Fachbereich Erziehungswissenschaft und Psychologie

der Freien Universität Berlin

Which processes govern the emergence of CBCA criteria in witness statements?

An integrated and theory-driven approach for analysis

Dissertation

zur Erlangung des akademischen Grades

Doktor der Philosophie (Dr. phil.)

vorgelegt von

Benjamin G. Maier (M.Sc.)

Berlin, 2021

Erstgutachterin:

Prof. Dr. Renate Volbert

Zweitgutachter:

Prof. Dr. Stefan Krumm

Tag der Disputation:

09.12.2022

Abstract

In its original version, Criteria-Based Content Analysis (CBCA) comprises a systematic set of 19 content criteria that are expected to occur more frequently or stronger in experience-based than fabricated statements. This Thesis was designed to address frequently criticized aspects of CBCA, namely the poor theoretical footing behind the compilation of criteria in substantiating truthfulness and the absence of a weighting system to sort the individual criteria in relation to their diagnostic value. Referring to the notion that creative and strategic demands are the driving forces behind the differences in statement quality between false and true reports, **study I** experimentally manipulated the presence of these two demands to examine the effects on participants' (N = 30) CBCA sum scores. Other than expected, the pairing of creative and strategic demands yielded fabricated statements with higher CBCA sum scores than observed from any of the other experimental conditions. Possibly then, participants may be able to produce fabricated statements in much higher quality than commonly assumed. Future investigations that provide a more objective point of reference for the comparison of scores are needed, however, before more definite conclusions can be drawn. The subsequent two studies tested whether the three-dimensional structure of the revised CBCA model of Volbert and Steller (2014) would contain criteria that within each set yielded comparable motivational (study II) and cognitive (study III) properties. Pertaining to a fictitious scenario, **study II** examined how laypersons (N = 135) assess the strategic meaning of the criteria when deceiving. Overall, participants were inclined to produce memory-related criteria (set 1) but tended to avoid script-deviant (set 2) and strategy-based (set 3) criteria. The findings thus supported the suitability of the revised model in representing motivational - and diagnostically relevant - differences between the criteria. **Study III** tested to what degree coaching enabled participants (N = 60) to simulate script-deviant or strategy-based criteria when reporting. If compared to a control group of uninformed participants, coaching participants had no effects on the CBCA sum scores of their fabricated statements, suggesting that both script-deviant and strategy-based criteria are cognitively difficult to simulate. In turn, implementing coaching-restricted scale scores yielded tentative, but not conclusive evidence for strategy-based criteria being easier to integrate into false statements than script-deviant criteria. In summary, study II provided clearly interpretable as well as practically valuable results about the motivational properties of the criteria sets, whereas the in part counterintuitive (study I) or somewhat ambiguous (study III) findings from the other two studies call for further investigations to elucidate.

Keywords: Criteria-Based Content Analysis (CBCA), creative and strategic demands, individual deception strategies, motivational versus cognitive properties, coaching, diagnostic value, mundane realism

Table of Contents

General Introduction

Criteria-Based Content Analysis (CBCA) as a truth-detection tool

Allegations of sexual offenses often pertain to highly intimate situations, rendering defendant and accuser the only witnesses available. If the former denies the allegation and no physical evidence is accessible, the task of assessing the truthfulness of the accuser's incriminating testimony becomes paramount in legal proceedings. In criminal courts of Germany and several other European countries (Volbert & Steller 2014), psychological expert-witnesses undertake this task by referring to a diagnostic assessment procedure termed Statement Validity Assessment (SVA). The multiple-stage approach first takes the individual competencies of the statement provider as well as the case-specific conditions into account, and if certain prerequisites are met, focuses on the semantic content of the statement by utilizing its subcomponent Criteria-Based Content Analysis (CBCA).

The underlying rationale of CBCA holds that the content of experience-based accounts is qualitatively higher than the content of fabricated statements (the so-called Undeutsch-Hypothesis; Undeutsch, 1967). In its original version, it comprises a systematic set of 19 content criteria that practitioners and scholars had deemed suitable to substantiate truthfulness (Steller & Köhnken, 1989; see Table 1). To date, a multitude of laboratory and field studies have confirmed that experience-based accounts yield higher quality than fabricated statements and in this way corroborated the overall validity of CBCA as a *truth*-detection tool (for recent meta-analysis see Amado et al., 2015; Amado et al., 2016; Oberlader et al., 2016; 2021).

Table 1: Original compilation of CBCA criteria (Steller & Köhnken, 1989).

General Characteristics
1. Logical consistency
2. Unstructured Production
3. Quantity of details
Specific Contents
4. Contextual embedding
5. Description of interactions
6. Reproduction of conversation
7. Unexpected complication during the incident
Peculiarities of content
8. Unusual details
9. Superfluous details
10. Accurately reported details not comprehended
11. Related external associations
12. Accounts of subjective mental state
13. Attribution of perpetrator's mental state
Motivation-related content
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 elements
19. Details characteristic of the offense

Empirical criticism about CBCA (and CBCA research)

Clearly though, CBCA is not without its critics. Apart from concerns about the tool's susceptibility to coaching (e.g. Vrij et al., 2004) or characteristics such as event familiarity (Blandon-Gitlin et al., 2005) and fantasy-proneness (Schelleman-Offermans & Merckelbach, 2010), scholars primarily criticized that the compilation of criteria bears no underlying theoretical foundation and lacks a formal weighting system (e.g. Porter & ten Brinke, 2010). That is, equal diagnostic value is given to each criterion when assessing a statement's truthfulness, albeit it seems plausible (Steller & Köhnken, 1989; Volbert & Steller, 2014) and indeed has been shown (Hommers, 1997; Vrij, 2005) that some criteria are more likely than

others to appear in true statements only. Naturally, the relevant question for gauging the diagnostic value of any criterion is not how likely the criterion is to occur in true (i.e. experience-based) statements, but how likely it is to occur in true *relative* to false (i.e. fabricated) statements (Maier et al., 2018).

Against this background, it seems worth pondering that the empirical findings about the validity of CBCA are based on studies that typically summed up the individual criteria scores to one comprehensive (total) CBCA sum score, which subsequently served as the relevant variable for further analysis (e.g. Akehurst et al., 2001; Welle et al., 2016). As Maier et al. (2018) pointed out, such a simplistic approach may underestimate the actual utility of CBCA, since it ignores the possibility that some criteria could be more sensitive to truthfulness and hence bear higher diagnostic value than others. There is also controversy within the traditional deception detection literature to what degree results from experimental studies are transferable to real-world forensic situations (e.g. Wright Whelan et al., 2015). While the former typically introduce participants to low-stake situations in which the act of lying remains largely inconsequential (e.g. Burgoon, 2015), the latter arguably carry severe consequences if being found guilty of lying. As the potential outcome for the liar can highly affect his or her performance (Porter & ten Brinke, 2010), the rather trivial contexts of laboratory studies may have failed to evoke cues from participants that in forensic situations might saliently emerge (i.e. Frank & Svetieva, 2012).

Recent developments in regards to CBCA

It has to be noted however that the scientific debate about the strengths and weaknesses of CBCA (e.g. Ruby & Brigham, 1997; Porter & ten Brinke, 2010) primarily pertains to the tool's original version (Steller & Köhnken, 1989) developed several decades ago. In reference to the rationale that the content of experience-based accounts is qualitatively higher than the content

of fabricated accounts as specified by the Undeutsch Hypothesis (Undeutsch, 1967), Steller and Köhnken (1989) indeed based their compilation of criteria on pragmatic rather than theoretical considerations (Volbert & Steller, 2014). Contending that a lying person needs to put more effort in inventing information (*creative* demands or efforts related to *primary deception*) and in presenting the fabricated event in a credible manner (*strategic* demands or efforts related to *secondary deception*), Köhnken (1990) subsequently classified these criteria into cognitive versus motivational. Both classes of criteria are thought to occur more frequently in true statements, albeit for different reasons: While cognitive criteria should be typically too difficult to produce when fabricating, motivational criteria should be avoided out of strategic considerations when lying.

In 2014, Volbert and Steller introduced a revised CBCA model that pays greater attention than before to theoretical considerations of what processes govern the emergence of criteria in true statements. Taking previous work of Niehaus (2008a) into account, the authors decided to distinguish the class of cognitive criteria further as these pertain to characteristics of either *episodic autobiographical memory* (set 1, criteria such as *temporal* or *spatial details*) or *script-deviant information* (set 2, criteria such as *unusual details* or *unexpected complications*). Motivational criteria related to *efforts of strategic self-presentation*, such as *admitting lack of memory* or *spontaneous corrections*, comprise the third set of the revised model, displayed in Table 2. For higher ease of readability, the three criteria sets will be referred to in this Thesis by the following labels: *Memory-related criteria* (set 1), *script-deviant criteria* (set 2) and *strategy-based criteria* (set 3).

Table 2**.** Revised system of CBCA criteria[1] (Volbert & Steller, 2014).

Autobiographic memory versus script information		*Strategic self-presentation*
Criteria related to episodic autobiographical memory (set 1)	*Criteria related to script-deviant information* (set 2)	*Criteria related to efforts of positive strategic self-presentation* (set 3)
Information about everyday life routines [C]	Unexpected complications [C]	Spontaneous corrections [M]
Spatial information [C]	Superfluous details [C]	Admitting lack of memory [M]
Temporal information [C]	Unusual details [C]	Efforts to remember [M]
Description of interactions [C]	Related external associations [C]	Expressing uncertainty [M]
Reproduction of conversations [C]	Accurately reported details not comprehended [C]	Reality controls [M]
Emotions and feelings [C]		Raising doubt about one's own testimony [M]
Own thoughts [C]		Raising doubts about one's own person [M]
Sensory Impressions [C]		Self-deprecation [M]
Attribution of perpetrator's mental state [C]		Pardoning the perpetrator [M]
Personal implications [C]		Unstructured production [M]

[C] = Cognitive criteria; [M] = Motivational criteria (according to the originally binary classification).
Note. Adapted from "The Strategic Meaning of CBCA Criteria from the Perspective of Deceivers" by Maier, B.G. et al., 2018, *Frontiers in Psychology, 9:855.*

Considerations about the diagnostic value of the criteria

While the binary classification of cognitive versus motivational criteria (Köhnken, 1990) implies that the diagnostic value of a criterion is to be derived from either its cognitive or motivational component, Niehaus et al. (2005) pointed out that considerations of the underlying

[1] Volbert and Steller (2014) understand their allocation of specific criteria to be exemplary rather than irrevocably. For illustration purposes, the structure presented in this article thus differs slightly from the version originally presented by the authors. For instance, the criterion unstructured production was classified as motivational and added to the third set by the authors of the current article; originally, it was classified as cognitive and allocated to a separate "statement as a whole" category (see Maier et al., 2018, for more details).

motivational properties need to be applied to cognitive criteria as well, and vice versa. Detailed insight about the motivational component should thus provide a first hint toward the criterion's diagnostic value (Maier et al., 2018): If the deceiver considers the criterion to be strategically detrimental to his or her efforts to appear credible, the likelihood of its emergence in fabricated accounts is generally lower. In turn however, if the deceiver is inclined to produce a criterion because he or she believes the criterion to be strategically favorable to his or her credibility, a higher likelihood for its occurrence does not necessarily follow: Whether or not the criterion will emerge should then hinge on its cognitive properties, that is, how difficult it is for the deceiver to integrate the criterion in his or her statement (*differential controllability*, Köhnken, 1990).

In briefer words then, two considerations require clarification when assessing a criterion's diagnostic value: (1) To what extent is the deceiver *inclined* to produce the criterion and (2) to what extent would the deceiver *be capable* of doing so.

Defining the scope of this Thesis

To reiterate, the empirical findings available to date imply that overall CBCA validly discriminates true from fabricated statements. There are, however, theoretical concerns as well as practical reservations among scholars about the utility of CBCA in substantiating truthfulness. The research that was conducted within the scope of this Thesis was designed to address some of the more prominent points frequently criticized in the literature, in particular the alleged absence of a theoretical footing for the compilation of the criteria and the lack of a weighting system that would sort the individual criteria according to their level of diagnosticity.

Regarding the first main point of concern, Köhnken (1990) introduced the concept of creative and strategic demands and in this way did offer a broad theoretical explanation of why the criteria in general are more likely to appear in true rather than in false statements.

Admittedly though, Köhnken laid out his theoretical considerations only after the original version of CBCA (Steller and Köhnken, 1989) had already been published, rendering it difficult to determine to what extent these considerations had influenced the authors in compiling the criteria. Even if the concept of creative and strategic demands may not have been acknowledged then (i.e. when the compilation of criteria was developed), creative and strategic demands may nonetheless validly explain why the criteria in general are more likely to appear in true rather than false statements. At least this appears to be a question worth examining: To the author's knowledge, the idea of creative and strategic demands as the driving forces behind CBCA-related differences between truthtellers and liars has not been empirically tested so far.

To be more precise, experimental CBCA studies typically required participants to either report truthfully from memory or to fabricate events that were in fact never experienced and in this way did manipulate the presence of creative demands. The presence of strategic demands, on the other hand, is thought to vary with the level of motivation to appear credible (Vrij, Fisher, et al., 2008). Systematically varying the incentives for being believed should hence manipulate the presence of strategic demands. Most experimental CBCA studies appear to have, however, neglected to provide such incentives to begin with, and the few studies which did provide incentives tended to keep their distribution identical across trials. The experimental paradigm of the first study (study I) therefore included two different interview settings (T1 and T2), in each of which participants had to tell a true and false autobiographical event. While the first setting was designed to manipulate creative demands only, the second setting compelled participants to additionally deal with strategic demands. In sum then, the presence of creative and/or strategic demands should consistently differ across the different interview conditions, allowing to test whether the two types of task demands affect participants CBCA sum scores in ways that correspond to Köhnken's (1990) proposition.

The second and third study (study II and study III) refer to potential differences in the diagnostic value of the criteria, which if illuminated, could equip practitioners with valuable insight for assessing the truthfulness of statements as accurately as possible. To accomplish this endeavour, the Thesis builds on the notions that the diagnostic value of any criterion is to be derived from both its motivational and cognitive component (Niehaus et al., 2005) and that the compilation of criteria can be grouped more meaningfully into three different sets of criteria as suggested by the revised CBCA model (Volbert & Steller, 2014). More precisely, the two separate studies were carried out to test whether the three-dimensional structure of the revised model would contain criteria that within each set yielded comparable motivational (i.e. ascribed strategic meaning; study II) and cognitive (i.e. difficulty associated with their production; study III) properties.

In order to examine the motivational properties of the criteria, study II inquired via an online survey about the individual content-related deception strategies of laypersons. Participants were asked to assume the perspective of a protagonist who within a brief story outline had to come up with a convincing excuse to successfully deceive his interlocutor. Based on the fictitious scenario, participants had then to indicate for each CBCA criterion whether its presence in the protagonist's excuse would rather weaken or strengthen his credibility. The relevant question for analysis was then to what extent the obtained patterns of strategic ratings would correspond to the structure of the revised model, or phrased differently, to what degree the composition of each of the three criteria sets would contain criteria that on the motivational level are compatible with each other.

Referring to the findings from study II as starting point and building on the experimental paradigm of study I, study III subsequently investigated the cognitive properties of the criteria. The experimental paradigm entailed the two interview settings already described for study I. Other than before, the participants were coached between the two sessions about the sets of

criteria that had received largely negative strategic ratings in study II. This coaching manipulation ensured that participants were propelled to produce the criteria in the second setting, and thus allowed to test how well participants were capable of doing so by comparing the presence of the criteria in statements obtained from the first setting (T1; baseline conditions) to the presence of criteria in statements obtained from the second setting (T2; target conditions). As baseline and target conditions also differed in regards to their evocation of strategic demands (absent versus present) that may also have affected participants' CBCA scores, the results obtained from participants in study I were further used as point of reference for meaningfully discerning the effects of coaching. That is, the magnitude of the change in CBCA scores from baseline (T1) to target (T2) condition was again compared to the change in CBCA scores observed for participants who had received no coaching, but otherwise provided their statements under identical conditions (the same participants already recruited as the experimental group for study I, now serving as the control group in study III). Figure 1 illustrates the experimental design of study I and study III as well as their different paths of analysis in relation to the underlying research questions:

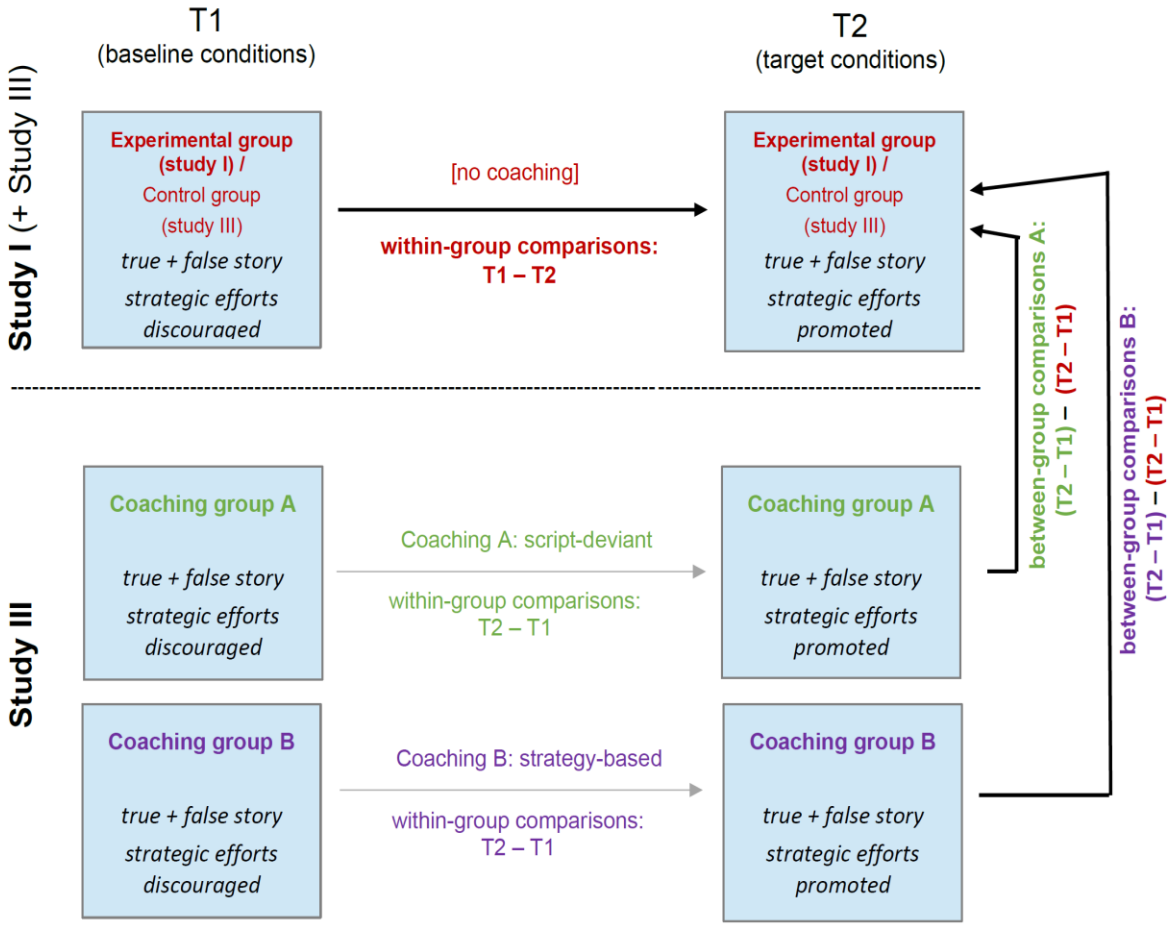

Figure 1. Illustration of the experimental design pertinent to study I and study III as well as their underlying paths of analysis.

In summary, the primary research questions of this Thesis pertain to the theoretical footing of CBCA criteria in substantiating truthfulness (study I) and to the possibility that some criteria might be diagnostically superior to others (study II and study III). The revised three-dimensional model from Volbert & Steller (2014) served as the underlying foundation for analysis, which promises to be particularly relevant for study II and study III: If their results should yield clear differences in motivational and/or cognitive properties between the criteria sets of the revised model, their collective findings would then allow to weigh the criteria sets according to their diagnostic value. For illustration, sets whose criteria consistently carried negative strategic meaning (motivational component) and were difficult to produce (cognitive component) would be diagnostically most valuable, whereas sets whose criteria carried positive strategic meaning and were comparatively easy to produce would be diagnostically least valuable.

Study I: When lying requires more efforts than just to fabricate: The effects of creative and strategic demands on statements' content quality

Study I has not been published; the following sections, therefore, describe the study in full detail.

Introduction

Creative and strategic demands as the key components of cognitive load

Lie detection approaches based on cognitive theory, such as the *traditional* or *innovative* cognitive load account (see Vrij, Fisher, et al., 2008), appear most fitting to explain why honestly reporting witnesses produce statements with more CBCA criteria than lying witnesses (Volbert & Steller, 2014). They rest on the assumption that lying tends to be cognitively more demanding than truth-telling, a proposition for which empirical evidence from various sources had been reported (for an overview, see Vrij & Granhag, 2012). There are two ubiquitous and cognitively demanding aspects that apply to truth-tellers but not to lying witnesses: Truth-tellers report from actual memory, and in doing so are convinced that the event in question had happened as reported (Volbert & Steller, 2014). A lying person, on the other hand, needs to put (more) deliberate effort in inventing information that substitutes the missing memory of the alleged experience (creative demands or efforts related to primary deception; Köhnken, 1990) and in masking the discrepancy between statement and belief by presenting the fabricated event in a credible manner (strategic demands or efforts related to secondary deception; Köhnken, 1990). With cognitive demands increasing, liars are more likely to revert to simpler strategies that may conserve their mental resources but compromise their performance on more complex tasks, such as fabricating a statement in a detailed and consistent fashion (Blandon-Gitlin et al., 2014). Consequently, and in line with the basic notion that statement content varies as a function

of the cognitive processes involved in its production (Undeutsch, 1954, as cited in Oberlader et al., 2016), experience-based statements should be more elaborate than fabricated statements (Köhnken, 1990).

Vrij et al. (2010) further refined the description of task demands by proposing six principles that make lying cognitively more demanding (Table 3). They fit neatly into the distinction of creative versus strategic demands, as most of the (sub)principles either refer to the aspect of *what* content the liar is producing (creative demands) or address *how* this content is delivered (strategic demands).

Table 3. Six principles that make lying cognitively more demanding (Vrij et al., 2010).

Principle	Primarily addresses:
1. Formulating the lie	
a. Inventing a story	creative demands
b. Monitoring one's fabrications to maintain plausibility and consistency	creative demands
2. Monitoring and controlling one's own behavior	strategic demands
3. Monitoring the interviewer's reaction carefully	strategic demands
4. Reminding oneself to act and roleplay	strategic demands
5. Suppressing the truth when fabricating	/
6. Intentionally and deliberately activating the lie	/

Previous experimental CBCA studies and their potential limitations

Elaborating on the notion that lying is not per se mentally more taxing than telling the truth (Burgoon, 2015; McCornack, 1997), Vrij, Fisher, et al. (2008) emphasized that lying is cognitively more demanding to the degree that the six principles – or their corresponding creative or strategic task demands, respectively – are in effect. As experimental CBCA studies

generally require participants to fabricate stories that were not previously experienced, the presence of creative demands appears to be comparable to the ones deceivers in forensic situations are facing: In both situations, the interviewee typically must construct his or her statement from cognitive scripts (e.g. Schank & Abelson, 1997) to substitute for the lack of experience-based memories.[2] Strategic demands on the other hand are likely to come into play only if the interviewee is sufficiently motivated to appear credible (Vrij, Fisher, et al., 2008).

Among the 39 CBCA laboratory studies that were included in the recent meta-analysis of Oberlader et al. (2016), only 11 investigations[3] appear to have provided incentives to motivate participants to lie or tell the truth well. Being already the exception rather than the rule, these incentives entailed monetary gains that in most cases were rather trivial than large in nature (see O'Sullivan et al., 2009).[4] Their effectivity may thus have been limited, given that the size of the incentives or consequences constitutes an important feature to consider (Ekman, 1985/2001). Furthermore, in less than half of these studies any motivational measures beyond offering monetary gains were undertaken to make the task emotionally more engaging or personally more relevant for participants, such as introducing the prospect of negative consequences if not being believed (Vrij & Mann, 2006; Vrij et al., 2007) or presenting the interviewer as being experienced with the detection of deception (Janka, 2003; Rutta, 2001; Wrege, 2004). Pertaining to the decades-long controversy of how well findings from laboratory

[2] If applicable though, statement providers may incorporate previously experienced elements in their false story, both in actual (Colwell et al., 2015) and experimental settings (Leins et al., 2013). Dependent on the extent to which these truly experienced elements can be relied on, such strategy may arguably mitigate the strength of creative demands associated with fabricating.

[3] Oberlader et al. (2016) classified 12 studies that provided incentives to motivate participants. For one study (Vrij, Edward, et al., 2000) the author of this Thesis could not find any descriptions in the original article that would hint at any incentives being provided for telling the truth or lying well.

[4] In only 2 of the 11 studies, the reward for a convincingly delivered report exceeded the value of 10€ (Vrij & Mann, 2006; Vrij et al., 2007).

CBCA studies transfer to real-life situations (e.g. Arntzen, 2011; Steller & Köhnken, 1989), it thus appears that previous CBCA studies may have widely neglected to evoke strategic task demands.

Theoretical considerations about the effects of creative and strategic demands

Volbert and Steller (2014) consequently speculated that "differences between truthful and fabricated statements due to variations in cognitive load or self-presentation strategy should be smaller [in experimental studies] than in real forensic situations" (p. 210). In terms of cognitive difficulty, strategic demands put an additional strain on the cognitive resources available to the interviewee and thus should impede his or her verbal performance. This detrimental effect on verbal performance should be markedly stronger when fabricating, since simulating memories is already in itself cognitively more difficult than truthfully reporting them, leaving less cognitive resources available to the interviewee to cope with the additional strategic demands. Truth-tellers in turn may also be prompted to behave strategically – i.e. according to Vrij and Granhag (2012), high-stake situations may affect the behavior of truth-tellers and liars in a similar way – and hence may likewise experience higher cognitive load when reporting. Yet, their verbal performance should remain largely unaffected: Being able to simply report from genuine memory, the coping with deliberate strategic efforts to appear credible should hardly compromise their ability to tell the event in a detailed and consistent fashion. In brief then, the presence of strategic task demands seems unlikely to impair verbal performance unless paired with creative demands. If being equally introduced to truth-tellers and liars, strategic demands should hence further pronounce the content-related differences between true (strategic demands) and false (creative *and* strategic demands) statements.

Aims and hypotheses

The goal of study I is to empirically test these assumptions by contrasting two different interview settings, with each setting being comprised of a true (reporting from memory) and false (fabricating) condition: The first setting (T1; labelled $SE_{[-]}$ in study I) was intended to evoke creative demands only and hence precluded any incentives to engage in strategic efforts (i.e. strategic efforts being *discouraged*). The second setting (T2; labelled $SE^{[+]}$ in study I) was designed to compel participants to additionally cope with strategic demands (i.e. strategic efforts being *promoted*). Therefore, measures were undertaken to create a life-like set-up that resembled actual forensic situations as closely as possible, along with the implementation of significant monetary rewards for being believed. Rather than just fabricating a story devoid of any concerns to appear credible as required in the false condition of the first setting ($SE_{[-]}$), the second setting ($SE^{[+]}$) thus required participants to not only fabricate but to successfully deceive in the false condition.

As illustrated in Table 4, the degree to which creative and/or strategic demands are present should therefore differ across the four interview conditions of our experimental paradigm: Creative demands should solely be present when fabricating (false conditions of $SE_{[-]}$ and $SE^{[+]}$), whereas the presence of strategic demands should depend on participants' motivation to appear credible (true and false condition of $SE^{[+]}$).

Table 4. Conditions of the experimental paradigm and the assumed presence of task demands.

	True	*False*
$SE_{[-]}$	-	creative demands
$SE^{[+]}$	strategic demands	creative + strategic demands

As pointed out above, the presence of creative and/or strategic demands should both add to the magnitude of cognitive load experienced. If these assumptions were to be confirmed, the amount of perceived cognitive load hence needed to be lowest for the true condition of $SE_{[-]}$ (neither creative nor strategic demands are present) and highest for the false condition of $SE^{[+]}$ (both creative and strategic demands are being evoked).

Based on the stipulation that creative demands, as well as strategic demands if paired with creative demands, impair verbal performance, the following four hypotheses regarding differences in statements' content quality between conditions were predicted:

H1: Within interview setting $SE_{[-]}$, true statements have higher CBCA scores than fabricated statements ($SE_{[-]true} > SE_{[-]false}$).

H2a: Comparing the true conditions, the CBCA scores of statements do not differ between the two interview settings ($SE_{[-]true} = SE^{[+]}_{true}$).

H2b: Comparing the false conditions, statements obtained in interview setting $SE_{[-]}$ have higher CBCA scores than statements obtained in $SE^{[+]}$ ($SE_{[-]false} > SE^{[+]}_{false}$).

H3: Within interview setting $SE^{[+]}$, true statements have higher CBCA scores than fabricated statements ($SE^{[+]}_{true} > SE^{[+]}_{false}$). In comparison to interview setting $SE_{[-]}$, the magnitude of this difference between true and fabricated statements is further intensified.

Methods

Participants and Design

A total of 30 participants[5] (16 females) were recruited through advertisement at an online market platform (*eBay Kleinanzeigen*; *n* = 24) or via mailing lists for students enrolled at universities in Berlin. For recruiting, the study was announced as an experiment about "the relationship between creativity and linguistic expression in describing personal events"[6] with the possibility of earning between 25€ and 50€. Only participants above the age of eighteen, with high proficiency in German and no previous knowledge about SVA, were eligible to participate. Their age ranged from 18-57 old and their average age was *M* = 35.97 (*SD* = 11.84). Most participants were working professionals from various fields (*n* = 11), followed by university students (*n* = 10), apprentices (*n* = 3) and persons being unemployed (*n* = 3). Participants received compensation of either 50€ or, if applicable, 25€ in combination with student credit for participation. After initial participation, a total of seven individuals had to be replaced by new participants for the following reasons: Refusal to be interviewed at the forensic institute (1), no further appearance after the preparatory (5) or first interview session (1).

The study involved a 2 (truth status: true versus false) x 2 (strategic efforts: discouraged versus promoted) within-subjects design.

For the selection of the appropriate sample size an a priori estimation was conducted using the G*Power software package (version 3.1.9.2; see Faul et al., 2007). The estimation

[5] In total, 90 participants had been recruited. Among this sample, 30 participants were retained for further analysis in study I, as the remaining 60 participants underwent experimental manipulations that go beyond measuring the effects of creative and strategic demands on statements' content quality (see study III). The subset of participants relevant for the study I were randomly distributed within the whole sample.

[6] The study was conducted exclusively in German. Any literal descriptions of study instructions or questionnaire items in this Thesis were thus directly translated from German language.

for a repeated-measures ANOVA was limited to the variable truth status, with the effect size *d* being based on the study of Oberlader et al. (2016) who reported an overall effect size of *d* = 0.95 for CBCA scores to discriminate between true and false reports. The power analysis (with the following parameters: $d = 0.95$, $\alpha = .05$, power = .95) indicated that a sample size of at least $N = 21$ was required to test for differences between true and false reports. Since strategic efforts (as additional within-subject variable) were predicted to pronounce the expected differences between true and false reports, a sample size of $N = 30$ was deemed clearly sufficient.

Materials and Procedure

General Procedure

The procedure consisted of three phases, with each participant undergoing a preparatory session followed by two separate interview sessions ($SE_{[-]}$ and $SE^{[+]}$). Within each interview session, participants were to report one event based on actual experience (true condition) and one event based on fabrication (false condition). The two true and two false events selected by the participant in the preparatory session were randomly assigned[7] to $SE_{[-]}$ and $SE^{[+]}$. Similarly, the order of event sequence within $SE_{[-]}$ and $SE^{[+]}$ was previously determined in a randomized controlled fashion[8]. Considering that in forensic situations prospective interviewees are typically informed about the interview well in advance, a time gap of minimum three days (maximum 23 days) between any consecutive sessions was maintained to maximize ecological

[7] For both true and both false events, each participant was equally likely to either report the event chosen first in $SE_{[-]}$ and the event chosen last in $SE^{[+]}$, or to report the event chosen last in $SE_{[-]}$ and the event chosen first in $SE^{[+]}$.

[8] Half of the participants were instructed to start with the true event in $SE_{[-]}$ and then to start with the false event in $SE^{[+]}$; the other half of participants were instructed to start with the false event in $SE_{[-]}$ and then to start with the true event in $SE^{[+]}$.

validity. The preparatory and the first interview session ($SE_{[-]}$) took place at one of the seminar rooms of a private University (*Psychologische Hochschule Berlin*), the second interview session ($SE^{[+]}$) took place in the interview room of an institute (*Zentrum für Aussagepsychologie Berlin*) shared by forensic psychologists to carry out credibility assessments mandated by German courts.

Preparatory Session

In the preparatory session, the study investigator repeated the general purpose of the study as previously outlined for recruiting. After participants had signed an informed consent form an instruction sheet was handed out and read aloud by the study investigator, pointing out to participants the following three requirements to be considered when choosing suitable events for the upcoming reports: The events in question should not date back more than ten years in time (1), the nature of the events should be unique and bear personal significance rather than pertain to trivial everyday-life occurrences (2), and the unfolding of the events should have lasted several minutes and entail own actions as well as interactions with at least one other person (3).[9]

Regarding true reports, it was stressed that the events needed to be told as happened, without containing any exaggerations or fictional elements. Concerning false reports, it was emphasized that the events' core elements must be the sole product of one's own imagination in order to thwart the adoption of strategies that would allow participants to remain close to

[9] These requirements were implemented to control for potential event-related factors that might affect the content quality of either true or fabricated accounts, such as the fading of memories with the increase of time-length between event and interview. Furthermore, the requirements should ensure that specific events characteristics (e.g. interactions with other persons) were present as otherwise the emergence of CBCA criteria related to such occurrences (e.g. reproduction of conversation) would have been precluded a priori.

truth-telling (i.e. by relying on original memories when fabricating; Leins et al, 2013). Next, the following instruction informed participants about their prospective tasks:

> For each event, you should try to provide a convincing and credible account. You are supposed to create the impression in listeners that you have actually experienced your truthfully-reported as well as your fabricated events. It is particularly important to report the events as detailed and extensive as possible. The report of each event should roughly be five minutes in length.

A preselection of 12 topics was then provided, among which participants had to select two different topics for their true events and two different topics for their false events. Endorsing Steller's (1989) recommendation to tailor CBCA study events towards cases of sexual abuse allegations, the topics available for selection were intended to simulate the emotional and experiential valence of real-life forensic situations related to such cases. Therefore, only situations were included in the final topic selection in which the statement provider would be likely to be directly involved, to be negatively emotionally aroused, and to feel a loss of control (see Steller et al., 1992), such as being the victim of a criminal act, being attacked by an animal or having a serious accident. Table 5 displays the entire range of topics available for selection.

Table 5. The selection of topics as presented to participants and their frequency of being reported.[10]

(1)	Being a victim of an attempted or committed criminal offense (e.g. burglary, theft, robbery, rape, molestation, blackmailing, etc.). [*n* = 16]
(2)	Suffering an accident with subsequent medical assistance/treatment (e.g. in traffic, at sports, at work, at home, etc.). [*n* = 20]
(3)	Being caught of and sanctioned for an illicit or embarrassing act (e.g. receiving a monetary fine or warning). [*n* = 12]
(4)	Experiencing existential feelings of fear or panic related to one's life or health. [*n* = 8]
(5)	Getting lost in nature or wildness (e.g. when hiking). [*n* = 11]
(6)	Losing a significant amount of money in the course of a risky endeavor (such as gambling). [*n* = 3]
(7)	Disclosure of an affair previously kept secret by yourself or by your (former) partner. [*n* = 6]
(8)	A sudden and unexpected event of death within your close family or peer network. [*n* = 6]
(9)	Failing in a personally important task or assignment. [*n* = 17]
(10)	Following someone else's advice with devastating consequences. [*n* = 1]
(11)	Being attacked by an animal or another person. [*n* = 10]
(12)	Experiencing a high-risk situation due to one's own or other person's negligence. [*n* = 9]

[10] The number of times each topic was reported by participants is provided in square brackets. The total number of reports amounts to $N = 119$, since one report from the total sample of $N = 120$ had to be excluded.

Only after suitable topics had been selected, the study investigator conveyed to participants that forensic psychologists professionally trained in credibility assessment were interested to test whether individuals with high story-telling abilities could successfully deceive them. Therefore, if in the upcoming interview session the participant's story-telling abilities were judged to be sufficiently high, he or she would be invited to have a second interview session at a "forensic institute"[11], with the prospect of higher monetary gains. In detail, the study investigator addressed the participants in the following way:

> Just recently we happened to form a collaboration with psychologists being employed at a forensic institute. The psychologists, who are professionally trained in assessing the credibility of statements and provide expert testimony at court, consider the research focus of our study relevant for their work. They are particularly interested in whether individuals with high story-telling abilities would be able to successfully lie to them. Therefore, in the upcoming interview session, after having listened to your two stories, I will briefly evaluate your story-telling abilities. If they seem sufficiently high, I will invite you to have the final interview session at the forensic institute. In this case, the final interview session would take the same amount of time, but you would be interviewed by a forensically trained psychologist and would have the possibility to earn up to twice as much (50€). Otherwise, if you do not qualify in the upcoming interview session or decide then to reject the invitation, our study will simply proceed as originally planned. The final interview session then will take place here again, I will again be your interviewer, and you will receive the fixed amount of 25€ upon completion.[12]

[11] Strictly speaking, the term forensic is not an accurate description for the institute. For practical purposes, the study investigator used this term when providing details of the institute to participants, with the intention to keep instructions brief. Technically, the institute reflects a collaboration of freelancing psychologists trained in credibility assessment.

[12] In case participants had asked about the prospect of additional earnings before the selection of topic events was finished, the study investigator deliberately refused any information to minimize the likelihood that participants would choose events strategically (i.e. choosing topics that may be easier to fabricate and thus increase one's chances of achieving higher monetary rewards, etc.).

Subsequently, participants filled out several questionnaires that inquired about their demographic and socio-economic background. At the end of the preparatory session, participants received a take-home sheet on which the date and location for the first interview session ($SE_{[-]}$) were noted. The sheet also repeated the previously outlined requirements regarding event characteristics and task instructions and further indicated which two events participants were to tell first and second in the next session.

First Interview Session ($SE_{[-]}$)

On average six days later (M = 6.34; SD = 3.7), the first interview session ($SE_{[-]}$) took place at the same location (*Psychologische Hochschule Berlin)* as the preparatory session. In the beginning, the study investigator repeated the task requirements and motivated participants to report both events in a detailed and comprehensive way. At the same time, the study investigator sought to minimize any incentives to engage in strategic efforts by highlighting that he was already aware of the event's underlying truth status, rendering any explicit attempts of deception irrelevant. In the false condition, participants should hence be motivated to fabricate well (creative demands) but be discouraged to engage in any strategic efforts to deceive (creative and strategic demands). In detail, the study investigator addressed the participants in the following way (illustrated here for participants who, based on the randomly assigned order of event-sequence, were instructed to start with the true event in $SE_{[-]}$):

> As I had indicated last week on your take-home sheet, you will now first tell me your true event and then tell me your false event. As pointed out before, depending on the way you deliver the two events, you may qualify to have your final interview session at a forensic institute, with the prospect of earning higher rewards. For both events, I will first let you freely report and then will ask you a few more detailed questions. Unless you have further questions, I will now start the recording device and begin with the first question about the true event.

The average numbers of words uttered by the participants were *M* = 1107.63 (*SD* = 763.81) for the true and *M* = 1067.00 (*SD* = 560.30) for the false condition. After the interviews about the true and the false event were finished, participants filled out a questionnaire that among other things inquired about their prior preparation efforts and their subjective experiences during the interviews.

The study investigator then indicated that the participant would fulfill the necessary criteria for the second interview session to take place at the forensic institute and provided further details about the prospect of winning additional monetary rewards: Apart from the fixed amount of 25€, 20€ could be further gained if the participant succeeded in making the interviewer believe that both his or her true and false reports were based on actual experiences. By requiring that both events had to be judged as true participants should be deterred from deliberately lowering the quality of their true account to make their fabricated report appear more convincing in comparison. With the intention to promote strategic efforts related to the principle "carefully monitoring the interviewer's reaction" as stated by Vrij et al. (2010; see Table 3), participants were further instructed that additional 5€ could be gained if they correctly predicted how the interviewer would rate the truth status of their true and false account. Similarly, having the (sub)principle "monitoring one's fabrication to maintain plausibility and consistency" in mind, participants were informed that at the end of the session a second interviewer might appear to elaborate further on some aspects they had been previously outlined to the first interviewer, in case the first interviewer needed assistance in determining the truth status of their reports. A take-home sheet was handed out again at the end of the first interview session ($SE_{[-]}$). The sheet was largely identical to the sheet provided at the end of the preparatory session, but this time also stressed the need to conceal the truth status of the two reports from the interviewer.

Second Interview Session ($SE^{[+]}$)

On average eight days later (*M* = 8.5, *SD* = 2.57), the second and final interview session ($SE^{[+]}$) took place at the forensic institute. The study investigator received the participants at the street entrance and led them into the interview room. After repeating the task requirements, he again motivated participants to report both events in a detailed and comprehensive way. In contrast to the previous interview session ($SE_{[-]}$), his instruction this time was designed to motivate both intrinsically (i.e. task of deceiving a psychologist allegedly trained in credibility assessment) and extrinsically (i.e. prospect of higher monetary rewards) to not only fabricate but to engage in strategic efforts in order to successfully deceive in the false condition (creative and strategic demands). In detail, the study investigator addressed the participants in the following way:

> Remember, you will only receive additional 20€ if the interviewer - a forensic psychologist trained in credibility assessment - considers both your reports to be true. After the interviews, you will be asked to predict your success rate for both your true and false event on a questionnaire. If you are correct, you will gain additional 5€. Also, keep in mind there is the possibility that yet another interviewer will elaborate on some aspects of your reports, in case your first interviewer is unable to reach a clear decision for either event. Otherwise, I will directly return to this room and let you know how the first interviewer has decided. Okay, then I will leave the room now and tell the interviewer that you are ready. I wish you the best of luck!

The study investigator then left the interview room, and one of the confederates entered. The confederate briefly introduced him- or herself to the participant, thereby expanding on his role of being a forensic psychologist experienced with assessing the credibility of statements. Referring to the preassigned versions of the interview protocol (refer to section "Interview structure" for more detail), he or she conducted the interview in the same way as the study investigator had done during the first interview session ($SE_{[-]}$). The average numbers of words spoken by the participants were *M* = 1103.30 (*SD* = 649.10) for the true and *M* = 1239.23 (*SD* = 599.49) for the false condition. After participants had finished with both reports, the

confederate handed over a questionnaire and left the room. The questionnaire required participants to predict how the interviewer would rate the truth status of their reports but otherwise was largely identical to the one handed out in $SE_{[-]}$. Some minutes later, the study investigator reentered the room and handed over another questionnaire that inquired to which degree the experimental manipulations were successful (i.e. whether participants believed that the interviewer had in fact been a trained forensic psychologist). Afterwards, participants were fully debriefed about the true identities of the confederates and the actual purpose of the study. All participants received the full amount of financial compensation (50€), regardless of performance and prediction accuracy.[13]

Interviewers

All interviews during $SE_{[-]}$ were conducted by the study investigator (and author of this Thesis; male, age 29), and all interviews during $SE^{[+]}$ were conducted by one of three confederates. The confederates (2 females, age 26 and 27; 1 male, age 29) were graduate students in psychology. The number of participants interviewed was roughly equally distributed among the confederates (range 8-12).

Interview structure

In each interview, the interviewers started with an open-ended question to initiate free narratives and subsequently proceeded with increasingly more narrowly phrased follow-up questions. Sequence and phrasing of the follow-up questions were varied across four versions of a standardized interview protocol. The specific versions were preassigned to each interview

[13] As done by Vrij & Mann (2006), participants were told that both their stories had been judged true by the interviewer. Without this pretense, participants would have realized that they would have been paid the full amount regardless of performance, and the 25€ bonus might no longer have been an effective incentive for subsequent participants.

condition in a randomized controlled fashion. This manipulation was intended to keep the structure of the interviews as similar as possible across conditions while reducing potential training effects that were likely to arise if the interviews' structure had been kept identical throughout the experiment.

Prior to the start of the experiment, the study investigator trained the confederates to apply each version of the interview protocol consistently and systematically. The training was conducted in one session of three hours, in which each confederate performed a test interview, with the study investigator serving as the interviewee and feedback provider.

The following three paragraphs provide a complete description of the standardized interview protocol and its four versions:

Each version of the interview protocol started with the question *"About what event are you reporting now? Please indicate the topic."* Next, free reports were initiated by providing the following instruction: *"Please report to me what had happened. Please describe the event as extensive as possible".* When participants appeared to have exhausted their free reports, two subsequent retrieval attempts were made. First, the interviewer identified and briefly rephrased the key moment (i.e. climax) of the interviewee's report and prompted the interviewee to describe this moment again ("*Can you describe this one more time; how did you experience this situation?*"). Second, the interviewer identified and briefly rephrased the element of the report that he found most complex or difficult to comprehend, asking the interviewee to elaborate further for clarification (*"Can you explain this to me again please?"*).

While the previous questions were kept identical across all protocol versions, the order and exact phrasing of the following three questions were systematically varied: One question inquired about the consequences of the event (i.e. "*Have you learned something from this experience?"; "When did you last think of this event?; "Did the event have longer-lasting consequences for you?"; "Does the experience of the event still affect you today?*"), one

question referred to the context of the event, with each protocol version stating three options among which the interviewer had to choose the one currently best applicable (i.e. "*What did you do immediately before the event?"; "Why were you at the location where the event happened?"; What did you do immediately after the event?"*) and one question was supposed to address elements that the interviewee was unlikely to have anticipated. This "unanticipated" question should be phrased in such a way that simple Yes- or No-answers would be precluded, appear innocuous to the interviewee, and relate to the broader context of the participant's report. If the interview was unable to come up with a suitable question, the protocol provided the following options among which the one appearing most promising was to be selected: *"How were the weather conditions? What clothes did you wear? What exactly was your position relative to the other person(s)? How exactly did the person/object/environment look like?"*.

The final question was kept identical again in all protocol versions and inquired whether there was anything left the interviewee wanted to add before the interview was to be concluded ("*Is there something else that you would like to add?")*.

Questionnaires

After participants had finished their true and false reports, questionnaires were presented in $SE_{[-]}$ and $SE^{[+]}$ pertaining to the following domains:

Perceived cognitive load

Identical to the approach of Vrij and Mann (2006), cognitive load was measured for each of the four interview conditions by collecting participants' responses to the following three items: "The interview required a lot of thinking", "The interview was mentally difficult", and "During the interview, I had to concentrate a lot". The answers could be provided on a

seven-point Likert scale ranging from (1) certainly not to (7) certainly and were clustered into the "cognitive load" variable (Cronbach's alpha = 0.94).[14]

Motivation to perform well

For the first interview session ($SE_{[-]}$) participants' *motivation* was assessed for the true and false condition collectively by the following item: "How important is it for you to qualify for the next interview session at the forensic institute?" (the answer could be given on a seven-point Likert scale ranging from (1) not at all important to (7) highly important). For the second interview session ($SE^{[+]}$) participants' motivation was assessed separately for the true and false condition by the items "How motivated were you to convince the interviewer that your true event really happened?" (true condition) and "How motivated were you to convince the interviewer that your fabricated event really happened?" (false condition). The answers could again be given on a seven-point Likert scale ranging from (1) not at all motivated to (7) highly motivated. Furthermore, participants were asked whether aspects other than the prospect of higher monetary rewards contributed to their motivation, and if yes, to briefly elaborate.

Efforts of preparation

To assess the extent to which participants prepared themselves before the interviews, participants had to indicate for each interview condition whether they had taken notes about relevant story details ("Did you take written notes about the content of the story prior to the interview?"), whether they had practiced their story presentation in any way ("Did you practice your presentation of the story in any way prior to the interview, such as presenting

[14] In comparison, Vrij and Mann (2006) reported a Cronbach's alpha value of 0.71 after clustering the three items into one variable.

the story to a friend?") and whether they had prepared in any ways other than previously outlined ("If not already addressed by any of the previous questions, did you prepare the presentation of your story in any other way?"). For the false conditions only, an additional fourth question inquired about the use of external sources for inventing the story ("Did you use any sources of information, such as books or movies, when inventing your story?").

Subsequently, participants' preparation efforts were classified into (0) low versus (1) high, by using the following rating scheme: A score of "0" was assigned unless at least two of the three or four (false condition) questions had been marked with "yes".

Predictions about the interviewer's ratings ($SE^{[+]}$)

In $SE^{[+]}$, participants had to predict whether the interviewer would rate the report's underlying truth status ("How do you think the forensic psychologist will rate your true/false story?") as either true or false.

All participants (n = 30) provided predictions for both their reports. Regarding their true report, most participants (83.3%, n = 25) predicted their story to be rated as true. In turn, less than half of the participants (43.3%, n = 13) predicted such an outcome for their false report.

Perceived authenticity of the second interview setting ($SE^{[+]}$)

To check the success of the experimental manipulations, the questionnaire handed out in $SE^{[+]}$ entailed the following additional items: "Did you believe that the interviewer in front of you was, in fact, a forensic psychologist trained in assessing the credibility of statements?" and "Did you believe that after the interviews a second interviewer might appear to ask you further questions?". Possible answer options were "no", "yes, but I had some doubts", or "yes". Also, participants were further asked to indicate the extent to which the prospect of being questioned again by a second interviewer had affected them during their true and false

reports (the answers could be given on seven-point Likert scales ranging from (1) not at all to (2) very strongly).

CBCA Analysis

Compilation of CBCA criteria

The set of content characteristics used in the current study differed in some ways from the traditional CBCA criteria as described by Steller and Köhnken (1989), since the revised CBCA model as proposed by Volbert and Steller (2014) was used as reference. To this model some modifications were applied to accommodate the purpose of this study. Instead of 19 criteria, the final set used for the study's rating procedure comprised 25 criteria. The criteria *logical consistency* and *quantity of details* were erased because their coding would have required a different coding method (scale rates; refer to section "Coding of the interview transcripts" for more detail). The criteria *reality control* and *raising doubts about one's own person* were added, and the following criteria of the traditional version were modified: *Contextual embedding* was divided into three separate criteria, namely *temporal information*, *spatial information*, and *information about everyday-life routines*. Similarly, the criterion *account of mental subject state* was divided into the criteria *sensory impressions, emotions and feelings,* and *own thoughts*[15]. Furthermore, criteria that referred exclusively to criminal contexts were adjusted to extend their applicability to non-criminal contexts (e.g. the criterion *details characteristic of the offense* was transformed into *details characteristic of the event).*

From this final set of 25 criteria, the criteria *accurately reported details misunderstood*, *reality control*, *raising doubts about one's own person* and *pardoning the perpetrator* were coded, but subsequently excluded from further analysis because of their low overall occurrence ($n < 1.5$).

[15] For a concise description of the newly added or modified criteria see Niehaus (2001), pp. 121-130.

Rater Training

The interviews were transcribed verbatim (total number of transcripts: 119)[16], and copies were provided to two CBCA-trained raters. The first rater was a forensic psychologist with experience both in research as well as in the forensic application of CBCA, the second rater was a graduate student in psychology who carried out the ratings as part of her master's thesis project. Prior to the rating of the transcripts, the second rater received extensive training in CBCA scoring by the first (expert) rater to ensure a consistent and standardized approach when rating each criterion. The training included the reading and discussing of literature that entailed detailed descriptions and examples of the criteria. Both raters also assessed the presence of criteria in several transcripts not contained in the current study on a 4-point Likert Scale (ranging from (0) not present to (4) strongly present) until they had reached an at least fair level of interrater-reliability ($ICC \geq .4$; see Cicchetti, 1994).

Coding of the interview transcripts

Each transcript was rated independently by both raters who were blind to the specific experimental conditions (i.e. they had no knowledge about the truth status or the presence of strategic demands). For each CBCA criterion, the raters first counted how often the criterion occurred in the transcript (frequency counts). Following the common procedure, repeated information was not counted twice (i.e. Vrij & Mann, 2006). Second, the raters estimated how *strongly* each frequency count related to the criterion by assigning weights to each identified occurrence (1 = weakly present, 2 = strongly present). Third, the weighted frequency counts were summed up to form the final score of the individual criterion. Most CBCA studies used

[16] One report (false condition, $SE_{[-]}$) was excluded from the total sample of $N = 120$, as the participant had provided the report in such a way that its event condition (truth status) would have been easily discernible for the CBCA raters.

3-point (e.g. Blandon-Gitlin et al., 2009) or 5-point (e.g. Akehurst et al., 2011) Likert scales to rate the presence of each criterion in the entire transcript as coding method (scale rates). More importantly though, weighted frequency counts combine the methodological advantages of frequency counts over scale rates in evaluating the presence of content characteristics (see Nahari, 2016) with the benefits of scale rates in taking the criterion's strength of appearance into account. The raters, therefore, used weighted frequency counts and carried out this procedure for all transcripts in multiple intervals of 20 to 30 transcripts each. Between such intervals, the raters gathered repeatedly to reappraise the consistency of their individual ratings in order to enhance reliability. More specifically, they made sure that their frequency counts referred to the same text passage (an essential aspect which is often ignored in the literature; see Sporer, 2012).

Reliability analysis

Pearson product-moment correlations for the final ratings of each criterion by each rater were conducted to obtain consistency estimates as a realistic measure of inter-rater reliability (see Akehurst et al., 2001). For all 21 criteria the ratings were significantly and positively correlated and except for *sensory impressions* ($r = .27$) all criteria yielded at least fair too good inter-rater reliability scores of $r > .40$ (see Goedert et al., 2005), ranging from $r = .54$ (*temporal information*) to $r = .94$ (*description of interactions, external related associations*). All criteria but *sensory impressions* were thus retained for further analysis, resulting in a total set of 20 criteria used for final analysis. Table 6 provides a complete display of the obtained Pearson correlation values.

Table 6. Pearson's r values for individual criteria (and for the CBCA sum score).

#	Criterion	Pearson's *r*
1	Information about everyday-life routines	.55*
2	Spatial information	.62*
3	Temporal Information	.54*
4	Descriptions of Interactions	.94*
5	Reproduction of conversations	.83*
6	Emotions and feelings	.72*
7	Own thoughts	.74*
8	Sensory impressions	.27*
9	Attribution of other person's mental state	.74*
10	Personal implications	.55*
11	Unexpected complications	.86*
12	Superfluous details	.85*
13	Unusual details	.74*
14	Related external associations	.94*
15	Details characteristic of the offense	.80*
16	[Accurately reported details not comprehended]	-
17	Unstructured production	.68*
18	Spontaneous corrections	.72*
19	Admitting lack of memory	.87*
20	Efforts to remember	.70*
21	[Reality controls]	-
22	Raising doubts about one's own testimony	.93*
23	[Raising doubts about one's own person]	-
24	Self-deprecation	.78*
25	[Pardoning the perpetrator]	-
	CBCA sum score	.87*

Note:* $p < .01$.
[ ] Excluded from final analysis ($n < 1.5$).

Based on the set of 20 criteria, the two raters' final (weighted frequency) criterion scores were averaged for each criterion to form the mean criterion score. These mean criterion scores were then summed up to form the mean CBCA sum score (i.e. the sum of the criteria ratings for each transcript averaged across raters), which served as the dependent variable for further

analysis. Also, the final criterion scores for each rater were summed up separately to obtain rater-specific CBCA sum scores. These rater-specific CBCA sum scores correlated significantly and revealed a product-moment coefficient (r = .87) that can be taken to indicate excellent inter-rater reliability (see Goedert et al., 2005).

Results

Manipulation check

Perceived cognitive load

To test whether creative (within each interview setting, participants experience higher cognitive load when fabricating) and strategic demands (irrespective of truth status, participants experience higher cognitive load in $SE^{[+]}$ than in $SE_{[-]}$) both add to the amount of cognitive load perceived by participants, a 2 (truth status) x 2 (strategic efforts) repeated measures ANOVA was conducted with cognitive load as the dependent variable. There were no significant outliers and data were normally distributed as assessed by boxplots. The main effects for truth status, $F(1,29) = 22.91$, $p < .001$, $\eta_p 2 = 0.44$ and strategic efforts, $F(1,29) = 10.3$, $p = .003$, $\eta_p 2 = 0.26$, showed significant differences in the amount of cognitive load perceived between trials, which confirmed our notion that both creative and strategic demands lead to an increase in the amount of cognitive load perceived. The interaction between these two factors was statistically not significant, $F(1,29) = 0.01$, $p = .946$, $\eta_p 2 < 0.01$, implying that incentives to engage in strategic efforts were equally effective in adding to cognitive load perceived, regardless of whether participants were lying or telling the truth. Correspondingly, a paired sample-test (based on the difference scores of $SE^{[+]}_{false} - SE_{[-]false}$ and $SE^{[+]}_{true} - SE_{[-]true}$) revealed that the mean increase in cognitive load from $SE_{[-]}$ to $SE^{[+]}$ for the false conditions (M = 0.78, SD = 1.55) did not differ significantly from the mean increase in cognitive load for the true conditions (M = 0.80, SD =

1.68), M_{Diff} = -0.02, 95% CI [-0.69, 0.64], $t(29) = -0.068$, $p = .946$, $d = 0.01$. Figure 2 illustrates the effects of truth status and strategic efforts on experienced cognitive load graphically.

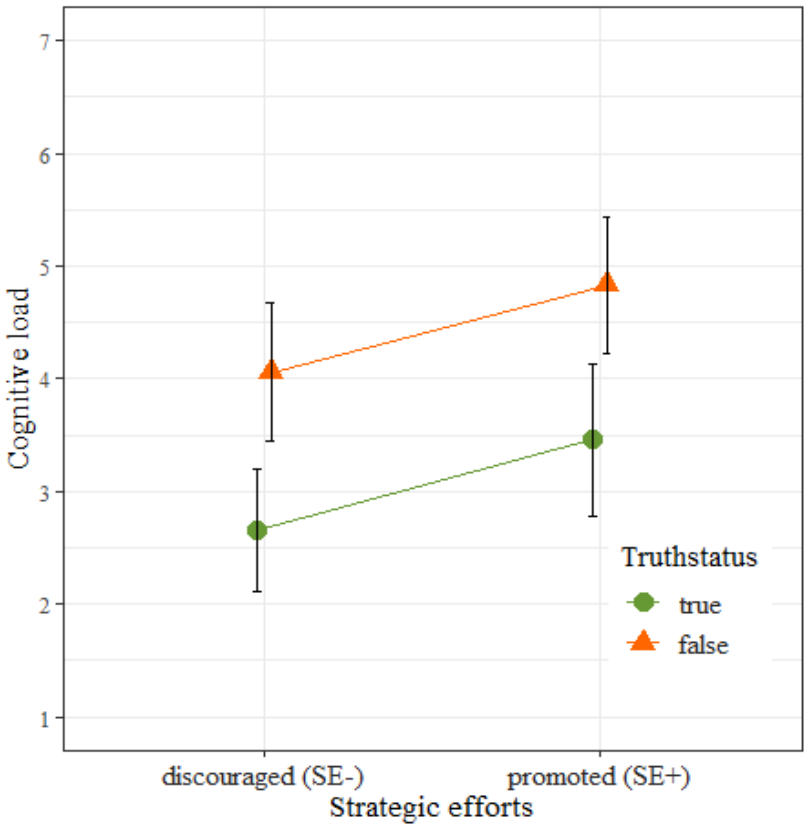

Figure 2: Amount of perceived cognitive load in relation to truth status and strategic efforts, with 95% confidence intervals.

Motivation to perform well

Since participants were instructed that qualification for $SE^{[+]}$ was dependent on their "storytelling" performance as displayed in $SE_{[-]}$, participants' indications of how important they deemed qualifying for the second interview session were interpreted as an indirect measure of their motivation to tell both events in $SE_{[-]}$ in a convincing way. More than half of the participants (56.7%, $n = 17$) indicated that they considered it to some extent important (score > 4 on the seven-point Likert scale) to qualify for the second interview session ($SE^{[+]}$) at the

forensic institute (M = 4.47, SD = 1.59). Regarding $SE^{[+]}$, the vast majority of participants reported to have been to some extent motivated (scores > 4 on the seven-point Likert scales) to convince the interviewer that the event had actually happened in both the true (76.7%, n = 24) and the false (90%, n = 27) condition. This motivation to convince the interviewer in $SE^{[+]}$ did not differ significantly between the true (M = 5.87, SD = 1.25) and false (M = 6.13, SD = 1.01) event condition, M_{Diff} = -0.27, 95% CI [-0.68, 0.15], $t(29) = 1.313$, $p = .199$, d = -0.24. When averaging the obtained scores of the true and false condition for $SE^{[+]}$ into one motivational variable[17], a comparison of motivational scores between $SE_{[-]}$ (M = 4.47, SD = 1.59) and $SE^{[+]}$ (M = 6.00, SD = 0.99) revealed significant higher scores for $SE^{[+]}$, M_{Diff} = 1.53, 95% CI [0.95, 2.11], $t(29) = 5.407$, $p < .001$, d = 0.99.

Regarding financial compensation, most participants (80.0%, n = 24) stated that apart from the prospect of earning additional success-related monetary rewards, financially unrelated aspects had also propelled them to perform at their best during the interviews, such as finding out about one's own lie abilities under challenging conditions (n = 13) or exhibiting a more general interest towards the experiment's area of research (n = 5).

Preparation efforts

Table 7 displays the number of participants whose preparation efforts prior to the interview were classified as low versus high (= at least two of the relevant three or four questions had been positively answered) in the respective interview condition as well as their associated mean CBCA sum scores and standard deviations. As can be seen from Table 7, in each interview condition the number of participants with low preparation efforts was larger than the number of participants with high preparation efforts. Because of the small and unequally distributed

[17] With the benefit of hindsight, it appears that a methodologically superior approach would have been to inquire about participants' motivation in $SE_{[-]}$ separately for the true and false condition.

sample sizes, testing for statistical differences in CBCA sum scores between groups seemed unwarranted.

Table 7. Number of participants with *low* versus *high* preparation efforts and the associated mean CBCA sum scores and std. deviations for each condition.

Strategic efforts	*Truth status*	*low preparation efforts*			*high preparation efforts*		
		n	*M*	*SD*	*n*	*M*	*SD*
discouraged	true	22	17.98	12.10	8	20.12	12.07
discouraged	false	21	13.43	5.12	8	18.19	7.98
promoted	true	23	14.02	5.76	7	17.93	7.42
promoted	false	20	17.37	6.52	10	23.10	13.23

Perceived authenticity of the second interview setting ($SE^{[+]}$)

The experimental manipulations in $SE^{[+]}$ (e.g. interviewer being introduced as forensic psychologists trained in credibility assessment, emphasizing the prospect of additional success-related financial rewards) were designed to bring the features of the underlying interview setting closer to forensic situations, with the intention to encourage participants to engage in strategic efforts. The results suggest that these manipulations had been successful: All but one participant (96.7%, n = 29) indicated that they had conceived their interviewer in $SE^{[+]}$ as a forensic psychologist professionally trained in credibility assessment (among those, 37.9% specified to have had experienced some doubts about the interviewer's true identity, but overall still had believed that the interviewer was in fact who he or she pretended to be). Similarly, most participants (66.7%, n = 20) expressed that they had believed that a second interviewer could emerge at the end of $SE^{[+]}$ to ask further questions (among those, 10% reported to have had experienced some doubts). Despite this high rate, participants predominantly stated that the prospect of being questioned by a second interviewer had only weakly affected them when

reporting, both for the true (76.7%, n = 23) and the false (80.0%, n = 24) condition (scores < 3 on the seven-point Likert scales; M_{true} = 1.79, SD_{true} = 1.40; M_{false} = 1.80, SD_{false} = 1.35).

Preliminary considerations

The found differences in perceived cognitive load between conditions indicate that the experimental design successfully manipulated cognitive difficulty by varying creative and strategic demands when reporting. Yet, the experimental measures in SE[+] designed to evoke strategic demands could also have affected participants' motivation in a more general sense. For example, the prospect of higher monetary rewards may have (also) motivated participants to perform well during the interviews or to prepare for them thoroughly, in ways unrelated to any strategic efforts. Both participants' general motivation and preparational efforts may in turn have affected their experience of cognitive load and/or verbal performance during the interviews: Vrij and Mann (2006) for instance found that CBCA scores were positively correlated with the extent to which participants were keen to perform well, while Burgoon (2015) concluded that communicators are better able to approximate credible communication patterns with advanced preparation. For these reasons, the variables motivation and preparation efforts were entered into the final model of the main analysis, to account for their potentially confounding effects on the mediating (cognitive load) and/or dependent (CBCA sum score) variable.

Main Analysis

R (R Core Team, 2019) and *lme4* (Bates et al., 2015) were used to perform a linear mixed-effects analysis of the relationship between the response variable CBCA sum score and the independent predictor variables strategic efforts and truth status. As fixed effects, strategic efforts and truth status (with interaction term), as well as motivation and preparation efforts

(without interaction term), were entered into the model. As random effects, an intercept for subjects and a by-subject random slope for preparation efforts as well as their covariance were included to represent individual differences in the overall level of achieved content quality (as measured by the CBCA sum score) and in the effect of preparation efforts. Based on theoretical considerations and tested by comparing reduced models via maximum likelihood-ratio tests, the model thus included the maximal random effect structure justified by the data (i.e. Barr et al., 2013).

Visual inspection of residual plots did not reveal any obvious deviations from homoscedasticity or normality. Calculation of variance inflation factor (VIF) values for the full model indicated collinearity to be absent between any combination of fixed factors (VIF < 3; see Zuur et al., 2010).

Tests of the parameter estimates and associated *p*-values were obtained via restricted maximum likelihood estimation in conjunction with Satterthwaite's degrees of freedom method using *lmerTest* (Kuznetsova et al., 2017). The variable motivation was scaled with -1 to render its zero-point equivalent to the lowest value on the 7-point Likert Scale as provided on the questionnaire (1 = "not at all motivated"). With the interaction between the two categorical factors (strategic efforts and truth status) found to be significant in our model, orthogonal sum-to-zero contrasts rather than treatment contrasts were used to obtain more meaningful parameter estimates (i.e. *main effects* rather than *simple effects*)[18], presented in Table 8. Note that the value of the intercept corresponds to the grand mean of the response variable, with the lower-order effects for strategic efforts and truth status being estimated at the level of the grand mean. To

[18] Under conditions in which higher-order effects are present (i.e. interaction), treatment contrasts yield parameter estimates that do not accurately represent lower-order effects (e.g. main effects), as these lower-order effects are estimated at the level of the baseline and hence represent simple effects rather than main effects (for more detail, see Singmann & Kellen, 2019).

illustrate: For models with categorical factors and higher-order effects, orthogonal contrasts allow for the interpretation of both higher- and lower-order effects. In such contrast schemes, the intercept corresponds to the grand mean and lower-order effects are estimated at the level of the grand mean. In the subsequent analysis, parameter estimates were obtained via *effects coding*. For factors with only two levels, the effect-coded parameter value is equal to half of the difference between the two conditions (for more detail, see Singmann & Kellen, 2019).

Table 8. Multilevel model (effect coded) parameter estimates for mean CBCA sum score.

	Mean CBCA sum score			
Predictors	*Estimates*	*SE*	*95% CI*	*p*
Fixed Effects				
(Intercept)	16.85	2.69	11.59 – 22.12	**<0.001**
Strategic efforts: Promoted	0.18	0.66	-1.12 – 1.47	0.787
Truth status: False	-0.05	0.52	-1.08 – 0.97	0.918
Motivation	0.11	0.52	-0.91 – 1.14	0.827
Preparation efforts: High	1.50	1.17	-0.79 – 3.79	0.198
Strategic efforts x Truth status	2.01	0.52	1.00 – 3.03	**<0.001**
Random Effects				
Residual variance (σ^2)	30.68			
Subject intercept variance (τ_{00})	58.60			
Subject Preparation variance (τ_{11})	9.21			
Intercept, Preparation covariance (ρ_{01})	0.87			
Observations	119			

The analysis yielded only a significant strategic efforts x truth status interaction (b = 2.01, SE = 0.52, $p < .001$), suggesting that discouraging versus promoting strategic efforts affected participants' CBCA sum scores differently, depending on the statement's underlying truth status. The control variables motivation and preparation efforts had only weak effects on participants' CBCA sum scores, with each increase in the level of motivation being associated with a 0.11-point increase in the mean CBCA sum score and with high preparation efforts yielding reports that on average were 3.00 points higher relative to reports provided under low preparation efforts. To get a better understanding of the significant interaction between the two factors strategic efforts and truth status, estimated marginal means based on the output of the linear mixed-effects model were computed for each combination of factor levels separately, using *emmeans* (Lenth, 2018). The results are presented in Table 9.

Table 9. Estimated marginal means, standard errors and 95% confidence intervals.

Strategic efforts	*Truth status*	*Margin*	*SE*[*]	*95% CI*
discouraged	true	19.23	1.93	15.29 - 23.16
discouraged	false	15.09	1.95	11.13 - 19.06
promoted	true	15.56	1.90	11.70 - 19.42
promoted	false	19.48	1.88	15.65 - 23.31

Note: [*] Degree of freedoms calculated via Satterthwaite's method.

The estimated marginal means show the mean CBCA sum score for each level combination of strategic efforts and truth status, adjusted for the control variables motivation and preparation efforts. Figure 3 depicts the estimated marginal means and their associated error bars:

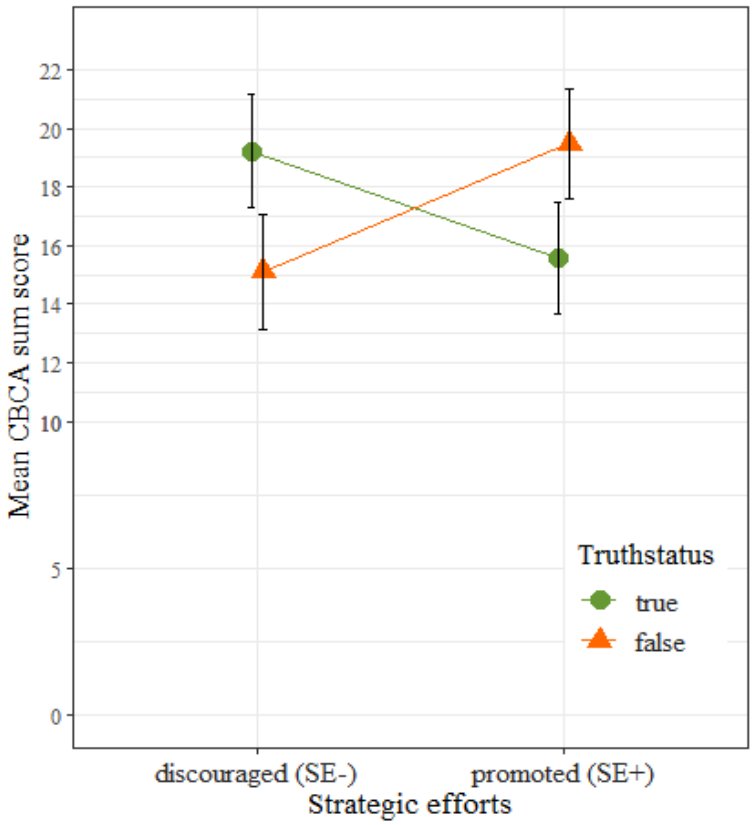

Figure 3: Mean CBCA sum score by strategic efforts and truth status, adjusted for participants' motivation and preparation efforts, with error bars (±1 SE).

Hypotheses testing

As can be deduced from the pattern shown in Figure 3, no overall effect of either strategic efforts or truth status was present, but there was a cross over interaction. That is, the effect of strategic efforts on mean CBCA sum scores was opposite, depending on the underlying truth status but irrespective of participants' motivation and preparational efforts. Such pattern seems to accord with the generally proposed notion that the effects of strategic efforts are moderated by truth status but clearly contradicts the more specifically formulated hypotheses, which speculated that participants' strategic efforts would impair the content quality of their false, but not true statements. To address each of the hypotheses individually, pairwise comparisons were run to test for significant differences among the conditions; presented in Table 10.

Table 10. Pairwise comparisons and their mean differences, standard errors, 95% confidence intervals, effect sizes *d*, t-values and associated p-values.

Strategic efforts/Truth status	M_{Diff}	*SE*	*95% CI*	*d*	*t*	*p*
(discouraged/true) vs (discouraged/false)	4.13	1.47	[1.25 - 7.01]	0.39	2.81	**0.031**
(discouraged/true) vs (promoted/true)	3.67	1.63	[0.47 - 6.86]	0.35	2.25	0.117
(discouraged/true) vs (promoted/false)	-0.25	1.71	[-3.60 - 3.10]	-0.02	-0.15	0.999
(discouraged/false) vs (promoted/false)	-4.38	1.73	[-7.77 - -0.99]	-0.42	-2.53	0.061
(promoted/true) vs (discouraged/false)	0.46	1.66	[-2.79 – 3.71]	0.04	0.28	0.992
(promoted/true) vs (promoted/false)	-3.92	1.47	[-6.80 - -1.04]	-0.38	-2.66	**0.046**

Table 10 indicates that if strategic efforts were being discouraged ($SE_{[-]}$), the CBCA scores of true statements were on average significantly higher than the CBCA scores of false statements (M_{Diff} = 4.13, 95% CI [1.25, 7.01], t = 2.81, p = .031, d = 0.39), supporting Hypothesis 1 (H1: $SE_{[-]true} > SE_{[-]false}$).

Seemingly in line with the prediction regarding true statements and with Hypothesis 2a (H2a: $SE_{[-]true} = SE^{[+]}_{true}$), the discouragement ($SE_{[-]true}$) versus promotion ($SE^{[+]}_{true}$) of strategic efforts did not yield significant differences in CBCA scores (M_{Diff} = 3.67, 95% CI [0.47, 6.86], t = 2.25, p = .117, d = 0.35). Yet, without testing for *equivalence* (Wellek, 2010) a nonsignificant test result should not be interpreted as evidence for the absence of a true effect or difference (Lakens, 2017). Using the R package *TOSTER* (Lakens, 2017), the "two one-sided tests" (TOST) procedure (Schuirmann, 1987) was thus carried out, a simple equivalence testing approach that can be used to statistically reject the presence of effects large enough to be considered meaningful (for more details, see Lakens, 2017). First, a lower (-0.53) and upper

(0.53) equivalence bound was prespecified based on the smallest effect size[19] (d_z = 0.53) that the study would have sufficient power to detect.[20] As the two effect sizes (d_z = 0.51 and d_z = -0.48) that were associated with statistically significant differences between conditions fall within the equivalence bounds, any effect outside of this range can safely be considered meaningful. The subsequently performed equivalence test indeed indicated that the presence of effects more extreme than the equivalence bounds could not be rejected ($t(29) = -0.66$, $p > .05$), implying that the observed difference in CBCA scores between conditions is not close enough to zero to be practically equivalent (Seaman & Serlin, 1998). That is, albeit the mean difference between $SE_{[-]true}$ and $SE^{[+]}_{true}$ was statistically not significant, there may still be a meaningful difference between these conditions that is smaller than the study could reliably detect. Hence, Hypothesis 2a is rejected.

For false statements the discouragement ($SE_{[-]false}$) versus promotion ($SE^{[+]}_{false}$) of strategic efforts yielded CBCA scores that in $SE_{[-]}$ were lower than in $SE^{[+]}$, resulting in a difference that was marginally significant (M_{Diff} = -4.38, 95% CI [-7.77, -0.99], t = -2.53, p = .061, d = -0.42). This difference in scores between $SE_{[-]false}$ and $SE^{[+]}_{false}$ being directly opposite to the direction originally predicted, Hypothesis 2b (H2b: $SE_{[-]false} > SE^{[+]}_{false}$) is rejected.

Consequently, with strategic efforts being promoted ($SE^{[+]}$), the CBCA scores of true and false statements differed significantly, M_{Diff} = -3.92, 95% CI [-6.80, -1.04], t = -2.66, p =

[19] Unless otherwise specified, Cohen's *d* was calculated manually by using the standardized mean difference for within-subject comparisons, which is the difference between measures divided by their averaged standard deviation. Cohen's *d* ignores the correlation between the measures and is generally recommended for reporting effect sizes (Lakens, 2013).

Alternatively, another Cohen's *d* family effect size, Cohen's d_z for correlated measurements, was calculated by dividing the t-value by the square root of *n* (Rosenthal, 1991). Cohen's d_z is required for conducting equivalence tests for differences between dependent means (Lakens, 2017).

[20] A power analysis was run to calculate the smallest effect size our study can detect in a null effect significance test, with N = 30 (pairs), 80% desired power, and an α of .05.

.046, d = -0.38. Other than expected, this difference was not driven by true statements having higher scores than false statements, but by false statements having higher scores than true statements. Thus, Hypothesis 3 (H3: $SE^{[+]}_{true} > SE^{[+]}_{false}$) is rejected.

Discussion

The study examined how creative and strategic task demands affect participants' CBCA scores when reporting autobiographical events. Participants had to tell a true and false story each across two different interview settings. The first setting ($SE_{[-]}$) was designed to discourage strategic efforts to appear credible by highlighting that the interviewer knew the underlying truth status already in advance, requiring participants to only fabricate in the false condition (creative demands). In contrast, the second setting ($SE^{[+]}$) entailed a variety of experimental measures designed to promote additional strategic efforts in participants, meaning that apart from fabricating, participants needed to successfully deceive in the false condition (creative demands paired with strategic demands). As these measures may have affected participants' motivation to perform well in ways unrelated to their inclinations to behave strategically and, in a similar way, may have influenced their preparational efforts prior to the respective interviews, the potential effects of both "general" motivation levels and preparation efforts on CBCA sum scores were controlled for in the main analysis.

General findings

In reference to the premises of cognitive lie detection approaches, it was stipulated that the absence of creative demands would result in higher CBCA scores for true relative to false statements (Hypothesis 1). Within the first setting ($SE_{[-]}$) true reports were associated with lower cognitive load and, in line with the first hypothesis and corresponding to the findings of most previous CBCA studies, indeed yielded significantly higher CBCA scores than false reports,

resulting in an effect size of small strength (d = 0.39; according to the classification of Cohen, 1988). With creative demands being paired with strategic demands in the false condition of the second setting ($SE^{[+]}$), the difference in CBCA sum scores between true and fabricated accounts was predicted to further increase (Hypothesis 3). Regarding cognitive load, participants perceived both the true and false conditions to be cognitively more taxing when strategic efforts were being promoted, which lends support to the theory-driven assumption that resource-intensive strategic demands come into play once the interviewee is sufficiently motivated to appear credible (Vrij, Fisher, et al., 2008; see Table 3). The true and false reports differed significantly as well in $SE^{[+]}$, but quite strikingly, this difference was now driven by CBCA scores that for the false condition in- rather than decreased. As the CBCA scores of truthful reports in turn declined relative to $SE_{[-]}$ the difference between true and false reports was further exacerbated in $SE^{[+]}$, resulting in a negative effect size of small strength (d = - 0.38; according to the classification of Cohen, 1998). In this way, the prediction of significantly higher CBCA sum scores for true relative to false reports in $SE^{[+]}$ (Hypothesis 3) was turned upside-down, with the observed pattern of scores between conditions being contrary to the underlying rationale of CBCA. To the author's knowledge, a negative effect size of such magnitude in an experimental CBCA study is unprecedented.[21]

In summary then, the implementation of strategic task demands in $SE^{[+]}$ was associated with higher cognitive load both in the true and false condition but seemed to have affected the conditions' corresponding CBCA scores in opposite ways.

[21] For illustration, among their sample of 39 CBCA laboratory studies Oberlader et al. (2016) identified only three investigations with negative effect sizes, all of which were of small (Merckelbach, 2004; g = - 0.25) or lesser (Porter, Yuille & Lehmann, 1999; g = - 0.15; Willén & Strömwall, 2012; g = -0.08) strength.

Implications from the findings for true statements

For true reports ($SE^{[+]}_{true}$), the increase in cognitive load was associated with a tentative decrease in CBCA scores. Even though the resulting difference between $SE_{[-]true}$ and $SE^{[+]}_{true}$ was statistically not significant, equivalence testing showed their mean scores to possibly still differ in a meaningful way. While it was hypothesized that strategic demands would only exert detrimental effects on statements' content quality if being paired with creative demands (i.e. $SE^{[+]}_{false}$), the observed pattern may indicate that even in the absence of creative demands (i.e. $SE^{[+]}_{true}$), strategic demands negatively affect the content quality of statements.

Alternatively, it could be speculated that the tentative decline of CBCA scores in $SE^{[+]}_{true}$ reflects participants' inclinations to deliberately compromise the quality of their true report, with the intention to make their false story appear more convincing in comparison. Some of the participants had indeed mentioned during debriefing to have pursued this very strategy in $SE^{[+]}$, despite the experimental efforts to suppress strategic considerations of this kind by emphasizing that both the true and false story needed to be judged truthful by the (same) interviewer. Future studies could examine to what degree the assignment of a different interviewer person to each condition affects CBCA scores. From the perspective of participants, such a measure should render any efforts to intentionally lower their own performance futile and hence should minimize the likelihood of any such strategies being pursued. If the content quality of true statements were then still found to decline in the presence of strategic task demands, attempts to willfully perform worse under such circumstances could be ruled out as an underlying explanation.

Implications from the findings for false statements

In contrast, for false reports obtained in $SE^{[+]}$ the increase in cognitive load was associated with a tentative – marginally significant – increase in CBCA scores. In other words, (only) in the

condition in which participants were prompted to do their best to successfully deceive the interviewer ($SE^{[+]}_{false}$) cognitive load was positively associated with statements' content quality. It thus appears that participants were able to more than compensate for increased task demands when deceiving (i.e. $SE^{[+]}_{false}$, creative and strategic demands) and in this way produced statements that qualitatively surpassed their statements provided under lower task demands (i.e. $SE_{[-]false}$, creative demands only).

Maybe then, rather than reflecting how cognitively taxing participants perceived the task(s) to be, the amount of experienced cognitive load in $SE^{[+]}_{false}$ may primarily pertain to the cognitive efforts deceptive participants felt to have invested in the task(s). Against this background, it would be interesting to gauge whether the positive relationship between cognitive load and statements' content quality would cease to exist for deceiving participants at (even) higher thresholds of task difficulty. Several studies have had participants tell their narrative in reverse order (e.g. Vrij, Mann, et al., 2008; Evans et al., 2013) and found this technique to increase cognitive load as well as to improve deception detection (for an overview, see Vrij et al., 2017). Interestingly, Evans et al. (2013) included a sub-selection of seven CBCA criteria into their analysis of outcome variables and found the majority to discriminate between true and false stories in the higher cognitive load (reverse order) condition only. Since the authors provided no incentives to evoke strategic efforts in participants – rendering it questionable whether participants were trying their best to deceive – subsequent investigations could examine on a larger set of CBCA criteria whether implementing creative and strategic task demands in combination with the reverse order technique results in an outcome more in line with the current findings (i.e. significantly higher CBCA scores for false relative to true stories in the presence of strategic task demands) or more in the direction the findings of Evans et al. (2013) appear to point towards to (i.e. significantly lower CBCA scores for false versus true stories in the reverse order condition).

Limitations and further suggestions for future research

In this context, some important methodological limitations of the study need to be noted that future investigations should address. Except for the prospect of being interviewed by a second interviewer, the study asked participants to rate the general *success* of the experimental measures in $SE^{[+]}$ in terms of being perceived as authentic (e.g. "Did you believe that the interviewer was trained in credibility assessment?") but did not directly inquire about the measures' *effectiveness* in making them behave strategically. For this reason, one can only indirectly infer from the high authenticity generally attributed to these measures and the observed increase in cognitive load that strategic demands (or efforts) were effectively evoked in $SE^{[+]}$. Further research is needed to individually examine the effectiveness of these measures by implementing self-report procedures suitable for more direct assessment. In reference to the principles stated by Vrij et al. (2010; see Table 3), a questionnaire could, for instance, inquire how strongly participants had monitored their fabrications or their own behavior during the interviews.

Secondly, on a rather basic level, it remains to be established which of the different experimental measures in $SE^{[+]}$ (e.g. prospect of success-related monetary rewards, interviews taking place at a forensic institute, etc.) were the main driving forces behind the observed increase in both cognitive load and CBCA scores. As these measures were introduced in an "all at once" fashion, their individual effects on the outcome variables cannot be disentangled. Subsequent studies could, for example, vary the presence of certain measures (e.g. neglecting versus implementing success-related monetary rewards, introducing a neutral versus a forensically-related interview location) while keeping the presence of others constant (e.g. introducing the interviewer as experienced in credibility assessment and/or as being unaware of the underlying truth status).

Finally, it needs to be noted that the sample size of the study (N = 30) was rather small. Being of little concern for the study's main analysis in the form of within-subjects comparisons on mean CBCA sum scores, the reduced power and its associated risk of Type II errors would be problematic for the comparison of single CBCA criteria across conditions. Presumably, the statistical power of such computations would not be sufficient to detect all differences between conditions, rendering further elaboration or discussion impractical. Subsequent studies with larger sample sizes could meaningfully investigate how well single criteria discriminate between truthful and fabricated reports within different interview settings. From a theoretical perspective, such comparisons may yield important practical implications concerning the diagnostic value of individual CBCA criteria. For instance, several vignette studies found that when deceiving laypersons are inclined to integrate certain criteria while out of strategic considerations, they intend to avoid the expression of other criteria (for an overview, see Maier et al., 2018). As these studies only examined how laypeople rate the strategic meaning of the criteria in theory based on fictitious scenarios, investigations as outlined above could provide experimental settings appropriate to test in which ways content-related deception strategies translate to the practical level.

Conclusions

The aforementioned limitations notwithstanding, study I arguably represents a novel and valuable approach to test how creative and strategic task demands affect the content quality of statements. Relative to the first interview setting ($SE_{[-]}$), the study successfully promoted strategic efforts by approximating the structural features of the second setting ($SE^{[+]}$) to real-world forensic situations, thereby addressing the frequently accented need for investigations pertaining to settings high in mundane realism or ecological validity (e.g. Frank & Svetieva, 2012; Oberlader et al., 2016). If motivated to appear credible and hence to behave strategically

($SE^{[+]}$), participants were found to produce fabricated statements that had significantly higher CBCA sum scores than their truthful reports. In this way, the study's results starkly contrast with the extant CBCA literature. It can therefore be argued that under specific experimental circumstances individuals may be much better able to deal with the pairing of creative and strategic demands and produce fabricated statements of higher quality than commonly assumed (1). Furthermore, these findings appear to suggest that previous CBCA studies may have largely overlooked the aspect of promoting strategic efforts in participants and in this way may have neglected to differentiate between the acts of fabricating versus deceiving (2). Against this background, further research is needed to elucidate the specific conditions and contingencies under which fabricated statements tend to yield higher content quality than truthful ones, based on experimental settings that are attuned to the importance of integrating forensically relevant (i.e. creative and strategic) task demands.

Study II: The strategic meaning of CBCA criteria from the perspective of deceivers

Study II has been published in 2018. The following sections, therefore, describe the study in a rather concise and summarized fashion, with the full version of the study's published manuscript being displayed in Appendix A.

Introduction

The primary idea behind study II was to examine how laypersons assess the strategic meaning of CBCA criteria in the context of deception, that is, to elucidate whether their individual content-related deception strategies would prompt them to rather integrate (i.e. positive strategic meaning) or avoid (i.e. negative strategic meaning) the criteria when deceiving. Such insight about the motivational properties of the criteria promises to entail practical information about their diagnostic value, since criteria that laypersons tend to avoid are diagnostically superior to criteria that laypersons are inclined to integrate. Before study II was conducted, two investigations had been published that inquired about deceivers' verbal strategies in relation to CBCA (Niehaus et al., 2005; Niehaus, 2008b). Having applied a modified group structure to the original CBCA model, the authors of both studies found that participants rated most motivational criteria as strategically negative and that participants also ascribed (positive or negative) strategic meaning to cognitive criteria. Their modified CBCA model did not allow for intelligible group-based distinctions, however, since some of the newly devised groups contained criteria of both positive and negative ratings. Against this background, the relevant research question of study II was to test whether the three-dimensional structure of the revised model as introduced by Volbert and Steller (2014) would correspond better to the observed patterns of laypersons' strategic value ratings.

Based on the two previous studies (Niehaus et al., 2005; Niehaus, 2008b), it was predicted that participants would ascribe mostly positive strategic meaning to memory-related criteria (set 1) and predominantly negative strategic meaning to script-deviant (set 2) and strategy-based (set 3) criteria. Taken together, it was expected that for each of the three criteria sets the pattern of laypersons' strategic ratings would result in a degree of comparatively high "motivational" compatibility. Put differently, the strategic ratings should be either consistently negative or consistently positive within each criteria set.

Methods

A vignette was presented via an online questionnaire to 135 participants (M_{Age} = 28.6, Range 19-67; 32 men) to inquire about their content-related deception strategies. After having read the story outline in which a fictional protagonist was in need of delivering a convincing excuse to his boss, participants were asked to assume the perspective of the protagonist in order to assess the strategic meaning of 24 CBCA criteria on a five-point scale (ranging from -2 "No, this would strongly weaken my [the protagonist's] credibility" to +2 "Yes, this would strongly strengthen my [the protagonist's] credibility", with the neutral value 0 "My [the protagonist's] credibility would remain unchanged" being placed at the center of the scale).

Results

For each criterion, one-sample t-tests were computed to assess whether the criterion's average strategic rating differed significantly from the neutral value 0. Overall, participants considered 18 out of the 24 criteria to be strategically relevant (i.e. the strategic ratings of the criteria differed significantly from 0). In line with the predictions, all memory-related criteria (set 1) that participants regarded strategically relevant received positive value ratings. On the other

hand, and also as hypothesized, the strategic ratings for all script-deviant (set 2) and most strategy-based (set 3) criteria were of negative value.

Discussion

The findings of study II confirmed that laypersons consider most CBCA criteria strategically relevant and are inclined to either integrate (criteria with positive strategic meaning) or avoid (criteria with negative strategic meaning) them in their deceptive statements. Few exceptions notwithstanding, the individual predictions regarding the strategic meaning of each of the three criteria sets were widely confirmed: Participants indeed tended to ascribe rather positive strategic meaning to memory-related criteria (set 1) but tended to ascribe negative strategic meaning to script-deviant (set 2) and strategy-based (set 3) criteria. Overall, the three-dimensional structure of the revised model (Volbert & Steller, 2014) yielded a pattern of strategic value ratings that were largely homogenous and, in comparison to the models previously applied (Niehaus et al., 2005; Niehaus, 2008b), proved to be substantially higher in "motivational" compatibility.

These findings then provide valuable input for appraising potential differences in the diagnostic value of the criteria (sets). That is, script-deviant (set 2) and strategy-based (set 3) criteria appear to be diagnostically superior to the memory-based criteria from set 1, considering that the negative strategic meaning allocated to them should render their emergence in fabricated statements less likely. It is important to note, however, that such diagnostic evaluations are of rather heuristic than definitive nature since they only pertain to the motivational component of the criteria. Additional insight about the cognitive component is needed for more conclusive assessments of the criteria' diagnostic value.

Study III: Encouraging participants to integrate CBCA criteria into their statements: Different effects for different sets of criteria?

Study III has not been published; the following sections, therefore, describe the study in full detail.

Introduction

The cognitive component of CBCA criteria

Being built on the three-dimensional structure of the revised model (Volbert & Steller, 2014), study II (Maier et al., 2018) investigated the motivational properties of CBCA criteria and revealed consistent differences between the three criteria sets: While participants tended to ascribe rather positive strategic meaning to memory-related criteria (set 1), they tended to ascribe negative strategic meaning to script-deviant (set 2) and strategy-based (set 3) criteria. If one restricts diagnostic evaluations of the criteria to this "motivational" perspective, criteria from set 2 and set 3 appear to be diagnostically superior to the criteria from set 1 since their negative strategic meaning should lower their likelihood for emerging in fabricated statements.

Yet, the absence of criteria in a statement does not necessarily mean that the statement provider was reluctant to produce them. Vice versa, from a deceiver's sole willingness or predisposition to simulate certain criteria one cannot infer his or her actual capacity to do so (*differential controllability;* Köhnken, 1990). For a more conclusive assessment of the criteria' levels of diagnosticity additional insight from the "cognitive" perspective is needed, that is, empirical knowledge about the cognitive difficulty associated with the criterion's production is required (Maier et al., 2018). Assessing to what degree lying participants can simulate criteria is only feasible under conditions in which the participants consider efforts to produce them worthwhile. CBCA coaching studies that inform subjects about the positive strategic meaning

of the criteria promote such conditions and hence seem pertinent to address the question at hand. To the authors' knowledge, three studies that experimentally tested how coaching participants about CBCA criteria affects the content and quality of their statements have been published to date (Vrij, Kneller, et al., 2000; Vrij et al., 2002; 2004).

Previous CBCA studies about coaching participants

Vrij, Kneller, et al. (2000) found lying participants who were informed about some criteria to achieve similar CBCA scores than truth-tellers and significantly higher CBCA scores than uninformed liars. Against this background, the authors concluded that coaching statement providers may compromise the validity of CBCA but refrained from drawing definite conclusions yet as further, methodologically more advanced, studies were warranted.

The two subsequently conducted studies (Vrij et al., 2002; 2004) thus implemented a design that had several methodological advantages over the previous investigation, such as (also) informing truth-tellers about CBCA criteria and providing incentives for participants to appear credible. Both studies also recruited different age groups and, at least for the group of adult participants, largely replicated the previous findings. That is, the statements of informed liars yielded significantly higher CBCA scores than the statements of uninformed liars. Furthermore, CBCA scores between false and true statements differed significantly only if lying participants were not informed about the criteria. The authors thus concluded that coaching statement providers reduces the discriminatory power of CBCA assessments and consequently raised concerns about their use in forensic settings.

It is important to note, however, that the generalizability of the studies' results to actual forensic situations may be problematic: Rather than just informing participants about the criteria and how their occurrences increase the credibility of one's statements, the authors (e.g. Vrij et al., 2002; 2004) delivered detailed and explicit illustrations of how to integrate the criteria into

the statement to be provided. For instance, concerning the criterion *unexpected complications*, the following instruction was given: ´*Pretend that something unusual or unexpected happened. For example, pretend that you dropped some of your discs for the game on the floor*.´ Consequently, in order to accomplish the (re)production of the criterion in the subsequent interview, participants needed to solely restate the priorly given exemplary instruction (i.e. ´*I dropped some of my discs on the floor*.´) at a fitting moment. Such task of repeating unmodified phrases is markedly different and arguably less challenging than the task(s) statement providers had to deal with in forensic interviews (see Rutta, 2001), namely, to come up with suitable examples for the criteria on their own and to integrate them in a logically coherent and structurally consistent fashion into their narrative.

The current study

While Vrij, Kneller, et al. (2000) and Vrij et al. (2002, 2004) confined their coaching measures to about half of the CBCA criteria contained in the original version (Steller & Köhnken, 1989), study III paid heed to the established knowledge about individual deception strategies (Maier et al., 2018; see study II) and used the revised three-set model (Volbert & Steller, 2014) as the point of reference for determining the sets of criteria to be coached. That is, participants were informed either about script-deviant (coaching group A) or strategy-based (coaching group B) criteria, considering that the negative strategic meaning allocated to them should render coaching measures most feasible.

As underlying foundation the same experimental paradigm that was already described for study I was used: To briefly reiterate, participants had to provide a true and false statement about autobiographical events at a first interview session (T1; labelled $SE_{[-]}$ in study I), in which any concerns about being believed were rendered irrelevant by highlighting that the interviewer was already aware of the underlying truth status in advance (baseline conditions). The same

participants had then to provide a true and a false statement about different events at a second interview session (T2; labelled $SE^{[+]}$ in study I), whose structural features were approximated to real-world forensic situations to prompt participants to appear credible (target conditions). For illustration, apart from proposing significant monetary rewards for being believed, the study investigator had the interviews take place at a forensic institute and introduced the interviewers as psychologists professionally trained in credibility assessment.

To this study design, the experimental manipulation of coaching was added for study III. Thereby, when informing participants about the criteria (after T1 but before T2) the study investigator deliberately provided only generic examples to them. In this way, participants were required to devise their own examples and it was left to their discretion whether and how to integrate the criteria into their narratives later on. A time gap of at least three days was further ensured between informing participants about CBCA and conducting the subsequent interviews, to allow enough time for processing and preparing. Taken together, the design of study III allows for examining the effects of coaching within a forensically relevant setting (T2, target conditions), separately for true and false statements. Its primary question of interest is then to what extent the magnitude of changes in CBCA scores from the baseline (T1) to the target (T2) condition differs *between* informed (coaching group A or B) versus uninformed (control group) participants.

It is important to understand that this comparison of changes in CBCA scores from T1 to T2 *between* the three participant groups is crucial for validly measuring the effects of coaching, considering that for all three participant groups the interview settings differed systematically between T1 (strategic efforts being discouraged) and T2 (strategic efforts being promoted). Consequently, simply comparing the changes in scores from T1 to T2 for coaching group A or B separately would not allow for clear inferences about the strength of coaching effects, since the differences between the two interview settings may have likewise affected

participants' CBCA scores (see study I for more detail or Figure 1 for a graphical display of the experimental paradigm).

Hypotheses

Concerning the true statements of uninformed participants, the findings of study I indicated a tentative decline in CBCA scores from the first (T1) to the second (T2) interview session. It was hence speculated that in T2 participants may have deliberately compromised the quality of their true report with the intention to make their false report appear more convincing in comparison. This being the situation, it was expected for study III that informed participants would be less likely to intentionally lower the quality of their true report in T2: Their newly gained knowledge about either script-deviant or strategy-based criteria should render alternative strategies – such us deliberately integrating the criteria into their false report – more promising. Hence, the following two hypotheses were formulated:

H1: Regarding true statements, the change in CBCA scores from the baseline ($T1_{true}$) to the target ($T2_{true}$) condition differs significantly between participants informed about script-deviant criteria (coaching group A) and uninformed participants (control group). More specifically, the decline in CBCA scores from $T1_{true}$ to $T2_{true}$ is less pronounced for participants informed about script-deviant criteria than for uninformed participants.

H2: Likewise, the change in CBCA scores from $T1_{true}$ to $T2_{true}$ differs significantly between participants informed about strategy-based criteria (coaching group B) and uninformed participants (control group). Again, the decline in CBCA scores from $T1_{true}$ to $T2_{true}$ is less pronounced for participants informed about strategy-based criteria than for uniformed participants.

Concerning false statements, the study's question of interest is to what extent the coaching measures enable participants to simulate the respective criteria. From a theoretical perspective, this task should be comparatively easy to accomplish with strategy-based criteria (coaching group B): Primarily referring to simple expressions of memory-related deficits (e.g. *admitting lack of memory*) or of doubting one's own credibility (e.g. *raising doubts about one's own testimony*), simulating strategy-based criteria should require neither particular cognitive efforts nor sophisticated skills (Köhnken, 1990). Script-deviant criteria (coaching group A; e.g. *unusual details* or *unexpected complications*), on the other hand, should be cognitively more difficult to simulate, since they refer to event-related characteristics that go beyond the limitations of simplified, script-guided knowledge (Maier et al., 2018). When fabricating an event, participants cannot draw on actual memories to retrieve such script-deviant elements and instead would need to overcome the limited scope of their own imagination for being able to produce them (Köhnken, 1990). Consequently, the following two hypotheses were formulated:

H3: Regarding false statements, the change in CBCA scores from the baseline ($T1_{false}$) to the target ($T2_{false}$) condition does not differ between participants informed about script-deviant criteria (coaching group A) and uninformed participants (control group).

H4: In contrast, the change in CBCA scores from $T1_{false}$ to $T2_{false}$ differs significantly between participants informed about strategy-based criteria (coaching group B) and uninformed participants (control group). More specifically, the increase in CBCA scores from $T1_{false}$ to $T2_{false}$ is more pronounced for participants informed about strategy-based criteria than for uniformed participants.

Methods

Hint: As the experimental paradigm and the analysis performed for study III are partly built on the paradigm and data used for study I, some parts of the following sections have already been described in the Methods sections of study I. For ease of reading, methodological details that were deemed essential for comprehending the design and analysis of study III are provided in full detail in the following sections, irrespective of any previous descriptions. Methodological details that seemed less relevant for understanding the design and analysis of study III are described in abbreviated form, with footnotes indicating that a more elaborate description can be found in the Methods section of study I.

Participants and Design

A total of 90 participants (52 females) were recruited through advertisement at an online market platform (eBay Kleinanzeigen; n = 63) or via mailing lists for students enrolled at universities in Berlin. For recruiting, the study was announced as an experiment about "the relationship between creativity and linguistic expression in describing personal events"[22] with the possibility of earning between 25€ and 50€. Only participants above the age of eighteen, with high proficiency in German and no previous knowledge about forensic credibility assessment were eligible to participate. Their age ranged from 18-60 old and their average age was M = 33.76 (SD = 10.37). Most participants were working professionals from various fields (n = 40), followed by either university (n = 33) or high-school (n = 3) students, persons being unemployed (n = 10) and apprentices (n = 4). Participants received compensation of either 50€ or, if applicable, 25€ in combination with student credit for participation. After initial participation, a total of 17 participants had to be replaced in the course of the study for the

[22] The study was conducted exclusively in German. Any literal descriptions of study instructions or questionnaire items in this article were thus directly translated from German.

following reasons: Refusal to be interviewed at the forensic institute (1), no further appearance after the preparatory (15) or first interview session (1).

The study involved a 2 (truth status: true versus false) x 2 (interview session: T1 versus T2) x 3 (coaching: none, informing about script-deviant criteria, informing about strategy-based criteria) mixed-subjects design.

The selection of the appropriate sample size was based on the a priori estimation that was conducted for study I, indicating that a sample size of $N = 30$ was clearly sufficient for a single participant group to test for differences between true and false reports. Consequently, for study III an identical sample size was sought for each of the two additionally recruited participant groups (i.e. coaching group A and coaching group B), resulting in an overall sample size of $N = 90$.

Materials and Procedure

General Procedure

The procedure consisted of three phases, with each participant undergoing a preparatory session followed by two separate interview sessions (T1 and T2). Within each interview session, participants were to report one event based on actual experience (true condition) and one event based on fabrication (false condition). The two true and two false events selected by the participant in the preparatory session were randomly assigned[23] to T1 and T2. Similarly, the order of event sequence within T1 and T2 was previously determined in a randomized

[23] For both true and both false events, each participant was equally likely to either report the event chosen first in T1 and the event chosen last in T2, or to report the event chosen last in T1 and the event chosen first in T2.

controlled fashion[24]. Considering that in forensic situations prospective interviewees are typically informed about the interview well in advance, a time gap of minimum three days (maximum 23 days) between any consecutive sessions was ensured to maximize ecological validity. The preparatory and the first interview session took place at one of the seminar rooms of the University (*Psychologische Hochschule Berlin*), the second interview session took place in the interview room of an institute (*Zentrum für Aussagepsychologie Berlin*) shared by forensic psychologists to carry out credibility assessments mandated by German courts.

Preparatory Session

In the preparatory session, the study investigator repeated the general purpose of the study as previously outlined for recruiting. After participants had signed an informed consent form an instruction sheet was handed out and read aloud by the study investigator, pointing out to participants the following three requirements to be considered when choosing suitable events for the upcoming reports: The events in question should not date back more than ten years in time (1), the nature of the events should be unique and bear personal significance rather than pertain to trivial everyday-life occurrences (2), and the unfolding of the events should have lasted several minutes and entail own actions as well as interactions with at least one other person (3).[25]

[24] Half of the participants were instructed to start with the true event in T1 and then to start with the false event in T2; the other half of participants were instructed to start with the false event in T1 and then to start with the true event in T2.

[25] These requirements were implemented to control for potential event-related factors that might affect the content quality of either true or fabricated accounts, such as the fading of memories with the increase of time-length between event and interview. Furthermore, the requirements should ensure that specific events characteristics (e.g. interactions with other persons) were present as otherwise the emergence of CBCA criteria related to such occurrences (e.g. reproduction of conversation) would have been precluded a priori.

Regarding true reports, it was stressed that the events needed to be told as happened, without containing any exaggerations or fictional elements. Concerning false reports, it was emphasized that the events' core elements must be the sole product of one's own imagination in order to thwart the adoption of strategies that would allow participants to remain close to truth-telling (i.e. by relying on original memories when fabricating; Leins et al, 2013). Next, the following instruction informed participants about their prospective tasks:

> For each event, you should try to provide a convincing and credible account. You are supposed to create the impression in listeners that you have actually experienced your truthfully-reported as well as your fabricated events. It is particularly important to report the events as detailed and extensive as possible. The report of each event should roughly be five minutes in length.

A preselection of 12 topics was then provided, among which participants had to select two different topics for their true events and two different topics for their false events. Endorsing Steller's (1989) recommendation to tailor CBCA study events towards cases of sexual abuse allegations, the topics available for selection were intended to simulate the emotional and experiential valence of real-life forensic situations related to such cases. Therefore, only situations were included in the final topic selection in which the statement provider would be likely to be directly involved, to be negatively emotionally aroused, and to feel a loss of control (see Steller et al., 1992), such as being the victim of a criminal act, being attacked by an animal or having a serious accident. Table 11 displays the entire range of topics available for selection.

Table 11. The selection of topics as presented to participants and their frequency of being reported.[26]

(1)	Being a victim of an attempted or committed criminal offense (e.g. burglary, theft, robbery, rape, molestation, blackmailing, etc.). [*n* = 47]
(2)	Suffering an accident with subsequent medical assistance/treatment (e.g. in traffic, at sports, at work, at home, etc.). [*n* = 55]
(3)	Being caught of and sanctioned for an illicit or embarrassing act (e.g. receiving a monetary fine or warning). [*n* = 28]
(4)	Experiencing existential feelings of fear or panic related to one's life or health. [*n* = 29]
(5)	Getting lost in nature or wildness (e.g. when hiking). [*n* = 40]
(6)	Losing a significant amount of money in the course of a risky endeavor (such as gambling). [*n* = 21]
(7)	Disclosure of an affair previously kept secret by yourself or by your (former) partner. [*n* = 20]
(8)	A sudden and unexpected event of death within your close family or peer network. [*n* = 20]
(9)	Failing in a personally important task or assignment. [*n* = 32]
(10)	Following someone else's advice with devastating consequences. [*n* = 4]
(11)	Being attacked by an animal or another person. [*n* = 37]
(12)	Experiencing a high-risk situation due to one's own or other person's negligence. [*n* = 23]

[26] The number of times each topic was reported by participants is provided in square brackets. The total number of reports amounts to *N* = 356, since four reports had to be excluded from the total sample of *N* = 360.

Only after suitable topics had been selected, the study investigator conveyed to participants that forensic psychologists professionally trained in credibility assessment were interested to test whether individuals with high story-telling abilities could successfully deceive them. Therefore, if in the upcoming interview session (T1) the participant's story-telling abilities were judged to be sufficiently high, he or she would be invited to have a second interview session (T2) at a forensic institute[27], with the prospect of higher monetary gains.[28]

Subsequently, participants filled out several questionnaires that inquired about their demographic and socio-economic background. At the end of the preparatory session, participants received a take-home sheet on which the date and location for the first interview session were noted. The sheet also repeated the previously outlined requirements regarding event characteristics and task demands and further indicated which two events participants were to tell first and second in the next session.

First Interview Session (T1)

On average seven days later (M = 6.87; SD = 3.29), the first interview session took place at the same location (*Psychologische Hochschule Berlin)* as the preparatory session. In the beginning, the study investigator repeated the task requirements and prompted participants to report both events in a detailed and comprehensive way. At the same time, the study investigator sought to minimize deliberate efforts to appear credible by highlighting that he was already aware of the

[27] Strictly speaking, the term forensic is not an accurate description for the institute. For practical purposes, the study investigator used this term when providing details of the institute to participants, with the intention to keep instructions brief. Technically, the institute reflects a collaboration of freelancing psychologists trained in credibility assessment.

[28] For a literal display of the instruction provided, see the description of the Methods section (subsection: Preparatory Session) for study I.

event's underlying truth status, rendering any explicit attempts of deception irrelevant.[29]

The average numbers of words uttered by the participants were *M* = 1105.80 (*SD* = 585.25) for the true and *M* = 1039.70 (*SD* = 448.41) for the false condition. After the interviews about the true and the false event were finished, participants filled out a questionnaire that among other things inquired about their prior preparation efforts and their subjective experiences during the interviews.

The study investigator then indicated that the participant would fulfill the prerequisites for the second interview session to take place at the forensic institute and provided further details about the prospect of winning additional monetary rewards: Apart from the fixed amount of 25€, 20€ could be further gained if the interviewer believed that the participant's true and false report were both based on actual experiences. In addition, 5€ would be paid if the participant correctly predicted the interviewer's respective ratings (i.e. true versus false). The study investigator also pointed out that at the end of the session a second interviewer might appear to elaborate further on some aspects participants had been previously outlined to the first interviewer, in case the first interviewer needed assistance in determining the truth status of their reports.

Coaching Measures

All participants were then told that reporting the events as detailed and extensive as possible would increase their chances of being believed, since the interviewer(s) were professionally

[29] For a literal display of the instruction provided, see the description of the Methods section (subsection: First Interview Session) for study I.

trained in assessing the credibility of statements by focusing on the statement's content and quality.[30]

For participants assigned to the control group, a take-home sheet – largely identical to the sheet provided at the end of the preparatory session – was handed out next and the session concluded.

For participants assigned to one of the two coaching groups (coaching group A or B), the study investigator handed out a (take-home) coaching sheet and read aloud the general instructions:

> Forensic psychologists focus on a variety of aspects when assessing the credibility of witness statements as mandated by the court. By far, however, the content and quality of the statement are most important to them. That is, they analyze how and in which ways the event in question is being described by referring to so-called content criteria. These are specific content-related criteria of which lay-persons typically assume that they would reduce or constrain the credibility of a statement. In fact, however, forensic psychologists attribute positive meaning to these criteria. That is, the presence of these criteria substantiates rather than disproves a statement's truthfulness. In the following, these content criteria are briefly described and illustrated by means of examples.

Next, dependent on group membership, the sheet provided descriptions and generic examples for either five script-deviant criteria (coaching group A*)* or five strategy-based criteria (coaching group B). Table 12A displays the coaching sheet as provided to participants from coaching group A:

[30] In their study, Vrij et al. (2002) labelled the control condition the "light coaching" (as opposed to "heavy coaching") condition, arguing that participants in this condition were told that they would increase their chances of being believed if they reported the event in a detailed fashion. From this perspective, the control condition in study III could also be regarded as a "light coaching" (rather than "no coaching") condition.

Table 12A. Coaching sheet with detailed descriptions of script-deviant criteria as provided to participants in coaching group A.

In the following, these content criteria are briefly described and illustrated by means of examples:

1. **Unexpected complications**

   Unexpected complications refer to descriptions of unsuccessful, interrupted or uncompleted actions that occur in the course of the event. The complications may be invoked by unexpected difficulties during the event sequence or by externally caused disturbances.

   Example: *And I was just about to secretly copy the information from the sheet – as suddenly a fierce wind gust swirled up all the papers.*

2. **Unusual details**

   This criterion is present if the statement provider reports details that at first glance seem unusual or eccentric and which are not necessarily to be expected from the event in question. At the same time, however, such details must not be unrealistic.

   Example: *And I then saw how the robber was storming to the waiting car – instead of driving off right away, the driver however stalled the engine first and had to restart it.*

3. **Superfluous details**

   Superfluous details are details that are reported by the statement provider, even though these details are unnecessary to render the actual event sequence comprehendible. If such details were lacking, the description of the event would still appear conclusive. Example: *Standing in front of a TV screen, the man suddenly pulled out his pistol and demanded that nobody should move. I still remember how on the screen the current news about the weather was being shown.*

[continues on next page]

4. **Related external associations**

   This criterion refers to elements of the statements which content-wise are similar to sequences of the main event. They are, however, not directly related to the main event and happened at a different time and with different persons.

   Example: *Upon confronting my partner, he denied everything and claimed that the situation was completely different. Once, my best friend had also caught her boyfriend cheating on her, and her boyfriend had reacted just in the very same way.*

5. **Accurately reported details not comprehended**

   This criterion is present if a section of the event is described correctly, even though the statement provider did not grasp its actual meaning (retrospectively, i.e. when providing the statement).

   Example: *And after a few minutes, the sound of the engine increasingly softened, and I could only accelerate fitfully. I then parked the scooter on the side of the road. I was quite angry – the rental agent had explicitly told me that the scooter was in perfect condition. Obviously, this was a lie.*

   [The statement provider is correctly describing how he or she rented a scooter that stopped working after some time. Other than inferred, however, the scooter stopped working because the statement provider had missed refueling gas.]

Table 12B displays the coaching sheet as provided to participants from coaching group B:

Table 12B. Coaching sheet with detailed descriptions of strategy-based criteria as provided to participants in coaching group B.

In the following, these content criteria are briefly described and illustrated by means of examples:

1. **Unstructured production**

   This criterion is present if parts of the event are reported in a chronologically unstructured fashion. In other words, the order in which the sequences of the event are being reported deviates from the actual chronological order of the event. For instance, actions that occurred at the beginning of the event are reported towards the end of the interview or vice versa. The statement overall needs to be still conclusive, however, meaning that its various components need to form a coherent event sequence if combined together.

   Example: *And then I hid in the room because I very well knew that I could not escape through the door. How did I know that? Oh right, I totally forgot to mention this earlier: When I entered the room at the very beginning, I had observed how the employee locked the door with a key.*

2. **Spontaneous corrections**

   This criterion is present if the statement provider specifies or corrects parts of his or her statement spontaneously, that is, without being prompted to do so.

   Example: *I saw the car – I mean, the autobus – passing by.*

3. **Admitting lack of memory**

   When reporting the event, the statement provider admits to gaps in his or her memory or knowledge, or else indicates some level of uncertainty.

   Example: *And then I saw the person running past – I believe she had dark hair. But this moment I have some difficulties remembering well; I am not really sure whether her hair was indeed dark.*

[continues on next page]

4. **Raising doubts about one's own testimony**

   This criterion is present if the statement provider questions the credibility of his or her statement (alternatively: of his or her own person). At the same time, he or she remains convinced that the event sequence took place as reported.

   Example: *What I was observing there appeared to be highly unlikely to me; when I was seeing it, I immediately thought to myself: No one will believe you this! It did, however, happen precisely as I said!*

5. **Reality controls**

   The statement provider describes that he or she temporarily doubted the "trueness" of his or her perceptions. For instance, she or he reports to have been unsure whether a sequence of the event was indeed taking place or whether the sequence of the event was rather the product of his or her imagination.

   Example: *As I saw this happening in front of my eyes, I was not sure if I was really experiencing it or whether I was not rather dreaming it.*

   In this context, the statement provider may describe active efforts to reappraise the "trueness" of his or her perceptions.

   Example: *I then closed my eyes for a brief moment, and when I opened them again after a few seconds, it was clear to me: This is no dream, this is really happening!*

For each criterion, the study investigator ensured that the participant correctly understood its description and respective example before proceeding. Afterwards, the study investigator told participants that they should try to integrate the criteria if they sought to maximize their chances of being believed by the interviewer. Finally, another take-home sheet (identical to the one distributed to the control group) was handed out and the session concluded.

Second Interview Session (T2)

On average eight days later (M = 7.57 SD = 2.61), the second and final interview session (T2) took place at the forensic institute. After repeating the task requirements, the study investigator again prompted participants to report both events in a detailed and comprehensive way. In contrast to the previous interview session (T1), his instruction this time was designed to motivate participants to appear credible and hence to successfully deceive in the false condition.[31]

The study investigator then left the interview room, and one of the confederates entered. The confederate briefly introduced him- or herself to the participant, thereby expanding on his role of being a forensic psychologist experienced with assessing the credibility of statements. Referring to the preassigned versions of a standardized interview protocol (see subsection "Interview structure" for more details), he or she conducted the interview in the same way as the study investigator had done during the first interview session. The average numbers of words spoken by the participants were M = 1137.63 (SD = 560.77) for the true and M = 1236.17 (SD = 535.88) for the false condition. After participants had finished with both reports, the confederate handed over a questionnaire largely identical to the one handed out in T1 and left the room. If assigned to coaching group A or B, participants received an additional questionnaire that inquired about their efforts to integrate the criteria they had been previously informed about. Some minutes later, the study investigator reentered the room and fully debriefed participants about the true identity of the confederate and the actual purpose of the

[31] For a literal display of the instruction provided, see the description of the Methods section (subsection: Second Interview Session) for study I.

study. All participants received the full amount of financial compensation (50€), regardless of performance and prediction accuracy.[32]

Interviewers

All interviews during T1 were conducted by the study investigator (and first author of this article; male, age 29), and all interviews during T2 were conducted by one of three confederates. The confederates (2 females, age 26 and 27; 1 male, age 29) were graduate students in psychology. The number of participants interviewed was nearly equally distributed among the confederates (range: 29–31).

Interview structure

In each interview, the interviewers started with an open-ended question to initiate free narratives and subsequently proceeded with increasingly more narrowly phrased follow-up questions. Sequence and phrasing of the follow-up questions were varied across four versions of a standardized interview protocol[33]. The specific versions were preassigned to each interview condition in a randomized controlled fashion. This manipulation was intended to keep the structure of the interviews as similar as possible across conditions while reducing potential practice effects that were likely to arise if the interviews' structure had been kept identical throughout the experiment.

[32] As done by Vrij & Mann (2006), participants were told that both their stories had been judged true by the interviewer. Without this pretense, participants would have realized that they would have been paid the full amount regardless of performance, and the 25€ bonus might no longer have been an effective incentive for subsequent participants.

[33] For a more detailed display of the four versions of the standardized interview protocol, see the description of the Methods section (subsection: Interview structure) for study I.

Prior to the start of the experiment, the study investigator trained the confederates to apply each version of the interview protocol consistently and systematically. The training was conducted in one session of three hours, in which each confederate performed a test interview, with the study investigator serving as the interviewee and feedback provider.

Questionnaires

After participants had finished their true and false reports, questionnaires were presented in T1 and T2 pertaining to the following domains[34]:

Motivation to perform well

For the first interview session participants' *motivation* for the true and false condition was assessed collectively by the following item: "How important is it for you to qualify for the next interview session at the forensic institute?" (the answer could be given on a seven-point Likert scale ranging from (1) not at all important to (7) highly important). For the second interview session participants' motivation was assessed separately for the true and false condition by the items "How motivated were you to convince the interviewer that your true event really happened?" (true condition) and "How motivated were you to convince the interviewer that your fabricated event really happened?" (false condition). The answers could again be given on a seven-point Likert scale ranging from (1) not at all motivated to (7) highly motivated. Furthermore, participants were asked whether aspects other than the prospect of higher monetary rewards contributed to their motivation, and if yes, to briefly elaborate.

[34] In addition to the domains described below, participants from the control group were also asked about the perceived cognitive load during the interviews in T1 and T2 (see study I for more details).

Efforts of preparation

To assess the extent to which participants prepared themselves before the interviews, participants had to indicate for each interview condition whether they had taken notes about relevant event details ("Did you take written notes about the content of your report prior to the interview?"), whether they had practiced their report in any way ("Did you practice the presentation of your report in any way prior to the interview, such as reporting to a friend?") and whether they had prepared in any ways other than previously outlined ("If not already addressed by any of the previous questions, did you prepare your report in any other way?"). For the false conditions only, an additional fourth question inquired about the use of external sources for inventing the event in question ("Did you use any sources of information, such as books or movies, when inventing the content of your report?").

Subsequently, participants' preparation efforts were classified into (0) low versus (1) high, by using the following rating scheme: A score of "0" was assigned unless at least two of the three or four (false condition) questions had been marked with "yes".

Predictions about the interviewer's ratings (T2)

In T2, participants had to predict (on a separate sheet provided prior to the commencement of the two interviews) whether the interviewer would rate their report's underlying truth status ("How do you think the forensic psychologist will rate your true/false story?") as either true or false.

Perceived authenticity of the second interview setting (T2)

To gauge how authentic participants perceived the second interview setting, the questionnaire handed out in T2 entailed the following additional items: "Did you believe that the interviewer in front of you was, in fact, a forensic psychologist trained in assessing

the credibility of statements?" and "Did you believe that after the interviews a second interviewer might appear to ask you further questions?". Possible answer options were "no", "yes, but I had some doubts", or "yes". Also, participants were further asked to indicate the extent to which the prospect of being questioned again by a second interviewer had affected them during their true and false reports (the answers could be given on seven-point Likert scales ranging from (1) not at all to (2) very strongly).

Participants' efforts to integrate previously coached criteria (T2)

If assigned to coaching group A or B, participants received an additional questionnaire that presented the following items, separately for the true and false condition: "Regarding the interview session today, did you try to integrate the content criteria which you had been previously informed about in your true/false report?". Possible answer options were "no", "yes, partly", or "yes, completely (all criteria)".

CBCA Analysis

CBCA compilation

The array of content characteristics used in study III differed in some ways from the traditional CBCA criteria as described by Steller & Köhnken (1989), since the revised system of content characteristics as proposed by Volbert & Steller (2014) was applied. In addition, further modifications were made to accommodate the purpose of study III. Instead of 19 criteria, the final array used for the study's rating procedure comprised 25 criteria. The criteria *logical consistency* and *quantity of details* were erased from the traditional version because their coding would have required a different coding method (scale rates; see the section below). The criteria *reality control* and *raising doubts about one's own person* were added, and the following criteria of the traditional version were modified: *Contextual embedding* was divided into three

separate criteria, namely *temporal information*, *spatial information*, and *information about everyday-life routines*. Similarly, the criterion *accounts of subjective mental state* was divided into the criteria *sensory impressions, emotions and feelings*, and *own thoughts.*[35] Furthermore, criteria that referred exclusively to criminal contexts were adjusted to extend their applicability to non-criminal contexts (e.g. the criterion *details characteristic of the offense* was transformed into *details characteristic of the event).*

From this set of 25 criteria, the criterion *raising doubts about one's own person* was coded, but subsequently excluded from further analysis because of its low overall occurrence ($n < 1.5$).

Rater Training

The interviews from the interview session T1 and T2 were transcribed verbatim (total number of transcripts: 356)[36], and copies were provided to two CBCA-trained raters. The first rater was a forensic psychologist with experience both in research as well as in the forensic application of CBCA, the second rater was a graduate student in psychology. Prior to the rating of the transcripts, the second rater received extensive training in CBCA scoring by the first (expert) rater to ensure a consistent and standardized approach. The training included the reading and discussing of literature that entailed detailed descriptions and examples of the content criteria. Both raters also rated the presence of criteria in several transcripts not contained in the current study on a 4-point Likert Scale (ranging from (0) not present to (4) strongly present) until they had reached satisfactory interrater reliability ($ICC \geq .4$; see Cicchetti, 1994).

[35] For a concise description of the newly added or modified criteria see Niehaus (2001), pp. 121-130.

[36] In total, four reports from the total sample of $N = 360$ were excluded for the following reasons: Lack of event-related information to render CBCA analysis feasible (T2, true condition); the truth status would have been easily discernible for the CBCA raters (T1; false condition); a previously experienced event was reported as fabricated (T2, false condition) and vice versa (T2, true condition).

Coding of the interview transcripts

Each transcript was rated independently by both raters who were blind to the specific experimental conditions (i.e. they had no knowledge about truth status, interview session and coaching condition). For each CBCA criterion, the raters first counted how often the criterion occurred in the transcript (frequency counts). Following the common procedure, repeated information was not counted twice (i.e. Vrij & Mann, 2006). Second, the raters estimated how *strongly* each frequency count related to the criterion by assigning weights to each identified occurrence (1 = weakly present, 2 = strongly present). Third, the weighted frequency counts were summed up to form the final score of the individual criterion. Most CBCA studies used 3-point (e.g. Blandon-Gitlin et al., 2009) or 5-point (e.g. Akehurst et al., 2011) Likert scales to rate the presence of each criterion in the entire transcript as coding method (scale rates). More importantly though, weighted frequency counts combine the methodological advantages of frequency counts over scale rates in evaluating the presence of content characteristics (see Nahari, 2016) with the benefits of scale rates in taking the criterion's strength of appearance into account. The raters, therefore, used weighted frequency counts and carried out this procedure for all transcripts in multiple intervals of 20 to 30 transcripts each. Between such intervals, the raters gathered repeatedly to reappraise the consistency of their individual ratings to enhance reliability. More specifically, they made sure that their frequency counts referred to the same text passage (an essential aspect which is often ignored in the literature; see Sporer, 2012).

Reliability analysis

Pearson product-moment correlations for the final scores of each criterion by each rater were conducted to obtain consistency estimates as a realistic measure of interrater reliability (see Akehurst et al., 2001). For all 24 criteria the ratings were significantly and positively correlated,

and all criteria revealed at least fair too good interrater reliability scores of $r > .40$ (see Goedert et al., 2005), ranging from $r = .45$ (*spatial information*) to $r = 1.00$ (*pardoning the perpetrator*). All 24 criteria were thus retained for further analysis. Table 13 provides a complete display of the obtained Pearson correlation values.

Table 13. Pearson's r values for individual criteria (and for the CBCA sum score).

#	Criterion	Pearson's *r*
1	Information about everyday-life routines	.47*
2	Spatial information	.45*
3	Temporal Information	.46*
4	Descriptions of Interactions	.83*
5	Reproduction of conversations	.81*
6	Emotions and feelings	.63*
7	Own thoughts	.74*
8	Sensory impressions	.46*
9	Attribution of other person's mental state	.74*
10	Personal implications	.62*
11	Unexpected complications	.81*
12	Superfluous details	.78*
13	Unusual details	.69*
14	Related external associations	.75*
15	Details characteristic of the event	.79*
16	Accurately reported details not comprehended	.89*
17	Unstructured production	.62*
18	Spontaneous corrections	.75*
19	Admitting lack of memory	.87*
20	Efforts to remember	.70*
21	Reality controls	.94*
22	Raising doubts about one's own testimony	.84*
23	[Raising doubts about one's own person]	-
24	Self-deprecation	.83*
25	Pardoning the perpetrator	1.00*
	CBCA sum score	.82*

* $p < .01$.
[ ] Excluded from final analysis ($n < 1.5$).

Based on this array of 24 criteria, the two raters' final (weighted frequency) criterion scores were averaged for each criterion to form the mean criterion score. These mean criterion scores were then summed up to form the mean CBCA sum score (i.e. the sum of the criteria ratings for each transcript averaged across raters). Also, the final criterion scores for each rater were summed up separately to obtain rater-specific CBCA sum scores. These rater-specific

CBCA sum scores correlated significantly and revealed a product-moment coefficient (r = .82) that can be taken to indicate excellent interrater reliability (see Goedert et al., 2005).

Results

Questionnaires and manipulation checks

Motivation to perform well

Since participants were instructed that qualification for T2 was dependent on their "storytelling" performance as displayed in T1, participants' indications of how important they deemed qualifying for the second interview session were interpreted as an indirect measure of their motivation to tell both events in T1 in a convincing way. More than half of the participants (55.6%, n = 50) indicated that they considered it to some extent important (score > 4 on the seven-point Likert scale) to qualify for the second interview session at the forensic institute (M = 4.48, SD = 1.68). Viewed for each participant group separately (i.e. control group, coaching group A, coaching group B), the scores did not differ between any of the three groups (p > .05).

Regarding T2, the majority of participants reported to have been to some extent motivated (scores > 4 on the seven-point Likert scales) to convince the interviewer that the event had actually happened in both the true (68.9%, n = 62) and the false (88.9%, n = 80) condition. This motivation to convince the interviewer in T2 was overall significantly higher in the false (M = 5.90, SD = 1.03) than in the true (M = 5.35, SD = 1.37) event condition, M_{Diff} = 0.55, 95% CI [0.26, 0.84], $t(88) = 3.804$, $p < .001$, d = 0.46. Within the false condition, the motivational scores did not differ between any of the three participant groups (p > .05); within the true condition, the motivational scores differed significantly between the control group (M = 5.87, SD = 1.25) and coaching group A (M = 4.83, SD = 1.36), M_{Diff} = 1.04, 95% CI [0.36, 1.72], $t(57) = 3.049$, p = .003, d = 0.79.

Efforts of preparation

For each participant group, Table 14 displays the number of participants whose preparation efforts prior to the interview were classified as low versus high (= at least two of the relevant three or four questions had been positively answered) in the respective interview condition as well as their associated mean CBCA sum scores and standard deviations. For T2, chi-square tests of independence showed that there was no significant relationship between coaching and preparation efforts, regardless of whether participants were reporting true events, X^2 (2, N = 88) = 1.14, p = .566, or false events, X^2 (2, N = 89) = 2.05, p = .36. Because of the small and unequally distributed sample sizes, testing for statistical differences in CBCA sum scores within or between groups was deemed unwarranted.

Table 14. Number of participants with low versus high preparation efforts and the associated mean CBCA sum scores and std. deviations for each participant group and experimental condition.

Participant group	*Interview session*	*Truth status*	*low preparation efforts*			*high preparation efforts*		
			n	*M*	*SD*	*n*	*M*	*SD*
Control Group	T1	true	22	18.64	13.08	8	21.19	12.98
	T1	false	21	14.24	5.24	8	19.50	9.07
	T2	true	23	14.56	6.28	7	18.21	7.23
	T2	false	20	18.10	7.06	10	23.85	13.36
Coaching group A (script-deviant)	T1	true	22	17.27	9.14	8	20.50	8.15
	T1	false	17	12.41	7.06	13	15.38	3.54
	T2	true	22	16.91	7.50	7	21.93	7.90
	T2	false	14	17.46	6.35	15	19.53	7.17
Coaching group B (strategy-based)	T1	true	23	16.17	6.94	7	16.14	4.76
	T1	false	15	12.97	4.60	15	18.00	10.20
	T2	true	19	17.74	6.23	10	19.75	11.12
	T2	false	17	14.44	6.57	13	22.23	10.27

Predictions about the interviewer's ratings (T2)

All but one participant (from coaching group A) provided predictions of how the interviewer would rate their true report (n = 89) and all participants provided predictions of how the interviewer would rate their false report (n = 90). Table 15 displays the proportions of these predictions, separately for each participant group.

Table 15. Proportions of participants who predicted that their true/false report would be rated as either true or false, separately for each participant group (T2 only).

Participant group	*Truth status*	*"rated as true" prediction*	*"rated as false" prediction*
Control group	true	83,3%	16,7%
	false	43,3%	56,7%
Coaching group A (script-deviant)	true	69,0%	31,0%
	false	56,7%	43,3%
Coaching group B (strategy-based)	true	83,3%	16,7%
	false	53,3%	46,7%

Chi-square tests of independence showed that there was no significant relationship between coaching and the nature of participants' predictions, regardless of whether their predictions pertained to true events, X^2 (2, N = 89) = 2.40, p = .301, or false events, X^2 (2, N = 90) = 1.16, p = .561.

Perceived authenticity of the second interview setting (T2)

The experimental manipulations in T2 (e.g. interviewer being introduced as forensic psychologists trained in credibility assessment, emphasizing the prospect of additional success-related financial rewards) were designed to bring the features of the underlying interview setting closer to forensic situations. The results suggest that these manipulations had been successful: Most participants (92.2%, n = 83) indicated that they had conceived their interviewer in T2 as a forensic psychologist professionally trained in credibility assessment (among those, 32.5% specified to have had experienced some doubts about the interviewer's true identity, but overall

still had believed that the interviewer was in fact who he or she pretended to be). A chi-square test of independence revealed that there was no significant relationship between coaching and participants' beliefs about the interviewer's professional background, X^2 (4, N = 89) = 4.14, p = .387. Similarly, most participants (76.7%, n = 69) expressed that they had believed that a second interviewer could appear at the end of T2 to ask further questions (among those, 23.2% reported to have had experienced some doubts). Again, a chi-square test of independence indicated that there was no significant relationship between coaching and participants' beliefs about the appearance of a second interviewer, X^2 (6, N = 90) = 5.31, p = .505.

Concerning financial compensation, the majority of participants (77.8%, n = 70) stated that apart from the prospect of earning additional success-related monetary rewards, financially unrelated aspects had also propelled them to perform at their best during the interviews, such as finding out about one's own lie abilities under challenging conditions (n = 35) and/or exhibiting a more general interest towards the experiment's area of research (n = 12).

Participants' efforts to integrate previously coached criteria (T2)

Regarding *true* statements, less than half of the participants informed about script-deviant criteria (43.3%, n = 13) indicated that they had tried to integrate the respective criteria. Among the participants informed about strategy-based criteria a similar proportion (40.0%, n = 12) provided such response. The proportion of participants who reported that they had attempted to integrate the previously coached criteria did not differ between the two coaching conditions, X^2 (1, N = 58) = 0.07, p = .791.

For *false* statements, most participants informed about script-deviant criteria (83.3%,

n = 25) reported that they had attempted to integrate the respective criteria.[37] Similarly, the majority of participants informed about strategy-based criteria (80.0%, n = 24) responded in this way. Again, the proportion of participants who reported that they had attempted to integrate the previously coached criteria did not differ between the two coaching conditions, X^2 (2, N = 59) = 1.38, p = .501.

Main Analysis

For analysis, R (R Core Team, 2019) and *lme4* (Bates et al., 2015) were used to perform a linear mixed-effects analysis of the relationship between the response variable CBCA sum score and the independent predictor variables interview session, truth status and coaching. As fixed effects, interview session, truth status and coaching (with interaction terms) were entered into the model as well as motivation and preparation efforts (without interaction terms) to account for their potentially confounding effects on the dependent variable (e.g. Burgoon, 2015; Vrij & Mann, 2006). The variable motivation was scaled with -1 to render its zero-point equivalent to the lowest value on the 7-point Likert Scale as provided on the questionnaire (1 = "not at all motivated"). As random effects, an intercept for subjects and a by-subject random slope for preparation efforts as well as their covariance were included to represent individual differences in the overall level of achieved CBCA sum scores and in the effect of preparation efforts. Based on theoretical considerations and tested by comparing reduced models via maximum likelihood-ratio tests, the applied model thus includes the maximal random effect structure justified by the data (i.e. Barr et al., 2013). Visual inspection of residual plots did not reveal any obvious deviations from homoscedasticity or normality. Tests of the parameter estimates were

[37] Among those responses, only one participant provided the "yes, completely" option. In all other cases of positive responses (i.e. for true and false statements), participants from either coaching group marked the "yes, partly" option.

obtained via restricted maximum likelihood estimation in conjunction with Satterthwaite's degrees of freedom method using *lmerTest* (Kuznetsova et al., 2017) and are presented in Table 16.

Table 16. Effect coded parameter estimates for mean CBCA sum score.

	Mean CBCA sum score		
Predictors	*Estimates*	*SE*	*95% CI*
Fixed Effects			
(Intercept)	15.59	1.73	12.21 – 18.97
Interview session: (T2; target condition)	-4.96	1.58	-8.07 – -3.03
Truth status: False	-3.94	1.55	-6.98 – 0.90
Coaching: Script-deviant (A)	-0.94	2.02	-4.91 – 3.03
Coaching: Strategy-based (B)	-2.90	2.03	-6.88 – 1.08
Motivation	0.83	0.28	0.27 – 1.39
Preparation efforts: High	2.82	1.12	0.63 – 5.01
Interview session x Truth status	7.95	2.18	3.68 – 12.22
Interview session x Coaching: Script-deviant (A)	4.45	2.20	0.13 – 8.76
Interview session x Coaching: Strategy-based (B)	6.53	2.22	2.18 – 10.88
Truth status x Coaching: Script-deviant (A)	-1.07	2.21	-5.39 – 3.25
Truth status x Coaching: Strategy-based (B)	2.08	2.24	-2.31 – 6.47
Interview session x Truth status x Coaching: Script-deviant (A)	-3.95	3.11	-10.04 – 2.13
Interview session x Truth status x Coaching: Strategy-based (B)	-7.99	3.11	-14.09 – -1.88
Random Effects			
Residual variance (σ^2)	34.26		
Subject intercept variance (τ_{00})	21.51		
Subject Preparation variance (τ_{11})	19.75		
Intercept, Preparation covariance (ρ_{01})	0.33		
Observations[38]	350		

[38] One participant from coaching group A and two participants from coaching group B did not provide their motivational scores at T1 (pertaining to both the true and false condition). These observations were therefore excluded from the multilevel model analysis.

Subsequently, for each participant group estimated marginal means based on the output of the linear mixed-effects model were computed for each combination of interview session and truth status, using *emmeans* (Lenth, 2018). The results are presented in Table 17.

Table 17. Estimated marginal means, standard errors and 95% confidence intervals of mean CBCA sum scores for each participant group and experimental condition.

Participant group	*Interview session*	*Truth status*	*Margin*	*SE*	*95% CI*
Control Group	T1	true	20.37	1.49	17.42 - 23.31
	T1	false	16.43	1.51	13.46 – 19.41
	T2	true	15.40	1.50	12.45 – 18.36
	T2	false	19.42	1.52	16.43 - 22.41
Coaching group A (script-deviant)	T1	true	19.43	1.52	16.44 - 22.41
	T1	false	14.42	1.52	11.41 – 17.42
	T2	true	18.91	1.50	15.95 – 21.87
	T2	false	17.90	1.54	14.86 – 20.94
Coaching group B (strategy-based)	T1	true	17.47	1.52	14.47 – 20.47
	T1	false	15.61	1.54	12.56 – 18.65
	T2	true	19.03	1.51	16.06 - 22.00
	T2	false	17.14	1.51	14.16 - 20.12

Note: Degree of freedoms calculated via Satterthwaite's method.

For each of the three participant groups, the estimated marginal means show the mean CBCA sum score for each level combination of interview session (T1 vs T2) and truth status, adjusted for the control variables motivation and preparation efforts. Figure 4 depicts the estimated marginal means and their associated error bars visually, illustrating for each

participant group how the CBCA scores of participants' true and false statements changed from T1 to T2.

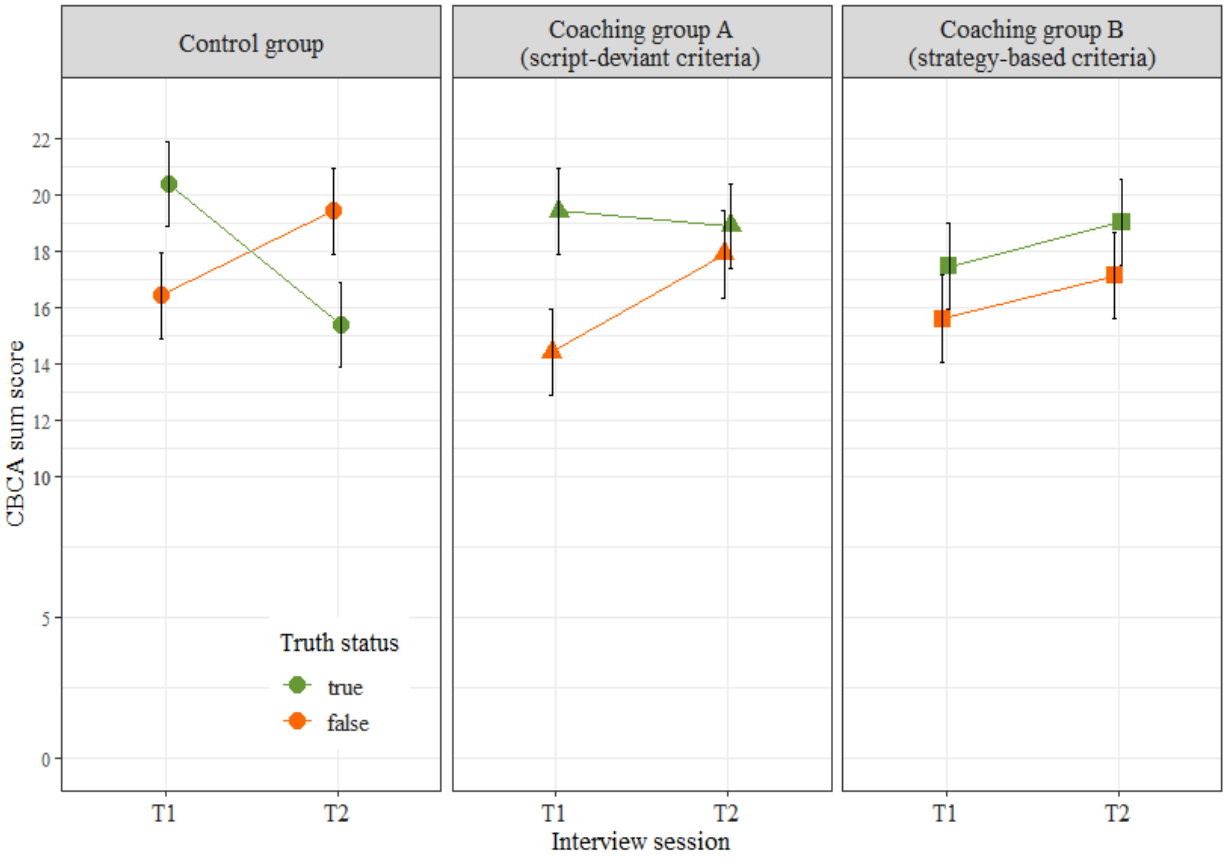

Figure 4: Mean CBCA sum scores for each participant group and experimental condition, adjusted for participants' motivation and preparation efforts, with error bars (±1 SE).

Hypotheses testing

As the first step to follow up on the changes of CBCA scores from T1 to T2, pairwise comparisons were run to test for significant differences between true and between false statements *within* each participant group. Among these comparisons displayed in Table 18, only for the true statements of uninformed participants (control group) a statistically significant difference emerged between $T2_{true}$ and $T1_{true}$ (M_{Diff} = -4.96, 95% CI [-8.07, -1.86], t = -3.13, p = .010, d = -0.61).

Table 18. Pairwise comparisons and their mean differences, standard errors, 95% confidence intervals, effect sizes *d*, t-values and associated p-values for each participant group.

Participant group	*Interview session/ Truth status*	M_{Diff}	*SE*	*95% CI*	*d*	*t*	*p*
Control group	T2/true vs T1/true	-4.96	1.58	[-8.07 – -1.86]	-0.61	-3.13	**.010**
	T2/false vs T1/false	2.99	1.62	[-0.19 – 6.17]	0.36	1.84	.255
Coaching group A (script-deviant)	T2/true vs T1/true	-0.52	1.57	[-3.60 – 2.56]	-0.06	-0.33	.988
	T2/false vs T1/false	3.48	1.64	[0.27 – 6.70]	0.42	2.12	.149
Coaching group B (strategy-based)	T2/true vs T1/true	1.56	1.60	[-1.56 – 4.69]	0.19	0.98	.760
	T2/false vs T1/false	1.53	1.62	[-1.64 – 4.70]	0.19	0.95	.779

Note: Degree of freedoms were calculated via Satterthwaite's method. P-values were adjusted for multiple comparisons via the Tukey method.

The difference in scores between the first and second interview session for true and for false statements ($T2_{true}$ vs $T1_{true}$ and $T2_{false}$ vs $T1_{false}$) *within* each participant group form the foundation for testing the hypotheses, that is, to examine to what extent those difference scores (M_{Diff}) differ *between* the three participant groups (M_{Diff}^{*}). Remember that only these particular comparisons *between* the participant groups allow for measuring the actual strength of coaching effects, as otherwise the systematic differences between the two interview settings (T1 versus T2) could not be ruled out as a potentially confounding variable (see Figure 1, p. 10, for illustration).

Differences between groups for true statements (Hypotheses 1 and 2)

The change in scores from $T1_{true}$ to $T2_{true}$ differed significantly between coaching group A and the control group (M_{Diff}^{*} = 4.45, *SE* = 2.20, 95% CI [-0.13, 8.76], *p* = .043, *d* = 0.73), with the decline in scores being less pronounced for participants informed about script-deviant criteria (M_{Diff} = -0.52) than for uninformed participants (M_{Diff} = -4.96). Hence, Hypothesis 1 is supported.

The change in scores from $T1_{true}$ to $T2_{true}$ also differed significantly between coaching group B and the control group (M_{Diff}^{*} = 6.53, *SE* = 2.22, 95% CI [2.18, 10.88], *p* = .003, *d* = 1.07). The decline in scores was only present for uninformed participants (M_{Diff} = -4.96) and in this way less pronounced for participants informed about strategy-based criteria (M_{Diff} = 1.56). Thus, Hypothesis 2 is supported as well.

Differences between groups for false statements (Hypothesis 3 and 4)

It was predicted that there was no difference in the change of CBCA scores from $T1_{false}$ to $T2_{false}$ between participants informed about script-deviant criteria (coaching group A) and uninformed participants (control group). Seemingly in line with this prediction, the increases in scores did indeed not differ (M_{Diff}^{*} = 0.49, *SE* = 2.21, 95% CI [-3.85, 4.83], *p* = .824, *d* = 0.08) between participants informed about script-deviant criteria (M_{Diff} = 3.48) and uninformed participants (M_{Diff} = 2.99). Yet, without testing for *equivalence* (Wellek, 2010) a nonsignificant test result should not be interpreted as evidence for the absence of a true effect or difference (Lakens, 2017). Using the R package *TOSTER* (Lakens, 2017), the "two one-sided tests" (TOST) procedure (Schuirmann, 1987) were carried out, a simple equivalence testing approach that can be used to statistically reject the presence of effects large enough to be considered meaningful (for more details, see Lakens, 2017). First, a lower (-0.73) and upper (0.73) equivalence bound were prespecified based on the smallest effect size (*d* = 0.73) that was associated with a

statistically significant ($p < .05$) and hence meaningful difference when comparing other conditions (see Hypothesis 1). The subsequently performed equivalence test indicated that the presence of effects more extreme than the equivalence bounds could be rejected ($t(56) = 2.480$, $p = .008$), implying that the observed difference in CBCA difference scores between the two groups is close enough to zero to be equivalent (Seaman & Serlin, 1998). In other words, it can be concluded that the observed difference ($M_{Diff}^{*} = 0.49$) is statistically not different from zero and statistically equivalent to zero, clearly supporting Hypothesis 3.

In contrast, the change in CBCA scores from $T1_{false}$ to $T2_{false}$ was expected to significantly differ between participants informed about strategy-based criteria (coaching group B; $M_{Diff} = 1.53$) and uninformed participants (control group; $M_{Diff} = 2.99$). However, again there was no statistically significant difference in the increases of scores between the two groups ($M_{Diff}^{*} = -1.46$, $SE = 2.22$, 95% CI [-5.81, 2.98], $p = .657$, $d = -0.23$), with the direction of the effect being even opposite to the prediction. Consequently, Hypothesis 4 is rejected.

Subsequent considerations about false statements

For false statements then, providing participants with information about either script-deviant or strategy-based criteria did not result in an increase of their CBCA sum scores that differed from the increase observed for uninformed participants. Considering that the sum score as the dependent variable comprised mostly criteria about which participants of either coaching group remained uninformed (19 out of 24 criteria), however, the applied measure may have been limited in its sensitivity to capture potential coaching effects. For this reason, the identical analysis described above was performed separately for participants from coaching group A and B, with *coaching-restricted scale scores* rather than sum scores serving as the dependent variable in each case.

For illustration, for each coaching group all criteria were deleted from the sum score except the ones for which information was being provided, resulting in a *script-deviant scale score* applied to participants from coaching group A and a *strategy-based scale score* applied to participants from coaching group B. For both coaching-restricted scale scores the raters' respective CBCA scores correlated significantly and yielded product-moment coefficients (script-deviant scale score: $r = .83$; strategy-based scale score: $r = .83$) that represent excellent interrater reliability

Coaching-restricted scale scores as the dependent variable

To follow up on the changes of coaching-restricted scale scores from T1 to T2, again pairwise comparisons were run first to test for significant differences between false statements *within* each participant group. Figure 5 illustrates how for each participant group the mean script-deviant and strategy-based scale scores of false statements shifted from T1 to T2. Remember that after T1 but before T2 participants from coaching group A were informed about five script-deviant criteria, while participants from coaching group B were informed about five strategy-based criteria (see Figure 1, p. 10, for illustration).

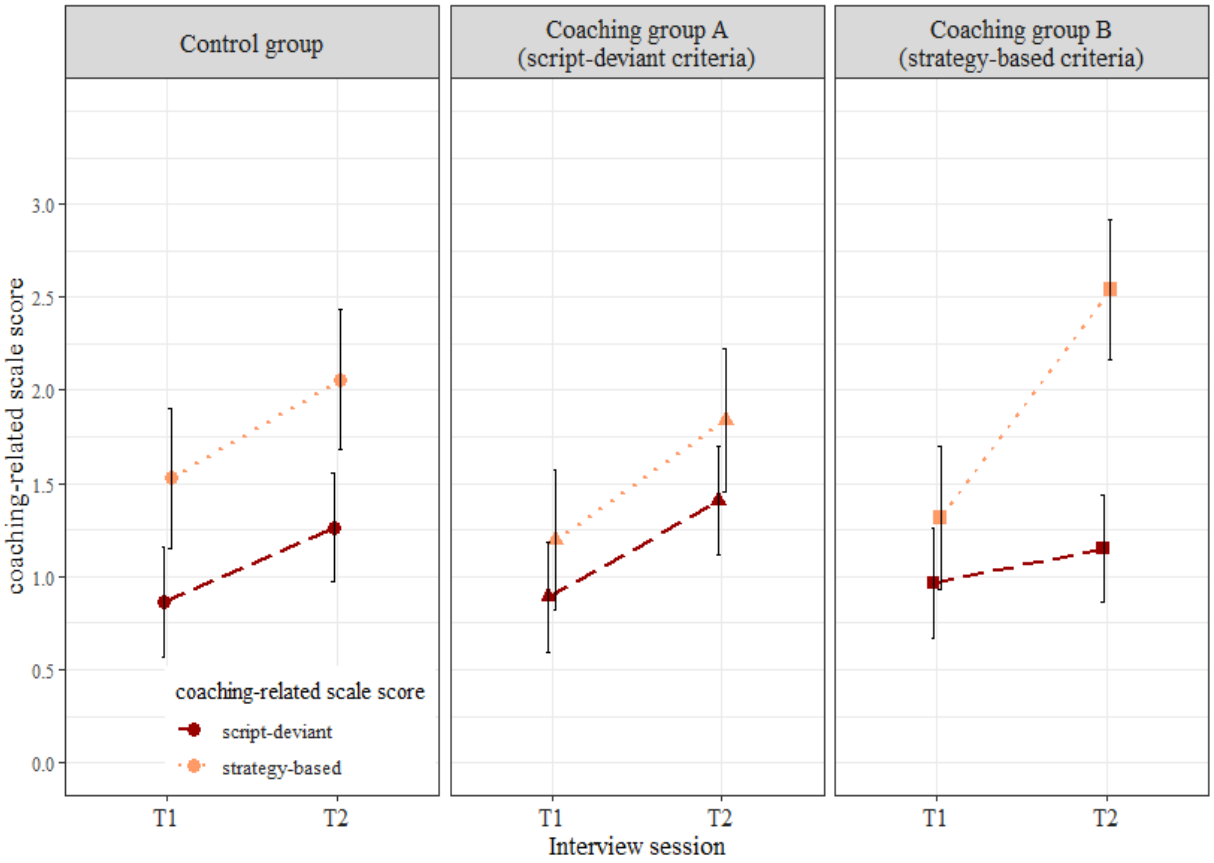

Figure 5. Mean *script-deviant* and *strategy-based scale scores* for each participant group in the false conditions, adjusted for participants' motivation and preparation efforts, with error bars (±1 SE).

Significance testing for the script-deviant scale score

As displayed in Table 19, regarding script-deviant scale scores none of the pairwise comparisons (M_{Diff}) yielded a statistically significant difference. Pertaining to hypothesis 3, there was also no difference in the increase of script-deviant scale scores (M_{Diff}^{*} = 0.12, *SE* = 0.52, 95% CI [-0.90, 1.14], p = .822, d = 0.08) *between* participants informed about script-deviant criteria (coaching group A; M_{Diff} = 0.52) and uninformed participants (M_{Diff} = 0.40).

Table 19. Pairwise comparisons and their mean differences, standard errors, 95% confidence intervals, effect sizes *d*, t-values and associated p-values for each participant group, based on the *script-deviant scale scores* (rather than CBCA sum scores) of false statements.

Participant group	*Interview session/ Truth status*	M_{Diff}	*SE*	*95% CI*	*d*	*t*	*p*
Control group	T2/false vs T1/false	0.40	0.38	[-0.35 – 1.15]	0.25	1.05	.719
Coaching group A	*T2/false vs T1/false*	*0.52*	*0.38*	*[-0.23 – 1.27]*	*0.33*	*1.36*	*.525*
Coaching group B	T2/false vs T1/false	0.19	0.38	[-0.55 – 0.93]	0.12	0.49	.960

Significance testing for the strategy-based scale score

As shown in Table 20, the pairwise comparisons (M_{Diff}) yielded a statistically significant difference between $T2_{false}$ and $T1_{false}$ exclusively for participants informed about the respective (strategy-based) criteria (M_{Diff} = 1.22, 95% CI [0.30, 2.15], t = 2.59, p = .049, d = 0.60). Yet, if compared to the control group as specified by Hypothesis 4, this increase in strategy-based scale scores for coaching group B was again not sufficient in strength to statistically differ (M_{Diff}^{*} = 0.70, SE = 0.65 , 95% CI [-0.58, 1.97], p = .284, d = 0.39) from the increase observed for uninformed participants (M_{Diff} = 0.53).

Table 20. Pairwise comparisons and their mean differences, standard errors, 95% confidence intervals, effect sizes *d*, t-values and associated p-values for each participant group, based on the *strategy-based scale scores* (rather than CBCA sum scores) of false statements.

Participant group	*Interview session/ Truth status*	M_{Diff}	*SE*	*95% CI*	*d*	*t*	*p*
Control group	T2/false vs T1/false	0.53	0.47	[-0.40 – 1.45]	0.26	1.12	.680
Coaching group A	T2/false vs T1/false	0.64	0.48	[-0.29 – 1.58]	0.31	1.35	.533
Coaching group B	*T2/false vs T1/false*	*1.22*	*0.47*	*[0.30 – 2.15]*	*0.60*	*2.59*	***.049***

Discussion

Study III examined to what extent providing information about sets of criteria affects the CBCA scores of participants' true and false statements. In the baseline conditions (T1), participants had to tell a true and false event while any deliberate attempts to appear credible were rendered extraneous. Subsequently, some of the participants were informed about either script-deviant (coaching group A) or strategy-based criteria (coaching group B), and then all participants had to tell a true and false event again in the target conditions (T2). The setting of T2 entailed various experimental manipulations to make participants concerned about being believed, allowing to test the effects of coaching under forensically relevant conditions. While previous studies showed that laypersons tend to avoid the production of both script-deviant and strategy-based criteria when deceiving, theoretical considerations imply that upon willingness the latter criteria should be easier to simulate than the former. Against this background, the study's main

question of interest is to what extent the effects of coaching differ between the two criteria sets. More specifically, the change in CBCA scores from baseline (T1) to target (T2) condition served as the relevant outcome variable, which then was compared between uninformed participants (control group) and informed participants from either coaching group A or B as specified by the study's hypotheses. In this way, potentially confounding effects that may have emerged from the differences in the two interview settings were held constant across all participant groups (see Figure 1, p. 10, for illustration).

Findings regarding true statements

For true statements, the CBCA sum scores of uniformed participants (control group) significantly declined from baseline (T1) to target (T2) condition. Based on a similar observation derived from study I[39], the author of this Thesis speculated that in the absence of any conceivable strategy to make their false statement appear credible, participants may have deliberately compromised the quality of their true report. For the current study, it was contended that such self-handicapping inclinations should be less likely in participants who had previously acquired detailed knowledge about CBCA, as alternative and seemingly more promising strategies were made available to them. Indeed, the sum scores of participants informed about either script-deviant (coaching group A) or strategy-based (coaching group B) criteria did not significantly decline from T1 to T2. More importantly, if compared between the participant groups, the magnitude of these changes differed as predicted between coaching group A and

[39] While the sample of participants in study I was identical to the (sub)sample used as control group in study III, the formation of the CBCA sum score as dependent variable differed slightly between the two studies (i.e. being comprised of 21 versus 24 criteria). Consequently, testing for significance between conditions yielded partly different results in regards to the (sub)sample. For instance, in study I CBCA sum scores decreased only tentatively from $T1_{true}$ to $T2_{true}$, while in the current study this decline was statistically significant.

the control group (Hypothesis 1; $d = 0.73$) as well as between coaching group B and the control group (Hypothesis 2; $d = 1.07$), yielding effect sizes of medium to large strength (according to the classification of Cohen, 1988).

Taken together, these findings lend support to the notion that statement providers may compromise the quality of their true report if – and only if – no alternative strategies are conceivable to render their false report more credible in comparison. Further experimental investigations are needed, however, that scrutinize these findings within slightly modified settings - such as assigning a different interviewer person to each experimental condition - before more definite conclusions can be drawn (see study I for more details).

Findings regarding false statements

CBCA sum score as the dependent variable

A different pattern of changes in CBCA sum scores emerged for false statements: Irrespective of group membership, the sum scores tentatively increased from the baseline (T1) to the target (T2) condition. In line with the study's prediction (Hypothesis 3) and further substantiated after testing for equivalence, the change in sum scores from T1 to T2 did thus not differ between participants informed about script-deviant criteria and uninformed participants ($d = 0.08$). Obviously then, informing participants about script-deviant criteria had no noticeable effect on their CBCA sum scores when reporting false events.

In turn, it was hypothesized that providing information about strategy-based criteria would result in statements with higher sum scores relative to the statements of uninformed participants. Yet and in contrast to this prediction (Hypothesis 4), the change in CBCA sum scores from T1 to T2 did not differ between participants informed about strategy-based criteria and uninformed participants either ($d = -0.23$). In fact, as indicated by the negative effect size *d*, the direction of the effect was opposite to the prediction, meaning that the increase in CBCA

sum scores was tentatively more pronounced for uninformed participants than for participants informed about strategy-based criteria.

Since participants from both coaching groups predominately indicated to have attempted to integrate the set of criteria they had previously been informed about, it appears that within the study's experimental setting both script-deviant and strategy-based criteria were difficult to simulate for participants, despite their deliberate efforts to do so. While this finding lends support to the assumption about the cognitive difficulty associated with the production of script-deviant criteria when fabricating, it contrasts with the notion that simulating strategy-based criteria should be comparatively easy to accomplish.

Coaching-restricted scale scores as the dependent variable

It needs to be noted, however, that the study's findings about the (absent) effects of coaching on the quality of false statements were obtained when using the global CBCA sum score as the underlying measure. Being comprised of 24 individual criteria, the sum score may not have fully captured potential coaching effects, since the applied coaching measures pertained to only small subsets of the criteria pool. Therefore, the analysis was performed again for false statements from coaching group A and B, with the underlying CBCA sum scores being reduced to the five criteria participants in each group were informed about.[40]

For script-deviant criteria scale scores, once more no differences were observed in the change of scores between participants informed about the criteria (coaching group A) and uninformed participants, with the resulting effect size of $d = 0.08$ being identical to the one obtained from the main analysis. It can therefore be concluded that encouraging participants to

[40] For illustration, the formation of coaching-restricted scale scores resulted in a script-deviant scale score (i.e. comprised of five script-deviant criteria) for coaching group A and a strategy-based scale score (i.e. comprised of five strategy-based criteria) for coaching group B.

integrate script-deviant criteria had no effect on the CBCA scores of their false statements, irrespective of whether these scores were measured on a global (sum score) or coaching-restricted (scale score) level.

For strategy-based scale scores, however, within-group comparisons showed the scores to significantly increase from baseline (T1) to target (T2) condition only if participants had been previously informed about the criteria (coaching group B). This finding differs from the results derived from the main analysis - i.e. *within* any participant group, no significant differences in the CBCA sum scores of false statements were found between T1 and T2 - and seems to suggest that for strategy-based criteria coaching effects might indeed be present but be detectable only if measured on a less global and more coaching-sensitive (scale score) level. Caution is warranted, though, when interpreting the significant increase of scores found *within* coaching group B: Because of the study's design, the group's change in scores from baseline (T1) to target (T2) condition becomes a meaningful measure for potential coaching effects only when contrasted with the change in scores observed for uninformed participants (control group), who received no coaching but otherwise provided their statements in identical interview conditions. Thereby, for coaching group B the increase from T1 to T2 did still not yield a statistically significant difference if compared to the (statistically not significant) increase observed for uniformed participants. On the other hand, the direction of the effect reversed (from d = -0.23 as obtained from the main analysis to d = 0.39) and in this way aligned with the direction of the study's original prediction (Hypothesis 4).

In conclusion then, for script-deviant criteria no evidence was found for any coaching effects on the quality of participants' false statements, regardless of the measure's sensitivity. In turn, if measured on the coaching-restricted scale rather than CBCA sum score, tentative (i.e. significant increase in scores from T1 to T2 exclusively for coaching group B) but not yet conclusive (i.e. no significant difference in the increase of scores if compared to the control

group, but a positive effect size of small strength) evidence was obtained that informing participants about strategy-based criteria results in higher-quality statements.

Limitations and possible directions for future studies

Apart from its underlying truth status, an array of personal and situational variables may affect the CBCA score of a statement (for an overview of relevant research findings, see Volbert and Steller, 2014). On the one hand, the study's approach of measuring the relative difference between the target (T1) and baseline (T2) conditions to deduce the outcome variable (M_{Diff}) for each participant controls well for the potential effects of personal variables on statements' content and quality (see also Schemmel et al., 2020).[41] On the other hand, however, the study's design, with its emphasis on maximising ecological validity, inherently restricted experimental efforts to keep situational variables equal between trials.

For illustration, previous studies about coaching had participants provide their statement about a videotaped (Vrij, Kneller, et al., 2000) or staged (Vrij et al., 2002, 2004) event and in this way kept key characteristics, such as the event's complexity or the time interval between event and interview, identical. In contrast, participants in the current study were to report fully fabricated and previously experienced events, whose natural variety in characteristics could only be broadly confined by specifying rather general guidelines and parameters for topic selection (see Methods section). Striving to keep the emotional and experiential valence of the study events comparable to real-life forensic situations (Steller, 1989), for instance, only events in the topic selection were included in which participants would be likely to be directly involved, to be negatively emotional aroused, and to feel a loss of control (see Steller et al., 1992). In practice, however, it turned out that some of the topics open to selection (i.e. *getting*

[41] In addition, participants were deliberately recruited from various professional fields to minimize selection bias.

lost in nature or wildness, disclosure of an affair, failing in a personally important task or assignment) were likely to evoke reports from participants that appeared to be rather trivial in nature and typically lower in complexity than the reports prompted by some other topics (i.e. *being a victim of an attempted or committed criminal offense, suffering an accident with subsequent medical assistance/treatment, experiencing existential feelings of fear or panic*). Similarly, the study's requirement that the events to be chosen should not date back more than ten years in time resulted in time intervals between event and interview that could considerably differ between trials, ranging from few days to several years. Intending to address the frequently accented need for investigations pertaining to settings high in mundane realism or ecological validity (e.g. Frank & Svetieva, 2012; Oberlader et al., 2016), the author of this Thesis argues, however, that these concessions to the study's degree of experimental control reflect the necessary costs intrinsically linked to such endeavor.

Built on the experimental design of the current study, future studies could, nonetheless, introduce stricter guidelines and parameters for topic selection and associated event characteristics to control better for the influence of situational variables. Implementing stricter guidelines and parameters, such as requiring time intervals of at least several years in length for the event in question, will increase the costs associated with conducting the study even further since selecting suitable topics would become more difficult for participants. Therefore, future researchers would need to find the right balance between the study's degree of experimental control and its practicality. Arguably delicate to accomplish, such studies would not only provide valuable insight about the replicability of the current findings but could also follow up on the tentatively supported speculation that strategy-based criteria might be easier to simulate (and hence be diagnostically less valuable) than script-deviant criteria.

Conclusions

The limited control over situational variables notwithstanding, study III arguably represents a novel and valuable approach for testing the effects of coaching on statements' CBCA scores under forensically relevant conditions. In summary, three major implications can be deduced from its findings:

Regarding true statements, the pattern of CBCA sum scores obtained from uninformed participants reversed in T2, meaning that their false statements yielded higher scores than true statements. Such pattern markedly contrasts with the underlying rationale of CBCA and, if considered in a stand-alone fashion, could be taken as evidence that if motivated to appear credible participants are able to increase the quality of fabricated statements to an extent that renders CBCA assessment futile. The pattern of scores obtained from participants informed about CBCA provides valuable context, however, as no such reversal of scores was observable in the target conditions (T2). Against this background, it seems that the reversal of scores for uninformed participants was primarily driven by their inclination to deliberately lessen the quality of their true report to render their false report more credible in comparison. The author of this Thesis hence speculates that such "self-handicapping" tendencies may reflect a lack of alternative strategies conceivable to participants and encourages future studies to elucidate.

Second, in contrast to previous results (Vrij, Kneller, et al., 2000; Vrij et al., 2002; 2004) the findings of study III revealed that overall informing participants about CBCA had no effects on the quality of their false statements if compared to the statements of uniformed participants. While the way participants were coached in the current study arguably bears higher similarity to actual forensic situations, further methodological differences are present between study III and the three previous investigations that need to be acknowledged (such as the degree of experimental control, the number of criteria being coached and the absence or presence of baseline versus target conditions). Consequently, again subsequent studies are needed to

determine the underlying reasons for the discrepancy in findings. Until then, it can be argued that the commonly expressed notion about CBCA being susceptible to coaching (e.g. Porter & ten Brinke, 2010) should be expressed with (more) caution.

Third and final, relating to the diagnostic value of script-deviant and strategy-based criteria, Maier et al. (2018) showed their value to be high from a motivational perspective since lay-persons tend to avoid their production when deceiving. The results of the current study further corroborated their diagnostic value in a more general sense, since the obtained pattern of CBCA scores suggest that both criteria groups were (cognitively) difficult to simulate for participants even if being encouraged to do so. Thereby, the prediction that upon coaching strategy-based criteria would be cognitively less difficult to simulate than script-deviant criteria was not supported if tested with the CBCA sum score as the underlying measure. Yet, when restricting the analysis to coaching-restricted scale rather than sum scores, there was tentative evidence for this prediction. Since assessing the diagnostic value of the criteria (sets) is of high value for practitioners (Oberlader et al., 2016), the author of this Thesis hopes that the recommendations for future studies higher in experimental control will inspire other researchers to elaborate on the robustness of the current study's findings.

General discussion

The principal goal of this Thesis was to address frequently criticized aspects of CBCA, namely the poor theoretical footing behind the compilation of criteria in substantiating truthfulness and the absence of a weighting system to sort the individual criteria in relation to their diagnostic value. Three separate albeit conceptually linked studies were conducted: Referring to the (theoretical) notion that creative and strategic demands are the driving forces behind the differences in statement quality between false and true reports (Köhnken, 1990), study I experimentally manipulated the presence of these two demands to examine the effects on CBCA sum scores. Study II and study III investigated to what degree the criteria differ from each other in terms of their diagnostic value, which can be gauged for any criterion by taking both its motivational and cognitive properties into account (Niehaus et al., 2005): Study II thus inquired to what extent participants were inclined to produce the different criteria (motivational component) while study III examined to what extent participants were capable of doing so (cognitive component). Thereby, study II and study III specifically relied on the revised three-dimensional model from Volbert & Steller (2014) that, based on conceptual and theoretical considerations, introduced three main sets for grouping the individual criteria. Furthermore, study I and study III were partly built on the same experimental design, meaning that participants from study I also served as a subsample in study III. The following sections discuss the broader implications that can be derived from the findings of the three studies, particularly in relation to the main research questions of this Thesis. Because not all of these questions could be unambiguously solved, suggestions for future investigations are also outlined.

The theoretical footing of CBCA (study I)

The observed relationship between cognitive load and CBCA scores

The findings of study I demonstrated that both creative and strategic demands added to the amount of cognitive load as perceived by participants. Regarding strategic demands, participants further indicated both the true and false condition to be cognitively more taxing with strategic efforts being promoted, which supports the assumption that not only liars but also truth-tellers may be prompted to behave strategically (Vrij and Granhag, 2012). Against this background, the theoretical concept of creative and strategic demands appears pertinent for explaining why lying (i.e. creative and strategic demands are present) tends to be cognitively more demanding than truth-telling (only strategic demands might be present).

The cognitive load account (e. g. Vrij, Fisher, et al., 2008) further stipulates a negative relationship between the amount of cognitive load perceived and one's ability to produce statements of high quality and in this way offers a theoretical explanation for why CBCA criteria substantiate truthfulness: Conceived to be indicative of different forms of statement quality, it follows that the criteria overall should be more likely to saliently emerge in true rather than false statements due to differences in perceived cognitive load.

Regarding this proposed relationship between cognitive load and CBCA sum scores, the observation that with strategic efforts being discouraged (T1; labelled $SE_{[-]}$ in study I), false statements (i.e. creative demands being present) yielded lower CBCA sum scores than true statements (i.e. creative demands being absent) corroborated the assumption of the cognitive load account. That is, the presence of creative demands appears to have resulted in both increased cognitive load and, presumably as a consequence, impaired statement quality. In stark contrast, however, with strategic efforts being promoted (T2; labelled $SE^{[+]}$ in study I), the pairing of creative and strategic demands revealed a positive rather than negative relationship

between cognitive load and statement quality and yielded higher CBCA sum scores than observed for any of the other three experimental conditions. To illustrate once more: Even though the perceived cognitive load was highest, false statements in T2 (creative and strategic demands being present) yielded higher rather than lower CBCA sum scores, both if compared to true statements from T2 (only strategic demands being present) or to false statements from T1 (only creative demands being present).

The need for an external point of reference when interpreting the findings

In the Discussion section of study I, different explanations were provided that might account for the seemingly incongruous findings. For instance, concerning the finding of higher CBCA sum scores for false rather than true statements in T2, it was speculated that participants may have deliberately compromised the quality of their true report, intending to make their false statement appear more convincing in comparison. If one accepts this line of arguing, it follows that regrading the second interview session (T2), the obtained CBCA sum scores for false statements were not necessarily "high" per se, but only higher relative to the (intentionally lowered) CBCA sum scores of true statements. While further empirical investigations as outlined in the Discussion section of study I are required to scrutinize if participants in T2 did indeed willingly compromise the quality of their true statements, it is important to note here that the problem of "relativity" represents a general problem when interpreting the findings of study I and may also extend to its other experimental conditions and their respective comparisons. Put differently, a more meaningful or "objective" point of reference would be needed from which the obtained CBCA scores for each condition could be truly judged as comparatively low or high.

One might argue that the first interview session (T1) could be used as such point of reference since the interview setting was designed to discourage strategic efforts in participants

and in this way may have resembled the typical experimental settings of previous CBCA studies (see Discussion section of study I, in which it is pointed out that the majority of experimental CBCA studies neglected to sufficiently motivate participants to appear credible and hence may have failed to evoke strategic demands). At a closer look though, such resemblance may only be applicable to some of the measures associated with the (failed) promotion of strategic efforts, such as (the lack of) monetary rewards substantial enough to render them meaningful for participants. To further maximise the difference in the strength of strategic demands between T1 and T2 in study I, the interviewer in T1 also made clear to participants that he was already aware of their reports' underlying truth status before prompting the reports from them. To the author's knowledge, this form of experimental manipulation is unique to study I and markedly differs from the procedures of other CBCA studies, in which participants were usually told that the interviewer was unaware of the truth status of their report. It seems reasonable to suspect that highlighting that the interviewer knows the truth status in advance represents a rather powerful measure for minimizing any concerns to appear credible and may have had considerable effects on participants when reporting. The comparability of findings between the first interview session of study I (T1) and other experimental CBCA studies may thus be limited.

Similar to the comparison of true statements, caution is therefore warranted when interpreting the change in scores for false statements as observed in study I. Based on the increase of CBCA sum scores from T1 (creative demands being present) to T2 (creative and strategic demands being present) it was concluded in the Discussion section of study I that participants may be much better able to deal with the pairing of creative and strategic demands and produce false statements of higher quality than commonly assumed. While this line of arguing appears conclusive if restricted to the findings of study I alone, the conclusions may not hold up if extended to the settings of other CBCA studies. For instance, within study I the

CBCA sum scores that were obtained for false statements in T2 are clearly higher if compared to the scores obtained in T1. Yet, technically, the scores obtained from T2 could overall still be lower if compared to the scores that other CBCA studies observed under experimental settings not implemented by study I (i.e. settings with strategic demands widely neglected but, at the same time, with the interviewer being unaware of the reports' truth status).

Implications for the experimental design of future studies

As a variety of methodological differences between CBCA studies (including the way criteria are counted and the CBCA sum score is computed, if at all) prevent meaningful comparisons of their scores between studies, it seems promising for future investigations to bring the experimental setting of T1 closer to the setting of previous CBCA studies. For illustration, while monetary rewards and other incentives that may motivate participants to appear credible should still be kept low in T1, prior knowledge about the reports' underlying truth status could be concealed from the interviewers. Compared to study I, the applied procedures to discourage participants to appear credible in T1 would then be undoubtedly weaker but may still be sufficient for (experimentally) separating the process of fabricating (creative demands only, T1) from the process of deceiving (creative and strategic demands, T2). If so, such experimental setup would resemble the setups of other CBCA studies much closer and in this way would comprise a more meaningful reference point against which the obtained scores from T2 (i.e. a setting with high monetary rewards and other incentives to promote strategic efforts in participants) could be compared against.

Curiously though, when serving as study investigator in study I the author of this Thesis gained the impression that many participants were strongly motivated to provide convincing reports in the first interview setting (T1), irrespective of the interviewer's prior knowledge about their reports' truth status. In the second interview setting (T2), on the other hand, many

participants seemed to be particularly impressed by the fact that the interviews were taking place in an actual forensic institute and were conducted by an (allegedly) professionally trained psychologist. Based on this anecdotal evidence, it appears particularly interesting to investigate whether studies as outlined above (i.e. investigations whose experimental settings, if compared to study I, bear higher comparability to previous CBCA studies in regards to T1 but whose settings for T2 remain largely unchanged) would really yield different results that align better with the prediction of the cognitive load account.

Until such studies are being conducted, the findings of study I provide clear evidence for the cognitive costs associated with creative and strategic demands but remain ambiguous for how this increase in cognitive costs may affect the statement quality of true and false statements. Phrased differently, strategic demands were associated with increased cognitive load both in the true and false condition but had affected participants' CBCA sum scores in opposite and, in relation to the premises of the cognitive load account, counterintuitive ways. The ambiguity of the results' implications notwithstanding, the author of this Thesis contends that study I represents a valuable first step to elaborate further on the theoretical footing of CBCA.

The diagnostic value of individual CBCA criteria (study II & study III)

Other than study I which investigated how theoretically relevant task demands affect the presence of CBCA criteria in general (i.e. in terms of CBCA sum scores), the subsequent two studies examined the motivational and cognitive properties of criteria on the individual and/or group-based level. To reiterate: Two considerations require clarification when assessing a criterion's diagnostic value, namely (1) to what extent the deceiver is *inclined* to produce the criterion and (2) to what extent the deceiver would *be capable* of doing so. In other words then, the diagnostic value of any criterion is to be derived from both its motivational and cognitive

component (Niehaus et al., 2005). Based on the idea that the compilation of individual criteria can be grouped more meaningfully into three different sets of criteria as suggested by the revised CBCA model (Volbert & Steller, 2014), the two studies tested whether the three-dimensional structure of the revised model would contain criteria that within each set yielded comparable motivational (i.e. ascribed strategic meaning; study II) and cognitive (i.e. difficulty associated with their production; study III) properties.

The motivational component of CBCA criteria (study II)

Findings about the motivational properties of the three criteria sets

The findings of study II demonstrated that laypersons consider most CBCA criteria strategic relevant and are inclined to either integrate or avoid them in their deceptive statements. More importantly, the patterns of strategic value ratings were widely consistent within each of the three criteria sets and, in this way, corroborated the "motivational" compatibility of the revised CBCA model: While participants tended to ascribe positive strategic meaning to memory-related criteria (set 1), they tended to ascribe negative strategic meaning to script-deviant (set 2) and strategy-based (set 3) criteria. Study II, therefore, provided insight about the motivational properties of the criteria (sets) that allow for quite straightforward conclusions: Based on the homogenous strategic value ratings that were obtained by the three-dimensional structure of the revised model, it can be inferred that script-deviant (set 2) and strategy-based (set 3) criteria are diagnostically superior to memory-based criteria (set 1).

Methodological limitations

It needs to be noted, however, that the findings of study II were derived from rather small subject groups and pertained to fictitious scenarios that were quite specific in context. As pointed out by Sporer et al. (2020), additional studies should therefore test to what extent the

results apply to different case vignettes and populations. From a broader perspective, one may also argue that the self-report questionnaires used in study II represented an (undesired) experimental manipulation in itself, as participants had been made explicitly aware of the existence of the various criteria before being asked to rate them. Strictly speaking, the obtained results, therefore, indicate only how participants assess the strategic meaning of the criteria after the criteria had been introduced to them through illustrative examples. The results do not, however, allow for inferences about how participants intuitive behavior in actual forensic situations would have been prior to their participation in the study: It simply cannot be ruled out that participating in the study changed participants' awareness of the criteria and, as a consequence, their respective inclinations to avoid or produce them.

Practical implications

Despite some methodological limitations that are associated with study II, distinct practical guidelines for appraising the diagnostic value of the criteria can be formulated: The occurrence of script-deviant (set 2) or strategy-based (set 3) criteria in a statement is diagnostically relevant, as deceivers would typically refrain from simulating such contents. In contrast, no such avoidance tendencies are to be expected for memory-based (set 1) criteria, implying that the mere presence of these criteria does not automatically support a statement's truthfulness. These generalized guidelines are based solely on the motivational component of the criteria. Thus, they can only be of heuristic value until the combined insight from both the motivational and the cognitive perspective allows for better-founded diagnostic assessments (Niehaus, 2008).

The cognitive component of CBCA criteria (study III)

Findings about the cognitive properties of script-deviant and strategy-based criteria

Designed to shed light on the cognitive component of script-deviant (set 2) and strategy-based

(set 3) criteria, the results obtained from study III turned out less straightforward to interpret than the "motivational" findings of study II. Based on theoretical assumptions, it was predicted that when deceiving script-deviant criteria would be cognitively more difficult to simulate than strategy-based criteria. However, coaching participants about script-deviant and strategy-based criteria had no effects on their CBCA sum scores if compared to the corresponding scores of the control group. As CBCA sum scores may have been an insensitive measure for detecting potential coaching effects, the analysis was repeated with coaching-restricted scale scores. Again, between-group comparisons showed that providing participants with information about either script-deviant or strategy-based criteria did not result in an increase of their coaching-restricted scale scores that would have differed from the increase observed for uniformed participants. Interestingly though, on the level of the preliminary conducted within-group comparisons, the respective coaching-restricted scale scores increased significantly only for participants that had been previously informed about strategy-based criteria.

Taken together, the results thus appear to be somewhat ambiguous: If measured on the CBCA sum scores scale, the findings about false statements suggested that both sets of criteria were equally difficult to simulate for coached participants. At the same time, if the measurements were based on the more sensitive coaching-restricted scale rather than CBCA sum scores, some tentative evidence emerged that upon being coached strategy-based criteria are easier to simulate than script-deviant criteria when deceiving.

Methodological limitations

Study III shared the experimental design and one of the participant groups (control group) with study I. To reiterate: Both studies entailed two separate interview settings, which differed in their discouragement (T1 or baseline condition) versus promotion (T2 or target condition) of

strategic demands. For study I the analysis was focused on the change in CBCA scores from T1 to T2 (within-group comparisons) to examine the effects of strategic demands. For false statements, the promotion of strategic efforts yielded unexpected and counter-intuitive results in form of CBCA sum scores that descriptively in- rather than decreased. Earlier in the General Discussion section, the problem of "relativity" was pointed out in regards to study I, meaning that apart from the difference in scores between T1 and T2 an additional "objective" reference point would be required for more meaningful interpretations of the scores.

In study III, the change in scores from T1 to T2 served only as a preliminary outcome variable, in order to subsequently contrast the change in scores as observed for the control group with the change in scores obtained from either one of the two coaching groups (between-group comparisons). Because of these between-group comparisons, any effects that may have emerged from the differences in the two interview settings were held constant across all participant groups, allowing to examine the effects of coaching under conditions in which strategic demands were being (equally) promoted. Consequently, since between-group rather than within-group comparisons constituted the relevant outcome variable, the problem of "relativity" or the lack of an objective reference point seems to be less pressing for the design of study III. On the other hand, considering that the change in scores of uniformed participants (control group) served as the point of reference, it might be problematic that their scores unexpectedly increased once strategic efforts were being promoted (T2): Because of this increase from T1 to T2, one cannot rule out the possibility that when conducting the respective between-group comparisons, the positive effect of strategic demands on CBCA scores may have blurred the (presumably also positive) effects of coaching.

Implications for the experimental design of future studies

In future studies it would therefore be conducive to test whether the coaching manipulation as carried out in study III would yield larger - and possibly, statistically significant - differences in the change of CBCA scores between uninformed and coached participants if the two interview settings did not differ as much in their presence of strategic demands.

In this context, it seems worth pointing out that controversy exists within the traditional deception literature to what degree results from experimental studies are transferrable to real-world forensic situations (e.g. Wright Whelan et al., 2015). While the former typically introduce participants to low-stake situations in which the act of lying remains largely inconsequential (e.g. Burgoon, 2015), the latter arguably carry severe consequences if being found guilty of lying. As the potential outcome for the liar can highly affect his or her performance (Porter & ten Brinke, 2010), the rather trivial contexts of laboratory studies may fail to evoke cues from participants that in forensic situations might saliently emerge (i.e. Frank & Svetieva, 2012). Addressing the frequently accented need for investigations pertaining to interview settings high in ecological validity or mundane realism (e.g. Frank & Svetieva, 2012; Oberlader et al., 2016), elaborate efforts were therefore undertaken in study III to approximate the structural features of the target condition (T2) to real-world forensic situations as closely as possible. Building on the theoretical framework described in study I, particular emphasis was given to expose participants to the task demands that are likely to emerge in forensic contexts (i.e. strategic demands related to secondary deception; Köhnken, 1990). The author of this Thesis argues that this deliberately chosen and – compared to previous experimental CBCA studies – novel design of T2 represents an important feature that allowed for examining the cognitive properties of the criteria in an integrated and theory-driven fashion. Furthermore, the approach of measuring the relative difference between the target (T1) and baseline (T2) condition to deduce the outcome variable for each participant promises to control well for potential effects of personal variables

on statements' content and quality. In light of the unexpected effects of strategic demands on participants' CBCA scores that may have blurred potential coaching effects, however, future studies should nonetheless bring the structural design of the baseline condition (T1) closer to the (elaborate) set-up of the target condition (T2).

Apart from adjustments to the structural design of the two interview settings, subsequent investigations could introduce a third coaching group in which participants are being informed about memory-related (set 1) criteria. Such coaching manipulation was not entailed in study III: Considering that laypersons predominantly ascribed positive strategic meaning to memory-related criteria in study II, it was speculated that informing about the criteria would have little effect on participants' behavior as they should be naturally inclined to produce the criteria anyways, irrespective of any coaching manipulations. Yet, as pointed out above, while the results from study II shed light on how laypersons assess the strategic meaning of the criteria after the criteria had been introduced to them, the findings do not allow for inferences about how their intuitive behavior in (actual or simulated) forensic situations would have been before they participated in the study. Against this background, it seems worthwhile to also examine to what degree coaching about memory-related criteria changes participants' CBCA scores. Such findings would then provide an additional point of reference against which the findings about the simulatability of script-deviant and strategy-based criteria could be compared against.

Summary and final conclusions

A multitude of both experimental and field studies have demonstrated the overall validity of CBCA in discriminating between true and false witness statements. This Thesis contributes to previous research in two ways: First, it empirically tested the assumption that creative and strategic demands negatively affect participants' CBCA sum scores to elaborate on the theoretical footing of the criteria. Simply put, the findings confirmed the predicted relationship between creative demands and CBCA sum scores. The evocation of strategic demands,

however, yielded unexpected results that warrant further investigation. Second, based on the revised CBCA model of Volbert and Steller (2014) and pertaining to the idea that some criteria might be diagnostically superior to others, this Thesis examined the motivational and cognitive properties of the three criteria sets. Regarding the motivational properties, the results indeed suggest that script-deviant (set 2) and strategy-based (set 3) criteria are diagnostically more valuable than memory-based criteria (set 1). The findings about the cognitive properties were less straightforward to interpret but overall indicated that both script-deviant and strategy-based criteria are rather difficult to simulate for participants, which can be taken as further evidence for their diagnostic value. In summary, this Thesis represents a novel step in addressing important questions associated with the utility of CBCA and provides concrete practical implications as well as suggestions for future studies to elaborate on the in part counterintuitive (study I) or ambiguous (study III) results.

References

Akehurst, L., Köhnken, G., & Höfer, E. (2001). Content credibility of accounts derived from live and video presentations. *Legal and Criminological Psychology*, *6*(1), 65–83. https://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. https://doi.org/10.1002/acp.1669

Amado, B. G., Arce, R., & Fariña, F. (2015). Undeutsch hypothesis and Criteria Based Content Analysis: A meta-analytic review, *7*, 3–12. https://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*(2), 201–210. https://doi.org/10.1016/j.ijchp.2016.01.002

Arntzen, F. (2011). *Psychologie der Zeugenaussage: System der Glaubhaftigkeitsmerkmale* [*Psychology of providing witness testimony: System of credibility criteria*] (5th ed.). Munich, Germany: Beck.

Barr, D.J., Levy, R., Scheepers, C., & Tily, H.J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language, 68,* 255–278. https://doi.org/10.1016/j.jml.2012.11.001

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67(1),* 1-48. https://doi.org/10.18637/jss.v067.i01

Blandon-Gitlin, I., Fenn, E., Masip, J., & Yoo, A. H. (2014). Cognitive-load approaches to detect deception: searching for cognitive mechanisms, *Trends in Cognitive Science, 18(9),* 441–444. https://doi.org/10.1016/j.tics.2014.05.004

Blandon-Gitlin, I., Pezdek, K., Lindsay, D. S., & Hagen, L. (2009). Criteria-Based Content Analysis of true and suggested accounts of events. *Applied Cognitive Psychology*, 23, 901–917. https://doi.org/10.1002/acp.1504

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*(2), 187–197. https://doi.org/10.1007/s10979-005-2417-8

Burgoon, J. K. (2015). When is deceptive message production more effortful than truth-telling? A baker's dozen of moderators. *Frontiers in Psychology*, *6*(DEC), 1–9. https://doi.org/10.3389/fpsyg.2015.01965

Cicchetti, D. V. (1994). Guidelines, Criteria, and Rules of Thumb for Evaluating Normed and Standardized Assessment Instruments in Psychology. *Psychological Assessment*, *6*(4), 284–290. https://doi.org/10.1037/1040-3590.6.4.284

Cohen, J. (1988). Statistical Power Analysis for the Behavioral Sciences (2nd ed.). London, UK: Routledge.

Colwell, K., James-Kangal, N., Hiscock-Anisman, C., & Phelan, V. (2015). Should police use ACID? Training and credibility assessment using transcripts versus recordings. *Journal of Forensic Psychology Practice*, *15*(3), 226–247. https://doi.org/10.1080/15228932.2015.1035187

Ekman, P. (1985/2001). *Telling Lies*. New York, NY: Norton.

Evans, J. R., Michael, S. W., Meissner, C. A., & Brandon, S. E. (2013). Validating a new assessment method for deception detection: Introducing a Psychologically Based Credibility Assessment Tool. *Journal of Applied Research in Memory and Cognition*, *2*, 33–41. https://doi.org/10.1016/j.jarmac.2013.02.002

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*, 175–191.

Frank, M. G., & Svetieva, E. (2012). Lies worth catching involve both emotion and cognition. *Journal of Applied Research in Memory and Cognition*, *1*(2), 131–133. https://doi.org/10.1016/j.jarmac.2012.04.006

Goedert, 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. https://doi.org/10.1348/135532505X52680

Hommers, W. (1997). "Die aussagepsychologische Kriteriologie unter kovarianzstatistischer und psychometrischer Perspektive [CBCA criteria from the perspective of covariance statistics and psychometrics]," in Psychologie der Zeugenaussage. Ergebnisse der Rechtspsychologischen Forschung, eds L. Greuel, T. Fabian, and M. Stadler (Weinheim: Psychologie Verlags Union), 87–100.

Hothorn, T., Bretz, F., & Westfall, P. (2008). Simultaneous inference in General Parametric Models. *Biometrical Journal, 50(3),* 346–363. https://doi.org/10.1002/bimj.200810425

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

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

Kuznetsova, A., Brockhoff, P.B., & Christensen, R.H.B. (2017). lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software, 82,* 1–26. https://doi.org/10.18637/jss.v082.i13

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. Frontiers in Psychology, 4, 1–12. https://doi.org/10.3389/fpsyg.2013.00863

Lakens, D. (2017). Equivalence tests: A practical primer for t-tests, correlations, and meta-analyses. *Social Psychological and Personality Science, 8(4),* 355–362. https://doi.org/10.1177/1948550617697177

Leins, D. A., Fisher, R. P. & Ross, S. J. (2013). Exploring liars' strategies for creating deceptive reports. *Legal and Criminological Psychology, 18,* 141-151. https://doi.org/10.1111/j.2044-8333.2011.02041.x

Lenth, R. (2018). emmeans: Estimated Marginal Means, aka Least-Squares Means. R package version 1.3.0. https://CRAN.R-project.org/package=emmeans

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

McCornack, S. A. (1997). "The generation of deceptive messages: laying the groundwork for a viable theory of interpersonal deception," in Message Production: Advances in Communication Theory, ed. J. O. Greene (Mahwah, NJ: LEA), 91–126.

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

Nahari, G. (2016). When the long road is the shortcut: A comparison between two coding methods for content-based lie-detection tools. *Psychology, Crime & Law, 2744*(August), 1–35. https://doi.org/10.1080/1068316X.2016.1207770

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

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

Niehaus, S. (2008b). Täuschungsstrategien von Kindern, Jugendlichen und Erwachsenen [Deception strategies of children, adolescents and adults]. Forensische Psychiatrie Psychol. Kriminol. 2, 46–56. https://doi: 10.1007/s11757-008-0059-7

Niehaus, S., Krause, A., and Schmidke, J. (2005). Täuschungsstrategien bei der Schilderung von Sexualstraftaten [Deception strategies when reporting sexual offences]. Zeitschrift Für Sozialpsychol. 36, 175–187. https://doi.org/10.1024/0044-3514.36.4.175

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*(4), 440–457. https://doi.org/10.1037/lhb0000193

Oberlader, V. A., Naefgen, C., Koppehele-Gossel, J., Quinten, L., Banse, R., & Schmidt, A. F. (2019). Correction to Oberlader et al. (2016). *Law and Human Behavior*, *43*(2), 165.

Oberlader, V. A., Quinten, L., Banse, R., Volbert, R., Schmidt, A. F., & Schönbrodt, F. D. (2021). Validity of Content-Based Techniques for Credibility Assessment–How Telling is an Extended Meta-Analysis Taking Research Bias into Account? *Applied Cognitive Psychology*, *35*, 393–410. https://doi.org/10.1002/acp.3776

O'Sullivan, M., Frank, M. G., Hurley, C. M., & Tiwana, J. (2009). Police Lie Detection Accuracy: The Effect of Lie Scenario, *33*(6), 530–538. https://doi.org/10.1007/s10979-008-9166-4

Porter, S., & Brinke, L. (2010). The truth about lies: What works in detecting high-stakes deception? *Legal and Criminological Psychology*, *15*(1), 57–75. https://doi.org/10.1348/135532509X433151

Porter, S., Yuille, J. C., & Lehman, D. R. (1999). The nature of real, implanted, and fabricated memories for emotional childhood events: Implications for the recovered memory debate. *Law and Human Behavior*, *23*, 517–537. https://doi.org/10.1023/A:1022344128649

R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Rosenthal, R. (1991). *Meta-analytic procedures for social research.* Newbury Park, CA: SAGE Publications, Incorporated. https://doi.org/10.4135/9781412984997

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*(4), 705–737. https://doi.org/10.1037/1076-8971.3.4.705

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

Schank, R. C., & Abelson, R. P. (1977). Scripts, plans, goals and understanding: An inquiry into human knowledge structures. Oxford, UK: Erlbaum.

Schelleman-Offermans, K., & Merckelbach, H. (2010). Fantasy proneness as a confounder of verbal lie detection tools. *Journal of Investigative Psychology and Offender Profiling*, *7*(3), 247–260. https://doi.org/10.1002/jip.121

Schemmel, J., Maier, B.G., & Volbert, R. (2020). Verbal baselining: Within-subject consistency of CBCA scores across different truthful and fabricated accounts. *The European Journal of Psychology Applied to Legal Context, 12,* 35–42. https://doi.org/10.5093/ejpalc2020a4

Schuirmann, D.J. (1987). A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics, 15,* 657–680. https://doi.org/10.1007/BF01068419

Seaman, M. A., & Serlin, R. C. (1998). Equivalence confidence intervals for two-group comparisons of means. *Psychological Methods, 3,* 403-411. https://doi.org/10.1037/1082-989X.3.4.403

Singmann, H. & Kellen, D. (2019). An Introduction to Mixed Models for Experimental Psychology. In D. H. Spieler & E. Schumacher (Eds.), *New Methods in Neuroscience and Cognitive Psychology* (pp. 4–31). Hove, UK: Psychology Press. http://singmann.org/download/publications/singmann_kellen-introduction-mixed-models.pdf

Sporer, S. L. (2012). Making the subjective objective? Computer-assisted quantification of qualitative content cues to deception. In E. Fitzpatrick, J. Bachenko, & T. Fornaciari (Eds.), *Proceedings of the Workshop on Computational Approaches to Deception Detection* (pp. 78- 85). Stroudsberg, PA: Association for Computational Linguistics.

Sporer, S. L., Manzanero, A. L., & Masip, J. (2020). Optimizing CBCA and RM research: recommendations for analyzing and reporting data on content cues to deception. *Psychology, Crime & Law*. https://doi.org/10.1080/1068316X.2020.1757097

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

Steller, M. & Köhnken, G. (1989). Criteria-based statement analysis. Credibility assessment of children's statements in sexual abuse cases. In D. C. Raskin (Ed.), Psychological methods for investigation and evidence (pp. 217–245). New York: Springer.

Steller, M., Wellershaus, P., & Wolf, T. (1992). Realkennzeichen in Kinderaussagen: Emprische Grundlagen der Kriterienorientierten Aussageanalyse [Credibility criteria for children's testimonies: Emprical foundations for a a criterion-oriented analysis of testimonies]. *Zeitschrift für Experimentelle und Angewandte Psychologie, 39(1),* 151–170.

Undeutsch, U. (1967). Beurteilung der Glaubhaftigkeit von Zeugenaussagen [Assessing the credibility of witnesses' testimony]. In U. Undeutsch (Ed.), Handbuch der Rechtspsychologie (Vol. 11, pp. 26–181). Göttingen, Germany: Hogrefe.

Volbert, R., & Steller, M. (2014). Is This Testimony Truthful, Fabricated, or Based on False Memory? *European Psychologist, 19*(3), 207–220. https://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, and Law*, *11*(1), 3–41. https://doi.org/10.1037/1076-8971.11.1.3

Vrij, A., Akehurst, L., Soukara, S., & Bull, R. (2002). Will the truth come out? The effect of deception, age, status, coaching, and social skills on CBCA scores. *Law and Human Behavior*, *26*(3), 261–283. https://doi.org/10.1023/A:1015313120905

Vrij, A., Akehurst, L., Soukara, S., & Bull, R. (2004). Let me inform you how to tell a convincing story: CBCA and reality monitoring scores as a function of age, coaching, and deception. *Canadian Journal of Behavioural Science / Revue canadienne des sciences du comportement*, *36*(2), 113–126. https://doi.org/10.1037/h0087222

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. https://doi.org/10.1023/A:1006610329284

Vrij, A., Fisher, R. P., & Blank, H. (2017). A cognitive approach to lie detection: A meta-analysis. *Legal and Criminological Psychology*, *22*(1), 1–21. https://doi.org/10.1111/lcrp.12088

Vrij, A., Fisher, R., Mann, S., & Leal, S. (2008). A cognitive load approach to lie detection. *Journal of Investigative Psychology and Offender Profiling*, *5*(1–2), 39–43. https://doi.org/10.1002/jip.82

Vrij, A., & Granhag, P. A. (2012). Eliciting cues to deception and truth: What matters are the questions asked. *Journal of Applied Research in Memory and Cognition*, *1*(2), 110–117. https://doi.org/10.1016/j.jarmac.2012.02.004

Vrij, A., Granhag, P. A., Mann, S., & Leal, S. (2010). Outsmarting the Liars: Toward a Cognitive Lie Detection Approach. *Current Directions in Psychological Science*, *20*(1), 28–32. https://doi.org/10.1177/0963721410391245

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*, 57–70. https://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. https://doi.org/10.1080/10683160500129007

Vrij, A., Mann, S. A., Fisher, R. P., Leal, S., Milne, R., & Bull, R. (2008). Increasing cognitive load to facilitate lie detection: The benefit of recalling an event in reverse order. *Law and Human Behavior*, *32*, 253–265. https://doi.org/10.1007/s10979-007-9103-y

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 Hum Behav*, *31*, 499–518. https://doi.org/10.1007/s10979-006-9066-4

Willén, R. M., & Strömwall, L. (2012). Offenders' uncoerced false confessions: A new application of statement analysis? *Legal and Criminological Psychology, 17*, 346–359. https://doi.org/10.1111/j.2044-8333.2011.02018.x

Welle, I., Berclaz, M., Lacasa, M. J., & Niveau, G. (2016). A call to improve the validity of criterion-based content analysis (CBCA): Results from a field-based study including 60 children's statements of sexual abuse. *Journal of Forensic and Legal Medicine, 43*, 111–119. https://doi.org/10.1016/j.jflm.2016.08.001

Wellek, S. (2010). *Testing statistical hypotheses of equivalence and noninferiority* (2nd ed.) Boca Raton, FL: CRC Press.

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.

Wright Whelan, C., Wagstaff, G. F., & Wheatcroft, J. M. (2015). Subjective cues to deception/honesty in a high stakes situation: an exploratory approach. *The Journal of Psychology*, *149*(5), 517–534. https://doi.org/10.1080/00223980.2014.911140

Zuur, A. F., Ieno, E. N., & Elphick, C. S. (2010). A protocol for data exploration to avoid common statistical problems. *Methods in Ecology and Evolution, 1*, 3–14. https://doi.org/10.1111/j.2041-210X.2009.00001.x

Acknowledgements

"So it's been kind of a long road, but it was a good journey altogether." [Sidney Poitier]

I would like to express my gratitude to my supervisor Prof. Dr. Renate Volbert, who continuously encouraged and supported me throughout all stages linked to this project, starting with my efforts to secure a scholarship and ending with revising the final draft of this Thesis. The road was indeed long and certainly not free of occasional disappointments and setbacks, but thanks to your genuine interest and skilful advice, the journey could always continue.

Furthermore, I am grateful to Jonas Schemmel and Olivia Baum, who carried out CBCA ratings, as well as to Dilara Isbara, Marie Haemmerling and Maximilian F. Schwarz, who acted as confederates in study I and III. Also, I would like to thank Prof. Dr. Johannes Zimmermann for his advice on statistical analysis questions.

My work was individually supported by the Elsa-Neumann-Scholarship of Berlin. Otherwise, this research received no specific grant from any funding agency in the public, commercial, or non-profit sectors.

Appendix A: Publications

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

frontiers in Psychology

ORIGINAL RESEARCH
published: 08 June 2018
doi: 10.3389/fpsyg.2018.00855

The Strategic Meaning of CBCA Criteria From the Perspective of Deceivers

Benjamin G. Maier[1], *Susanna Niehaus*[2], *Sina Wachholz*[3] *and Renate Volbert*[1,3]*

[1] Psychologische Hochschule Berlin, Berlin, Germany, [2] Lucerne University of Applied Sciences and Arts, Lucerne, Switzerland, [3] Charité - Universitaetsmedizin Berlin, Institute of Forensic Psychiatry, Berlin, Germany

In 2014, Volbert and Steller introduced a revised model of Criteria-Based Content Analysis (CBCA) that grouped a modified set of content criteria in closer reference to their assumed latent processes, resulting in three dimensions of *memory-related*, *script-deviant* and *strategy-based criteria*. In this model, it is assumed that deceivers try to integrate memory-related criteria—but will not be as good as truth tellers in achieving this—whereas out of strategic considerations they will avoid the expression of the other criteria. The aim of the current study was to test this assumption. A vignette was presented via an online-questionnaire to inquire how participants (n = 135) rate the strategic value of CBCA criteria on a five-point scale. One-sample t-tests showed that participants attribute positive strategic value to most memory-related criteria and negative value to the remaining criteria, except for the criteria *self-deprecation* and *pardoning the perpetrator*. Overall, our results corroborated the model's suitability in distinguishing different groups of criteria—some which liars are inclined to integrate and others which liars intend to avoid—and in this way provide useful hints for forensic practitioners in appraising the criteria' diagnostic value.

Keywords: criteria-based content analysis, strategic self-presentation, content-related deception strategies, beliefs about verbal cues of deception, primary vs. secondary deception, cognitive vs. motivational component

OPEN ACCESS

Edited by:
Agata Debowska,
University of Sheffield,
United Kingdom

Reviewed by:
Dominic Willmott,
University of Huddersfield,
United Kingdom
Annelies Vredeveldt,
VU University Amsterdam,
Netherlands

***Correspondence:**
Renate Volbert
r.volbert@
psychologische-hochschule.de

Specialty section:
This article was submitted to Forensic and Legal Psychology, a section of the journal Frontiers in Psychology

Received: *07 March 2018*
Accepted: *11 May 2018*
Published: *08 June 2018*

Citation:
Maier BG, Niehaus S, Wachholz S and Volbert R (2018) The Strategic Meaning of CBCA Criteria From the Perspective of Deceivers. Front. Psychol. 9:855. doi: 10.3389/fpsyg.2018.00855

INTRODUCTION

The Empirical Footing of CBCA

The underlying rationale of Criteria-Based Content Analysis (CBCA) holds that the content of experienced-based accounts is qualitatively higher than the content of fabricated statements (the so-called Undeutsch Hypothesis; Undeutsch, 1967). After identifying content characteristics that practitioners and scholars deemed suitable to substantiate truthfulness, Steller and Köhnken (1989) compiled a systematic set of 19 CBCA criteria (see **Table 1**). Since then, a multitude of both field and laboratory studies have confirmed that experience-based accounts indeed yield higher content quality than fabricated statements, in this way corroborating the overall validity of CBCA as a *truth*-detection tool (for recent meta-analyses see Amado et al., 2016; Oberlader et al., 2016). Most of these studies summed up the individual criteria scores to one comprehensive (total) CBCA score, which subsequently served as the relevant variable for further analysis (e.g., Akehurst et al., 2001; Welle et al., 2016). Such an approach, however, may be too simplistic and may underestimate the actual utility of CBCA, since it ignores the possibility that some criteria might be more sensitive to truthfulness and hence bear higher diagnostic value than others. For instance, after having

TABLE 1 | Original compilation of CBCA criteria (Steller and Köhnken, 1989).

General characteristics
1. Logical consistency
2. Unstructured Production
3. Quantity of details
Specific contents
4. Contextual embedding
5. Description of interactions
6. Reproduction of conversation
7. Unexpected complication during the incident
Peculiarities of content
8. Unusual details
9. Superfluous details
10. Accurately reported details not comprehended
11. Related external associations
12. Accounts of subjective mental state
13. Attribution of perpetrator's mental state
Motivation-related content
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 elements
19. Details characteristic of the offense

identified children who were able to achieve a higher total CBCA score in their fabricated than in their truthful accounts, Hommers (1997) showed that the criteria *accurately reported details not comprehended*, *unexpected complications*, and *related external associations* still discriminated between true and fabricated statements. This suggests that particular weights should be assigned to these criteria as predictors of the veracity of statements. More empirical knowledge is essential however, if one intends to advance the prospect of weighting beyond the stage of mere suggestion. To date, CBCA still lacks a weighting system, despite the fact that numerous researchers have criticized its absence and stressed the need for implementation to increase the method's accuracy and sensitivity (e.g., Vrij, 2005; Porter and ten Brinke, 2010).

Theoretical Considerations About the Diagnostic Value of CBCA Criteria

The diagnostic value of a criterion refers to its validity in discriminating between self-experienced and fabricated statements. For identifying a criterion as diagnostically valuable, its occurrence in true statements is necessary, but by no means sufficient (i.e., Greuel et al., 1998): The relevant question is not how likely a criterion is to occur in true statements, but how likely it is to appear in true *relative* to fabricated statements. As a first step toward inferring the diagnostic value of a criterion, theoretical considerations of what processes govern its emergence loom necessary: Why exactly is a criterion expected to occur in true accounts, but not in fabricated ones? From a psychological perspective, two universal aspects apply to a truth teller but not to a lying person (Volbert and Steller, 2014): Truth tellers report from actual memory, and in doing so are convinced that the event in question had happened as reported. Therefore, in the statement of an honestly reporting person content criteria are likely to occur naturally, as they reflect phenomena associated with genuine memories and feelings of sincerity. A lying person on the other hand needs to put (more) deliberate effort in inventing information that substitutes the missing memory of the alleged experience (creative demands related to *primary deception*; Köhnken, 1990) and in masking the discrepancy between statement and belief by presenting the fabricated event in a credible manner (strategic demands related to *secondary deception*; Köhnken, 1990).

In accordance with the premise of primary vs. secondary deception, Köhnken (1990) distinguished between two forms of CBCA criteria by classifying them as being either cognitively- or motivationally-related. Both kinds of criteria indicate true statements, albeit for different reasons: The former relate to creative (or cognitive) demands; typically, they should be too difficult to produce when fabricating. Motivational criteria refer to how a witness presents a statement; typically, they should be avoided out of strategic considerations when lying. While such categorization suggests that a criterion's diagnostic value is to be derived from either its cognitive or motivational aspects, Niehaus et al. (2005) pointed out that both components need to be taken into account. That is, considerations of the underlying motivational component should also be applied to criteria originally regarded as purely cognitively-related, and vice versa. In summary then, two considerations require clarification if the diagnostic value of a criterion is to be deduced: (1) To what degree is the deceiver inclined to produce the criterion and (2) to what degree would the deceiver be capable of doing so. Insight about the motivational component should hence provide a first hint toward the criterion's diagnostic value: If the deceiver considers the criterion to be strategically detrimental to his or her self-presentation efforts, the likelihood of its emergence in fabricated accounts is generally lower. In turn however, if the deceiver ascribes positive strategic value to a criterion and thus is inclined to produce it, a higher likelihood for its occurrence does not necessarily follow: Whether or not the criterion will emerge should then crucially depend on the cognitive component, that is, how difficult it is for the deceiver to integrate the criterion in his or her statement (*differential controllability*, Köhnken, 1990).

Previous Research About the Motivational Component

The concept of strategic self-presentation is firmly established within the literature, stating that liars are typically more concerned with appearing credible than truth tellers (DePaulo, 1992; DePaulo et al., 2003). Inquiring about suspects' strategies to appear credible during police interrogations, Hartwig et al. (2007) correspondingly found that lying participants seemed to be more prone to adopting verbal[1] and non-verbal strategies

[1]In total, lying participants reported five categories of strategies regarding the verbal content of their statements, with "providing a detailed story" and "avoid

than truthfully reporting participants. Further scientific efforts to examine deceivers' verbal strategies are hardly existent however. While there are several articles that did elaborate on beliefs of lay people about verbal cues of deception, these studies primarily investigated which kind of contents are believed to be indicative for detecting lies in somebody else's statement (e.g., Granhag and Strömwall, 2000; Vrij et al., 2006; Bogaard et al., 2016). Consequently, insights derived from their findings do not necessarily elucidate on the actual content-related strategies of deceivers, since the results were gained from the perspective of the to-be-deceived rather than from the perspective of the deceiver (Niehaus et al., 2005). That is, while people's beliefs of how lies can be spotted in others are likely to affect their strategies in detecting them (Ryan et al., 2018), these beliefs must not necessarily govern their own way of acting when being deceptive. To our knowledge then, only two studies (Niehaus et al., 2005; Niehaus, 2008) exist which quantitatively examined how laypersons assess the strategic value of CBCA or other content-related criteria in the context of deception. This is certainly surprising, considering that the CBCA criteria classified as motivationally-related are only valid if the forensic assumptions—predicting that laypersons would ascribe negative strategic meaning to these criteria and try to avoid their production when deceiving—are correct. We, therefore, aim to elaborate further on the motivational component of CBCA criteria by building on the findings of Niehaus et al. (2005) and Niehaus (2008) about content-related deception strategies. Because the two previous studies are available in German language only, they are first introduced in more detail.

Studies About Content-Related Deception Strategies

Both studies asked subjects to take the perspective of a fictitious protagonist, who for personal reasons needed to convincingly lie about a certain type of event. The first investigation (Niehaus et al., 2005; $n = 120$; $M_{age} = 29.6$; all females) presented a scenario in which the protagonist's friend felt unable to press charges against her neighbor, who had previously raped her. Therefore, the protagonist decided to claim that she herself had been raped by the neighbor so that the perpetrator would still receive punishment. The story outline in the second study (Niehaus, 2008; $n = 50$; $M_{age} = 28.6$; 16 men) entailed less severe ramifications: The protagonist needed to find a convincing explanation why he had shown up late at work, as otherwise he would be fired by his boss. In need of an excuse, he wrongly accused his neighbor to have had him trapped in a cellar. After the presentation of the story outline, a standardized questionnaire was handed out on which each CBCA criterion was described in a discrete way as well as illustrated by means of an example embedded into the story outline. Subjects then had to indicate on a 5-point (Niehaus, 2008) or 7-point (Niehaus et al., 2005) scale whether they would rather integrate or avoid the criterion if they aimed to deliver the false statement as convincingly as possible. Negative ratings signified that the criterion was considered to weaken one's credibility, while a positive value indicated that the criterion was believed to promote deception efforts. If a criterion was considered to be strategically irrelevant, the neutral value "0" was to be assigned. For analysis purposes one-sample *t*-tests were conducted to reveal whether the averaged value ratings differed significantly from "0," indicating that subjects ascribed strategic meaning to the criterion. **Table 2** summarizes the obtained results from both studies. Note that some CBCA criteria, as well as their classification, differ from the original compilation (Steller and Köhnken, 1989) since the authors deduced five higher-level strategic goals that deceivers would pursue in the context of sexual rape allegations, and structured the criteria accordingly (for more detail see Niehaus, 2001; Niehaus et al., 2005).

As depicted in **Table 2**, the results give rise to two major implications: First, most motivational criteria were rated as strategically negative and second, cognitive criteria were found to carry strategic meaning as well. The pattern showing that subjects generally ascribed negative strategic meaning to motivational criteria is paramount, given that the discriminatory value of these criteria rests largely on the presumption that laypersons would tend to avoid them in deception contexts. The findings also corroborated the postulation of Niehaus et al. (2005), stating that for each criterion—regardless of its original classification—both motivational and cognitive aspects needed to be considered if its diagnostic value is to be deduced. In regards to the modified structure of the CBCA model (**Table 2**) however, group I (competency), IV (content-related inconspicuousness) and V (formal inconspicuous) contained criteria of both positive and negative valences. Consequently, the model's structure does not allow for clear group-based distinctions on the motivational level. Considering further that purely motivational aspects governed the classification of the criteria into the five different groups, no information about the cognitive component can be deduced from its criteria groups.

The Revised Model of CBCA Criteria

In 2014, Volbert and Steller introduced a revised CBCA model, which is based on theoretical considerations of what processes govern the emergence of criteria in statements (Volbert and Steller, 2014). The model still distinguishes criteria pointing to the differences in the cognitive processes of liars and truth tellers (cognitive criteria) from criteria referring to the aspects of strategic self-presentation (motivational criteria). Other than before, cognitive criteria are distinguished even further in the model, resulting in two main groups of cognitively-related criteria: The first group entails details characterizing *episodic autobiographical memory*, such as specific spatiotemporal and self-related information. When deceivers fabricate statements, these characteristics are likely to occur as well since they provide essential information (i.e., *temporal* or *spatial* details) without which any delivered account would appear incomplete. Because liars cannot draw on actual episodic memory however, the criteria are assumed to be expressed in a less elaborate way than in experience-based statements. The second main group of cognitive criteria comprise *script-deviant details*, such as *unusual details* or *unexpected complications during the incident*. Criteria

saying things that were not true" being the categories most frequently mentioned (each being mentioned by 22% of the relevant sample). As pointed out by Hartwig et al. (2007), these findings were limited to qualitative analyses and thus are best to be interpreted as a starting point for future research.

TABLE 2 | Strategic value ratings for CBCA criteria[c] (Niehaus et al., 2005; Niehaus, 2008).

Strategic goal	CBCA criterion	Classification
I. Competency (of the deceiver)	**Spontaneous corrections[a,b]**	M (–)
	Admitting lack of memory	M (–)
	Efforts to remember[a,b]	M (–)
	Expressing uncertainty[a,b]	M (–)
	Reality controls[a]	M (–)
	Justifying memory gaps/uncertainties[b]	M (+)
	Spontaneous clarifications	M (+)
II. Moral impeccability (of the deceiver)	**Raising doubts about one's own person[a,b]**	M (–)
	Self-deprecation[a]	M (–)
III. Deprecation (of the accused person)	Pardoning the perpetrator[a]	M (–)
IV. Content-related inconspicuousness (of the statement)	Unexpected complications[b]	C (–)
	Unusual details[a,b]	C (–)
	Information about everyday-life routines (context)[b]	C (+)
	Spatial information (context)[b]	C (+)
	Temporal information (context)[b]	C (+)
	Description of interactions[a]	C (+)
	Reproduction of conversations[a]	C (+)
	Emotions and feelings[a,b]	C (+)
	Attribution of perpetrator's mental state	C (+)
	Personal implications[a,b]	C (+)
V. Formal inconspicuousness (of the statement)	**Plausibility[a,b]**	C (+)
	Logical consistency[a,b]	C (+)
	Quantity of details[a,b]	C (+)
	Unstructured production[a,b]	C(–)
	Superfluous details[a,b]	C (–)
	Raising doubts about one's own testimony[a]	M (–)

$p < 0.05$: [a]Study of Niehaus et al. (2005); [b]study of Niehaus (2008).
Criteria in green font color received significant positive value ratings in at least one of the two studies, while criteria in red font color symbolize significant negative ratings. Additional bold font was applied if these criteria received significant ratings in both studies.
The third row shows whether a criterion was originally classified as cognitively ("C")- or motivationally-related ("M"), based on the work of Köhnken (1990). Furthermore, Niehaus et al. (2005) referred to the impression-management account to predict in advance whether laypersons would ascribe positive ("+") or negative ("–") strategic meaning to a criterion.
[c]To allow meaningful comparisons, criteria that were investigated in only one of the two studies are not presented (self-related/victim-related/neutral associations, attribution of negative traits, clichés, repetitions).

from this group should rarely occur in fabricated accounts, considering that a lying person must construct his or her statement from cognitive scripts (e.g., Schank and Abelson, 1977) to substitute for the lack of experience-based memories. Cognitive scripts reflect the liar's subjective assumptions of what properties typically look like for the event in question (Köhnken, 1990). Script-deviant criteria, on the other hand, refer to characteristics that go beyond the very limitations of such simplified, script-guided knowledge. If the statement giver cannot draw on actual memories providing a potential source for script-deviant elements, he or she should face great difficulties in deliberately producing them as he or she would need to overcome the limited scope of his or her own imagination (Köhnken, 1990). Finally, the third main group refers to motivational criteria, thereby addressing *efforts of positive strategic self-presentation* (see section Theoretical Considerations About the Diagnostic Value of CBCA Criteria for an explanation of why motivational criteria are expected to appear in true rather than fabricated statements). **Table 3** depicts the revised model, with the original binary classification of cognitively- vs. motivationally-related criteria presented in brackets behind each criterion.

Aim of the Present Study

The present study inquired about the content-related deception strategies of laypersons like done by Niehaus et al. (2005) and Niehaus (2008) before. However, their categorization of criteria into 5 groups resulted in value ratings that were of opposite valence within some of the groups (i.e., positive *and* negative value ratings across criteria of the same group), thereby rendering group-based distinctions impractical. Against this background, the main goal of the current study was to examine whether the theoretically-driven structure of the revised model would correspond better to the observed pattern of strategic value ratings. Put differently, the relevant question was to what degree the composition of each of the three criteria groups would contain criteria that on the motivational level are compatible with each other (i.e., containing criteria that consistently carry either negative or positive strategic meaning). Derived from the findings of Niehaus et al. (2005) and Niehaus (2008), we expected participants to rate the memory-related (group 1) criteria predominantly positive. In contrast, we expected participants to generally attribute negative strategic value to script-deviant (group 2) and strategy-based (group 3) criteria. Overall, for each of the three criteria groups, the predicted pattern of strategic ratings should result in a degree of compatibility higher than observed for any previous models, meaning that within each group the strategic value ratings should be either consistently negative *or* consistently positive.

METHODS

Participants

A total of 135 participants (M_{age} = 28.6, SD = 9.8, Range 19–67; 32 men) filled out a questionnaire[2] inquiring about content-related deception strategies. The sample consisted mostly of students (n = 66) or working professionals (n = 55). Participants were recruited via an online participation system of

[2]The study was conducted in German language.

TABLE 3 | Modified system of content characteristics[a] (Volbert and Steller, 2014).

Autobiographic memory vs. script information		Strategic self-presentation
Criteria related to episodic autobiographical memory (Group 1)	**Criteria related to script-deviant information (Group 2)**	**Criteria related to efforts of positive strategic self-presentation (Group 3)**
Information about everyday life routines [C]	Unexpected complications [C]	Spontaneous corrections [M]
Spatial information [C]	Superfluous details [C]	Admitting lack of memory [M]
Temporal information [C]	Unusual details [C]	Efforts to remember [M]
Description of interactions [C]	Related external associations [C]	Expressing uncertainty [M]
Reproduction of conversations [C]	Accurately details not comprehended [C]	Reality controls [M]
Emotions and feelings [C]		Raising doubt about one's own testimony [M]
Own thoughts [C]		Raising doubts about one's own person [M]
Sensory Impressions [C]		Self-deprecation [M]
Attribution of perpetrator's mental state [C]		Pardoning the perpetrator [M]
Personal implications [C]		

[C], Cognitive criteria; [M], Motivational criteria.

[a] *Volbert and Steller (2014) understand their allocation of specific criteria to be exemplary rather than irrevocably. For illustration purposes, the structure presented in this article thus differs slightly from the version originally presented by the authors: In the original version, a separate category "statement as a whole" addresses criteria that can only be evaluated if the statement is analyzed in its entirety, as opposed to scoring the same criterion multiple times at different parts of the statement ("single characteristics"). For reasons of clarity, we exclusively focused on single characteristics and rejected all "statement as a whole" criteria (namely reconstructability of the event, vividness of the event, quantity of details, unstructured production and spontaneous supplementing).*

the University of Potsdam (Germany) or through advertisement in public Facebook groups related to psychological topics. Prior to participation, all participants were assured that their data would be treated confidentially, and all participants gave written informed consent. Upon request, credit points were awarded for participation.

Procedure and Material

The survey was administered online by using the platform www.soscisurvey.de. The first items of the questionnaire asked subjects to provide information about age, gender, and occupation. Next, the story outline was presented, with the instruction to assume the perspective of the protagonist. The story closely resembled the one devised by Niehaus (2008), with only minor modifications to preclude potential misunderstandings. In brief, the story[3] described a protagonist who is at risk of losing his highly valued job, unless he would be able to deliver a convincing explanation to his boss for his (repeatedly) belated arrival at work. After having read the story outline, 27 CBCA criteria derived from the model of Volbert and Steller (2014) were presented on the questionnaire, with the sequence of criteria presentation being adjusted to the model's structure. For each criterion, we gave a short abstract description illustrated by an example embedded into the story outline. For instance, we first described the criterion *spontaneous corrections* by phrasing the question in the following way: "Without being asked, would you refute parts of the information you had already provided at an earlier stage and revise them?" Subsequently, the criterion was illustrated by means of the following story-related example: "Oh no, that was wrong what I had said earlier. In fact, I was holding the folder already in my hands at this time point." Before the first criterion was presented, we further instructed participants to not pay too much attention to the upcoming examples, but to indicate whether in principle they would rather integrate or avoid the criterion in question. Thereby, we asked participants to also consider to what degree they would feel confident to integrate the criterion in their fabricated accounts. With this instruction, we intended to strengthen participants' motivation to assess the strategic meaning of the criteria in a thorough and attentive way. For each criterion, participants should indicate their strategic assessment on a five-point scale (−2: "No, this would strongly weaken my credibility"; −1: "No, this would weaken my credibility"; 0: "It does not matter. My credibility would remain unchanged"; +1: "Yes, this would strengthen my credibility" +2: "Yes, this would strongly strengthen my credibility").

[3] See Appendix 1 for a display of the entire story as presented to participants.

RESULTS

Identical to the investigations of Niehaus et al. (2005) and Niehaus (2008), the goal of our analysis was to compare participants' strategic value ratings for each criterion to the neutral value "0" (= no relevant strategic meaning). Significant differences between any such pair of values would indicate that participants were inclined to either avoid or integrate the respective criterion, dependent on the valence of the criterion's rating (negative vs. positive). We therefore conducted one-sample t-tests to assess for each criterion whether its average strategic rating differed significantly at the $p < 0.05$ level from the value "0." To account for the inflated Type I error rate due to multiple t-tests, the Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995) was performed, which controls the expected proportion of falsely rejected null hypotheses (false positives). In consideration of the rather explanatory nature of our study, we preferred this method over Bonferroni corrections, since the latter greatly increase the probability of a Type II error (Narum,

2006). With the false discovery rate set at 5%, no false positives were detected. Consequently, we can safely conclude that at least 95% of the criteria ratings that were found to be significant were correctly identified as such. The results are presented in **Table 4**, with the criteria structured according to the revised model of Volbert and Steller (2014).

Viewed irrespective of criteria dimension or classification, the strategic value ratings for 18 of the 24 criteria differed significantly from "0," indicating that participants considered most criteria to be strategically relevant when deceiving. The effect sizes of the significantly rated criteria ranged from $d = 0.19$ to $d = 1.28$. For 12 of these 18 criteria the value ratings were negative, while the remaining 6 criteria received positive value ratings. In sum then, the criteria can be generally categorized as strategically relevant vs. strategically not relevant from the perspective of deceivers. As further hypothesized, the strategically relevant criteria were found to carry strategic meaning of either positive or negative valence.

Regarding cognitive criteria, the results showed that participants tended to ascribe strategic meaning of different valence to them. This finding, in fact, supports our predictions since at closer inspection a clear difference between the two dimensions of cognitive criteria was indeed visible: Taking only statistically significant value ratings into account, all ratings for memory-related (group 1) criteria were of positive valence. Among these criteria the effect sizes were of medium strength ($d > 0.5$; Cohen, 1988), except for *temporal information* ($d = 0.19$). The value ratings for 6 of the 10 memory-related criteria were statistically non-significant, among which the strategic ratings of 3 criteria (*own thoughts, sensory impressions, attribution of perpetrator's mental state)* were tentatively negative. Put another way, if participants attributed strategic meaning to a criterion from group 1 (in terms of value ratings that differed significantly from "0"), the valence was exclusively positive. In sharp contrast, for script-deviant (group 2) criteria all ratings assumed significant negative values, with the strength of their effect sizes ranging from small ($d > 0.2$; Cohen, 1988) to large ($d > 0.8$; Cohen, 1988).

Regarding motivational criteria (strategy-based criteria; group 3), the value ratings of all criteria differed significantly from "0." For 7 of the 9 criteria, these ratings were negative—with predominantly large effect sizes—and hence in line with our predictions. Other than predicted and deviating from the forensic assumptions about the strategic meaning of motivational criteria, *self-deprecation* $\{M = 0.36, SD = 1.18, t_{(134)} = 3.57, d = 0.31,$ 95% CI [0.16, 0.56], $p < 0.001\}$ and *pardoning the perpetrator* $\{M = 0.44, SD = 1.05, t_{(134)} = 4.93, d = 0.42,$ 95% CI [0.27, 0.62] $p < 0.001\}$ obtained significant positive values, suggesting that participants intended to integrate rather than avoid these contents.

DISCUSSION

Our findings indicate that on the motivational level, most CBCA criteria are clearly distinguishable in criteria that liars are inclined to integrate vs. criteria that liars are inclined to avoid in their statements. For half of the 18 strategically relevant criteria (as defined by value ratings differing significantly from "0") we obtained effect sizes of at least medium strength, which may be interpreted as evidence for the practical significance of these findings. Furthermore, even though the three-group structure of the revised model goes beyond simply dichotomizing the criteria in bearing positive vs. negative valence, it nonetheless yielded a widely homogenous pattern of strategic value ratings. In comparison to the 5-group classification used in the two previous studies (Niehaus et al., 2005; Niehaus, 2008), a considerably higher degree of compatibility within each criteria group was observable. For illustration purposes, **Table 5** applies the results of all three studies to the structure of the revised model, enabling viewers to graphically examine to what extent the studies' results tally with each other.

Inspecting the valence of the strategic ratings on group level, the degree of compatibility appears to be lowest for memory-related criteria of group 1. That is, in our study the strategic ratings of more than half of the criteria in this group failed to reach significance and in part showed a directional tendency opposite to the direction of the strategically relevant criteria. If collectively examined however, *attribution of perpetrator's mental state* remains the only criterion to which no (positive) strategic meaning was attributed in any of the three studies[4] (see **Table 5**). Nonetheless, the overall picture for criteria of group 1 remains less consistent than for the criteria of the other two groups, as positive ratings that were significant in *each* of the three studies (2 out of 7 criteria) were the exception rather than the rule.

Viewed the other way around however, it is crucial to note that none of the three investigations yielded significantly negative ratings for memory-related criteria, a pattern which sharply contrasts with the results found for script-deviant and strategy-based criteria. Motivational considerations are therefore of little avail in ascribing diagnostic value to memory-related criteria, as the underlying rationale requires that liars would typically avoid such contents. Instead, only considerations on the cognitive level may explain why these specific criteria are more likely to occur in true than in fabricated statements, implying that when lying certain contents are more difficult to produce than when telling the truth.

Regarding script-deviant criteria of group 2, our results showed that they consistently carry negative strategic meaning for laypersons. These results correspond well with the findings previously reported by Niehaus et al. (2005) and Niehaus (2008), as illustrated by **Table 5** (significant negative ratings across all three studies[5] for 2 out of 3 criteria). In contrast to memory-related criteria, motivational considerations thus appear relevant when assessing the diagnostic value of script-deviant criteria, considering that deceivers typically intend to avoid their production.

[4]The criteria *own thoughts* and *sensory impressions* were only investigated in the present study, so no comparisons to the other two studies can be made.

[5]The criteria *related external associations* and *accurately reported details misunderstood* were only investigated in the present study, so no comparisons to the other two studies can be made.

TABLE 4 | Mean value, standard deviation, effect size (Cohen's d), 95% confidence interval values and results of the one-sample *t*-test (test value = 0) for each criterion.

	M	*SD*	*d*	95% LCL	95% UCL	*t**	*p*
MEMORY-RELATED CRITERIA [C]							
Information about everyday-life routines	0.54	0.87	0.62	0.39	0.69	7.22	0.001
Spatial information	0.07	1.01	0.07	−0.11	0.24	0.77	0.222
Temporal information	0.20	1.07	0.19	0.02	0.38	2.17	0.016
Descriptions of interactions	0.01	0.97	0.08	−0.16	0.17	0.9	0.465
Reproduction of conversations	0.00	1.07	0.00	−0.18	0.18	0.00	0.500
Emotions and feelings	0.71	0.98	0.72	0.54	0.88	8.40	0.001
Own thoughts	−0.03	1.08	−0.03	−0.21	0.15	−0.32	0.750[A]
Sensory impressions	−0.16	1.06	−0.15	−0.34	0.03	−1.70	0.092[A]
Attribution of perpetrator's mental state	−0.11	0.90	−0.12	−0.26	0.04	−1.44	0.152[A]
Personal implications	0.46	0.90	0.51	0.31	0.61	5.96	0.001
SCRIPT-DEVIANT CRITERIA [C]							
Unexpected complications	−0.34	1.05	−0.33	−0.52	−0.16	−3.79	0.001
Superfluous details	−0.96	0.97	−1.00	−1.13	−0.80	−11.59	0.001
Unusual details	−0.97	1.15	−0.85	−1.17	−0.78	−9.84	0.001
Related external associations	−0.27	1.07	−0.25	−0.45	−0.09	−2.91	0.002
Accurately reported details not comprehended	−0.33	1.06	−0.31	−0.51	−0.15	−3.66	0.001
STRATEGY-BASED CRITERIA [M]							
Spontaneous corrections	−0.92	1.09	−0.84	−1.10	−0.73	−9.76	0.001
Admitting lack of memory	−0.21	1.14	−0.19	−0.41	−0.02	−2.20	0.015
Efforts to remember	−0.85	1.08	−0.79	−1.03	−0.67	−9.20	0.001
Expressing uncertainty	−0.41	1.17	−0.35	−0.61	−0.21	−4.06	0.001
Reality controls	−1.02	1.09	−0.94	−1.21	−0.84	−10.91	0.001
Raising doubts about one's own testimony	−0.23	1.20	−0.19	−0.43	−0.03	−2.23	0.014
Raising doubt about one's own person	−1.24	0.96	−1.28	−1.40	−1.07	−14.92	0.001
Self-deprecation	0.36	1.18	0.31	0.16	0.56	3.57	0.001[A]
Pardoning the perpetrator	0.44	1.05	0.42	0.27	0.62	4.93	0.001[A]

*$N = 135$; $df = 134$.

LCL, Lower confidence limit; UCL, Upper confidence limit; [C], Cognitive criteria; [M], Motivational criteria.

Criteria in green font color received significant positive value ratings, while criteria in red font color symbolize significant negative ratings.

[A]*For results contradictory to our hypotheses the p-values from two-tailed tests were reported. Otherwise, the p-values were derived from one-tailed tests, since the testing hypotheses were one directional.*

As pointed out before, both memory-related (group 1) and script-deviant (group 2) criteria were originally classified as being cognitively-related. Interestingly, even though the newly-made distinction of memory-related vs. script-deviant criteria was originally deduced from considerations on the cognitive level—assuming that different cognitive processes underlie their production—, the hypothesized differences seem to translate to the motivational level as well. In this way, the obtained pattern of strategic value ratings clearly corroborated the utility of distinguishing between two groups of cognitive criteria.

Other than group 1 and group 2 criteria, strategy-based (group 3) criteria are classified as motivational criteria, which implies that their validity depends largely on the assumption that deceivers out of strategic reasons avoid producing them. While our results indeed showed consistent negative ratings for 7 of the 9 criteria (with 4 of them being rated significantly negative across all three studies; see **Table 5**), participants in our study attributed positive strategic meaning to the criteria *self-deprecation* and *pardoning the perpetrator*. Participants in the investigation of Niehaus et al. (2005) on the other hand had rated the same criteria significantly negative. Possibly, the discrepant findings between the two investigations might be attributable to their variations in context: While the story outline of the current study revolved around a rather ordinary every-day work situation, the context in Niehaus et al.'s (2005) study bore graver ramifications and entailed false allegations of sexual rape. Such relationships between context and valence of the rating would, in fact, correspond well with the proposition of Niehaus et al. (2005), predicting that the negative strategic meaning of *self-deprecation* and *pardoning the perpetrator* is dependent on sufficient contextual gravity to render elements related to self-criticism unconceivable. Within less severe contexts on the other hand (i.e., scenarios typically used in laboratory studies, such as accusations of minor theft or insurance fraud) laypersons would ascribe positive strategic value to these elements, believing to make them appear more amiable and trustworthy. It is not clear however whether the negative strategic ratings reported by Niehaus et al. (2005) primarily reflect the sexual connotation of the context or rather the severe ramifications associated with it, leaving open the question under which specific circumstances the criteria could be valid indicators for truthfulness. Furthermore, empirical findings from several laboratory studies appear to

TABLE 5 | Strategic value ratings as obtained from all three studies[d].

Autobiographic memory vs. script information		Strategic self-presentation
Criteria related to episodic autobiographical memory [C]	**Criteria related to script-deviant information [C]**	**Criteria related to efforts of positive strategic self-presentation [M]**
Information about everyday life routines[b,c]	Unexpected complications[b,c]	**Spontaneous corrections[a,b,c]**
Spatial information[a]	**Superfluous details [a,b,c]**	Admitting lack of memory[c]
Temporal information[b,c]	**Unusual details [a,b,c]**	**Efforts to remember[a,b,c]**
Description of interactions[a]	Related external associations[c I]	**Expressing uncertainty[a,b,c]**
Reproduction of conversations[a]	Accurately details not comprehended[c I]	Reality controls[a,c]
Emotions and feelings[a,b,c]		Raising doubt about one's own testimony[a,c]
Own thoughts[I]		**Raising doubts about one's own person[a,b,c]**
Sensory Impressions[I]		
Attribution of perpetrator's mental state		Self-deprecation[a,c]
Personal implications[a,b,c]		Pardoning the perpetrator[a,c]

[C], Cognitive criteria; [M], Motivational criteria.

[I]Investigated only in the current study.

$p < 0.05$: [a]study of Niehaus et al. (2005); [b]study of Niehaus (2008); [c]current study.

Criteria in green font color received significant positive ratings in at least one of the three studies (and no significant negative ratings in any of the other studies). Vice versa, criteria in red font color received significant negative ratings (and no significant positive ratings in any of the other studies). If a criterion received both significant positive and negative value ratings across different studies, a blue font color was applied. Additional bold font highlights that the criterion was rated significantly across all three studies.

*[d]For reasons of clarity, the structure presented in this article differs slightly from the version originally presented by Volbert and Steller (2014); see footnote 5 for details. Also, **Table 5** excludes criteria that were previously investigated by Niehaus et al. (2005) and Niehaus (2008), but that the authors of the revised model had allocated to the "statement as whole" category (quantity of details, unstructured production) or not adopted at all (justifying memory gaps/uncertainties, spontaneous clarifications).*

dispute their validity in indicating true testimony. Amado et al. (2016) for instance identified in their meta-analysis *self-deprecation* and *pardoning the perpetrator* as the only criteria that failed to discriminate between true and fabricated statements, while Vrij (2008) even found that in two out of ten studies, *self-deprecation* appeared significantly more often in fabricated than in true statements. Crucially though, the design and nature of laboratory studies typically vary in important aspects from forensic interrogation settings (Volbert and Steller, 2014), including but not limited to the gravity of the context in which the interview takes place (Burgoon, 2015). Inferring from these findings that *self-deprecation* and *pardoning the perpetrator* are by or in themselves unsuitable in indicating true testimony may therefore be premature. Instead, further investigation seems warranted to explicate the exact contingencies under which laypersons are inclined to avoid rather than promote their production.

At least to a weaker degree, if viewed collectively the studies' results may hint at additional criteria that in their strategic meaning are context-dependent. Concerning memory-related (group 1) criteria for instance, context-dependency may explain why across studies participants rated the first five criteria uniformly within the two studies that introduced a nearly identical story outline, but differently so in the investigation that referred to a scenario with considerably graver ramifications (Niehaus et al., 2005). From a purely theoretical perspective, it seems further possible that the strategic meaning of strategy-based (group 3) criteria may depend on the underlying context as well. For instance, some of these criteria pertain to contents that in true accounts reflect the expression of genuine mnemonic processes, such as *admitting lack of memory* or *efforts to remember*. From their own experience liars may be well aware that memories fade with time, and thus may ascribe rather positive strategic meaning to these contents when the event in question dates back enough years in time. Since all three studies introduced scenarios in which only brief time periods lay between statement and the event to be imagined, their paradigms would be unsuitable for detecting such forms of context-dependency. Future studies could examine this issue by implementing scenarios that differ in regards to the length of time that had passed between event and statement deliverance.

Limitations

Some important limitations of our study deserve attention. As we made use of a questionnaire, we cannot be sure whether participants correctly understood every example that we provided for illustrating the criteria. Furthermore, our conclusions about content-related deception strategies are based on averaged findings that may not apply equally well across individual cases. For instance, Niehaus (2008) found that the specific content-related deception strategies vary between different age groups, suggesting that the developmental stage of a person may mediate the strategic meaning he or she ascribes to a criterion[6]. Most importantly, we only investigated how

[6]For explanatory purposes, we tested whether the strategic value ratings in our study differed between students ($n = 66$) and working professionals ($n = 55$). While the difference in age between students ($M_{age} = 23.70$, $SD = 3.70$) and working professionals ($M_{age} = 35.00$, $SD = 11.98$) was significant, $M = -11.30$, $t_{(62.60)} = -6.73$, $p < 0.001$, none of the one-way ANOVAS conducted for each criterion yielded a significant difference in the value ratings between groups. Presumably, age differences in the assignment of strategic value to certain CBCA criteria are linked to different developmental stages (i.e., adults versus children; see Niehaus et al., 2005). Despite the difference in age, both groups in our study consisted of adults, rendering the existence of different developmental stages unlikely.

lay people rate the strategic meaning of the criteria in theory but did not test in which ways their content-related deception strategies translate to the practical level. Considering that the potential outcome for the liar can highly affect his or her behavior associated with deception (Porter and ten Brinke, 2010), it seems reasonable to assume that participants' hypothetical use of content criteria in fictitious scenarios may differ from their actual verbal performance in real-life forensic settings. Future research would first need to explore or even manipulate the deception strategies of participants (i.e., pointing out to them the strategic meaning of criteria from the perspective of forensic practitioners), and subsequently motivate participants to successfully deceive within an ecologically valid, high stakes interrogation setting[7]. Such an approach would allow examining the relationship between the strategic meaning that statement providers ascribe to a criterion and the criterion's subsequent occurrence in their fabricated statements.

Conclusions

The current study demonstrated that CBCA criteria differ in their strategic meaning and that the three-dimensional structure of the revised model of Volbert and Steller (2014) is suitable for representing these differences. Few exceptions (such as *self-deprecation* and *pardoning the perpetrator* carrying strategic value opposite to the valence of their respective group) notwithstanding, our group-based predictions regarding the strategic meaning of the criteria were largely confirmed. That is, laypersons tended to rather ascribe positive strategic meaning to criteria related to episodic autobiographical memory (group 1) but tended to ascribe negative strategic meaning to criteria related to script-deviant information (group 2) and efforts of strategic self-presentation (group 3).

In practical terms, our results then provide valuable input for forensic practitioners in appraising the diagnostic value of the criteria: The fact that deceivers typically intend to refrain from simulating script-deviant (group 2) or strategy-based (group 3) criteria strengthens their validity in indicating true statements. In contrast, no such avoidance inclinations are to be expected for memory-related (group 1) criteria, implying that the mere presence of these criteria does not automatically support a statement's truthfulness. Such generalized guidelines can only be of heuristic value however, since positive strategic meaning is rather a prerequisite than actual indication for a criterion's emergence– whether the criterion occurs in the fabricated statement then depends on the statement giver's ability to produce such content (primary vs. secondary deception; Köhnken, 1990). More elaborate assessments of a criterion's diagnostic value thus necessitate additional insight about the cognitive component; above all, knowledge about the cognitive difficulty associated with the criterion's production is required. Such insight combined with our established knowledge about the strategic meaning of the criteria would then constitute a solid foundation for optimally assessing their diagnostic value.

[7]For definitions of high stakes, experimentally realistic lie detection scenarios as well as their effects on lie detection accuracy see O'Sullivan et al. (2009).

DATA AVAILABILITY

The raw data supporting the conclusions of this manuscript will be made available by the authors, without undue reservation, to any qualified researcher.

ETHICS STATEMENT

This study was carried out in accordance with the Ethical Guidelines of the German Psychological Society (DGPs). All subjects gave informed consent in accordance with these guidelines. Ethical approval was not deemed necessary as there was no foreseeable risk of harm or discomfort for participants.

AUTHOR CONTRIBUTIONS

RV, SN, and SW: contributed to conception and design of the study; SW: carried out the experiment and organized data acquisition; BM and SW: performed the statistical analysis; BM: wrote the first draft of the manuscript; BM and RV: wrote sections of the manuscript; RV, BM, and SN: contributed to manuscript revision. All authors read and approved the submitted version.

FUNDING

The work of BM was individually supported by the Elsa-Neumann-Scholarship of Berlin. Otherwise, this research received no specific grant from any funding agency in the public, commercial, or non-profit sectors.

REFERENCES

Akehurst, L., Köhnken, G., and Höfer, E. (2001). Content credibility of accounts derived from live and video presentations. *Legal Criminol. Psychol.* 6, 65–83. doi: 10.1348/135532501168208

Amado, B. G., Arce, R., Fari, F., and Vilari, M. (2016). Criteria-based content Analysis (CBCA) reality criteria in adults: a meta-analytic review. *Int. J. Clin. Health Psychol.* 16, 201–210. doi: 10.1016/j.ijchp.2016.01.002

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B.*57, 289–300.

Bogaard, G., Meijer, E. H., Vrij, A., and Merckelbach, H. (2016). Strong, but wrong: lay people's and police officers' beliefs about verbal and nonverbal cues to deception. *PLoS ONE* 11:e0156615. doi: 10.1371/journal.pone.0156615

Burgoon, J. K. (2015). When is deceptive message production more effortful than truth-telling? A baker's dozen of moderators. *Front. Psychol.* 6:1965. doi: 10.3389/fpsyg.2015.01965

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences, 2nd Edn.* Hillsdale, NJ: Lawrence Erlbaum Associates.

DePaulo, B. M. (1992). Nonverbal behavior and self-presentation. *Psychol. Bull.* 111, 203–243. doi: 10.1037/0033-2909.111.2.203

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

Granhag, P. A., and Strömwall, L. A. (2000). Effects of preconceptions on deception detection and new answers to why lie-catchers often fail. *Psychol. Crime Law* 6, 197–218. doi: 10.1080/10683160008409804

Greuel, L., Offe, S., Fabian, A., Wetzels, P., Fabian, T., Offe, H., et al. (1998). *Glaubhaftigkeit der Zeugenaussage [Credibility of Witness Testimony]*. Weinheim: Psychologie Verlags Union.

Hartwig, M., Granhag, P. A., and Strömwall, L. A. (2007). Guilty and innocent suspects' strategies during police interrogations. *Psychol. Crime Law* 13, 213–227. doi: 10.1080/10683160600750264

Hommers, W. (1997). "Die aussagepsychologische Kriteriologie unter kovarianzstatistischer und psychometrischer Perspektive [CBCA criteria from the perspective of covariance statistics and psychometrics]," in *Psychologie der Zeugenaussage. Ergebnisse der Rechtspsychologischen Forschung*, eds L. Greuel, T. Fabian, and M. Stadler (Weinheim: Psychologie Verlags Union), 87–100.

Köhnken, G. (1990). *Glaubwürdigkeit. Untersuchungen Zu Einem Psychologischen Konstrukt [Credibility. Investigations About a Psychological Construct]*. Munich: Psychologie Verlags Union.

Narum, S. R. (2006). Beyond Bonferroni: Less conservative analyses for conservation genetics. *Conserv. Genet.* 7, 783–787. doi: 10.1007/s10592-005-9056-y

Niehaus, S. (2001). *Zur Anwendbarkeit inhaltlicher Glaubhaftigkeitsmerkmale bei Zeugenaussagen Unterschiedlichen Wahrheitsgehalts [Applicability of CBCA Criteria in Statements of Varying Truth Content]*. Frankfurt am Main: Peter Lang.

Niehaus, S. (2008). Täuschungsstrategien von kindern, jugendlichen und erwachsenen [deception strategies of children, adolescents and adults]. *Forensische Psychiatrie Psychol. Kriminol.* 2, 46–56. doi: 10.1007/s11757-008-0059-7

Niehaus, S., Krause, A., and Schmidke, J. (2005). Täuschungsstrategien bei der schilderung von sexualstraftaten [deception strategies when reporting sexual offences]. *Zeitschrift Für Sozialpsychol.* 36, 175–187. doi: 10.1024/0044-3514.36.4.175

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

O'Sullivan, M., Frank, M. G., Hurley, C. M., and Tiwana, J. (2009). Police lie detection accuracy: the effect of lie scenario. *Law Hum. Behav.* 33, 530–538. doi: 10.1007/s10979-009-9191-y

Porter, S., and ten Brinke, L. (2010). The truth about lies: what works in detecting high-stakes deception? *Legal Criminol. Psychol.* 15, 57–75. doi: 10.1348/135532509X433151

Ryan, S., Sherretts, N., Willmott, D., Mojtahedi, D., and Baughman, B. M. (2018). The missing link in training to detect deception and its implications for justice. *Safer Commun.* 17, 33–46. doi: 10.1108/SC-07-2017-0027

Schank, R. C., and Abelson, R. P. (1977). *Scripts, Plans, Goals and Understanding: An Inquiry into Human Knowledge Structures*. Oxford: Erlbaum.

Steller, M., and Köhnken, G. (1989). "Criteria-based statement analysis. credibility assessment of children's statements in sexual abuse cases," in *Psychological Methods for Investigation and Evidence*, ed D. C. Raskin (New York, NY: Springer), 217–245.

Undeutsch, U. (1967). "Beurteilung der glaubhaftigkeit von zeugenaussagen [Assessing the credibility of witnesses' testimony]," in *Handbuch der Rechtspsychologie*, Vol. 11, ed U. Undeutsch (Göttingen: Hogrefe), 26–181.

Volbert, R., and Steller, M. (2014). Is this testimony truthful, fabricated, or based on false memory? *Eur. Psychol.* 19, 207–220. doi: 10.1027/1016-9040/a000200

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

Vrij, A. (2008). *Detecting Lies and Deceit: Pitfalls and Opportunities, 2nd Edn*. New York, NY: Wiley.

Vrij, A., Akehurst, L., and Knight, S. (2006). Police officers', social workers', teachers' and the general public's beliefs about deception in children, adolescents and adults. *Legal Criminol. Psychol.* 11, 297–312. doi: 10.1348/135532505X60816

Welle, I., Berclaz, M., Lacasa, M. J., and Niveau, G. (2016). A call to improve the validity of criterion-based content analysis (CBCA): results from a field-based study including 60 children's statements of sexual abuse. *J. Forensic Legal Med.* 43, 111–119. doi: 10.1016/j.jflm.2016.08.001

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

APPENDIX 1: DISPLAY OF THE ENTIRE STORY AS PRESENTED TO PARTICIPANTS (TRANSLATED FROM GERMAN)

Dear participants,

we are interested in your strategies of how to make a fabricated story appear credible. There are no "wrong" answers; we will not assess your answers on moral grounds, nor do we want to know if you would pursue such strategies in real-life. Please read the instructions carefully and please answer all questions in their order of appearance. For your responses, please put yourself into the perspective outlined in the following story:

For 6 weeks you have been holding a job position you had desired before for a long time. However, within these first few weeks some things went wrong: Twice already you have shown up to work late. Your boss is quite irritated about this, and since you are still on probation you should by no means show up late a third time. Your boss made clear that if this should happen again, you will lose your position. Yet, it happens again: You are oversleeping in the morning! You hurry to the tram, where you happen to meet your neighbor Mr Schneider, who is quite an irritable fellow. Once he had even punctured the tires of your bike. Running into this guy was the last thing you needed now!

During the tram ride you are contemplating how you can explain yourself best to your boss, and looking at your maliciously smirking neighbor you come up with the following story:

"This morning, right before I was about to leave for work, I went down to the cellar to fetch a folder with work-related documents. They contain my personal notes that I intended to use for work. So, here is what happened: I was unlocking the door to the cellar, and as always, I left the key in the lock. Upon entering the cellar, I recognized some noise behind me. As I turned around, Mr. Schneider - my neighbor, who constantly causes trouble - was standing outside. He quickly moved forward and locked up the door to the cellar. I had been calling and pounding on the door for quite a while until finally an unfamiliar person showed up and unlocked the door. Luckily, the key was still left on the outside lock. The unfamiliar person moved on, and I quickly made my way to work."

After having told this story to your boss, your boss remarks: "Oh, Mr Schneider, I do know this guy! He works as doorman for the company of my wife! Maybe I should talk to him, what he just did is hardly acceptable." Then, your boss instructs you to tell him again what had happened, as precisely as possible. He lets you know that he will not talk to Mr Schneider if he finds your story set out convincingly. Otherwise however, in case he will have doubts about your story, he will talk to Mr Schneider for clarification. Clearly, it is crucial now that your boss believes your story. You certainly do not want to lose your job! You even had wrongly accused another person- Mr Schneider would, of course, deny the allegations, so your only hope is to convince your boss with your story. Otherwise, you will lose your job for sure.

Appendix B: Zusammenfassung in deutscher Sprache (Abstract in German language)

Der Katalog der merkmalsorientierte Inhaltsanalyse enthält in seiner originären Fassung 19 sogenannte Realkennzeichen, anhand derer der Wahrheitsgehalt von Zeugenaussagen substantiiert werden kann. Den Realkennzeichen liegt die Annahme zu Grunde, dass diese in erlebnisbasierten Aussagen quantitativ oder qualitativ stärker auftreten als in erfundenen Aussagen. Die vorliegende Dissertationsarbeit zielt darauf ab, in der Forschungsliteratur häufig kritisierte Aspekte der merkmalsorientierten Inhaltsanalyse zu adressieren, verbunden mit der Hoffnung, so praxisrelevante Implikationen zu gewinnen. Konkret beziehen sich die als Schwachstellen der merkmalsorientierten Inhaltsanalyse identifizierten Aspekte auf die mangelnde theoretische Fundierung der Realkennzeichen zur Bestimmung des Wahrheitsgehalts und auf das Fehlen einer systematisierten Gewichtungsstruktur, um die Realkennzeichen nach ihrer individuellen diagnostischen Bedeutsamkeit zu sortieren. Insgesamt setzt sich die vorliegende Arbeit aus drei unterschiedlichen Erhebungsstudien zusammen. **Studie I** wurde entwickelt, die Stichhaltigkeit der theoriegeleiteten Annahme, zwischen wahren und erfundenen Aussagen auftretende Qualitätsunterschiede seien vor allem auf kreative und strategische Prozessanforderungen zurückzuführen, näher zu eruieren. In einem Experiment wurde daher das Vorliegen der beiden Anforderungen systematisch manipuliert, um die Auswirkungen auf die inhaltliche Qualität (gemessen anhand von über die einzelnen Realkennzeichen hinweg gebildeten Summenwerte) der von den Probanden ($N = 30$) vorgebrachten Zeugenaussagen zu prüfen. Nach den Angaben der Probanden erhöhten zwar sowohl kreative als auch strategische Anforderungen die wahrgenommene kognitive Beanspruchung; anders als erwartet führte das zeitgleiche Evozieren der beiden Anforderungen bei erfundenen Aussagen aber zu einer höheren statt geringeren Aussagequalität. Der überraschende Befund legt nahe, dass Probanden unter bestimmten Umständen trotz erhöhter

kognitiver Beanspruchung in der Lage sind, erfundene Aussagen in weitaus höherer als gemeinhin angenommen Qualität zu produzieren. Allerdings scheinen weitere Untersuchungen angezeigt, da auf Grund methodischer Einschränkungen ein reines Vergleichen der Summenwerte zwischen den in Studie I vorliegenden Experimentalbedingungen keinen „objektiven" Aufschluss über die tatsächliche Güte der erzielten Aussagequalität liefern kann. Die beiden weiteren Erhebungsstudien gründen sich auf einen von Volbert und Steller (2014) überarbeiteten Merkmalskatalog. Dieser ähnelt der originären Fassung hinsichtlich der enthaltenen Realkennzeichen weitestgehend, fasst die Inhaltsmerkmale aber anhand theoriegeleiteter bzw. prozessbezogener Überlegungen in die drei Merkmalsgruppen *episodische Erinnerung* (Gruppe 1), *Schemaabweichung* (Gruppe 2) und *fehlende strategische Selbstpräsentation* (Gruppe 3) zusammen. Auf Grundlage dieser Strukturierung sollte untersucht werden, inwieweit die motivationalen (Studie II) und kognitiven (Studie III) Eigenschaften der Merkmale innerhalb der Gruppen homogen ausfallen. **Studie II** setzte den – in der Aussagepsychologie nicht geschulten – Probanden (N = 135) ein fiktives Fallszenario vor, um auf diese Weise zu untersuchen, welche strategische Bedeutung den Inhaltsmerkmalen im Täuschungskontext zugeschrieben wird. Insgesamt zeigte sich, dass die Merkmale der *episodischen Erinnerung* überwiegend als positiv bzw. täuschungsförderlich bewertet wurden, während die Merkmale aus den beiden anderen Gruppen bis auf wenige Ausnahmen als negativ bzw. täuschungshinderlich eingeschätzt wurden. Somit erlaubt die überarbeitete Merkmalsstrukturierung, auf motivationaler Ebene für jede der Merkmalsgruppen diagnostisch bedeutsame Erwartungen hinsichtlich ihrer Auftretenswahrscheinlichkeit zu formulieren: Nur die der Gruppe *episodische Erinnerung* zugeordneten Merkmale wären demnach grundsätzlich auch in einer Falschaussage zu erwarten; folglich sind diese den Merkmalen aus den beiden anderen Gruppen (S*chemaabweichung* bzw. *fehlenden strategischen Selbstpräsentation*) diagnostisch unterlegen. **Studie III** schließlich adressierte in Hinblick auf das Produzieren der

Merkmale nicht die Frage des Wollens (motivationale Ebene), sondern die Frage des Könnens (kognitive Ebene). Ein Teil der Probanden wurde daher über die positive bzw. täuschungsförderliche Bedeutung der Merkmale aus den Gruppen der *Schemaabweichung* oder der *fehlenden strategischen Selbstpräsentation* aufgeklärt. Anschließend wurde untersucht, inwieweit die Probanden (N = 60) in der Lage waren, die aufgeklärten Merkmale in ihre Aussagen zu integrieren. Im Vergleich zu den erfundenen Aussagen der aus unaufgeklärten Probanden bestehenden Kontrollgruppe ließen sich keine Effekte auf die Realkennzeichen-Summenwerte feststellen, unabhängig davon, ob über Merkmale der *Schemaabweichung* oder über Merkmale der *fehlenden strategischen Selbstpräsentation* aufgeklärt wurde. Wurde die Aussagequalität hingegen anhand aufklärungsbezogener Skalenwerte (statt der Summenwerte) gemessen, so deuteten die Ergebnisse in Teilen auf eine möglicherweise leichtere Simulierbarkeit der letztgenannten Merkmalsgruppe hin. Zusammenfassend kann konstatiert werden, dass durch die vorliegende Arbeit eindeutige und diagnostisch wertvolle Erkenntnisse über die motivationalen Eigenschaften der einzelnen Realkennzeichen bzw. der übergeordneten Merkmalsgruppen gewonnen wurden (Studie II). Die Untersuchung von den Täuschungsbemühungen mutmaßlich unterliegenden Prozessanforderungen (Studie I) sowie die Prüfung der kognitiven Eigenschaften der Merkmalsgruppen (Studie III) lieferte erste und teils vielversprechende Ergebnisse, die jedoch der weiteren Abklärung im Rahmen zukünftiger Studien bedürfen.

Appendix C: Eigenständigkeitserklärung (Declaration of Authenticity)

Hiermit erkläre ich, Benjamin Maier, dass ich die vorliegende Dissertation selbstständig verfasst und ohne unerlaubte Hilfe angefertigt habe.
Alle Hilfsmittel, die verwendet wurden, habe ich angegeben. Die Dissertation ist in keinem früheren Promotionsverfahren angenommen oder abgelehnt worden.

Weiter versichere ich, dass ich alle wörtlichen und sinngemäßen Übernahmen aus Quellen und anderen Werken nach bestem Gewissen also solche gekennzeichnet sowie vollständig aufgeführt habe.

Ort, Datum

Eigenhändige Unterschrift
Benjamin Maier

Bezüglich der gemeinsam mit Koautorinnen erstellten Publikation, die Teil der vorliegenden Dissertation ist (Studie II bzw. Maier et al., 2018), lassen sich die Anteile wie folgt quantifizieren: Renate Volbert, Susanna Niehaus und Sina Wachholz erstellten das Konzept und Design der Untersuchung; Sina Wachholz organisierte die Datenerhebung und führte die Untersuchung durch; Benjamin Maier und Sina Wachholz führten die statistische Auswertung durch; Benjamin Maier verfasste das Manuskript; Benjamin Maier, Renate Volbert und Susanne Niehaus beteiligten sich an den Überarbeitungen des Manuskripts.