38, 39, 42, 43, 46, 49, 50, 52
	19 CBCA criteria	1.49	[0.77, 2.21]	4.05***	92.36	12	422	3, 17, 20, 22, 30, 32, 33, 37, 40, 41, 44, 45
	14-item version	1.20	[0.47, 1.93]	3.24**	92.67	5	412	15, 18, 29, 31, 48

Table 4 (*continued*)

Moderator	Moderator category	*g*	95% CI	*Q* *z*	I^2 in %	*k*	*n*	Study IDs
Scoring criteria				1.07				
	Absence/ presence	1.26	[0.67, 1.85]	4.20***	92.63	14	770	7, 9, 11, 15, 18–20, 30, 31, 46, 47, 49–51
	Scoring on a scale	0.88	[0.61, 1.16]	6.37***	91.29	35	2,283	1–6, 8, 10, 12–14, 16, 17, 21–24, 26, 28, 29, 32, 33, 35–45, 48, 52
Decision basis				19.09***				
	Discriminant function	1.51	[1.09, 1.93]	7.10***	90.57	20	920	2–5, 9, 12–14, 21, 23, 26, 27, 29, 32–34, 39, 41, 47, 51
	Rater decision	1.32	[0.57, 2.08]	3.42***	86.58	7	260	8, 11, 17, 20, 30, 40, 45
	Mean comparison	0.51	[0.28, 0.74]	4.66***	85.52	25	2,018	1, 6, 7, 10, 15, 16, 18, 19, 22, 24, 25, 28, 31, 35–38, 42–44, 46, 48–50, 52
				0.02				
	DF without cross-validation	1.67	[1.19, 2.16]	6.75***	92.28	19	829	3, 4, 9, 12–14, 21, 23, 26, 27, 29, 30, 32–34, 39, 41, 47, 51
	DF with cross-validation	1.06	[0.67, 1.44]	5.39***	0	2	119	2, 5

Table 4 (*continued*)

Moderator	Moderator category	*g*	95% CI	*Q* *z*	I^2 in %	*k*	*n*	Study IDs
Study designs				3.76				
	Field studies	1.66	[0.73, 2.59]	3.51***	90.17	6	335	3, 13, 15, 18, 19, 30
	Laboratory studies	0.89	[0.66, 1.12]	7.44***	90.71	46	2,863	1, 2, 4–12, 14, 16, 17, 20–29, 31–52
				1.02				
	Within-subjects	0.81	[0.42, 1.19]	4.12***	92.00	18	866	4, 6, 8, 10, 16, 17, 23, 27–29, 31, 32, 37, 41, 44, 52
	Between-subjects	1.08	[0.77, 1.38]	6.87***	91.25	34	2,332	1–3, 5, 7, 9, 12–15, 18–22, 24–26, 30, 33–36, 38, 40, 42, 43, 45–51
Publication				2.57				
	Unpublished	1.45	[0.68, 2.21]	3.71***	89.16	9	234	16, 17, 20, 22, 25, 32, 37, 40, 41
	Published	0.89	[0.65, 1.13]	7.25***	91.00	43	2,964	1–15, 18, 19, 21, 23, 24, 26–31, 33–36, 38, 39, 42–52

Note. DF = discriminant function. *p < .05, **p < .01, ***p < .001.

DISCUSSION

HOW WELL DO VERBAL TOOLS FOR CREDIBILITY ASSESSMENT WORK?

Meta-analysis 1 found strong evidence that verbal tools for credibility assessment do work: The overall effect size estimation revealed a large point estimate with moderate to large effect sizes in the confidence interval. Further analyses showed that this finding applied for both CBCA and RM. The CLES indicated that in 76% or 73% of the cases experience-based statements had descriptively higher CBCA or RM scores than fabricated statements. Although these results are still far from a 100% hit rate, the application of verbal tools for credibility assessment is way better than any unstandardized judgment that hardly exceeds chance (e.g., Bond & De Paulo, 2006). On a closer look, however, it appears difficult to determine their exact performance.

First, the overall effect size decreased to a moderate point estimate with small to large effects in the confidence interval after controlling for publication bias using the R_0 estimator of the trim-and-fill method. Although, the L_0 estimator provided no indication of publication bias and we included unpublished studies that, in addition, did not differ from published studies, we must at least consider that there are further unpublished studies that would weaken the present effect size estimation. Yet, as laid out above, the symmetry logic of trim-and-fill is based on the assumption that study results that are unpublished deviate strongly from the point estimate in the negative direction. In the present research context, a negative study effect would mean that the application of a verbal tool for credibility assessment has led to an erroneous assessment of the truth status: For example, the application of the CBCA criteria would have indicated an experience-based statement, whereas the statement would actually have been a lie or vice versa. Criteria of verbal tools for credibility assessment, at least of CBCA and RM, are, however, based on the theoretically founded and empirically confirmed assumption that true and deceptive responses are cognitively differently demanding. It seems therefore rather unlikely that unpublished studies would show exclusively negative effects and that verbal tools for credibility assessment would mistakenly point in the exact opposite direction. Nevertheless, as recommended by the Duval and Tweedie (2000a), the result of the R_0 estimator should be interpreted as a lower threshold.

Second, I^2 statistics indicated high heterogeneity between studies that reduced statistical power of effect size estimation and its interpretability (Borenstein, Hedges, Higgins, & Rothstein, 2009). The amount of true variance between studies, which was reflected in a wide confidence interval, could neither be explained by separate analyses of the procedures nor by moderator analyses. Hence, the question arises whether the included studies are actually a sample of *one* population. Although all studies compared experience-based and fabricated statements using CBCA or RM,

they might have tapped into different psychological processes of truth telling or lying. Even though we found only two statistically significant moderators, individual studies differed in many respects for providing experience-based and fabricated statements. Some of these factors may not have become statistically significant due to the low statistical power in moderator subsets, and others may not have been identified. However, as stated at the outset, it is the paradigm that determines the psychological processes that a measure, here CBCA or RM criteria, reflects. Hence, it cannot be ruled out that the study designs addressed different underlying mechanisms of true or deceptive responding (e.g., response inhibition when concealing a mock crime or construction of a fabricated autobiographic event) or at least addressed them to varying degrees. With regard to practical application, a wide range of study settings is certainly useful, since verbal tools for credibility assessment must work under different conditions. Yet, in order to investigate their validity, it is important to first take a step back and to be clear about which paradigm addresses which processes and whether these processes are actually related to truth telling or lying. If these questions are specified, various study designs can be implemented to accommodate different conditions in the field. However, in the current state of research, the interpretability of meta-analytical data is limited by the high degree of heterogeneity. In terms of these data, the answer to the first research question of how well verbal tools work for credibility assessment is not conclusive: It ranges from better than chance with a moderate bias-corrected point estimate to at best accurate in about 75% of the cases.

WHAT ARE OPTIMAL BOUNDARY CONDITIONS?

NUMBER OF CBCA CRITERIA

Regarding the second research question on optimal boundary conditions, moderator analyses showed that the complete version of 19 CBCA criteria exceeded any incomplete set. The use of incomplete sets is usually explained by the fact that individual criteria could not occur in certain study designs (e.g., *pardoning the perpetrator* if no perpetrator exists). In such cases, the ecological validity is of course limited and could have contributed to smaller effect sizes. Moreover, with regard to classical test theory, longer tests have better validity, at least under certain psychometric boundary conditions. Effect sizes of studies using the 14-item version by Raskin and colleagues (1991), which leaves out motivational criteria, did not differ statistically significantly from effect sizes of studies using the complete or any incomplete set. However, the lack of statistically significant differences could be explained by the low statistical power in this subset, which contained only five studies.

DECISION BASIS

Furthermore, moderator analyses revealed that studies comparing CBCA/RM scores of experience-based and fabricated statements found smaller effects than studies classifying statements via statistical decisions. Studies that compare mean scores use verbal tools for credibility assessment as a checklist. Each criterion is either evaluated as present or absent or rated on a Likert scale and the number of points is summed up. Even if some of these studies measure the intensity of the criteria, individual criteria are all equally considered. However, if statements are classified based on a statistical decision, individual criteria are weighted by their diagnostic value. Discriminant analysis only considers criteria that differentiate between experience-based and fabricated statements and weight them according to their effect sizes. Such criteria weighting needs to be cross-validated as it is modeled and tested on the same data set. Only two of the 21 studies using discriminant analysis carried out cross-validation. Moderator analyses showed no statistically significant difference between studies that carried out cross-validation and those that did not, which could have been due to the low statistical power of the very small subset of studies performing cross-validation ($k = 2$). It is therefore highly likely that the effect size calculated from studies using classifications by discriminant analysis without cross-validation is overestimated. Kleinberg, Arnzt, and Verschuere (2019) demonstrated that even if verbal cues and truth status are not correlated, discriminant analyses could lead to accuracy rates of up to 84% when used without cross-validation. The authors simulated different levels of correlation between verbal cues and the binary outcome truth or lie ($r = 0.00$, $r = 0.10$, or $r = 0.124$) and calculated average accuracy rates for a different number of CBCA criteria (8, 12, or 19) and different sample sizes (n = 40–1,000, in steps of 10) with or without cross-validation. For sample sizes below 320, as typically given in this research context (average sample size in Meta-Analysis 1: $N_{statements} = 68$), the use of models that were not cross-validated resulted in overestimates of 12% on average (range 6–29%) in independent test sets. However, when cross-validated models were applied to independent test sets, accuracy rate deviations were not greater than 5%, even for small samples. In view of these findings, cross-validation, model testing on independent data sets, and larger samples are essential for future studies using discriminant analysis if meaningful results are to be achieved.

FURTHER MODERATORS

Further moderator analyses (i.e., on publication status, year of publication, participants' age, motivation, experience status, role in the interview, and training status, as well as on the type of lie, statement mode, experience of raters, and scoring of criteria) yielded no statistically significant results. For some of these variables, this might be due to low-powered (small) subsets and

heterogeneous moderator categories. Regarding the type of lie, for example, we only distinguished between outright and concealment lies. We did not distinguish whether people had experienced an event in the context of a laboratory experiment, whether they made use of autobiographical memories, or both. A clear distinction was either not described or not possible in all studies. Therefore, we cannot rule out that there are factors, other than those found in the present meta-analysis, that influence the validity of verbal tools for credibility assessment. Particularly with regard to the Validity Checklist, the fourth stage of the SVA, it must be assumed that there are several personal and situational factors that need to be taken into account when assessing the content quality of statements.

WHAT SHOULD BE CONSIDERED FOR FUTURE RESEARCH?

STANDARDIZATION OF PROCEDURES

The examination of studies on verbal tools for credibility assessment has shown that future research should work on the standardization of CBCA and RM. The coding of criteria leaves room of interpretation. This applies in particular for CBCA criteria: When does something count as a *detail* and when is a detail *unusual*, when *superfluous*, and when *unexpected*? There have been attempts to specify coding decisions (e.g., Arntzen, 2011; Greuel et al., 1998), but criteria ratings are still subject to individual operationalization. For example, Arnzten (2011) defined *descriptions of interactions* as a sequence of actions and reactions. However, this definition leaves further questions unanswered: How many actions and reactions does it take to identify a sequence and code the criterion as present? Are all types of sequences equally meaningful, independent of content and context? Future studies should define coding rules a priori and describe them in detail. This is the only way to examine the validity of different operationalization. Although there is less room of interpretation for RM criteria, it is important to standardize the criteria set. Currently, there are different versions that contain different (numbers of) criteria. This complicates the interpretation of their validity, especially in light of partially contradictory findings (see EMPIRICAL EVIDENCE FOR RM).

BOUNDARY CONDITIONS

Although we found only two statistically significant moderators, future research has to address factors that, in addition to the truth status, influence the quality of statements. The SVA takes such factors into account: The Validity Checklist comprises alternative hypotheses of whether a person could or would have made a particular statement if it were not experience-based. In the case of alleged child sexual abuse, CBCA criteria may be met if, for example, a child is familiar with the subject of sexual intercourse through consumption of pornographic material. Thus, the

quality of a statement could also be increased by observed events and not necessarily by a real experience.

The Validity Checklist contains several other personal and situational factors that could influence the quality of a statement: for example, motives to report, the interview style, and the context of the original disclosure. However, it is neither standardized nor empirically validated. First, there are different versions including different alternative hypotheses (Raskin & Esplin, 1991; Steller, 1989; Steller & Boychuk, 1992; Yuille, 1988) and second, there are no exact specifications on how the obtained information is integrated with the results of the CBCA. Consider the above example: What statement quality would be expected if a child had prior knowledge of sexual intercourse but no sexual abuse had taken place? Or the other way around: What statement quality is necessary to rule out that a statement was generated in a way other than based on experience? There are no standardized guidelines to answer these questions, the assessment is case-specific. Of course, it is difficult to impossible to determine norms for all potential factors that influence statement quality. However, it is necessary to expand research, which has so far been limited to very few aspects (e.g., age, interview style, coaching of the interviewee; see Vrij, 2005). Particularly, with regard to RM it is important to investigate more closely whether this procedure also works with young children and when reported events occurred far back in time.

Results of the present moderator analyses indicated that future studies should use discriminant analysis to investigate under which personal and situational conditions verbal tools for credibility assessment work (best). This way, it could be investigated which criteria differentiate between experience-based and fabricated statements if certain boundary conditions are present (e.g., prior knowledge/experiences, reporting motives, or interview style) and how these should be weighted. Contrary to previous practice, it is imperative that research using discriminant analysis cross-validate results on independent samples.

ECOLOGICAL VALIDITY OF STUDY DESIGNS

Furthermore, future research must focus on the ecological validity of studies. Even though moderator analyses showed no statistically significant differences between laboratory and field studies or between different characteristics of laboratory studies (motivation to report, event characteristics, interview role, or type of lie), ecologically valid operationalization is indispensable if study results are to be generalized to practical contexts. It is, of course, difficult to impossible to create real-life conditions in the laboratory that are comparable, for example, to the experience of sexual abuse. Nevertheless, it is still important to maximize cognitive demands and the motivation for reporting experience-based or fabricated statements, for instance, by gathering

statements orally in face-to-face situations, posing follow-up questions, repeating interviews, and verifying the claimed details to whatever extent possible.

However, most laboratory studies lack precisely these conditions. In contrast, participants typically give only one free narrative – sometimes even in written form – and are not asked any follow-up questions. Moreover, laboratory studies often rely on (fabricated) autobiographical statements that cannot be verified. Besides cognitive demands, the motivation to make a deceptive statement in laboratory studies also differs greatly from real case scenarios. A false allegation in real life could have detrimental outcomes for others (e.g., imprisonment), whether desired or not, which cannot be compared with motivational conditions in experimental settings. In addition, laboratory as well as field studies usually simplify the concept of truth and lie as dichotomous although it is actually rather continuous. Liars often use episodic memories to invent an account and only modify individual details. In the original context for which CBCA criteria were developed, namely children's statements on sexual abuse (Steller & Köhnken, 1989), this problem is less prevalent: Usually, children have no previous sexual experience and cannot recourse to episodic memory. However, in cases where previous experience (or vicarious knowledge from media depictions) cannot be ruled out, for example, if the issue of concern is not whether sexual intercourse took place but whether it was consensual, the use of episodic memories can no longer be excluded. Finally, in some laboratory designs cues that are not part of a procedure, but still confounded with the truth status, find their way into statements. For example, in experiments in which raters assess the credibility of a number of statements based on the same experimentally induced experience (e.g., Memon, Fraser, Colwell, Odinot, & Mastroberardino, 2010), they might infer that a specific detail mentioned by many participants (e.g., "then she brought up the blue bucket") is highly likely to be associated with truth status. This is certainly a diagnostic cue that could not be determined in a single-case content-based assessment.

These exemplary shortcomings illustrate that experimental scenarios deviate to a substantial degree from real-life conditions and thus limit the generalizability of the present meta-analytic results to practical contexts. In order to be able to answer the question of whether empirical results can be generalized, we urgently need studies that are internally and ecologically valid, make lying maximally difficult, and enhance motivation to provide a convincing statement. In addition, future field studies should use validation criteria that are largely independent of the quality of statements.

CONCLUSION

With a large point estimate and moderate to large effects in the confidence interval, verbal tools for credibility assessment offer great potential to distinguish between experience-based and fabricated statements. This finding applied for both CBCA and RM, despite the fact that (1) there was a high level of heterogeneity between studies that could not be resolved by moderator analyses and, (2) it cannot be ruled out that effect size estimates are biased and thus verbal tools for credibility assessment only work to a smaller extent. In the end, however, the question of whether CBCA and RM work well enough to be the core part of expert witness testimonies is only partially an empirical one. Whether a piece of evidence is considered valid and how it is weighted is ultimately a legal decision that is for example guided by the Daubert standards, which are largely fulfilled in the present case (see EMPIRICAL EVIDENCE OF VERBAL TOOLS OF CREDIBILITY ASSESSMENT). For comparison, the identification of faces by eyewitnesses that demonstrated a comparable effect size of $g = 0.82$ (Meissner & Brigham, 2001) is widely accepted evidence in court. This does not, of course, justify the use of CBCA and RM, but might serve as a benchmark to interpret the present effect size estimates with regard to the practical applicability.

Compared to other promising psychophysiological or behavioral measures of deception, verbal tools for credibility assessment show similar effect sizes (see for example meta-analyses by Leue & Beauducel, 2019, on P3; by Suchotzki et al., 2017, on different reaction time-based methods). Yet, they are easier to apply for the average psychological expert witness and do not require technical equipment. In addition, other measures of deception or, more precisely, other paradigms used to capture these measures have certain limitations concerning construct validity and/or application-specific boundary conditions.

For example, the Concealed Information Test (Lykken, 1959) measures not lying, but *crime-related knowledge*. Participants are asked crime-related questions and are given one relevant (e.g., a crime characteristic) and several neutral response options, all of which they should deny. It is supposed that the relevant alternative is significant/salient only for people with crime knowledge and causes different physiological or behavioral responses compared to the neutral alternatives. In addition to limited construct validity with respect to lying, problems arise in the application: In order to carry out the Concealed Information Test, details of the incident must be known to the diagnostician and the guilty respondent. Moreover, critical and neutral items need to be indistinguishable from the perspective of the innocent respondent. Therefore, the procedure cannot be applied if it is in doubt whether an event has taken place at all, if critical details are not encoded or remembered by the guilty respondent, or if they have been leaked.

These remarks illustrate once again the importance of *first* specifying the psychological processes of truth telling and/or lying a paradigm addresses. This call applies to all paradigms, no matter what measure they use (e.g., event-related potentials, reaction time, verbal cues). In the *next step*, the validity of different measures should be investigated under ecologically valid conditions. Recent study results by Gibbons, Schnuerch, Wittinghofer, Armbrecht, and Stahl (2018) indicated that hereby a combination of reaction time-based measures and event-related potentials seems promising. The combination with other approaches may also enhance the performance of verbal tools for credibility assessment. Although there are several important appeals to future research, we can state for now that, at least in empirical research settings, CBCA and RM are among the best empirically validated methods for assessing the credibility of statements.

INTERIM CONCLUSION

Meta-Analysis 1 demonstrated that verbal tools for credibility assessment have great potential for distinguishing between experience-based and fabricated statements: A REMA revealed a large point estimate with moderate to large effect sizes in the confidence interval. Although the R_0 estimator of trim-and-fill pointed to a downward correction, we gave less weight to this finding since other data indicated no concern of publication bias (i.e., L_0 estimator of trim-and-fill, moderator analysis on published and unpublished studies).

However, the scientific and social debate on the so-called *replication crisis* (re)intensified awareness of quality problems in (psychological) research – and thus for the impact of publication bias. It has long been known that studies that report statistically significant findings are more likely to get published than studies that report statistically non-significant results (e.g., Sterling, 1959). Kühberger, Fritz, and Scherndl (2014) provided an exemplary illustration of this problem: In a random sample of 1,000 studies from 2007, they found three times as many statistically significant as non-significant results (see also Fanelli, 2011). This disproportion is inconsistent with the generally low statistical power in psychological research (Bakker, van Dijk, & Wicherts, 2012). It is therefore conceivable that statistically non-significant study results are suppressed, either due the reluctance of researchers to publish such findings (Cooper, DeNeve, & Charlton, 1997) or their being disregarded by reviewers (Mahoney, 1977) and editors (Coursol & Wagner, 1986), or both. The *Open Science movement*, which emerged in response to replication crisis, has reemphasized the impact of publication bias and related problems with alarming data.

For example, the *Open Science Collaboration* (2015) investigated the reproducibility of psychological research and tried to replicate results of 100 experimental and correlational studies published in three high-ranking journals. This mammoth project showed that, depending on the criterion, only 36% to 68% of the original results could be replicated. It is reasonable to assume that remaining studies applied *questionable research practices* (QRPs; John, Loewenstein, & Prelec, 2012) and could therefore not be replicated. QRPs refer to researchers' degrees of freedom to collect and analyze data in different ways (e.g., different criteria for excluding outliers, choosing the dependent variable that works best, optionally increasing the sample size if results are not significant yet) in order to achieve statistically significant results. These analysis strategies are not necessarily chosen intentionally (Gelman & Loken, 2014). Researchers might find the expected results within the first analysis and stop analyzing. However, if they had not found the effect, they would probably have carried out further analyses. Hence, even if research strategies are unintentional, they can still be arbitrary and motivated by the results (Silberzahn et. al., 2018; Simonsohn, Simmons, & Nelson, 2015).

These findings are highly relevant for the present research on the validity of verbal tools for credibility assessment. They demonstrated that data collecting, analyzing, reporting, and/or publishing strategies dramatically reduce the likelihood that statistically non-significant studies will be accessible and integrated into meta-analyses. Therefore, if QRPs and publication bias are present, statistically significant studies are overrepresented in meta-analyses and, consequently, effect sizes will be overestimated. Although Meta-Analysis 1 did not clearly indicate cause of concern regarding bias in data, the impact of bias should not be downplayed, but rather taken into account if it cannot be completely ruled out. However, a simulation study by Carter, Schönbrodt, Gervais, and Hilgard (2019) pointed out that REMA and trim-and-fill, the methods used in Meta-Analysis 1, are not suitable for that purpose: Both meta-analytic methods produced an unacceptably high number of false-positive results when publication bias was present (Carter et al., 2019). To take the problem of validity-threatening biases seriously, we re-analyzed previous data with other bias-correcting meta-analytical methods in Meta-Analysis 2[5].

[5] Meta-Analysis 2 is under review at PLoS ONE as Oberlader, V. A., Quinten, L., Banse, R., Volbert, R., Schmidt, A. F., & Schönbrodt, F. D. (2019). How robust are meta-analytic findings on the validity of content-based credibility assessment? A comparison of six meta-analytic methods and recommendations for future research. Manuscript submitted for publication. For this reason I refer to "we" when reporting on Meta-Analysis 2.

META-ANALYSIS 2

The principle aim of Meta-Analysis 2 was to test the robustness of the previous meta-analytic findings on the validity of verbal tools for credibility assessment using different bias-correcting meta-analytical methods in addition to REMA and trim-and-fill. Based on the results of their simulation study, Carter et al. (2019) recommended, first, to run an a priori *method performance check* in order to assess the performance of individual meta-analytic methods under expected conditions in the data set and, second, to carry out a sensitivity analysis, which weights the results of all meta-analytic methods according to step one. Within this two-step procedure, we expanded the database and examined a third verbal tool for credibility assessment, the *Scientific Content Analysis* (SCAN). We sought to answer the following research question: Are the findings of our previous meta-analysis robust? More specifically, are CBCA, RM, and SCAN valid procedures to assess statement credibility according to the present empirical literature? Finally, what are optimal boundary conditions as well as implications for future research?

META-ANALYTIC METHODS

The informative value of a meta-analysis is limited by the quality of the data basis. For example, a meta-analysis will yield biased effect size estimates if the included studies themselves contain biased effects. To address this problem, several meta-analytic methods correct for biases in data sets. There are a number of simulation studies investigating the performance of these methods (e.g., Hedges & Vevea, 1996; Stanley & Doucouliagos, 2014; Simonsohn, Nelson, & Simmons, 2014; McShane, Böckenholt, & Hansen, 2016; Stanley, 2017). In short, results indicate that bias-correcting meta-analytic methods fail under certain conditions. Yet, due to the fact that each study examines only a limited number of methods and simulates only a limited subset of conditions, the studies disagree on which method to recommend under which conditions.

Against this background, Carter et al. (2019) conducted a comprehensive simulation study and investigated the performance of seven meta-analytic methods (REMA, trim-and-fill, PET-PEESE, *p*-curve, *p*-uniform, 3PSM, and WAAP) under various conditions. They systematically varied the severity of publication bias (none, medium, or high) and QRPs (none, medium, or high). In addition, the authors simulated different effect sizes (d = 0, 0.2, 0.5, or 0.8), degrees of heterogeneity between studies (τ = 0, 0.2, or 0.4), and numbers of studies included in meta-analyses (k = 10, 30, 60, or 100). This simulation study revealed that there is no single meta-analytic method that uniformly outperforms other meta-analytic methods under all conditions. Which meta-analytic method performs best depends on the research environment.

REMA

The REMA approach assumes that the true effects of individual studies are distributed around an average true effect, whereas variance is assigned to both sampling error and true variance between studies. This way, researchers can meta-analyze studies that investigate one phenomenon, but vary in their underlying effect due to study characteristics. Since studies on the validity of verbal tools for credibility differ in many respects, for example with regard to event characteristics, type of lie, and population, a REMA is, in principle, suitable in the present research context. However, Carter et al. (2019) showed that it leads to almost 100% false-positive rates (i.e., incorrect rejection of a true null effect) and overestimation of the true effect size if publication bias is present. Accordingly, the authors advise against using REMA if publication bias cannot be ruled out. Although we included all unpublished studies that we knew of and that were accessible at that time in Meta-Analysis 1, we cannot rule out that there are further unpublished studies we have not been able to take into account. Therefore, we used the trim-and-fill method to estimate the number of missing studies and to correct the effect size. However, trim-and-fill has its own limitations.

TRIM-AND-FILL

As described in Meta-Analysis 1, the trim-and-fill method is based on the graphical display of the effect sizes plotted against the standard error in a funnel plot (Duval & Tweedie, 2000a). Although the primary goal of trim-and-fill is to detect the presence of publication bias, it also provides an estimate of the true effect size and the number of missing studies. However, the symmetry logic is based on the notion that studies that were not published or disclosed deviate greatly from the mean value in the negative direction. However, as discussed in Meta-Analysis 1, it might be the case that the unpublished studies show null, and not negative effects. Moreover, the simulation study by Carter et al. (2019) demonstrated that the trim-and-fill method revealed almost 100% false-positive rates (i.e., incorrect rejection of a true null effect) as well as overestimation of the true effect size if publication bias was present.

PET-PEESE

The precision-effect test (PET; Stanley & Doucouliagos, 2014) is a meta-regression approach that corrects for the influence of small-study effects. This meta-analytic method makes use of the fact that significant effects that occur despite small sample sizes (e.g., as a result of QRPs) usually lead to negative correlations between effect size and sample size or, respectively, to positive correlations between effect size and standard error in the data set of meta-analyses. The PET method plots a regression line based on a weighted linear regression model, $d_i = b_0 + b_1 se_i + e_i$

with b_0 as intercept and b_1 as slope of the relation between the ith effect size estimate d_i and its standard error se_i. The estimated intercept, where a theoretically infinitely large sample has a standard error of zero, displays the bias-corrected effect size. However, this adjustment can be an overcorrection, as small-sample effects do not necessarily result from QRPs or publication bias.

The precision-effect estimate with standard error (PEESE; Stanley & Doucouliagos, 2014) is based on the same approach. Yet, in contrast to PET, PEESE calculates a quadratic regression line, as the influence of publication bias is assumed to be higher in low-precision studies with lower statistical power (i.e., small sample size, large standard error) than in high-precision studies with higher statistical power (i.e., large sample size, small standard error). Thus, the effect size is regressed on the squared standard error in a weighted least squares regression model, $d_i = b_0 + b_1 se_i^2 + e_i$ with b_0 as intercept and b_1 as slope of the relation between the ith effect size estimate d_i and its squared standard error se_i^2.

Simulation studies showed that PET outperforms PEESE when the true effect is zero, since PEESE overestimates the size of null effects, and vice versa, PEESE outperforms PET when the true effect is non-zero, since PET underestimates the size of non-zero effects (Stanley & Doucouliagos, 2014). Based on these findings it has been recommended to combine both meta-analytic methods into a conditional estimator called PET-PEESE: If the PET estimator produces a non-significant result, this estimator should be chosen. If the PET estimator produces a significant result, the PEESE estimator should be chosen. Carter et al. (2019) showed that false-positive rates of PET-PEESE increased with smaller sample sizes, higher heterogeneity, and fewer biases. Moreover, with decreasing sample size as well as increasing biases and heterogeneity, PET-PEESE revealed underestimation of the true effect size.

P-CURVE AND P-UNIFORM

A *p*-curve is the distribution of all statistically significant *p*-values ($p < .05$) across a set of studies. The shape of a *p*-curve is a function of the statistical power of the included studies, which is in turn a function of the effect size and sample size. If there is a null effect, the *p*-curve is flat since the probability that a *p*-value falls within a certain interval is uniformly distributed, i.e., the same number of *p*-values is expected between .00 and .01 and, for example, between .04 and .05. However, if a true effect exists, the shape of the *p*-curve is right-skewed, as there are relatively more low *p*-values than high *p*-values. Hereby the following applies: The bigger the effect size, the more right-skewed the *p*-curve. If QRPs are present, i.e., if study results were pressed below the significance threshold of $p = .05$, the right-skewness decreases. The *p*-curve is therefore suitable to make inferences on the presence of QRPs and/or publication bias. Furthermore, it

can be used to test the absence of a true effect and to estimate the true effect size (Simonsohn et al., 2014). However, effect size estimation using the *p*-curve is only based on statistically significant studies, thus excluding null effects, and, moreover, not providing confidence intervals.

Like the *p*-curve, *p*-uniform uses the distribution of *p*-values to test for publication bias, to test the absence of a true effect, and to estimate the effect size of statistically significant studies ($p < .05$). The same assumptions apply as for the *p*-curve. Technically, *p*-curve and *p*-uniform differ only in the algorithms used (McShane et al., 2016), however, *p*-uniform allows for computing confidence intervals. Carter et al. (2019) showed that with increasing heterogeneity both *p*-curve and *p*-uniform revealed increasing false-positive rates and overestimation of the true effect size.

SELECTION METHODS

The selection methods approach assesses and corrects for publication bias by modeling the conditions under which studies get published or not. Hedges (1984) introduced the first selection method, which included two models: A data model that assumes a fixed true effect size, and a selection model based on the assumption that only statistically significant studies get published. The specification of both models by selection methods allows assessment of the identifiability of model parameters and testing hypotheses for model parameters using maximum likelihood estimation. In addition, the extent of publication bias can be estimated and corrected. This rationale has been extended in other selection methods (Hedges & Vevea, 1996; Iyengar & Greenhouse, 1988). For example, Iyengar and Greenhouse (1988) based their selection method on the assumption that both statistically significant and statistically non-significant studies can get published. They integrated a weight function approach modeling the probability of statistically significant studies to get published (*p*-values < .025, one-tailed) compared to the probability of statistically non-significant studies to get published (.025 < *p*-values < 1, one-tailed). Since this method assumes homogeneous effect sizes, an extended version of this selection method considers three parameters (three-parameter selection method; 3PSM): an effect size parameter of the true average effect size, a weight parameter for the probability of publication for statistically significant and non-significant studies, and a parameter for the amount of heterogeneity between studies (McShane et al., 2016). Carter et al. (2019) showed that false-positive rates of 3PSM increased with smaller sample sizes, higher heterogeneity, and less biases. Moreover, with decreasing sample size as well as increasing biases and heterogeneity, 3PSM revealed underestimation of the true effect size.

WAAP

The weighted average of adequately powered studies (WAAP) is another method to reduce the influence of publication bias in meta-analyses (Stanley, Doucouliagos, & Ioannidis, 2017). WAAP only includes studies with a statistical power > 80%, i.e., studies that meet the standard of adequate power (Cohen, 1988) and are thus most informative. By excluding low-powered studies, WAAP is ideally less affected by potential biases. If no biases exist, the exclusion of low-powered studies means only a small loss of information and statistical power for meta-analytic effect size estimation. The standard error of each effect size estimate is compared to the division of the fixed-effects weighted average by 2.8. If the standard error is smaller than the quotient, a study has adequate power, as the true effect must differ at least 2.8 standard errors from zero when a 5% level of statistical significance and a power of 80% are assumed (1.96 [standard normal value for a significance value of 5%] + 0.84 [standard normal value for a statistical power of 80%]). The single effect size estimates of adequately powered studies are weighted by $1/se^2$ (Stanley et al., 2017). According to Carter et al. (2019), WAAP causes alarmingly high false-positive rates and overestimation of the true effect size when publication bias was present, although this effect was slightly reduced by higher heterogeneity and smaller sample sizes.

Although all presented meta-analytic methods fail under certain conditions, Carter et al. (2019) recommended using all of them and weighting their results according to an a priori method performance check based on expected conditions in the data set. This way it is transparent how conclusions would change if the research environment changed.

SCIENTIFIC CONTENT ANALYSIS

Updating the data set from Meta-Analysis 1 with further studies, we were able to additionally meta-analyze validation studies on SCAN, a verbal tool for credibility assessment developed by Avinoam Sapir, a former polygraph examiner (http://www.lsiscan.com/intro_to_scan.htm). SCAN is based on the assumption that experience-based and fabricated statements differ in language and structure. There is no standardized list that defines the number and operationalization of SCAN criteria. In the research context, sets of 12 or 13 criteria are commonly used (e.g., Bogaard, Meijer, Vrij, & Merckelbach, 2016; Nahari et al., 2012; Vanderhallen, Jaspaert, & Vervaeke, 2015). According to reports from SCAN course participants, some SCAN criteria are more likely to occur in experience-based statements and others in lies. Based on the criteria list provided by Nahari et al. (2012; Table 5), it seems that most criteria could be used both ways, either as indicators of experience-based or fabricated statements (e.g., use of pronouns: "*I* left the house" indicates the truth, whereas "Left the house" indicates deceit; Nahari et al., 2012).

Table 5

13 SCAN Criteria (Nahari et al., 2012)

Indicator for truth or lie	Criteria
Lie indicator	A change in terminology or vocabulary
Lie indicator	Emotions within the statement that are placed *before* the climax
Lie indicator	Omitting pronouns (e.g., “left the house” rather than “I left the house”)
Lie indicator	Lack of conviction and memory
Truth indicator	Denial of allegation
Lie indicator	Out of sequence information
Lie indicator	Ambiguous social introduction of involved persons (e.g., “we” instead of “me and my wife Lisa”)
Lie indicator	Spontaneous corrections (e.g., crossing out what has been written)
Truth indicator	Balanced structure of the statement (20% activities leading to the event, 50% actual event, 30% after the event)
Truth indicator	First-person singular past tense
Truth indicator	Correspondence of objective (actual duration of events) and subjective (amount of words used to describe these events) time
Lie indicator	Extraneous information
Lie indicator	Inclusion of words that indicate that some information is missing

Some SCAN criteria resemble CBCA criteria (e.g., *lack of conviction*, *spontaneous corrections*). Strikingly, according to SCAN, these criteria indicate a lie (and thus, remarkably resemble erroneous lay beliefs of which CBCA criteria should not be included if individuals aim to deceive; Maier et al., 2018), whereas according to CBCA they indicate an experience-based statement. This difference may be explained by the fact that SCAN is not grounded on explicitly formulated theoretical assumptions or empirical findings, whereas CBCA is based on the Undeutsch hypothesis (Undeutsch, 1967) from which an empirically-informed theoretical framework has been derived (Steller & Köhnken, 1989; Volbert & Steller, 2014).

So that the presence of SCAN criteria can be evaluated in a plain statement without interferences from an investigator, the suspect, witness, or alleged victim is asked to write down his/her version of what happened. Alternatively, the investigator can use the *Verbal Inquiry - the Effective*

Witness (VIEW) questionnaire that will, according to Sapir, "practically solve the case by itself" (http://www.lsiscan.com/intro_to_scan.htm). Unfortunately, there is no description of *how* the VIEW questionnaire actually achieves this challenging task. Although SCAN has repeatedly been criticized for its lack of scientific foundation and weak empirical evidence, it is widely used (e.g., Nahari et al., 2012). Sapir's homepage contains an impressively long list of SCAN course participants, ranging from the Federal Bureau of Investigation to the United States Department of Justice (http://www.lsiscan.com/id29.htm).

METHODS

DATABASE

INCLUSION AND EXCLUSION CRITERIA

We included unpublished and published English- and German-language studies that compared the quality of experience-based and fabricated statements using CBCA, RM, or SCAN. As in Meta-Analysis 1, we excluded laboratory studies that examined differences between experience-based and suggested statements.

KEYWORD SEARCH

We ran the keyword search in the following databases: PsycARTICLES, PsycINFO, and PSYNDEXplus Literature and Audiovisual Media. For English-language studies we used the following terms: "Criteria-Based Content Analysis", "CBCA", "Reality Monitoring", "RM", "Scientific Content Analysis", "SCAN", "Statement Validity Assessment", "SVA", OR "Validity Checklist"; AND "psychology of evidence", "statement analysis", "credibility", "credibility assessment", OR "deception". For German-language studies we used following keywords: "Kriterienbasierte Inhaltsanalyse", "CBCA", "Reality Monitoring", "RM", "Scientific Content Analysis", "SCAN", "Statement Validity Assessment", "SVA", OR "Validity Checklist"; AND "Aussagepsychologie", "Aussagebeurteilung", "Glaubhaftigkeit", OR "Glaubwürdigkeit". We did not translate some of the keywords into German, as the respective English technical terms have been established in the German literature. The keyword search in the databases was completed on November 21, 2017. In addition, we have contacted researchers on verbal credibility assessment and asked for their unpublished studies.

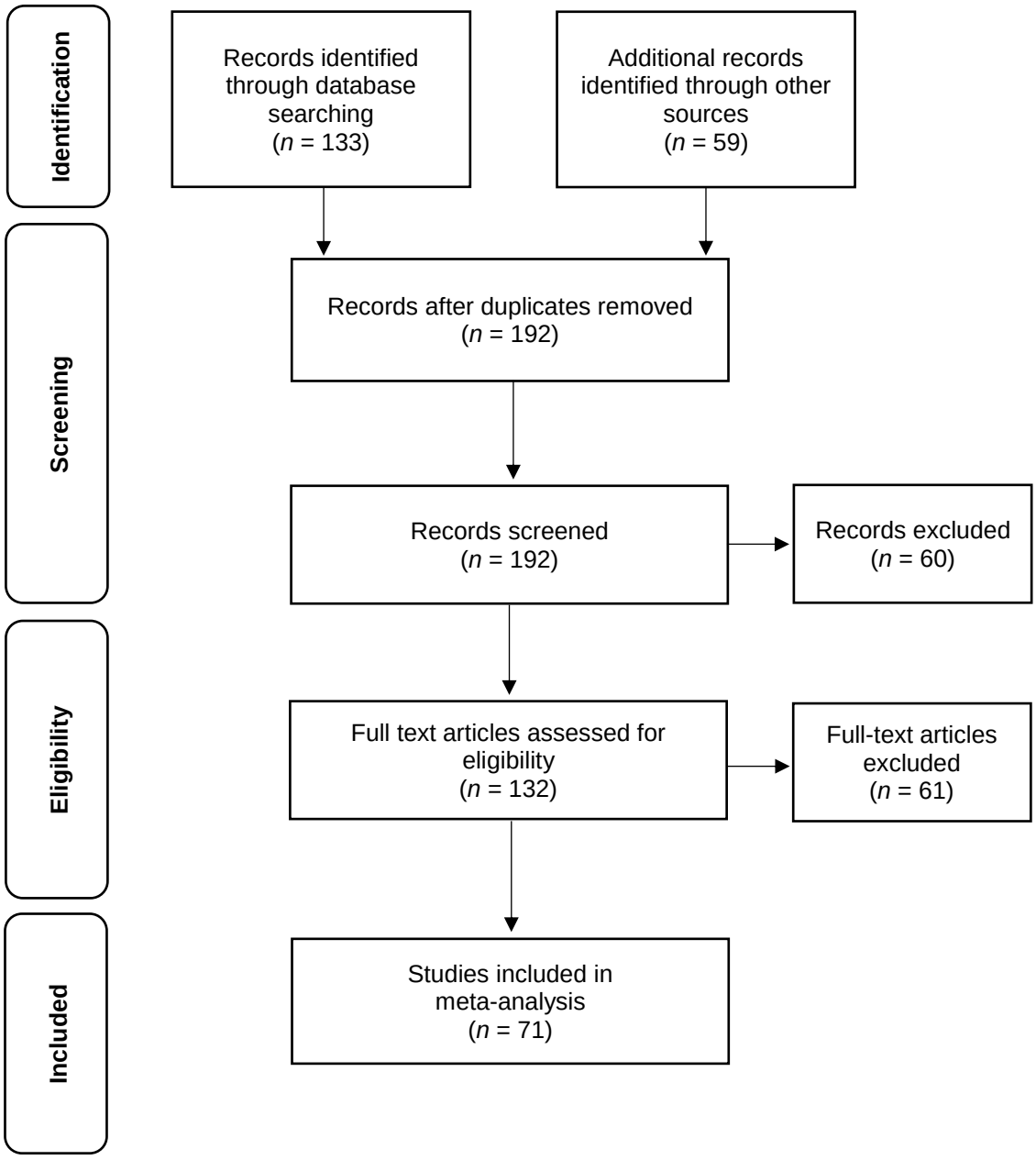

Figure 3. Full PRISMA diagram of the literature search of Meta-Analysis 2.

FINAL DATA SET

From a total of 192 identified studies, 71 matched the inclusion criteria (see Figure 3). We added 19 studies (study IDs 53-71) that were not incorporated in our first meta-analysis because they either had been published after our database search had been completed (54, 55, 56, 57, 58, 68) and/or because they investigated SCAN (57, 58, 59, 71). We also included studies that we had not identified in our first search because they were unpublished (61, 63, 64, 65, 66, 67, 68) or had not been included in any database (53, 60, 62, 70). In some studies, several comparisons were calculated based on one sample of experience-based and fabricated statements. To avoid the problem of dependent data, we made the following decision: If studies applied different verbal tools for credibility assessment to one data set, we included only one procedure per study to estimate the overall meta-analytic effect size and to run moderator analyses. That is, if studies investigated CBCA as well as RM and/or SCAN, we only included the effect size from CBCA results. If studies investigated RM and SCAN, we only included the effect size from RM results. In addition, we computed separate effect sizes for each procedure, each including all comparisons of one technique, to enhance statistical power. All other single case decisions can be found in the data table (column: description of effect size basis; see Appendix A).

MODERATOR VARIABLES

Moderator variables were identical to Meta-Analysis 1. To avoid repition, please refer to Meta-Analysis 1.

CODING PROCEDURE AND INTERRATER RELIABILITY

Two independent coders (first and second author of Oberlader et al., 2019) calculated effect sizes (Cohen's *d*, Hedges' *g*), standard errors, variances, inverse variance weights, *t*-values, and degrees of freedom. Based on a coding manual (see Appendix B), all moderator variables were rated. Intraclass correlation coefficients (two-way mixed, single measure) for continuous variables were 1.00 for all variables. Cohen's kappa for categorical variables ranged from .85 to 1.00. Cases of disagreement were discussed after computing the interrater reliability and a consent decision was made.

STATISTICAL ANALYSES

EFFECT SIZE MEASURE

The calcuation of effect sizes was identical to Meta-Analysis 1. To avoid repition, please refer to Meta-Analysis 1.

META-ANALYTIC METHODS

To investigate the robustness of the meta-analytical findings from Meta-Analysis 1, we used four bias-correcting meta-analytic methods in addition to REMA and trim-and-fill: PET-PEESE, *p*-uniform, 3PSM, and WAAP. To compare the reanalyzed results to our previous meta-analysis, REMA was calculated on the basis of *d*s and *g*s. For all other meta-analytic methods, we used only Hedges' *g*. Furthermore, we ran a *p*-curve analysis to test for publication bias and QRPs. For this purpose, we calculated the *t*-value for each study/effect size using the following formulas that we resolved to *t*: $d = t\sqrt{(1/n_1 + 1/n_2)}$ for between-subject designs and $d = t / \sqrt{n}$ for within-subjects designs (Lakens, 2013). However, we did not use *p*-curve for effect size estimation as it differs from *p*-uniform only in the estimation algorithm, but does not provide confidence intervals.

To illustrate the practical significance of the effect size estimates, we additionally calculated the CLES (McGraw & Wong, 1992) where applicable. Since the CLES requires the *n*s of the experience-based and fabricated statements and these were not available for the study subsets of some meta-analytic methods (trim-and-fill, *p*-uniform, and WAAP), CLES could only be calculated for REMA, PET-PEESE, and 3PSM.

In addition to effect size estimates, we provide information on the 95% CI, the z-statistic, the number of independent studies (k), and, where possible, the total number of statements (n). Moreover, we report z_m-statistic of moderation tests and τ as a measure of heterogeneity, which reflects the standard deviation of the distribution of true effect sizes under the assumption that true effect sizes are normally distributed.

TEST FOR OUTLIERS

The test of outliers was identical to Meta-Analysis 1. To avoid repition, please refer to Meta-Analysis 1.

METHOD PERFORMANCE CHECK

Using the Meta-Showdown Explorer (http://shinyapps.org/apps/metaExplorer/), we investigated the performance of the candidate methods under plausible conditions of the present data. According to Carter et al. (2019) results derived using any given method should be treated with caution if the Meta-Showdown Explorer showed poor performance. On the other hand, if a meta-analytical method showed good performance in all plausible research environments, this strengthens confidence in the conclusions drawn from such a result.

To this end, specifications must be selected for five areas: a) severity of publication bias, b) amount of heterogeneity, c) number of studies in meta-analysis, d) true effect size under H1, and e) QRPs environment. Since we actively searched for unpublished studies by requesting researchers in the field of verbal credibility assessment share their unpublished work and were thus able to include a larger number of unpublished studies, we assumed no to medium severity of publication bias. From our first meta-analysis we know that there is a high degree of heterogeneity among studies, so we chose the highest value ($\tau = 0.40$) from the available options. We included 71 studies in our meta-analysis. Thus, we selected the default options 60 and 100 for the number of studies. Our first meta-analysis resulted in an uncorrected estimate of $d = 1.00$ [95% CI [0.75, 1.25]). Since this result could be biased by potential publication bias and QRPs, we conservatively assumed a medium ($d = 0.50$) to high ($d = 0.80$) true effect under H1. Since the dependent variable, i.e., the score of verbal tools for credibility assessment, can be calculated in different ways (e.g., rating of absence/presence vs. rating of quality, inclusion of all criteria vs. a selection of criteria), a medium environment for QRPs was assumed.

After specifying the basic settings of the assumed research environment, criteria that define acceptable performance thresholds for a meta-analytical method must be chosen. We chose a maximum deviation of the average estimate from the true delta of 0.3 and a false-positive rate of $\leq 5\%$ as criteria for good performance. Based on this definition of good performance, the following performance characteristics were supposed: The performance of the REMA, PET-PEESE, *p*-uniform, and WAAP was poor under the H0 and under the H1. Trim-and-fill performed well under the H0 and H1 if there was no publication bias. Notably, 3PSM was the only bias-correction method that performed well under H0 and H1 in all plausible conditions. Hence, in the case that meta-analytical methods lead to conflicting conclusions, we give the strongest weight to the 3PSM results. In addition, we used only this method for moderator and subset analyses.

SOFTWARE

To calculate the interrater reliabilities, we used IBM SPSS Statistics 24. For the meta-analytical calculations, we used the following R packages in R Statistical Software (version 3.4.1; R Core Team, 2017): compute.es (AC Del Re, 2013), metafor (Viechtbauer, 2010), META (Schwarzer, 2007), *p*-uniform (van Aert, 2018), and weightr (Coburn & Vevea, 2017). In addition, we used the *p*-checker app (Schönbrodt, 2018) and the *p*-curve app 4.06 (http://www.p-curve.com/app4/) for *p*-curve analysis, and Meta-Showdown Explorer (http://shinyapps.org/apps/metaExplorer/) for the method performance check.

RESULTS

OVERALL EFFECT SIZE ESTIMATION

REMA

Table 6 displays the overall effect size estimates for the different meta-analytic methods. A REMA on the updated set of 71 studies showed that verbal tools for credibility assessment discriminated statistically significantly between experience-based and fabricated statements with a large point estimate and moderate to large effects in the confidence interval and high heterogeneity between studies. In about 75% of the cases, experience-based statements had descriptively higher scores than fabricated statements. No study was excluded as a statistical outlier.

Table 6

Overall effect size estimates of the different meta-analytic methods

Meta-analytic method	ES [95% CI]	Test statistic	*p*	τ	*k*	CLES (%)
REMA	0.96[a] [0.77, 1.15]	9.68[c]	< .001	0.76	71	75
	0.94[b] [0.75, 1.13]	9.71[c]	< .001	0.74	71	75
Trim-and-fill (R_0)	0.60[b] [0.35, 0.84]	4.74[c]	< .001	1.11	86	
PET-PEESE	-0.11[b] [-0.43, 0.22]	-0.65[d]	.516		71	47
p-uniform	1.04[b] [0.89, 1.20]	-8.51[e]	< .001		48	
3PSM	0.90[b] [0.61, 1.19]	6.32[c]	< .001	0.74	71	74
WAAP	0.47[b] [0.32, 0.62]	0.07[f]	< .001		24	

Note. Some values cannot be specified for all meta-analytic methods (cells are left empty). ES = effect size. [a]Based on Cohen's *d*. [b]Based on Hedges' *g*. [c]*z*-statistic. [d]*t*-statistic. [e]L.0 statistic. [f]WAAP test statistic.

TRIM-AND-FILL

The two trim-and-fill estimators yielded different results: The L_0 estimator indicated that no studies needed to be filled in. The R_0 estimator showed 15 missing studies. A REMA on the R_0-supplemented data set of 86 studies revealed a moderate point estimate for the bias-corrected effect size with small to large effects in the confidence interval and high heterogeneity between studies (Table 6).

PET-PEESE

As the PET intercept did not differ statistically significantly from zero, PET is preferable over PEESE, which revealed a small significant bias-corrected effect size (g = 0.39, 95% [0.21, 0.56], p < .001). Thus, the PET-PEESE estimator indicated that verbal tools for credibility assessment did not distinguish statistically significantly between experience-based and fabricated statements: The probability that experience-based statements had a higher score than fabricated statements was descriptively below the level of chance at 47% (Table 6). The correlation between effect size and standard error was r = .49, p < .001 (Figure 5).

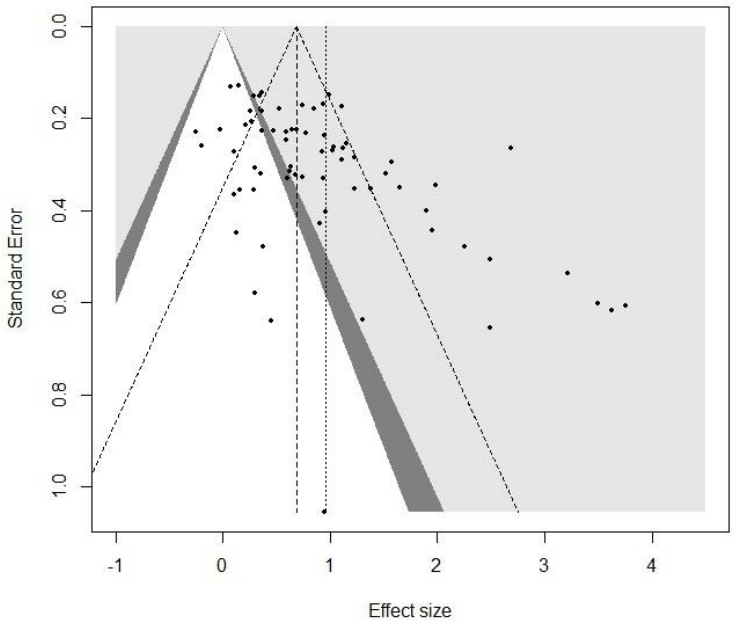

Figure 5. Scatterplot of effect size and standard error (r = .49, p < .001).

P-UNIFORM

Based on the statistically significant studies (k_{sig} = 48), p-uniform revealed a large point estimate with large effects in the confidence interval (Table 6).

3PSM

According to the adjusted 3PSM model, verbal tools for credibility assessment discriminated statistically significantly between experience-based and fabricated statements with a large point estimate and moderate to large effects in the confidence interval and high heterogeneity between studies. In about 74%, experience-based statements had descriptively higher scores than fabricated statements (Table 6). There were 46 studies in the one-tailed p-values < .025 interval and 25 studies in the .025 < one-tailed p-values < 1 interval.

WAAP

WAAP indicated that 24 studies had an adequate power > 80%. Hence, roughly a third of the updated data set could be used for effect size estimation that revealed a small point estimate and small to moderate effects in the confidence interval (Table 6).

P-CURVE ANALYSIS

The *p*-checker app showed a success rate of 74%. This means the rate of statistically significant studies was below the median observed power (87%). Accordingly, there was no inflation of statistically significant studies. Figure 6 shows the distribution of significant *p*-values (k = 50)[6]. The distribution was right-skewed, i.e., comparatively more studies fell into the range of low *p*-values.

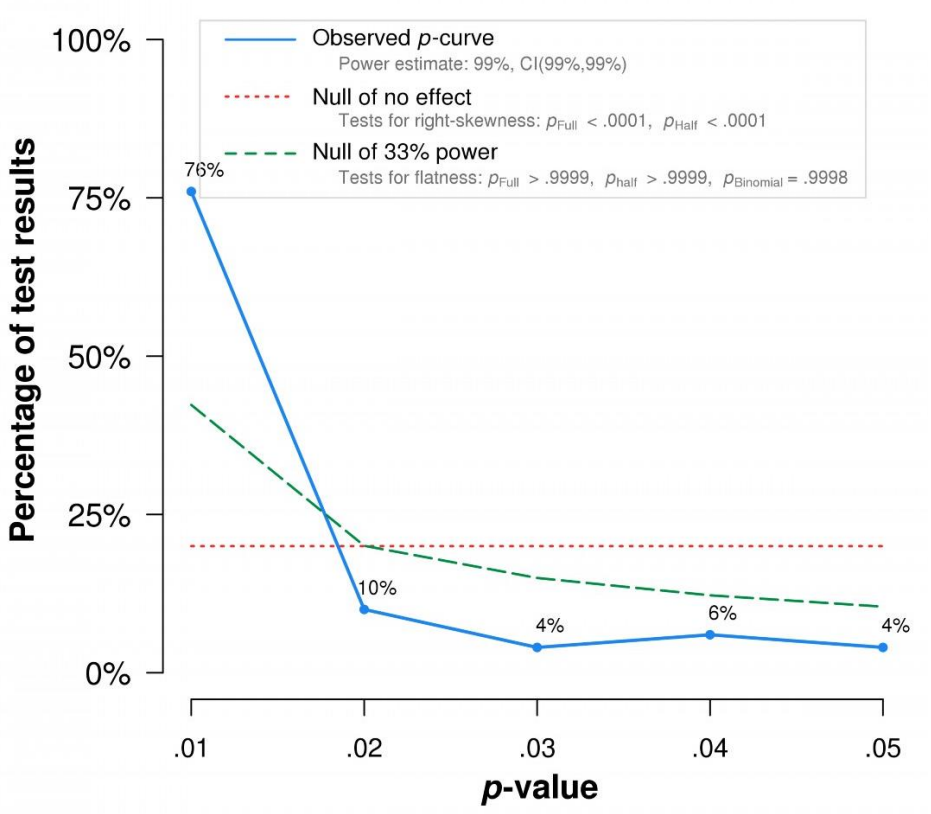

Figure 6. Graphical display of the *p*-curve. The observed *p*-curve includes 50 statistically significant (p < .05) results, of which 45 are p < .025.

[6] Due to different implementations in the computation of the *p*-value, two *p*-values were just significant in the *p*-curve analysis (Study ID 10 p = .049; Study ID 55 p = .047), leading to 50 included studies, and not significant in the *p*-uniform analysis (Study ID 10 p = .078; Study ID 55 p = .027, one-tailed), leading to 48 included studies.

EFFECT SIZE ESTIMATION PER PROCEDURE

Figure 4 displays the forest plot of the effect sizes, separate for CBCA, RM, and SCAN. Point estimates ranged from -0.25 to 3.66. Six point estimates were negative, i.e., in contrast to the hypothesis, but statistically not significantly different from zero; 28 confidence intervals included negative effect sizes.

CBCA

The adjusted 3PSM model revealed a large point estimate of $g_{adjusted} = 0.82$ (95% CI [0.48, 1.17], $z = 5.01$, $p < .001$, $\tau = 0.75$, $k = 55$, $N_{statements} = 3{,}008$). There were 36 studies in the one-tailed p-values < .025 interval and 19 studies in the .025 < one-tailed p-values < 1 interval. In about 73%, experience-based statements had descriptively higher scores than fabricated statements.

RM

The adjusted 3PSM model revealed a large point estimate of $g_{adjusted} = 0.73$ (95% CI [0.39, 1.06], $z = 4.54$, $p < .001$, $\tau = 0.46$, $k = 23$, $N_{statements} = 1{,}977$). There were 18 studies in the one-tailed p-values < .025 interval and five studies in the .025 < one-tailed p-values < 1 interval. In about 70%, experience-based statements had descriptively higher scores than fabricated statements.

SCAN

The adjusted 3PSM model revealed a large but statistically non-significant point estimate of $g_{adjusted} = 1.40$ (95% CI [-0.07, 2.87], $z = 1.87$, $p = .062$, $\tau = 0.46$, $k = 7$, $N_{statements} = 404$). In about 84%, experience-based statements had descriptively higher scores than fabricated statements. There were two studies in the one-tailed p-values < .025 interval and five studies in the .025 < one-tailed p-values < 1 interval. With only two studies in the one-tailed p-value < .025 interval, results should be interpreted with caution.

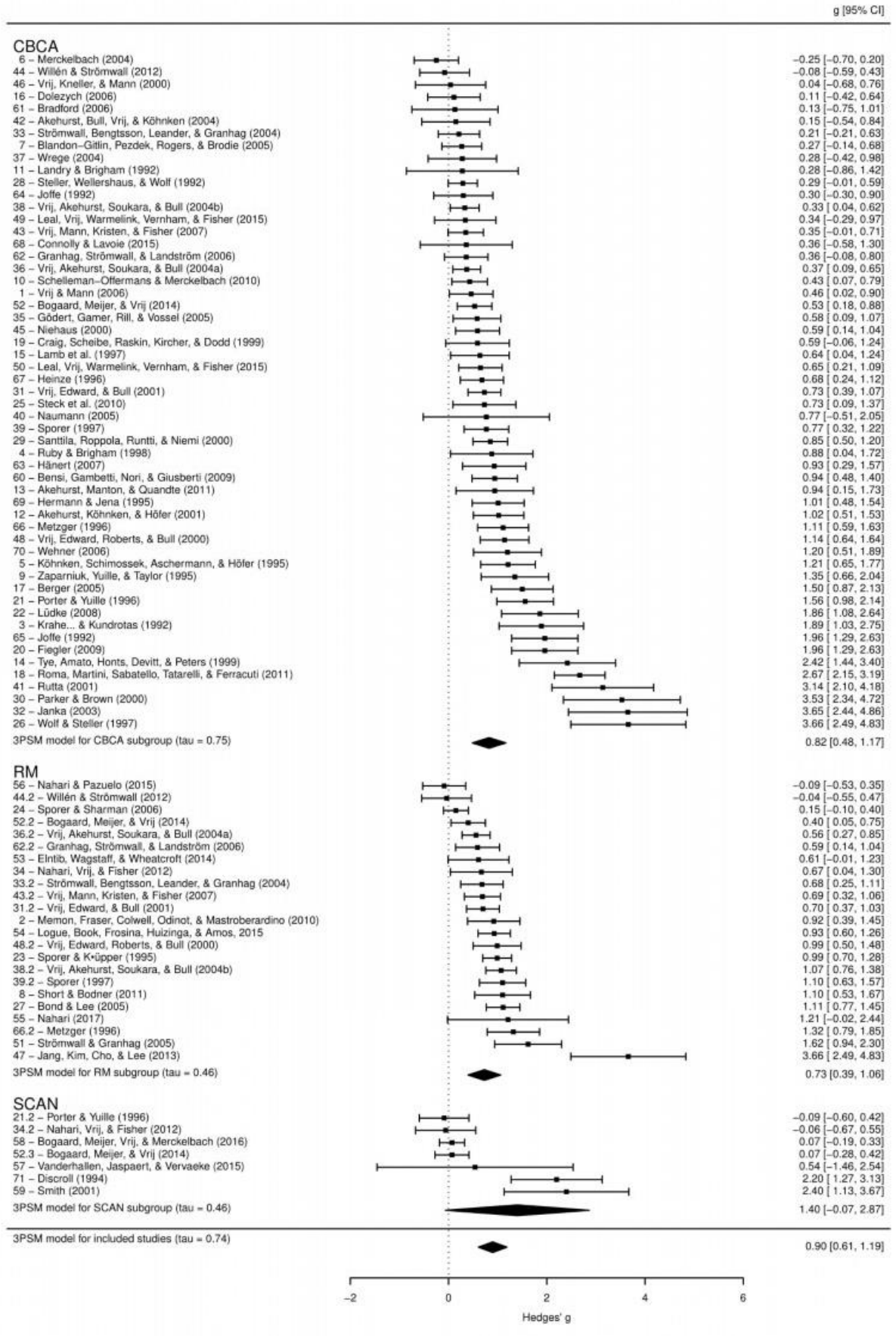

Figure 4. Forest plot separate for CBCA, RM, and SCAN including subset effect sizes and overall effect size estimation based on 3PSM.

MODERATOR ANALYSES

Table 7 displays the results of categorial moderator analyses with 3PSM. Despite the subsets *a priori decision rule*, *discriminant analysis without cross-validation*, and *discriminant analysis with cross-validation*, all moderator subsets showed statistically significant positive effect sizes. Although the moderator analyses was intended to explain true variance between studies, the heterogeneity within the subsets was still high ($\tau \geq 0.62$).

The χ_m-tests of moderation revealed three statistically significant results. First, moderator analyses with 3PSM showed statistically significantly larger effect sizes for field studies than for laboratory studies. In addition, effect size estimates based on mean differences between experience-based and fabricated statements were statistically significantly smaller than effect size estimates based on classifications by statistical or rater decisions. Studies that compared mean scores showed a moderate point estimate. Classifications based on statistical decisions or rater decisions revealed large effects. Studies that used a priori decision rules did not differ statistically significantly from the other moderator subsets. Finally, moderator analyses with 3PSM showed that the complete set of CBCA criteria outperformed any incomplete set. There was a large effect for studies using 19 CBCA criteria compared to a moderate effect for studies using incomplete versions. Studies using the 14-item version of the CBCA by Raskin et al. (1991) did not differ statistically significantly from the other moderator subsets.

All other categorical moderations were statistically non-significant. Only one of these results should be mentioned here, as it is a focal moderator closely related to the issue of publication bias: Effect sizes for published and unpublished studies yielded no statistically significant difference. Moreover, continuous moderators, the year of publication ($\chi = 1.78$, $p = .074$, $k = 71$) and the sex ratio in the sample ($\chi = 0.76$, $p = .445$, $k = 53$) were statistically non-significant.

Table 7

Results of moderator analyses with 3PSM

Moderator	Moderator categories	*g*	95% CI	z_m / *z*	τ	*k*	*n*	Study IDs
Age of participants				0.41				
	< 18	0.97	[0.52, 1.42]	4.24***	0.76	17	1,238	7, 14, 15, 18, 19, 28, 29, 33, 42, 45, 51, 62–65, 67, 68
	≥ 18	1.07	[0.73, 1.41]	6.13***		44	2,881	1–6, 8–11, 13, 16, 20–22, 24–27, 30–32, 34, 37, 39–41, 43, 44, 46–50, 52–58, 60, 61, 70
Motivating incentive				0.67				
	No	0.86	[0.58, 1.14]	6.00***	0.62	48	3,431	2, 4–12, 14, 17, 22–27, 29, 31–34, 39, 40, 42, 44–53, 56, 58, 60, 62–70
	Yes	0.71	[0.28, 1.14]	3.24**		13	1,171	1, 16, 21, 28, 35–38, 41, 43, 54, 55, 57
Experience status				0.68				
	Event not personally experienced	1.01	[0.47, 1.55]	3.67***	0.63	7	478	5, 9, 17, 31, 46, 48, 60
	Event personally experienced	0.82	[0.55, 1.10]	5.91***		56	4,194	1, 2, 4, 6–8, 10–12, 14, 16, 20–29, 32–45, 47, 49–58, 61–70
				1.07				
	Not accused	0.81	[0.53, 1.09]	5.70***	0.63	55	4,122	2, 4–12, 14, 16, 17, 20, 22–29, 31–33, 35–42, 44–46, 48–53, 55–58, 60, 62–69
	Accused	1.11	[0.57, 1.66]	3.99***		7	530	1, 21, 34, 43, 47, 54, 70
Participant training				0.95				
	No	0.89	[0.62, 1.17]	6.44***	0.63	55	4,097	1, 2, 4–12, 14, 16, 17, 20–29, 31–36, 39, 42–45, 47–49, 51, 53–55, 57, 58, 60–70
	Yes	0.62	[0.04, 1.20]	2.09*		7	492	37, 38, 40, 41, 46, 52, 56

Table 7 (*continued*)

Moderator	Moderator categories	*g*	95% CI	z_m / *z*	τ	*k*	*n*	Study IDs
Event characteristics (negative tone, personal involvement, loss of control)				0.60				
	At least one missing	0.88	[0.60, 1.16]	6.13***	0.63	47	3,644	1, 2, 5, 7–10, 12, 14, 20–27, 29, 31, 33–36, 38–40, 42–44, 46–52, 54, 55, 60, 62–66, 68–70
	All three met	0.75	[0.30, 1.19]	3.31**		14	928	4, 6, 11, 16, 17, 28, 32, 37, 41, 45, 53, 57, 58, 67
Type of lie				0.41				
	Concealment	0.93	[0.48, 1.39]	4.01***	0.64	12	781	2, 31, 34, 35, 47, 48, 54, 57, 60, 61, 67, 68
	Outright	0.84	[0.55, 1.12]	5.75***		49	3,791	1, 4–12, 14, 16, 17, 20, 22–24, 26–29, 32, 33, 36–46, 49–53, 55, 56, 58, 62–66, 69, 70
Statement mode				0.39				
	Oral	0.93	[0.62, 1.24]	5.88***	0.74	57	3,776	1–5, 7–9, 11–21, 25–33, 35–46, 48–51, 54, 55, 60–70
	Written	0.83	[0.32, 1.33]	3.22**		13	1,243	6, 10, 22–24, 34, 47, 52, 56–59, 71
Type of rater				all *z*s ≤ 1.66				
	Laypersons	0.65	[0.06, 1.23]	2.17*	0.65	7	495	6, 20, 22, 24, 34, 42, 61
	Trained participants	0.81	[0.51, 1.11]	5.26***		42	3,182	2, 4, 5, 8–13, 15, 19, 21, 23, 25, 26, 28, 31–33, 35, 36, 38–40, 43–46, 48, 51, 52, 55, 56, 58, 60, 62–65, 68–70
	Professionals	1.16	[0.77, 1.56]	5.79***		17	1,060	1, 3, 7, 16–18, 29, 30, 37, 41, 49, 50, 53, 54, 57, 59, 66

Table 7 (*continued*)

Moderator	Moderator categories	*g*	95% CI	z_m		τ	*k*	*n*	Study IDs
				z					
Set of CBCA criteria				1 < 2	2.58**				
				all other *z*s	≤ 1.41				
	Incomplete sets[1]	0.66	[0.31, 1.02]	3.69***			37	2,519	1, 4–7, 9–14, 16, 19, 21, 25, 26, 28, 35, 36, 38, 39, 42, 43, 46, 49, 50, 52, 60–68, 70
	19 CBCA criteria[2]	1.33	[0.85, 1.81]	5.45***		0.68	13	585	3, 17, 20, 22, 30, 32, 33, 37, 40, 41, 44, 45, 69
	14-item version	1.17	[0.50, 1.83]	3.44**			5	552	15, 18, 29, 31, 48
Scoring criteria				1.24					
	Absence/ presence	1.22	[0.78, 1.66]	5.97***			17	1,046	7, 9, 11, 15, 18–20, 30, 31, 46, 47, 49–51, 53, 54, 57
	Scoring on a scale	0.93	[0.61, 1.26]	5.68***		0.74	48	3,612	1–6, 8, 10, 12–14, 16, 17, 21–24, 26, 28, 29, 32, 33, 35–45, 48, 52, 55, 56, 58, 60–66, 68, 70, 71
Decision basis				1 > 3	3.84***				
				2 > 3	2.85**				
				all other *z*s	≤ 1.17				
	Discriminant function[1]	1.31	[0.99, 1.63]	8.06***			24	1,683	2–5, 9, 12–14, 21, 23, 26, 27, 29, 32–34, 39, 41, 47, 51, 58, 66, 67, 70
	Rater decision[2]	1.34	[0.83, 1.84]	5.21***		0.63	10	402	8, 11, 17, 20, 30, 40, 45, 59, 64, 65
	Mean comparison[3]	0.56	[0.25, 0.88]	3.54***			35	2,942	1, 6, 7, 10, 15, 16, 18, 19, 22, 24, 25, 28, 31, 35–38, 42–44, 46, 48–50, 52–56, 60–63, 68, 69
	A priori rule	0.73	[-0.21, 1.68]	1.52			2	34	57, 71

Table 7 (*continued*)

Moderator	Moderator categories	*g*	95% CI	z_m / *z*	τ	*k*	*n*	Study IDs
Decision basis	DF without cross-validation	0.73	[-0.51, 1.97]	1.16	1.21	22	1,508	3, 4, 9, 12–14, 21, 23, 26, 27, 29, 30, 32–34, 39, 41, 47, 51, 58, 66, 70
	DF with cross-validation	-0.07	[-2.53, 2.51]	0.05		2	119	2, 5
Study design				3.08**				
	Field studies	1.76	[1.20, 2.33]	6.11***	0.67	8	389	3, 13, 15, 18, 19, 30, 59, 71
	Laboratory studies	0.84	[0.56, 1.11]	5.97***		63	4,672	1, 2, 4–12, 14, 16, 17, 20–29, 31–58, 60–70
				1.54				
	Within-subjects	0.73	[0.34, 1.11]	3.67***	0.72	23	2,020	4, 6, 8, 10, 11, 16, 17, 23, 27–29, 31, 32, 37, 39, 41, 44, 52, 53, 58, 60, 61, 67
	Between-subjects	1.03	[0.73, 1.34]	6.63***		48	3,041	1–3, 5, 7, 9, 12–15, 18–22, 24–26, 30, 33–36, 38, 40, 42, 43, 45–51, 54–57, 59, 62–66, 68–71
Publication status				1.28				
	Unpublished	1.16	[0.71, 1.61]	5.08***	0.72	16	698	16, 17, 20, 22, 25, 32, 37, 40, 41, 61, 63–67, 69
	Published	0.87	[0.57, 1.17]	5.70***		55	4,363	1–15, 18, 19, 21, 23, 24, 26–31, 33, 34, 35, 36, 38, 39, 42–60, 62, 68, 70, 71

Note. DF = discriminant function. *p < .05, **p < .01, ***p < .001.

DISCUSSION

The informative value of a meta-analysis is limited by the quality of the underlying data. If a study set is biased, meta-analytic estimations of the true effect will be biased, too, unless the method is robust against biases. Within a comprehensive simulation study Carter et al. (2019) investigated the performance of different meta-analytic methods under varying conditions. Since results demonstrated that both methods used in Meta-Analysis 1, REMA and trim-and-fill, performed poorly when publication bias was present, we reanalyzed our data with further bias-correcting meta-analytic methods in an updated data set. A priori, we performed a method performance check to weight the results of individual methods according to their suitability for the present data set. So, what does the evidence tell us? Can we still claim that verbal tools for credibility assessment are valid procedures to distinguish between experience-based and fabricated statements? In short, this question can be affirmed. However, this affirmation needs to be qualified by the specific meta-analytic method, boundary conditions, and the examined procedures.

HOW ROBUST ARE META-ANALYTIC FINDINGS ACCORDING TO DIFFERENT META-ANALYTIC METHODS?

REMA

In line with our previous findings based on 52 studies, a REMA on the expanded data set of 71 studies revealed a large point estimate with moderate to large effects in the confidence interval. In other words, in 75% of the cases experience-based statements had descriptively higher scores than fabricated statements. However, according to the method performance check these results are only to be trusted if biases in the data set can be ruled out. Although the results of *p*-curve and moderator analyses gave no indication of publication bias, it is not entirely possible to exclude its presence. Therefore, on the basis of REMA no conclusion on the validity of verbal tools for credibility assessment is possible.

TRIM-AND-FILL

According to the R_0 estimator of trim-and-fill, verbal tools for credibility assessment distinguish between experience-based and fabricated statements with a moderate point estimate and small to large effects in the confidence interval. However, as discussed in Meta-Analysis 1, it seems rather unlikely that unpublished studies would show exclusively negative effects and that verbal tools for credibility assessment would mistakenly point in the exact opposite direction. Moreover, based on the method performance check, trim-and-fill would only be suitable if there were no

publication bias. Again, since we cannot entirely exclude publication bias, the informative value of the trim-and-fill estimate is limited.

PET-PEESE

The bias-correcting meta-regression approach PET-PEESE showed a null effect corroborating that verbal tools for credibility assessment cannot discriminate between experience-based and fabricated statements. Effect size and standard error were positively correlated (see Figure 5), i.e., studies with larger effect size estimates had less statistical power to actually find these effects. This relationship could result from QRPs, i.e., from (non-)intentional analysis strategies that incorrectly pushed *p*-values below the significance threshold of $p < .05$. However, there might be other reasons for this correlation. For example, studies with larger effects might have been of higher quality than studies with smaller effects. A higher quality study design might dovetail a more laborious (recruitment) procedure and, therefore, ultimately result in attenuated sample sizes. We coded the quality of the studies according to various criteria. For example: Was it an ecologically valid field study or a laboratory study? Were the users of the procedures experts? Was the event characterized by negative tone, personal involvement, and a certain loss of control? However, graphical inspection of scatterplots of effect size and standard error within moderator subsets demonstrated that none of these quality criteria provided an explanation for the positive correlation between both statistics: Even in subsets indicating a high study quality, positive associations occurred[7]. Since study designs differed in many aspects, it is conceivable that further quality differences exist. Apart from that, according to the method performance check, PET-PEESE performance was poor under all selected conditions. Consequently, the conclusion of a null effect, i.e., the possibility that verbal tools for credibility assessment cannot distinguish between experience-based and fabricated statements, should be viewed with skepticism.

P-UNIFORM

Based on the results of *p*-uniform it can be assumed that verbal tools for credibility assessment are able to discriminate between experience-based and fabricated statements to a large extent. However, this method only includes statistically significant studies. In our view, however, a meta-analysis needs to consider *all* performed studies. Otherwise, the estimate is upward biased. In addition, the method performance check showed that *p*-uniform performed poorly under all selected conditions. Hence, this estimate is of limited value.

[7] With regard to the small subsets of the moderator categories, correlation coefficients were not calculated.

WAAP

According to WAAP, it can be assumed that verbal tools for credibility assessment distinguish between experience-based and fabricated statements with a moderate point estimate and small to moderate effect sizes in the confidence interval. However, WAAP performance was poor under all conditions of the method performance check. Its interpretability is thus limited, at least under the presently supposed boundary conditions.

3PSM

According to the adjusted 3PSM model, verbal tools for credibility assessment distinguish between experience-based and fabricated statements with a large effect size and moderate to large effects in the confidence interval. In about 74%, experience-based statements had descriptively higher scores than fabricated statements. The method performance check showed that 3PSM performed well under all selected conditions.

INTEGRATION OF THE META-ANALYTIC FINDINGS

Now we have a multitude of results to answer the question of whether verbal tools for credibility assessment work. These results range from a null effect to a large point estimation. While REMA, trim-and-fill, *p*-uniform, WAAP, and 3PSM indicated that the procedures work more or less well, PET-PEESE showed a null effect. However, according to the method performance check with the Meta-Showdown Explorer, only 3PSM performed well under the preselected conditions, all other meta-analytic findings were of limited value. Based on this weighting, it can be assumed that verbal tools for credibility assessment are valid to a substantial degree. Yet, we must consider high heterogeneity between studies that could not be resolved by separate analysis per procedure or moderator analyses. This limits the interpretability of point estimates using 3PSM. The question arises whether the included studies are actually a sample of *one* population. As discussed in Meta-Analysis 1, differences in the study designs, although not identified by moderator analyses, might have addressed different underlying mechanisms of truth telling and lying, so that the same measures performed differently in individual studies.

ARE CBCA, RM, AND SCAN EQUALLY VALID?

SCAN

Based on a meta-analysis using 3PSM on the subset of seven SCAN studies, we conclude that this procedure cannot discriminate between experience-based and fabricated statements and should therefore not be used in practice. Although Sapir's homepage advertises that "SCAN will solve every case […] quickly and easily" (http://www.lsiscan.com/intro_to_scan.htm), the lack

of any theoretical and/or empirical basis indicates that SCAN cannot be considered to be a scientific instrument. Vrij (2015) concluded that the Daubert standards are not met for SCAN, except for the first question of whether its validity can be tested – however, the quality of research and its results are not sufficient. This assumption was supported by our findings: The adjusted 3PSM model revealed a large but statistically non-significant point estimate. The CLES of 84% is thus misleading. Only two studies fell into the interval of significant *p*-values (Driscoll, 1994; Smith, 2001). Both were field studies that used questionable criteria to establish ground truth. Driscoll (1994) used, among others, results of polygraph tests to determine the truth status of the statements. The polygraph tests have not been described, but since available polygraph tests have been heavily criticized (e.g., Meijer et al., 2016), the validity of this ground truth criterion is in question. Smith (2001) used, among others, the criterion *police dropped the case*. It is unclear to what extent the assessment of the credibility of statements (among other factors, such as unbalanced ground truth ratios) had influenced the police investigations. The assessment of a statement as deceptive could have led to dropping the case. In this circumstance, the ground truth criterion *police dropped the case* would not have been independent from the quality assessment of statements. Against this background, it is debatable whether the two positive study effects are valid at all. Either way, it must be stressed that the widespread use of SCAN is alarming if compared to the small number of empirical studies and their aggregated effect size.

CBCA

A meta-analysis using 3PSM on the subset of CBCA studies revealed a large point estimate with moderate to large effects in the confidence interval indicating that in 73% of the cases experience-based statements had descriptively higher scores than fabricated statements. Even though this rate is well below 100%, the CBCA is one of the best methods to distinguish between experience-based and fabricated statements (see comparison to other measures in CONCLUSION of Meta-Analysis 1). Furthermore, it should be noted that its application is only one part of the SVA. Beyond the assessment of statement quality, personal and contextual variables are taken into account (e.g., Steller, 1989), although the validity of this largely unstructured procedure has yet to be empirically demonstrated.

RM

As for CBCA, a meta-analysis using 3PSM on the subset of RM studies revealed a large point estimate with small to large effect sizes in the confidence interval indicating that in about 70% experience-based statements had descriptively higher scores than fabricated statements. The same applies here as to the CBCA: Even though this rate is far below 100%, RM is one of the best

methods available to discriminate between experience-based and fabricated statements (see comparison to other measures in CONCLUSION of Meta-Analysis 1). However, RM has so far only been investigated in laboratory studies, the examination in the field is still pending.

WHAT ARE OPTIMAL BOUNDARY CONDITIONS?

FIELD VS. LABORATORY STUDIES

Moderator analyses revealed that field studies outperformed laboratory studies. This finding corroborates the validity of verbal tools for credibility assessment for practical application: Experience-based and fabricated statements made in real cases could have been distinguished, at least to an acceptable extent. However, it should be noted again that in many cases the independence of ground truth criteria and statement quality is questionable and that the results may therefore be overestimated. On the other hand, this finding underpins the difficulty of experimentally imitating real cases in the laboratory. It is unethical to create scenarios resembling the experience of sexual abuse or similar offences of forensic interest. In order to approximate real case scenarios in an ethically acceptable manner, Steller (1989) recommended designing laboratory situations that affect participants personally, are perceived negatively, and are accompanied by a certain degree of loss of control (e.g., being witness of a wallet theft while playing a computer game, birth). Moderator analyses revealed no difference between studies that operationalized these criteria and those that did not. However, as the example "being witness of a wallet theft while playing a computer game" shows, even when these criteria are taken into account, it is difficult to create a situation in the laboratory that is even remotely as invasive as a real case scenario involving sexual victimization.

NUMBER OF CBCA CRITERIA

Moderator analyses on the extended data set confirmed the previous finding that the complete version of 19 CBCA criteria outperformed any incomplete criteria set. As explained above, this difference may be due to differences in the ecological validity of study designs. Studies that investigate the complete set of criteria usually provide settings in which all criteria could occur. On the other hand, studies that exclude individual criteria show less realistic settings. Independent of this, according to classical test theory the test length of the entire criteria set might have increased the validity.

DECISION BASIS

As in Meta-Analysis 1, we found statistically significant differences between studies comparing mean scores of experience-based and fabricated statements and studies classifying statements via

statistical decisions. Again, the latter worked better, which, as described in detail above, could be explained by the weighting of criteria in discriminant analysis. However, as previously noted, criteria weighting needs to be cross-validated, as it is modeled and tested on the same sample. Yet, only three of the 24 studies using discriminant analysis carried out cross-validation. Moderator analyses showed no difference between studies that carried out cross-validation and those that did not, which could have been due to the low statistical power of the subsets of studies carrying out cross-validation. It is therefore highly likely that the effect size of studies using statistical decisions is overestimated. In addition to these results, moderator analyses showed that classifications based on rater decisions outperformed mean score comparisons. Even though the underlying psychological processes of rater decisions are untraceable, it can be assumed that criteria are taken into account to varying degrees – either weighted by their diagnostic value (Maier et al., 2018) or according to personal and contextual aspects (e.g., Volbert & Steller, 2014) – or are even partially ignored. As with discriminant analysis, this could be the explanation for larger effect sizes.

FURTHER MODERATORS

As for the data set of 52 studies, further moderator analyses (i.e., on publication status, year of publication, participants' age, motivation, experience status, role in the interview, and training status, as well as on type of lie, statement mode, experience of raters, and scoring of criteria) yielded no statistically significant results. Again, for some of these variables, this might be due to low-powered (small subsets) and heterogeneous moderator categories, and others might not have been identified. However, this does not mean that there are no other moderators (see DISCUSSION of Meta-Analysis 1).

WHAT SHOULD BE CONSIDERED FOR FUTURE RESEARCH?

In the discussion of Meta-Analysis 1, we pointed to several aspects that future studies need to take into account in order to make progress in research on verbal tools for credibility assessment and to empirically legitimize their use: standardization of procedures, consideration of several boundary conditions, and design of ecologically valid studies. These calls do also apply on the basis of the extended data set in Meta-Analysis 2.

In addition, the method performance check demonstrated the importance to reduce publication bias and QRPs. Accordingly, the performance of most meta-analytic methods decreased dramatically when a certain amount of bias was preselected in the Meta-Showdown Explorer. Although the results of *p*-curve analysis and moderator analysis gave no indication of bias in the present data set, its existence cannot be entirely ruled out. In order to further reduce the

possibility of publication bias and QRPs, future studies should be pre-registered. If, for example, it is determined in advance how many statements are gathered, on what basis these are compared (i.e., which criteria are coded and how), and which analyses are performed, researchers' degrees of freedom decrease. This way, pre-registration protects against (un)intentional *p*-hacking. The prevention of validity-threatening biases of course also requires structural changes in the scientific community (e.g., changes in publication practice, changes in how researchers' performance is assessed). Moreover, future studies must include larger samples. The WAAP demonstrated that only 34% of the studies had a statistical power above 80%. This finding corresponds to the generally low statistical power in psychological research (Bakker et al., 2012) and should be a call for proper power analyses.

Moreover, the method performance check showed that a high amount of true variance between the studies limited the performance of most meta-analytic methods. Although the cause of heterogeneity could not be identified, it can be assumed that different paradigms addressed (different) underlying mechanisms of truth telling and lying to a varying extent. As discussed in Meta-Analysis 1, it is of course important to examine the validity of verbal tools for credibility assessment in various study settings. However, it needs to be specified in advance, as precisely as possible, which paradigm is used to address which processes. In this way, the comparability of studies could be increased and heterogeneity decreased.

CONCLUSION

We applied six meta-analytic methods to verify the robustness of the previous findings of Meta-Analysis 1. The results had a large range including conflicting conclusions from "verbal tools for credibility assessment do not work" to "verbal tools for credibility assessment do work". If we look at the weighted sum of these findings, there are, however, sound indications that CBCA and RM discriminate between experience-based and fabricated statements. In contrast, we strongly advise against using SCAN. Thus, on the basis of the method performance check, we gave the most weight to the results of 3PSM. Results of the other meta-analytic methods were of limited value, at least under the preselected conditions in the Meta-Showdown Explorer. The present selection was based on theoretical and empirical considerations concerning true effect size, severity of publication bias, QRPs environment, amount of heterogeneity, and number of studies in meta-analysis. However, it should be critically noted that these considerations are partly based on the results of Meta-Analysis 1, which are in turn of limited value according to the method performance check. Of course (with findings of future research) other preselections could be set. Criteria for good performance were chosen to minimize the Type I error, as it would be fatal to use methods that do not distinguish between experience-based and fabricated statements.

On the basis of their simulation study, Carter et al. (2019) pointed out that "researchers in psychology should not expect to produce a conclusive, debate-ending result by conducting a meta-analysis on an existing literature. Instead, we imagine meta-analyses may serve best to draw attention to the existing strengths and/or weaknesses in a literature […], and these results can then inspire a careful re-examination of methodology and theory, perhaps followed by large-scale, preregistered replication efforts" (p. 20). This suggestion applies here: The present results revealed a multitude of considerations on how future research could obtain (more) meaningful data in the field of verbal credibility assessment through pre-registered, adequately powered, and internally as well as ecologically valid studies.

OUTLOOK

The weighted sum of the present meta-analytical results demonstrated that CBCA and RM distinguish to a substantial degree between experience-based and fabricated statements, whereas SCAN does not. However, several boundary conditions of verbal tools for credibility assessment as well as methodical limitations of meta-analyses must be taken into account when interpreting the empirical evidence. In addition, our findings revealed a number of issues that need to be addressed in future research. Beyond these considerations based on the present data, I would like to refer to the calls of the first international workshop on verbal lie detection. In the course of the workshop, nine scientists and three practitioners gathered the major problems of research on verbal cues of credibility assessment and derived implications for future studies in a series of commentaries (Nahari et al., 2019).

Nahari (Commentary #2), for example, stressed the importance of establishing a strong theoretical foundation to a priori specify why certain cues work under what conditions. A study on RM showed that the difference in details between experience-based and fabricated statements decreased when statements were made after a certain time delay as compared to immediately after an event (Nahari, 2018). Although truth tellers still reported more details than liars, with a time lag of two weeks the number of true details decreased in both groups. However, liars compensated for this decrease with an increase of fabricated details, so that the difference in the total number of details declined. Since in practice statements are often made after a delay, Nahari (2018) called to examine this finding more closely and supplement the memory source monitoring approach with considerations that include strategic acting of liars (e.g., including details that cannot be verified, *Verifiability Approach*).

In Commentary #3, Taylor, Maroño, and Warmelink addressed the problem that study results based on group comparisons between scores of experience-based and fabricated statements say little about whether an individual's score indicates truth or lie. To determine the diagnostic power of verbal tools for credibility assessment, criteria are needed to decide at which point a statement is classified as experience-based or fabricated. Based on two criteria, the authors calculated classification rates for two study sets, which compared experience-based and fabricated statements in their number of details under different conditions. If one accepts a 50% false alarm rate, hit rates in both study sets did not exceed 75%. Without risking any false accusation, the average hit rate was only 40%. Hence, the authors do not consider the criterion *details* to be suitable for distinguishing between experience-based and fabricated statements in individual cases. This example illustrates that it is necessary to calculate classification rates under different criteria in order to determine the suitability of verbal cues.

In Commentary #4, Masip pointed out that future research needs to search for lie criteria. The CBCA contains only indicators of truth. The absence of criteria can be due to different reasons and cannot be clearly traced back to deception. If a statement is a little elaborate and detailed, one must conclude that the lie hypothesis cannot be rejected, but cannot be confirmed either. Although the RM approach contains one lie criterion (*cognitive operation*), its empirical support is limited and, thus, the performance for lie detection weak. Addressing this problem, Vrij, Leal, and Fisher (Commentary #5) presented an index composed of one truth criterion (*complications*) and two lie criteria (*common knowledge details*, *self-handicapping strategies*): *complications*/(*complications* + *common knowledge details* + *self-handicapping strategies*). A higher value indicates an experience-based statement, a lower value a fabricated statement. So far, the index has been examined in five studies (Vrij, Leal, Jupe, & Harvey, 2018; Vrij, Leal, et al., 2017; Vrij, Leal, Fisher, et al., 2018; Vrij, Leal, Mann, Fisher, Dalton, et al., 2018; Vrij, Leal, Mann, Fisher, Jo, et al., 2018) and outperformed the verbal cue *total detail*, which was considered to be the strongest cue for truthfulness to date (Amado et al., 2015). Hence, it seems worthwhile to follow this approach, even if truth tellers do not always use less *common knowledge details* and liars do not always use *self-handicapping strategies*. The authors considered how these differences could be amplified (e.g., by using certain interview protocols).

These comments are exemplary of a total of eleven problems that researchers and practitioners have identified (for further commentaries see Nahari et al., 2019). On the basis of their statements, the authors called for theoretically based and ecologically adaptable solutions for field challenges to be achieved by a *Theory-Protocol-Procedure*: Researchers should provide a theoretical and empirical basis for verbal tools for credibility assessment while practitioners should carry out their implementation in the field. In addition, the authors claimed not to pursue a one-size-fits-all approach, but rather to specify conditions under which verbal tools work or their use is limited. As Granhag concluded in Commentary #1, we should not leave it to others to contextualize our results. Finally, the authors referred to the *group to individual inference challenge* (G2i; Faigman, Monahan, Slobogin, 2014), i.e., the question of whether and how scientific knowledge derived from studying groups can be used in individual cases. In order to answer the question of how to convert group averages to individual classification (see Commentary #3), future research need to specify a cut-off criterion that applies not only to one study, but rather provide general acceptable sensitivity and specificty.

Verbal tools for credibility assessment make use of the fact that "At least, lying is thinking the truth" (Oliver Hassenkamp, translated from German). Lying requires recognizing that the social context demands certain information and producing a statement that is considered to be true. To accomplish these tasks, liars use strategic considerations that consume cognitive resources (Walczyk et al., 2014). Verbal tools for credibility assessment suppose that these psychological processes are reflected in verbal cues. The present data demonstrated the CBCA and RM belong to the most promising approaches. Nevertheless, both approaches have limitations that make their use difficult to impossible under certain conditions. There are intensive efforts to specify their validity more closely and to optimize the procedures. However, this research faces various methodical challenges. Since the use of verbal tools for credibility assessment has far-reaching consequences, it is essential that future studies actively address these challenges and meet the highest methodical standards in order to achieve meaningful results.

REFERENCES

References marked with an asterisk indicate studies included in the meta-analyses.

AC Del Re (2013). Compute.es: Compute Effect Sizes. R package version 0.2-2. Retrieved from http://cran.r-project.org/web/packages/compute.es

Abe, N. (2009). The neurobiology of deception: Evidence from neuroimaging and loss-of-function studies. *Current Opinion in Neurology*, *22*, 594–600. http://dx.doi.org/10.1097/WCO.0b013e328332c3cf

*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. http://dx.doi.org/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. http://dx.doi.org/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. http://dx.doi.org/10.1002/acp.1669

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

Alonso-Quecuty, M. L., Hernandez-Fernaud, E., & Campos, L. (1997). Child witnesses: Lying about something heard. In S. Redondo, V. Garrido, J. Perez, & R. Barbaret (Eds.), *Advances in psychology and law* (pp. 129–135). Berlin, Germany: Walter de Gruyter.

Amado, B. G., Arce, R., & Fariña, 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. http://dx.doi.org/10.1016/j.ejpal.2014.11.002

Amado, B. G., Arce, R., Fariña, F., & Vilariño, M. (2016). Criteria-based content analysis (CBCA) reality criteria in adults: A meta-analytic review. *International Journal of Clinical and Health Psychology*, *16*, 201–210. https://doi.org/10.1016/j.ijchp.2016.01.002

Arntzen, F. (1970). *Psychologie der Zeugenaussage: Einführung in die forensische Aussagepsychologie*. Göttingen, Germany: Verlag für Psychologie.

Arntzen, F. (2011). *Psychologie der Zeugenaussage: System der Glaubhaftigkeitsmerkmale* (5. Auflage). 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. http://dx.doi.org/10.1007/s10508-012-0022-8

Bakker, M., van Dijk, A., & Wicherts, J. M. (2012). The rules of the game called psychological science. *Perspectives on Psychological Science*, *7*, 543–554. http://doi.org/10.1177/1745691612459060

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

Ben-Haim, M. S., Williams, P., Howard, Z., Mama, Y., Eidels, A., & Algom, D. (2016). The emotional Stroop task: Assessing cognitive performance under exposure to emotional content. *Journal of Visualized Experiments*, (112), e53720. doi:10.3791/53720

*Berger, O. (2005). *Aspekte der Zeugenkompetenz und Validierung der Kriterienorientierten Aussageanalyse von Jugendlichen mit Intelligenzminderung* [Aspects of witnesses' competence and validation of criteria-based content analysis for adolescents with mental impairment] (Unpublished doctoral dissertation). Universität Regensburg, Regensburg, Germany. Retrieved from 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. http://dx.doi.org/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. http://dx.doi.org/10.1002/jip.1409

*Bogaard, G., Meijer, E. H., Vrij, A., & Merckelbach, H. (2016). Scientific Content Analysis (SCAN) cannot distinguish between truthful and fabricated accounts of a negative event. *Frontiers in Psychology*, *7*, 243. https://doi.org/10.3389/fpsyg.2016.00243

Bond, C. F., & DePaulo, B. M. (2006). Accuracy of deception judgments. *Personality and Social Psychology Review*, *10*, 214–234. http://dx.doi.org/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. http://dx.doi.org/10.1002/acp.1087

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

*Bradford, D. (2006). *Detection of deception in the confessional context* (Unpublished doctoral dissertation). University of New South Wales, Sydney, Australia.

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. https://doi.org/10.1016/S0193-3973(02)00107-7

Buller, D. B., & Burgoon, J. K. (1996). Interpersonal deception theory. *Communication Theory*, *6,* 203–242. https://doi.org/10.1111/j.1468-2885.1996.tb00127.x

Carter, E. C., Schönbrodt, F. D., Gervais, W. M., & Hilgard, J. (2019). Correcting for bias in psychology: A comparison of meta-analytic methods. *Advances in Methods and Practices in Psychological Science*, *2*. https://doi.org/10.17605/OSF.IO/9H3NU

Caso, L., Gnisci, A., Vrij, A., & Mann, S. (2005). Processes underlying deception: An empirical analysis of truth and lies when manipulating the stakes. *Journal of Investigative Psychology and Offender Profiling*, *2*, 195–202. https://doi.org/10.1002/jip.32

Coburn, K., & Vevea, J. L. (2017). *Estimating Weight-Function Models for Publication Bias.* Version 1.1.2

Cohen, J. (1988). *Statistical power analysis for the behavioral science* (2nd edition). Mahwah, NJ: Erlbaum.

*Connolly, D. A., & Lavoie, J. A. A. (2015). Discriminating veracity between children's reports of single, repeated, and fabricated events: A critical analysis of criteria-based content analysis. *American Journal of Forensic Psychology, 33*, 25–48.

Cooper, H., DeNeve, K., & Charlton, K. (1997). Finding the missing science: The fate of studies submitted for review by a human subjects committee. *Psychological Methods*, *2*, 447– 452. http://dx.doi.org/ 10.1037/1082-989X.2.4.447

Coursol, A., & Wagner, E. E. (1986). Effect of positive findings on submission and acceptance rates: A note on meta-analysis bias. *Professional Psychology: Research and Practice*, *17*, 136–137. http://dx.doi.org/10.1037/0735-7028.17.2.136

*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. http://dx.doi.org/10.1207/s1532480xads0302_2

Craik, F. I. M., & Bialystok, E. (2006). Cognition through the lifespan: Mechanisms of change. *Trends in Cognitive Sciences, 10*, 131–138. http://dx.doi.org/10.1016/j.tics.2006.01.007

Debey, E., De Schryver, M., Logan, G. D., Suchotzki, K., & Verschuere, B. (2015). From junior to senior Pinocchio: A cross-sectional lifespan investigation of deception. *Acta Psychologica*, *160*, 58–68. https://doi.org/10.1016/j.actpsy.2015.06.007

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

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

*Driscoll, L. N. (1994). A validity assessment of written statements from suspects in criminal investigations using the scan technique. *Police Studies*, *17*, 77–88.

*Dolezych, N. (2006). *Die Umsetzung von intuitiven Täuschungsstrategien in nicht erlebnisbasierten Aussagen* [The implementation of intuitive deceptive strategies within non-experience-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. http://dx.doi.org/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. http://doi.org/10.2307/2669529

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

*Elntib, S., Wagstaff, G. F., & Wheatcroft, J. M. (2015). The role of account length in detecting deception in written and orally produced autobiographical accounts using reality monitoring. *Journal of Investigative Psychology and Offender Profiling*, *12*, 185–198. https://doi.org/10.1002/jip.1420

Faigman, D. L., Monahan, J., & Slobogin, C. (2014). Group to individual (G2i) inference in scientific expert testimony. *The University of Chicago Law Review*, *81*, 417–480.

Fanelli, D. (2011). Negative results are disappearing from most disciplines and countries. *Scientometrics*, *90*, 891–904. https://doi.org/10.1007/s11192-011-0494-7

*Fiegler, S. (2009). *Zur Gültigkeit der Undeutsch-Hypothese unter Berücksichtigung der Schwierigkeit aus einer untrainierten Stichprobe* [On the validity of the Undeutsch-hypothesis with regard to difficulties in 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*, 66–82. http://dx.doi.org/10.2307/1359121

Furedy, J. J., Davis, C., & Gurevich, M. (1988). Differentiation of deception as a psychological process: a psychophysiological approach. *Psychophysiology, 25*(6), 683-688. http://dx.doi.org/10.1111/j.1469-8986.1988.tb01908.x

Gelman, A., & Loken, E. (2014). The statistical crisis in science. *American Scientist*, *102*, 460–465. https://doi.org/10.1511/2014.111.460

Gibbons, H., Schnuerch, R., Wittinghofer, C., Armbrecht, A. S., & Stahl, J. (2018). Detection of deception: Event-related potential markers of attention and cognitive control during intentional false responses. *Psychophysiology*, *55*, e13047. https://doi.org/10.1111/psyp.13047

*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. http://dx.doi.org/10.1348/135532505X52680

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

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.* Weinheim, Germany: Beltz.

*Hänert, P. (2007). *Die Validität inhaltlicher Glaubhaftigkeitsmerkmale unter suggestiven Bedingungen. Eine empirische Untersuchung an Vorschulkindern* [The validity of content-based credibility criteria under suggestive conditions. An empirical study study in a sample of preschool children] (Unpublished doctoral dissertation). Christian-Albrechts-Universität zu Kiel, Kiel, Germany.

Hartwig, M., & Bond, C. F., Jr. (2011). Why do lie-catchers fail? A lens model meta-analysis of human lie judgments. *Psychological Bulletin*, *137*, 643–659. https://dx.doi.org/10.1037/a0023589

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. https://doi.org/10.1177/1088868314556539

Hedges, L. V. (1984). Estimation of effect size under nonrandom sampling: The effects of censoring studies yielding statistically insignificant mean differences. *Journal of Educational and Behavioral Statistics*, *9*, 61–85. http://www.jstor.org/stable/1164832

Hedges, L. V., & Vevea, J. L. (1996). Estimating effect size under publication bias: Small sample properties and robustness of a random effects selection model. *Journal of Educational and Behavioral Statistics*, *21*, 299–332. http://dx.doi.org/10.2307/1165338

*Heinze, Y. (1996). *Inhaltliche Realkennzeichen in Aussagen Jugendlicher: Eine Simulationsstudie zur wissenschaftlichen Evaluation der inhaltsorientierten Aussageanalyse* [Content-based criteria in statements of adolescents: A simulation study for the scientific evaluation of content-based statement analysis] (Unpublished diploma thesis).Westfälische Wilhelms-Universität Münster, Münster, Germany.

Hernández-Fernaud, E., & Alonso-Quecuty, M. (1997). The cognitive interview and lie detection: A new magnifying glass for Sherlock Holmes? *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition, 11*, 55–68. https://doi.org/10.1002/(SICI)1099-0720(199702)11:1<55::AID-ACP423>3.0.CO;2-G

*Herrmann, M., & Jena, S. (1995). *Einzelfallexperimentelle Überprüfung inhaltlicher Realkennzeichen und Möglichkeiten einer am Einzelfall orientierten Kriterienentwicklung für die Glaubhaftigkeitsbegutachtung* [Experimental examination of content-based criteria on individual cases and possibilities of individual case-oriented development of criteria for credibility assessment] (Unpublished diploma thesis). Freie Universität Berlin, Berlin, Germany.

Higgins, J. P., Thompson, S. G., Deeks, J. J., & Altman, D. G. (2003). Measuring inconsistency in meta-analyses. *Bmj*, *327*, 557–560. https://doi.org/10.1136/bmj.327.7414.557

Horowitz, M. W., & Newman, J. B. (1964). Spoken and written expression: An experimental analysis. *The Journal of Abnormal and Social Psychology*, *68*, 640–647. http://dx.doi.org/10.1037/h0048589

Hyde, J. S., & Linn, M. C. (1988). Gender differences in verbal ability: A meta-analysis. *Psychological Bulletin, 104*, 53–69. http://dx.doi.org/10.1037/0033-2909.104.1.53

Iyengar, S., & Greenhouse, J. B. (1988). Selection models and the file drawer problem. *Statistical Science*, *3*, 109–117. Retrieved from https://www.jstor.org/stable/2245925

Jahn, M. (2001). *Grundlagen der Beweiswürdigung und Glaubhaftigkeitsbeurteilung im Strafverfahren.* Frankfurt a. M., Germany: Jura. Retrieved from https://www.jura.uni-frankfurt.de/55029767/Glaubhaftigkeitsbeurteilung.pdf

*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*, 18. http://dx.doi.org/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 quality of experience-based and fabricated statements] (Unpublished diploma thesis). Technische Universität Berlin, Berlin, Germany.

*Joffe, R. F. (1992). *Criteria-based content analysis: An experimental investigation with children* (Unpublished doctoral dissertation). The University of British Columbia, Vancouver, Canada.

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science*, *23*, 524–532. https://doi.org/10.1177/0956797611430953

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

Johnson, M. K., Foley, M. A., Suengas, A. G., & Raye, C. L. (1988). Phenomenal characteristics of memories for perceived and imagined autobiographical events. *Journal of Experimental Psychology: General*, *117*, 371–376. http://dx.doi.org/10.1037/0096-3445.117.4.371

Kellogg, R. T. (2007). Are written and spoken recall of text equivalent? *The American Journal of Psychology, 120*, 415–428. Retrieved from https://www.jstor.org/stable/20445412

Kleinberg, B., Arntz, A., & Verschuere, B. (2019). Being accurate about verbal credibility assessment. Retrieved from http://doi.org/10.31234/osf.io/h6pxt

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

Köhnken, G. (2004). Statement validity analysis and the detection of the truth. In P. A. Granhag & L. A. Strömwall (Eds.), *Deception detection in forensic contexts* (pp. 41–63). Cambridge, UK: Cambridge University Press.

*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. http://dx.doi.org/10.1037/0021-9010.80.6.671

*Krahé, B., & Kundrotas, S. (1992). Glaubwürdigkeitsbeurteilung bei Vergewaltigungsanzeigen: Ein aussagenanalytisches Feldexperiment [Judgment of the credibility of rape allegations: A content analytic field experiment]. *Zeitschrift für experimentelle und angewandte Psychologie*, *39*, 598–620.

Kühberger, A., Fritz, A., & Scherndl, T. (2014). Publication bias in psychology: A diagnosis based on the correlation between effect size and sample size. *PLoS ONE*, 9(9), e105825. https://doi.org/10.1371/journal.pone.0105825

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and ANOVAs. *Frontiers in Psychology*, *4*, 863. https://doi.org/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 & Neglect*, *21*, 255–264. http://dx.doi.org/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. http://dx.doi.org/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. http://dx.doi.org/10.1111/lcrp.12017

Leue, A. & Beauducel, A. (2019). A meta-analysis of the P3 amplitude in tasks requiring deception in legal and non-legal contexts. Manuscript submitted for publication. Retrieved from https://psyarxiv.com/rv77r/download/?format=pdf

Lipsey, M. W., & Wilson, D. B. (2000). *Practical meta-analysis (Applied social research methods).* Thousand Oaks, CA: Sage.

Loftus, E. F., & Pickrell, J. (1995). The formation of false memories. *Psychiatric Annals*, *25*, 720–724. https://doi.org/10.3928/0048-5713-19951201-07

*Logue, M., Book, A. S., Frosina, P., Huizinga, T., & Amos, S. (2015). Using reality monitoring to improve deception detection in the context of the cognitive interview for suspects. *Law and Human Behavior*, *39*, 360–367. http://dx.doi.org/10.1037/lhb0000127

*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 deceptive?] (Unpublished diploma thesis). Freie Universität Berlin, Berlin, Germany.

Lykken, D. T. (1959). The GSR in the Detection of Guilt. *Journal of Applied Psychology*, *43*, 385-388. http://dx.doi.org/10.1037/h0046060

Lykken, D. T. (1998). *A Tremor in the Blood: Uses and Abuses of the Lie Detector.* New York, NY, US: Plenum Press.

Mahoney, M. J. (1977). Publication prejudices: An experimental study of confirmatory bias in the peer review system. *Cognitive Therapy and Research*, *1*, 161–175. http://dx.doi.org/10.1007/BF01173636

Maier, B., Niehaus, S., Wachholz, S., & Volbert, R. (2018). The strategic meaning of CBCA criteria from the perspective of deceivers. *Frontiers in Psychology*, *9*, 855. https://doi.org/10.3389/fpsyg.2018.00855

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. http://dx.doi.org/10.1080/10683160410001726356

McGraw, K. O., & Wong, S. P. (1992). A common language effect size statistic. *Psychological Bulletin*, *111*, 361–365. http://dx.doi.org/10.1037/0033-2909.111.2.361

McShane, B. B., Böckenholt, U., & Hansen, K. T. (2016). Adjusting for publication bias in meta-analysis: An evaluation of selection methods and some cautionary notes. *Perspectives on Psychological Science*, *11*, 730–749. https://doi.org/10.1177/1745691616662243

Meijer, E., Verschuere, B., Gamer, M., Merckelbach, H., & Ben-Shakar, G. (2016). Deception detection with behavioral, autonomic, and neural measures: Conceptual and methodological considerations that warrant modesty. *Psychophysiology, 53,* 593–604. https://doi.org/10.1111/psyp.12609

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. http://dx.doi.org/10.1037/1076-8971.7.1.3

*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. http://dx.doi.org/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. http://dx.doi.org/10.1016/j.paid.2004.01.007

*Metzger, G. (1996). *Inhaltsgestützte Beurteilung der Glaubwürdigkeit von Zeugenaussagen* [Content-based assessment of the credibility of witness statements] (Unpublished diploma thesis). Christian-Albrechts-Universität zu Kiel, Kiel, Germany.

Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., Howerter, A., & Wager, T. D. (2000). The unity and diversity of executive functions and their contributions to complex "frontal lobe" tasks: A latent variable analysis. *Cognitive Psychology*, *41*, 49–100.

*Nahari, G. (2017). Top-down processes in interpersonal reality monitoring assessments. *Psychology, Public Policy, and Law*, *23*, 232–242. http://dx.doi.org/10.1037/law0000110

Nahari, G. (2018). Reality monitoring in the forensic context: Digging deeper into the speech of liars. *Journal of Applied Research in Memory and Cognition*, *7*, 432–440. https://doi.org/10.1016/j.jarmac.2018.04.003

Nahari, G., Ashkenazi, T., Fisher, R. P., Granhag, P. A., Hershkowitz, I., Masip, J., ... & Verschuere, B. (2019). 'Language of lies': Urgent issues and prospects in verbal lie detection research. *Legal and Criminological Psychology*, *24*, 1–23. https://doi.org/10.1111/lcrp.12148

*Nahari, G., & Pazuelo, M. (2015). Telling a convincing story: Richness in detail as a function of gender and information. *Journal of Applied Research in Memory and Cognition*, *4*, 363–367. http://dx.doi.org/10.1016/j.jarmac.2015.08.005

*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. http://dx.doi.org/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 non-experience-based statements with regard to different extents of pre-information] (Unpublished diploma thesis). Technische Universität Braunschweig, Braunschweig, Germany.

*Niehaus, S. (2000). *Zur Anwendbarkeit inhaltlicher Glaubhaftigkeitsmerkmale bei Zeugenaussagen unterschiedlichen Wahrheitsgehalts* [On the applicability of content-related credibility criteria in statements of different truth status] (Unpublished doctoral dissertation). Universität Osnabrück, Osnabrück, Germany.

Niehaus, S. (2008). Merkmalsorientierte Inhaltsanalyse. In R. Volbert, & M. Steller (Eds.), *Handbuch der Rechtspsychologie* (pp. 311–321). Göttingen, Germany: Hogrefe.

Oberlader, V. A., Naefgen, C., Koppehele-Gossel, J., Quinten, L., Banse, R., & Schmidt, A. F. (2016). Validity of content-based techniques to distinguish true and fabricated statements: A meta-analysis. *Law and Human Behavior*, *40*, 440–457. http://dx.doi.org/10.1037/lhb0000193

Oberlader, V. A., Quinten, L., Banse, R., Volbert, R., Schmidt, A. F., & Schönbrodt, F. D. (2019). How robust are meta-analytic findings on the validity of content-based credibility assessment? A comparison of six meta-analytic methods and recommendations for future research. Manuscript submitted for publication.

*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. http://dx.doi.org/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. http://dx.doi.org/10.1007/BF01498980

R Core Team (2017). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org/

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

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

*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. http://dx.doi.org/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. http://dx.doi.org/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. http://dx.doi.org/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 the credibility assessment of content quality in fabricated 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. http://dx.doi.org/10.1080/10683160008409802

Sartori, G., Agosta, S., Zogmaister, C., Ferrara, S. D., & Castiello, U. (2008). How to accurately detect autobiographical events. *Psychological Science*, *19*, 772–780. https://doi.org/10.1111/j.1467-9280.2008.02156.x

Schacter, D. L., Addis, D. R., Hassabis, D., Martin, V. C., Spreng, N., & Szpunar, K. K. (2007). The cognitive neuroscience of constructive memory: Remembering the past and imagining the future. *Philosophical Transactions of the Royal Society B, 362*, 773–786. https://doi.org/10.1098/rstb.2007.2087

*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. http://dx.doi.org/10.1002/jip.121

Schooler, J. W., Gerhard, D., & Loftus, E. F. (1986). Qualities of the unreal. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *12*, 171–181. http://dx.doi.org/10.1037/0278-7393.12.2.171

Schönbrodt, F. D. (2018). *p-checker: One-for-all p-value analyzer*. Retrieved from http://shinyapps.org/apps/p-checker/. The source code of this app is licensed under the open GPL-2 license and is published on Github.

Schwarzer, G. (2007). Meta: An R package for meta-analysis. R News, 40–45.

*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. http://dx.doi.org/10.1002/acp.1756

Silberzahn, E., Uhlmann, E. L., Martin, D. P., Anselmi, P., Aust, F. Awtrey, E., […], & Nosek, B. A. (2018). Many analysts, one data set: Making transparent how variations in analytic choices affect results. *Advances in Methods and Practices in Psychological Science.* https://doi.org/10.1177/2515245917747646

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). *p*-curve and effect size: Correcting for publication bias using only significant results. *Perspectives on Psychological Science*, *9*, 666–681. https://doi.org/10.1177/1745691614553988

Simonsohn, U., Simmons, J. P., & Nelson, L. D. (2015). Specification curve: Descriptive and inferential statistics on all reasonable specifications. http://dx.doi.org/10.2139/ssrn.2694998

*Smith, N. (2001). *Reading between the lines: An evaluation of the scientific content analysis technique (SCAN).* Home Office, Policing and Reducing Crime Unit, Research, Development and Statistics Directorate. London, UK.

Spence, S. A., Farrow, T. F. D., Herford, A. E., Wilkinson, I. D., Zheng, Y., & Woodruff, P. W. R. (2001). Behavioural and functional anatomical correlates of deception in humans. *Neuroreport, 12*, 2849–2853. http://dx.doi.org/10.1097/00001756-200109170-00019

*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. https://doi.org/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. A. Strömwall (Eds.), *The detection of deception in forensic contexts* (pp. 64–102). Cambridge University Press.

*Sporer, S. L., & Küpper, B. (1995). Realitätsüberwachung und die Beurteilung des Wahrheitsgehalts von Erzählungen: Eine experimentelle Studie [Reality monitoring and the judgment of the truth status of reports: 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. http://dx.doi.org/10.1002/acp.1234

Stanley, T. D., & Doucouliagos, H. (2014). Meta-regression approximations to reduce publication selection bias. *Research Synthesis Methods*, *5*, 60–78. https://doi.org/10.1002/jrsm.1095

Stanley, T. D., Doucouliagos, H., & Ioannidis, J. (2017). Finding the power to reduce publication bias. *Statistics in Medicine*, *36*, 1580–1598. https://doi.org/10.1002/sim.7228

*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, Germany. Retrieved from 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, NY, US: Kluwer/Plenum Press. http://dx.doi.org/10.1007/978-94-015-7856-1_8

Steller (2008). Psychophysiologische Aussagebeurteilung. In R. Volbert & M. Steller (Eds.), *Handbuch der Rechtspsychologie* (pp. 364–375). Göttingen, Germany: Hogrefe Verlag.

Steller, M., & Boychuk, T. (1992). Children as witnesses in sexual abuse cases: Investigative interview and assessment techniques. In H. Dent & R. Flin (Eds.), *Wiley series in the psychology of crime, policing and law. Children as witnesses* (pp. 47–71). Oxford, UK: John Wiley & Sons.

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, NY, US: 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.

Sterling, T. D. (1959). Publication decisions and their possible effects on inferences drawn from tests of significance - or vice versa. *Journal of the American Statistical Association*, *54*, 30–34. http://doi.org/10.2307/2282137

*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. http://doi.org/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. http://dx.doi.org/10.1375/pplt.12.2.345

Suchotzki, K., Verschuere, B., Van Bockstaele, B., Ben-Shakhar, G., & Crombez, G. (2017). Lying takes time: A meta-analysis on reaction time measures of deception. *Psychological Bulletin*, *143*, 428–453. http://dx.doi.org/10.1037/bul0000087

Szewczyk, H. (1973). Kriterien der Beurteilung kindlicher Zeugenaussagen. *Probleme und Ergebnisse der Psychologie*, *46*, 47–66.

Trankell, A. (1971). *Der Realitätsgehalt von Zeugenaussagen: Methoden der Aussagepsychologie*. Göttingen, Germany: Vandenhoeck & Ruprecht.

Trovillo, P. V. (1939). A history of lie detection. *Journal of Criminal Law and Criminology*, 29, 848–881.

*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. http://dx.doi.org/10.1207/s1532480xads0302_4

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

van Aert, R. C.M. (2018). *p*-uniform: Meta-analysis methods correcting for publication bias. R package version 0.1.0.

*Vanderhallen, M., Jaspaert, E., & Vervaeke, G. (2015). SCAN as an investigative tool. *Police Practice and Research*, *17*, 279–293. http://dx.doi.org/10.1080/15614263.2015.1008479

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, *36*, 1–48.Retrieved from http://www.jstatsoft.org/v36/i03/

Vincent, A., & Furedy, J. J. (1992). Electrodermal differentiation of deception: Potentially confounding and influencing factors. *International Journal of Psychophysiology*, *13*, 129–136. https://doi.org/10.1016/0167-8760(92)90052-D

Volbert, R., & Banse, R. (2014). Deception detection. *European Psychologist, 19,* 159–161. http://dx.doi.org/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. http://dx.doi.org/10.1027/1016-9040/a000200

Vrij, A. (2005). Criteria-based content analysis: A qualitative review of the first 37 studies. *Psychology, Public Policy, & Law*, *11*, 3–41. http://dx.doi.org/10.1037/1076-8971.11.1.3

Vrij, A. (2008). *Detecting lies and deceit: Pitfalls and opportunities*. Chichester, UK: Wiley.

Vrij, A. (2015). Verbal lie detection tools: Statement validity analysis, reality monitoring and scientific content analysis. In P. A. Granhag, A. Vrij, & B. Verschuere (Eds.), *Deception detection: Current challenges and cognitive approaches* (pp. 3–35). Chichester, UK: Wiley.

*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*, *36*, 113–126. http://dx.doi.org/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. http://dx.doi.org/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. http://dx.doi.org/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. http://dx.doi.org/10.1023/A:1006610329284

*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. http://dx.doi.org/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. http://dx.doi.org/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. http://dx.doi.org/10.1007/s10979-006-9066-4

Walczyk, J. J., Harris, L. L., Duck, T. K., & Mulay, D. (2014). A social-cognitive framework for understanding serious lies: Activation-decision-construction-action theory. *New Ideas in Psychology*, *34*, 22–36. https://doi.org/10.1016/j.newideapsych.2014.03.001

*Wehner, I. (2006). *Erhebung und Beurteilung von Tatverdächtigenaussagen* [Investigation and assessment of allegations of suspects]. Frankfurt a. M., Germany: Verlag für Polizeiwissenschaft.

*Willén, R. M., & Strömwall, L. (2012). Offenders' uncoerced false confessions: A new application of statement analysis? *Legal and Criminological Psychology*, *17*, 346–359. http://dx.doi.org/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. 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.

Yuille, J. C. (1988). The systematic assessment of children's testimony. *Canadian Psychology/Psychologie canadienne, 29*, 247–262. http://dx.doi.org/10.1037/h0079769

*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. http://dx.doi.org/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. https://doi.org/10.1016/S0065-2601(08)60369-X

Appendix A

Data table

Study ID	Authors (year)	Description of effect size basis	Page/ table
42	Akehurst, Bull, Vrij, & Köhnken (2004)	Acutal involvement vs. fabricated; calculation of CBCA score	p. 887, Table 3
12	Akehurst, Köhnken, & Höfer (2001)		p. 74, Table 3
13	Akehurst, Manton, & Quandte (2011)	Both raters combined	p. 240
60	Bensi, Gambetti, Nori, & Giusberti (2009)	Within-subjects design; overall CBCA score	p. 114, Table 2
17	Berger (2005)		p. 303, Table 4-75
7	Blandon-Gitlin, Pezdek, Rogers, & Brodie (2005)		p. 193
52	Bogaard, Meijer, & Vrij (2014)	Condition: Explanation; CBCA	p. 158, Table 3
52.2	Bogaard, Meijer, & Vrij (2014)	Condition: Explanation; RM	p. 158, Table 3
52.3	Bogaard, Meijer, & Vrij (2014)	Condition: Explanation; SCAN	p. 158, Table 3
58	Bogaard, Meijer, Vrij, & Merckelbach (2016)		p. 5
27	Bond & Lee (2005)		p. 322, Table 2
61	Bradford (2006)		p. 152
68	Connolly & Lavoie (2015)		p. 16
19	Craig, Scheibe, Raskin, Kircher, & Dodd (1999)		p. 82
71	Discroll (1994)		p. 84, Table 3
16	Dolezych (2006)	Calculation of CBCA score without the following criteria: *Lebensgewohnheiten, Erinnerungsbemühen, Zuschreiben negativer Eigenschaften, Klischees*	p. 75
53	Elntib, Wagstaff, & Wheatcroft (2014)	Before word count standardization; RM	p. 8, Table 1
20	Fiegler (2009)	Results of discriminant function 4	p. 101, Table 5.17
35	Gödert, Gamer, Rill, & Vossel (2005)	Witnesses vs. perpetrators	p. 235
62	Granhag, Strömwall, & Landström (2006)	CBCA	p. 91, Table 3
62.2	Granhag, Strömwall, & Landström (2006)	RM	p. 91, Table 3
63	Hänert (2007)	Group A vs. group C	p. 108, Table 13
67	Heinze (1996)		p. 90, Table 6
69	Hermann & Jena (1995)	Statements of both participants combined	p. 86, p. 90
47	Jang, Kim, Cho, & Lee (2013)	RM	p. 5, Table 2
32	Janka (2003)	Rater 1 and 2 combined	p. 111, Table 20
64	Joffe (1992)	Experienced vs. lightly coached	p. 119, Table 6
65	Joffe (1992)	Experienced vs. lightly coached	p. 116, Table 4
5	Köhnken, Schimossek, Aschermann, & Höfer (1995)		p. 679, Table 5

Appendix A (*continued*)

Study ID	Authors (year)	Description of effect size basis	Page/ table
3	Krahe & Kundrotas (1992)		p. 611, Table 6
15	Lamb et al. (1997)	Very likely and quite likely as well as very unlikely and quite unlikely combined	p. 260, Table 2
11	Landry & Brigham (1992)	Trained participants	p. 671, Table 3
49	Leal, Vrij, Warmelink, Vernham, & Fisher (2015)	Study 1	p. 137
50	Leal, Vrij, Warmelink, Vernham, & Fisher (2015)	Study 2	p. 141
54	Logue, Book, Frosina, Huizinga, & Amos (2015)	Calculation of RM score	p. 365, Table 1
22	Lüdke (2008)		p. 40, Table 11
2	Memon, Fraser, Colwell, Odinot, & Mastroberardino (2010)	RM version 1 and 2 combined	pp. 9
6	Merckelbach (2004)	High and low fantasy prone participants combined	p. 1379, Table 2
66	Metzger (1996)		p. 84, Table 10
66.2	Metzger (1996)		p. 88, Table 13
56	Nahari & Pazuelo (2015)	Information on informed truth tellers and liars via email	
55	Nahari (2017)	Forensic context; information on *SD* in false condition via email	pp. 236
34	Nahari, Vrij, & Fisher (2012)	Innocents vs. concealment, RM	p. 74, Table 3
34.2	Nahari, Vrij, & Fisher (2012)	Innocents vs. concealment, SCAN	p. 74, Table 3
40	Naumann (2005)	Controlgroup (mothers) vs. experimental group (trained non-mothers)	p. 94, Table 22
45	Niehaus (2000)	True statements vs. reproduction with experience	p. 285, Table 53
30	Parker & Brown (2000)	CBCA	p. 244, Table 1
21	Porter & Yuille (1996)		p. 451
21.2	Porter & Yuille (1996)		p. 450, Table 1
18	Roma, San Martini, Sabatello, Tatarelli, & Ferracuti (2011)	Total sample (male and female)	p. 617, Table 3
4	Ruby & Brigham (1998)	White and black speakers as well as first and second sample combined	p. 383, Table 5
41	Rutta (2001)		p. 78, Table 9
29	Santtila, Roppola, Runtti, & Niemi (2000)	CBCA	p. 175, Table 7
10	Schelleman-Offermans & Merckelbach (2010)	High and low fantasy prone participants combined	p. 255, Tale 2
8	Short & Bodner (2011)		p. 5
59	Smith (2001)	Experienced SCAN users	p. 18, Table 4
23	Sporer & Küpper (1995)		p. 183
24	Sporer & Sharman (2006)	Other ratings	p. 847, Table 4

Appendix A (*continued*)

Study ID	Authors (year)	Description of effect size basis	Page/ table
39	Sporer (1997)	CBCA	p. 383
39.2	Sporer (1997)	RM	p. 384
25	Steck et al. (2010)	Truth statements vs. fabricated statements	p. 13, Table 4
28	Steller, Wellershaus, & Wolf (1992)	Calculation of CBCA score	p. 165, Table 2
51	Strömwall & Granhag (2005)		p. 351
33	Strömwall, Bengtsson, Leander, & Granhag (2004)	CBCA	p. 662
33.2	Strömwall, Bengtsson, Leander, & Granhag (2004)	RM	p. 663
14	Tye, Amato, Honts, Devitt, & Peters (1999)		p. 101
57	Vanderhallen, Jaspaert, & Vervaeke (2015)	SCAN	pp. 6, Table 2, Table 3
1	Vrij & Mann (2006)	Interviewphase 1+2 combined	p. 345
36	Vrij, Akehurst, Soukara, & Bull (2004a)	CBCA total score	p. 24, Table 2
36.2	Vrij, Akehurst, Soukara, & Bull (2004a)	RM total score	p. 24, Table 2
8	Vrij, Akehurst, Soukara, & Bull (2004b)	CBCA	p. 121
38.2	Vrij, Akehurst, Soukara, & Bull (2004b)	RM	p. 121
31	Vrij, Edward, & Bull (2001)	CBCA	p. 905, Table 1
31.2	Vrij, Edward, & Bull (2001)	RM	p. 905, Table 1
48	Vrij, Edward, Roberts, & Bull (2000)	CBCA	p. 250, Table 1
48.2	Vrij, Edward, Roberts, & Bull (2000)	RM	p. 250, Table 1
46	Vrij, Kneller, & Mann (2000)	True vs. informed liars	p. 63, Table 1
43	Vrij, Mann, Kristen, & Fisher (2007)	All three interview styles combined; CBCA	p. 508, Table 1
43.2	Vrij, Mann, Kristen, & Fisher (2007)	All three interview styles combined; RM	p. 508, Table 1
70	Wehner (2006)	"Glaubhaftigkeitsindex (GI)"	p. 278, Table 21
44	Willén & Strömwall (2012)	CBCA total score	p. 353, Table 2
44.2	Willén & Strömwall (2012)	RM total score	p. 354, Table 3
26	Wolf & Steller (1997)		p. 129
37	Wrege (2004)	Trained participants	p. 80, Table 6a
9	Zaparniuk, Yuille, & Taylor (1995)	Event 1 and 2, coder 1, 2, and 3, and decision rule 1 and 2 combined	p. 348, Table 3

Note. A description of effect size basis is only provided if a study included several comparisons of experience-based and fabricated statements based on one sample.

Appendix B

Coding manual

Column name	Description	Coding
Study ID	Study IDs ranging from 1-71 Study IDs ending with XX.2 or XX.3 have been additionally used for subset meta-analyses of the individual verbal tools for credibility assessment	
Authors (year)	Authors and year of publication	
Jrnl	Journal name	
Pub	Type of publication	0 = unpublished 1 = published
VT	Verbal tools for credibility assessment	1= CBCA 2 = RM 3 = SCAN
Dsgn	Study design	0 = field study 1 = laboratory study
WithinBetween	Within-subjects study design: participants provide experience-based and fabricated statement Between-subjects study design: participants either provide experience-based or fabricated statement Only rated in laboratory studies	0 = within-subjects design 1 = between-subjects design
Male	Number of male participants	
Female	Number of female participants	
SexRatio	Ratio of female and male participants	0 = all men to 1 = all women
Ntotal	Total number of fabricated and experience-based statements	

Appendix B (*continued)*

Column name	Description	Coding
Nlie	Number of fabricated statements	
Ntrue	Number of experience-based statements	
Age	Age of participants	0 < 18 1 ≥ 18
Motivation	Financial or other motivating incentive to make a convincing statement Only rated in laboratory studies	0 = no incentive 1 = incentive
ConcealOut	Type of lie Concealment lie: statement is partly based on true aspects Outright lie: completely fabricated statement Only rated in laboratory studies	0 = concealment lie 1 = outright lie
Event	Experienced event is characterized by negative emotional tone, personal involvement, and loss of control Only rated in laboratory studies	0 = at least one criteria is missing 1 = all three criteria are met
Experienced	Experience status I Event was either not personally experienced, i.e., participants watched a movie, or personally experienced Only rated in laboratory studies	0 = not personally experienced 1 = personally experienced
Accused	Experience status II Participants either were accused or not accused of having done something Only rated in laboratory studies	0 = not accused 1 = accused

Appendix B (*continued*)

Column name	Description	Coding
CBCACriteria	Set of CBCA criteria	0 = any incomplete set of CBCA criteria 1 = complete set of 19 CBCA criteria (Steller & Köhnken, 1989) 2 = 14-item version (Raskin, Esplin, & Horowitz, 1991)
Scoring	Scoring of the criteria	0 = present/ absent 1 = rating on a Likert-scale
Training	Subjects were trained in providing criteria of verbal tools for credibility assessment; only rated in laboratory studies	0 = no 1 = yes
Statement	Mode of the statement	0 = oral 1 = written
Rater	Laypersons: received no training in the procedure used Trained raters: received training in the procedure used Professionals: work in the field of forensic statement assessment	0 = laypersons 1 = trained raters 2 = professionals
OverallES	Study used for overall effect size estimation or only for subset meta-analyses	0 = not used 1 = used
dbase	Decision basis for the classification of the statements as experience-based or fabricated	0 = discriminant function 1 = rater decision 2 = mean comparison of scores 3 = a priori decision rule
Crossval	Correction for potential overestimation of results based on discriminant analysis by cross validation	0 = no 1 = yes
d	Cohen's *d*	

Appendix B *(continued)*

Column name	Description	Coding
d_v	Variance of Cohen's *d*	
d_se	Standard error of Cohen's *d*	
g	Hedges' *g*	
g_v	Variance of Hedges' *g*	
g_se	Standard error of Hedges' *g*	
invarwe	Inverse variance weight	
t_obs	Observed *t*-value	