Of Liars and Deniers

The Credibility-Concealed Information Test as a

Response Time-Based Measure for Credibility Assessment in

Suspects and Witnesses

Kathrin Eickmeier

Of Liars and Deniers

The Credibility-Concealed Information Test as a Response Time-Based Measure for Credibility Assessment in Suspects and Witnesses

Inaugural-Dissertation
zur Erlangung der Doktorwürde
der
Philosophischen Fakultät
der
Rheinischen Friedrich-Wilhelms-Universität
zu Bonn

vorgelegt von
Kathrin Eickmeier

aus
Oberhausen

Bonn 2020

Gedruckt mit Genehmigung der Philosophischen Fakultät
der Rheinischen Friedrich-Wilhelms-Universität Bonn.

Zusammensetzung der Prüfungskommission:

PD Dr. Ina Grau
(Vorsitzende)

Prof. Dr. Rainer Banse
(Betreuer und Gutachter)

Prof. Dr. Henning Gibbons
(Gutachter)

PD Dr. Bernd Schlöder
(weiteres prüfungsberechtigtes Mitglied)

Tag der mündlichen Prüfung: 16.01.2020

Acknowledgements

The time I worked on this research project was an inspiring period filled by intense learning not only in the scientific field but also on a personal level. I want to thank Prof. Dr. Rainer Banse for offering me this opportunity, his valuable guidance, his meaningful suggestions, and his critical comments. Moreover, I am grateful for Dr. Alexander Schmidt, Jelena Rönspies, Verena Oberlader, and the entire team of the Department of Social and Legal Psychology who supported me with scientific discussions and encouraging words. The three years of working in the department were a great pleasure. Many thanks to my family and friends – especially to Elke, Achim, Fabian, Calin, Esther, and Simon – for their amazing sense of humor and enthusiasm even during difficult times. Last but not least, I want to thank the numerous students that were willing to act as participants for this research project as well as the cooperating schools and teachers.

Preface

The experimental procedures of this research project were approved by the ethics committee of the University of Bonn. Study 1, Study 2, and Study 3 were conducted in cooperation with Jelena Rönspies (Department of Social and Legal Psychology, University of Bonn). Jelena analyzed the participants' testimonies for content characteristics (reported in Rönspies, in prep.) but was not involved in the development of the Credibility-Concealed Information Test (CIT). The following bachelor's and master's students at the University of Bonn assisted in data acquisition: For Study 1, students of the M.Sc. Psychology course "Projektarbeit" directed the participants to fabricate stories or perform certain behavioral tasks and instructed them to respond to the CIT. For Study 2, Nina Schirrmacher, Tatjana Siepe, and Katrin Pucker directed the participants to fabricate stories or perform the behavioral tasks and instructed them to complete both the CIT and an IQ test. For Study 3, Selina Görner, Leona Landwehr, Johanna Lübken, and Anne Wertenbruch directed the participants to fabricate stories or perform the required tasks, interviewed the subjects, and instructed the participants to complete the CIT as well as an IQ test. Study 4 was conducted in cooperation with Robert Schnuerch and the Department of Motivation, Emotion, and Learning (University of Bonn). They contributed by providing the EEG. Multiple students assisted in data collection for Study 4. Eva Bux, Nawael Niazy, and Jessica Tepel directed the participants to invent and fulfill the required tasks and interviewed the subjects; Judith Ratayczak and Rafaela Warkentin prepared the EEG and instructed the participants to complete the CIT as well as the personality trait scales and cognitive function tests. I would like to thank Jelena, Robert, and the above-named students most warmly for their support and enthusiasm.

Abstract

There are numerous studies on polygraphs like the Concealed Information Test (CIT) that aim to detect whether a suspect possesses crime-related knowledge. In contrast, not a single psychophysiological or behavioral paradigm targets the credibility of a victim's accusation. For that purpose, a Credibility-CIT was developed in four studies. Participants experienced certain situations (truth tellers) or invented a story about these events (liars). Subsequently, the participants were interviewed about the alleged occurrences. The methodological innovation is that the given testimony served as an individual database for the items used in the Credibility-CIT: Short phrases that either referred to a lie or a true statement were selected from each participant's testimony. These idiographic items, called credibility probe items, were presented in the Credibility-CIT amongst several distractor items while response times and error rates were recorded.

Results of three studies indicate that liars respond more slowly than truth tellers to credibility probe items ($d_{\text{Study 1}} = 0.55$, $d_{\text{Study 2}} = 0.42$, $d_{\text{Study 4}} = 0.47$). Different encoding processes and emotional-motivational factors that induce varying orienting responses provide an explanation for the response time differences. In contrast, Study 3 did not indicate significant differences between truth tellers and liars. Methodological issues – especially regarding the selected response deadline – are discussed. In addition, EEG data were recorded in Study 4 but event-related potentials did not discriminate truth tellers from liars. Lastly, a meta-analysis was conducted over the four studies included in this research project. An overall effect size of $g = 0.40$ was observed for the Credibility-CIT's efficiency to discriminate truth tellers from liars based on response times. The results indicate that the Credibility-CIT has a meaningful potential as an indirect measure for credibility assessment.

Keywords: Credibility Assessment, Concealed Information Test, CIT, memory detection

Contents

Introduction

Liars and deniers are the two faces of deception. Both know the truth but claim something else – what they differ in is the content of this claim. When focusing on forensic contexts, liars simulate episodic memories about a crime and actively invent a story about that offense (i.e., an alleged victim/witness giving a false incrimination, or a suspect giving a false confession). In contrast, deniers negate or dissimulate any knowledge about an offense that they actually attended (i.e., a factual culprit, or a witness who protects the offender). Detecting any kind of deception is probably one of the oldest aims in civilized societies and of particular importance in forensic settings.

There is plenty of research on numerous "lie detector" paradigms – some corroborating and others challenging their respective validity (Lykken, 1998). What all "lie detectors" have in common is that the respondent is the suspect of a crime (i.e., a denier). Psychophysiological methods that scrutinize the credibility of a witnesses' statement (i.e., liars) do not exist yet; for this context of credibility assessment, only content-related approaches are applied. Even though lie detection and credibility assessment are methodologically different, they are not incompatible. Both approaches rely on a cognitive theory of deception and emphasize the role of memory processes and information processing in deception detection. This focus on cognition is also the reason why the term "lie detection" was replaced with the more accurate label "memory detection" (Verschuere & Meijer, 2014). Based on extensive work on memory-detection paradigms identifying deniers, this research project aims to develop a memory-detection paradigm for credibility assessment – that is, a response time-based procedure to debunk liars.

Credibility Assessment

In forensic contexts, a witness's – and particularly a victim's – statement about an offense constitutes important evidence. When there is no material evidence (like DNA material or similar), testimonies are often the only proof available. The Statement Validity Analysis (SVA; Köhnken & Steller, 1988) is the most established credibility assessment procedure in Germany and some other Western countries (Vrij, 2005). The SVA is a verbal approach that focuses on a testimony's content. The SVA's core component is the Criteria-Based Content Analysis (CBCA). The CBCA is based on the assumption that true statements have a higher quality than fabricated accounts (Undeutsch, 1989). As initially pointed out, truth tellers – in contrast to liars – can rely on episodic memory. Hence, true accounts should, for example, include more sensory information, unusual details, description of affective states, and contextual embedding (Volbert & Steller, 2014). Statements are analyzed for these "reality criteria". A high quantity and especially quality of these content characteristics are considered to be indicative of the credibility of a statement, but the absence of the characteristics does not indicate lying. Due to that, Rassin (2000) criticizes the CBCA as suffering from a "truth bias". Based on this focus on indicators of episodic memory, the SVA can, at its foundation, be classified as a memory-detection paradigm.

Cognitive Load Theory

A cognitive approach that stresses the relevance of memory processes builds the theoretical framework of credibility assessment. According to cognitive load theory, lying is more cognitively demanding than truth telling as it requires multitasking from the liar (Volbert & Steller, 2014; Vrij, Fisher, Mann, & Leal, 2006). Liars have to engage in several tasks simultaneously (Volbert & Steller, 2014): They have to fabricate a story solely based on cognitive scripts since they cannot rely on episodic memory about the incriminated event. Additionally, liars must avoid misinformation that can be revealed through a criminal investigation. Questions by the interviewer must be answered spontaneously and answers have to logically fit in the account given so far. If interrogated before, liars must also keep their previous statements in mind to avoid contradictions. These processes account for the high cognitive load of lying. Moreover, liars have to engage in strategic self-presentation in order to convey a trustworthy impression. Although this impression management is mainly determined motivationally, it still demands cognitive and especially working memory resources (Sporer, 2016).

Empirical evidence supports the cognitive load theory by indicating an influence of cognitive load on the form and content of lies. First and foremost, several studies and also meta-analyses indicate that lies have a lower quality as compared to recapitulating experiences (i.e., lies are less detailed and less embedded in a context; Volbert & Steller, 2014; for meta-analytical results see Amado, Arce, Fariña, & Vilariño, 2016; Oberlader et al., 2016). Moreover, the cognitive demand of lying is also shown on a neurological level. Prefrontal brain areas that are associated with executive control are activated during lying (Christ, Van Essen, Watson, Brubaker, & McDermott, 2009; Gamer, 2011). Additionally, the cognitive demand is reflected in behavioral changes. Since constructing the lie takes time and resources, responding deceitfully takes longer than responding truthfully in some reaction time paradigms (Suchotzki, Verschuere, Van Bockstaele, Ben-Shakhar, & Crombez, 2017).

Validity

A recent meta-analysis indicates the validity of the CBCA to discriminate liars from truth tellers with a large mean effect size of $g = 0.97$ (Oberlader et al., 2016). However, keeping in mind that the SVA is the only reliable method when there is no objective evidence, improving its detection accuracy seems highly desirable. Following a multi-method approach and by that dealing with the CBCA's truth bias (Rassin, 2000) would enhance the accuracy of court decisions. Until now, no complementary paradigms focusing on aspects other than content (e.g., psychophysiological or behavioral measures) have been developed for credibility assessment. From a theoretical perspective, the different memory processes underlying truth telling and lying (based on episodic memory vs. cognitive scripts) could be useful for diagnostic purposes. Hence, a latency-based memory-detection paradigm might be a promising new approach.

Response Time (RT) Approaches to Discriminate Truth Tellers from Liars

As already briefly mentioned, there is evidence that response time (RT) is a cue to deception (Walczyk, Roper, Seemann, & Humphrey, 2003). Responding deceptively to either a verbal question or a written stimulus takes longer than answering truthfully (Debey, De Houwer, & Verschuere, 2014). This finding led Walczyk and colleagues to develop the Activation-Decision-Construction-Action Theory (ADCAT; Walczyk, Harris, Duck, & Mulay, 2014; for a previous version of the model see Walczyk, Mahoney, Doverspike, & Griffith-Ross, 2009; Walczyk et al., 2005, 2003). This model explains the RT increase of deceptively answering questions in four (mostly) consecutive steps.

In the first step – called the activation component – the to-be answered question is encoded. Episodes or semantic memory content of the truth are activated in long-term memory. The truth is then automatically retrieved to working memory (Baddeley, 1992, 2000; Baddeley & Hitch, 1974; Sporer, 2016). Not only during truth telling but also during lying, the truth is usually activated and enters consciousness (i.e., working memory) in a first step. Other authors termed this assumption of the prepotent truth response "truth default theory" (Debey, De Houwer, & Verschuere, 2014; Verschuere & Shalvi, 2014).

According to ADCAT, in a second step (the decision component), the respondent decides to either lie or tell the truth. For evaluating the benefits and costs of truth telling motivational processes play a central role. This decision and evaluation process adds to RTs for liars.

In the third step (the construction component), the lie is invented. The truth is still activated in working memory and serves as a retrieval cue for lie construction. Liars can reduce the cognitive load by keeping the truth active and inventing an answer as close to this truth as possible (e.g., simply omitting or denying information). Inventing complex lies imposes a much higher load. For complex lies, semantic memory (scripts, schemata) and episodic memory (the truth and similar events) aids the construction of a convincing, consistent, and plausible answer (Sporer, 2016). This lie construction process will add to RT, too.

The last step concerns the verbal or habitual delivery of the lie (the action component). As the truth is the normative answer and active in working memory (see step 1), this honest response must be inhibited by the central executive. Moreover, liars engage in monitoring their behavior to appear credible. This also demands cognitive resources. Additionally, Walczyk et al. (2014) integrate the motivation to lie as an essential moderator in the ADCAT. They conceptualize motivation as the amount of cognitive resources the respondent is willing to invest in lying.

In order to test the ADCAT, Walczyk et al. conducted several studies that measured RTs between a question and the given answer using a voice key. Time between the last word of a question and the first utterance of the respondent's answer served as dependent variable (Walczyk et al., 2003, 2005, 2009). The studies indicated a trend that liars respond more slowly than truth tellers. However, results were inconsistent and the authors' explanations for insignificant findings not convincing. Moreover, meta-analytical results indicate that the latency between a question and the given answer is an invalid cue to deception (d = 0.02; DePaulo et al., 2003). On the other hand, studies using paradigms other than the voice key support the ADCAT's assumptions. Suchotzki et al. (2017) argue that it is not the RT effect itself that is invalid, but rather there are methodological issues in Walczyk et al.'s studies that induce the invalid RT results. Suchotzki et al. point out that for reliable RT-based memory detection paradigms, a computerized measurement, an instruction for fast responding, and a large number of valid trials are necessary. In their meta-analysis (Suchotzki et al., 2017), four paradigms met these criteria and proved to reliably discriminate truth and deceit: The autobiographical Implicit Association Test (Sartori, Agosta, Zogmaister, Ferrara, & Castiello, 2008), the Sheffield Lie Test (Spence et al., 2001), the Differentiation of Deception Paradigm (Furedy, Davis, & Gurevich, 1988), and the Concealed Information Test (Lykken, 1960). However, these paradigms have only been used in classic contexts of memory detection, mainly differentiating between a suspect of a crime (denier) and innocent subjects. None has ever been applied for credibility assessment, discriminating liars from truth tellers.

The Concealed Information Test (CIT)

The most theoretically founded memory detection approach is the Concealed Information Test (CIT; Lykken, 1960). The CIT is a computer-based paradigm that confronts a suspect with crime-related items while psychophysiological, neuronal, or behavioral measures are recorded (e.g., heart rate, skin conductance, event-related potentials [ERPs], RTs). Since suspects typically negate any offense-related knowledge, the classic CIT intends to distinguish deniers from innocents.

In the CIT, participants are confronted with several multiple-choice questions regarding previously undisclosed details of a certain crime (Lykken, 1959). For example, one question could concern the murder weapon ("What did the murderer use to bludgeon the victim's head?"). The question is then followed by the presentation of several potential answers (e.g., "vase, candleholder, rifle, cane, ashtray"), one of which corresponds to the actual fact. The basic rationale is that only guilty participants should recognize the correct response option (i.e., the actual weapon that was used to commit the murder) and show an orienting response. Therefore, their physiological response following the presentation of this option should be different from their response to the other options. This is traditionally assessed using peripheral measures of autonomous nervous activity (Ben-Shakhar & Elaad, 2003).

More recent laboratory versions of the CIT often involve the use of RTs and ERPs (see, e.g., Farwell & Donchin, 1991). In these variants, several words are presented consecutively on a screen. Among these are a few crime-related items (called *probe items*; e.g., considering the murder weapon: "knife") and a large amount of distractor items (called *irrelevant items*; e.g., "gun", "rope", "hammer", "bat"). Exactly as in the original version of the CIT, it is assumed that the probe items should stand out and trigger a special response only in culprits. In contrast, probes should not be distinguishable from irrelevant items to innocent participants. In order to warrant the participant's attention, a third item category is added, called *target items*. Target items have to be memorized before the beginning of the test. During the CIT, the question "Do you know this item?" is asked. Participants must tell whether the presented item is a memorized target item or not. Hence, target items must be confirmed (verbally or via button press: "Yes, I know this item"), but probe and irrelevant items have to be denied ("No, I don't know this item").

Crucial to the CIT is that probes fall in the familiar category for only guilty participants. As a consequence, accidentally confirming knowledge of the critical probe items (e.g., the murder weapon) would immediately expose the participant as culprit. Hence, guilty participants have to deal with the conflict between the familiarity of the probe stimulus and the urgent need to respond negating ("I don't know this item"). This stimulus-response incompatibility is central to the accuracy of the CIT and a lack of stimulus-response incompatibility is a large disadvantage of several other RT paradigms (De Houwer, 2008; Suchotzki, Verschuere, Crombez, & De Houwer, 2013; Verschuere, Crombez, Degrootte, & Rosseel, 2010). This significance of probe items is also referred to as oddball paradigm (Farwell & Donchin, 1991).

Orienting Response Theory

The orienting response theory builds the theoretical framework for the CIT. An orienting response (Sokolov, 1963) is elicited whenever a stimulus is novel, has a certain significance (i.e., has a signal value), or whenever changes in stimulation occur. In contrast, stimuli without significance result in habituation (Sokolov, 1963; Verschuere & Ben-Shakhar, 2011; Verschuere, Crombez, De Clercq, & Koster, 2004). Presentation of probe items in a CIT induces effects that share several characteristics with orienting responses. On a physiological level, orienting towards probe items leads to increased skin conductance (Ben-Shakhar & Elaad, 2003), heart rate deceleration (Verschuere et al., 2004), respiratory suppression (Gamer, 2011), and pupil dilatation (Lubow & Fein, 1996) – autonomic reactions that are also related to orienting responses. Neuronal activation of large amplitudes is elicited by familiar and meaningful stimuli (Rosenfeld, 2011) and therefore indicative of orienting responses. Specifically, probe presentation induces large P300 amplitudes in guilty participants (Gamer, 2011; Gamer, Klimecki, Bauermann, Stoeter, & Vossel, 2012). Moreover, an activation of the inferior frontal gyrus supports this theory as this brain region is active when unexpected events (like probe presentation) occur (Gamer, 2011). On a behavioral level, RTs and error rates increase when novel stimuli or probes are shown (Seymour & Kerlin, 2008; Seymour, Seifert, Shafto, & Mosmann, 2000; Verschuere et al., 2010).

Emotional-motivational factors. Notably, emotional-motivational factors can be integrated in this cognitive theory as they may influence the probe items' significance (Elaad & Ben-Shakhar, 1989; Verschuere & Ben-Shakhar, 2011). A high motivation to avoid being detected as a denier leads to a higher noteworthiness of the relevant probe items. In turn, this enhanced signal value of the probe items strengthens the orienting response and, hence, the CIT effect (Elaad & Ben-Shakhar, 1989).

Response Inhibition

Besides orienting towards the salient probe items, response inhibition plays a crucial role in concealing information. As already mentioned, the prepotent truth response has to be inhibited when lying (see ADCAT and truth default theory). The influence of inhibition is observable when the response conflict is removed from the CIT task (i.e., when participants admit knowledge about the critical probe items). In those overt deception tasks the RT-based CIT effect diminishes (Meijer, Smulders, Merckelbach, & Wolf, 2007; Suchotzki, Verschuere, Peth, Crombez, & Gamer, 2015). In contrast to RTs, two meta-analyses indicated that skin conductance does not depend on actively deceiving and, hence, on inhibiting the true response (Ben-Shakhar & Elaad, 2003; Meijer, Selle, Elber, & Ben-Shakhar, 2014). Conclusively, both, orienting response and response inhibition are relevant for the CIT effect with skin conductance reflecting orienting responses and RTs reflecting the role of inhibition processes.

Validity

A meta-analysis indicated that the CIT using skin conductance yields an effect size of d = 1.55 for discriminating deniers (guilty participants) from innocents (Ben-Shakhar & Elaad, 2003). Systematically comparing different CIT measures, RTs performed best (*d* = 1.97), followed by skin conductance (*d* = 1.46), heart rate (*d* = 1.07), and respiration (*d* = 0.85) (Verschuere et al., 2010). A threat to the CIT's validity is the vulnerability to countermeasures or faking (for a review see Ben-Shakhar, 2011). However, the RT-based CIT is less prone to countermeasures than physiological CITs, especially when a response deadline is added forcing the participant to react as quickly as possible (e.g., in less than 1,000 ms per item; Seymour et al., 2000). This speeded version of the RT-CIT seems to be a promising and thereby economic paradigm for memory detection. Besides its psychometric properties and practicability, the CIT is a theoretically founded indirect measure.

Fields of application. As initially pointed out, the CIT's classic field of application is the detection of culprits denying knowledge about a crime. Beyond the large number of laboratory studies, the CIT is practically applied on a daily base in Japan with about 5,000 examinations annually (Matsuda, Nittono, & Allen, 2012; Osugi, 2011). Moreover, the CIT can detect not only knowledge about crimes committed in the past, but also criminal intent (Meijer, Verschuere, & Merckelbach, 2010). Hence, the CIT's efficiency does not rely on actually enacting with the crime-related objects that later serve as probe items; rather, mentally dealing with these issues is sufficient to induce an orienting response. In other words, the CIT is sensitive to both, episodic and semantic memory content.

Additionally, the CIT can be applied not only in forensic settings, but also in clinical populations to examine memory processes in patients with memory deficits and to test for malingering (Allen, 2011). For example, the CIT is used with patients who have prosopagnosia or dissociative identity disorder. Patients with prosopagnosia are incapable of recognizing faces and are, in the CIT, confronted with pictures of familiar and unfamiliar faces (Bauer, 1984).

Regarding dissociative identity disorder, the focus lies on whether different "identities" can recall information from another ("inter-identity amnesia"; Allen & Movius, 2000). These applications of the CIT have in common that patients react systematically on probe items that they claim not to recognize explicitly (e.g., faces, knowledge of concurrent identities). The central and unresolved question is whether the prosopagnostic and dissociative identity patients faked their memory deficits or whether the CIT is actually capable of measuring implicit memory of which participants are unaware (Allen, 2011). At least for the prosopagnostic patients, malingering is rather unlikely as this deficit usually causes many complications in social life while benefits are negligible. However, it remains unclear whether the CIT is sensitive to implicit memory (Allen, 2011).

Overall, it is apparent that the CIT is sensitive to episodic, semantic, and possibly even to implicit memory. It is crucial that at least *some* knowledge about the subject matter exists. It seems that it is not so much the memory system itself, but rather the process of forming the memory content that should be consulted to complement the CIT's theoretical framework. The levels of processing theory ideally fits this demand.

Levels of Processing Theory

The levels of processing theory (Craik & Lockhart, 1972) integrates the processes of encoding and storage of information in memory. In contrast to the classic multicomponent models on memory (Atkinson & Shiffrin, 1968; Baddeley & Hitch, 1974), the levels of processing theory argues that semantic, visual, and acoustic stimuli are processed and stored similarly. The durability of a memory trace is rather a function of depth of encoding. Depth of encoding was initially considered to be a continuum of perceptual sensory processing (shallow) to semantic operations (deep). Later work (Craik & Tulving, 1975) emphasized that "depth" is an inadequate metaphor since processing is not sequential from perceptual to semantic. Rather, the terms “spread” or “elaboration” conceptualize the process as flexible. During elaborative processing associations between new information and information stored in memory are built. Hence, the levels of processing theory postulates that a recourse on memory is necessary for good retention. Elaborately encoded information will be maintained better and can be retrieved more easily (Craik & Tulving, 1975). In contrast, maintenance rehearsal (repeating information, learning by heart) only occurs at one processing level and, thus, represents shallow encoding that does not facilitate retention (Craik & Lockhart, 1972).

Levels of Processing and CIT

As initially outlined, unless participants have encoded the (crime-related) information, an item cannot gain the significance it needs to induce an orienting response in the CIT. Besides the general need for basic encoding, studies indicate that even the elaboration of encoding has an impact on the CIT effect. Carmel, Dayan, Naveh, Raveh, and Ben-Shakhar (2003) instructed participants to commit a mock crime. Half of them incidentally processed the relevant details while committing the crime (shallow encoding); the other half were instructed to pay close attention to certain details while committing the crime (elaborate encoding). During the CIT, participants intentionally memorizing and elaborately encoding the details had a higher skin conductance response to the relevant probe items than participants who encoded the information incidentally did.

A study by Seymour and Fraynt (2009) manipulated time and encoding effects using an RT-CIT. For the purpose of encoding, all participants were given a list of the crime-related probe items. Afterwards, half of the participants read an article that contained the same probe items (shallow encoding condition), while the other half had to complete four tasks involving these probe items (picture matching, word jumble, hand writing, and word shouting; deep encoding condition). The RT-CIT was conducted after a delay of 10 minutes, 24 hours, or 1 week. Results indicated that deeply (or better, elaborately) encoded items induce a larger RT-CIT effect than shallowly encoded items do (Figure 1). Seymour and Fraynt (2009) refer to the levels of processing theory for this effect, arguing that elaborately processed stimuli are better recalled and, hence, elicit stronger responses.

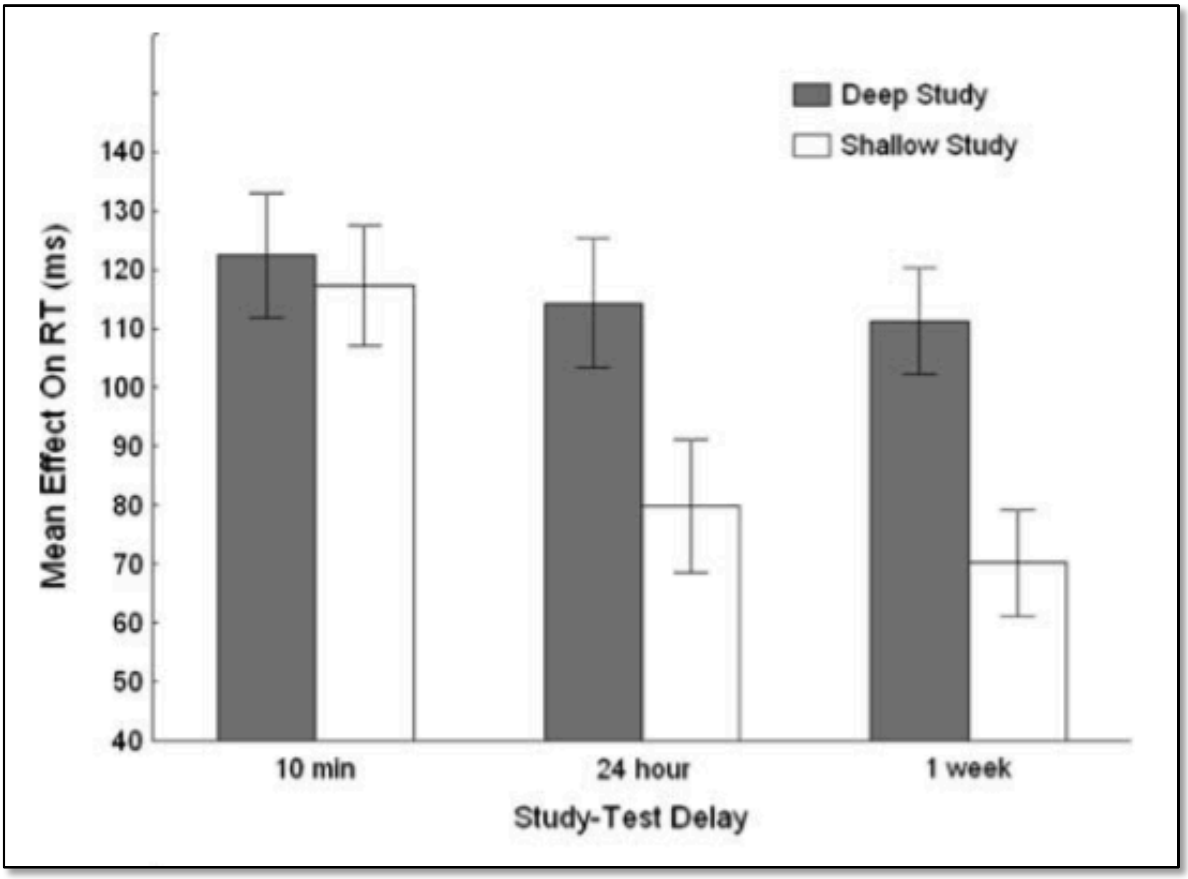

Figure 1. Time and encoding effects on the RT-CIT effect (RT on probes – RT on irrelevants) in the study by Seymour and Fraynt (2009). Deeply encoded information induced a stronger RT-CIT effect than shallowly studied information. Error bars represent +/- 1 SEM. Figure adopted from Seymour and Fraynt (2009).

Since elaborately encoded stimuli are generally better recalled (Craik & Lockhart, 1972), it remains unclear whether the CIT effects in the Seymour and Fraynt study rely only on better recognition in the deep encoding condition rather than on the impact of the encoding process itself. This relevance of recall effects for the CIT effect is indicated in a study by Gamer, Kosiol, and Vossel (2010): Participants had to commit a mock crime in which central details had to be processed to successfully commit the crime (e.g., memorizing a certain password). Encoding of peripheral details (i.e., the color of the stolen object) happened rather incidentally and was not guaranteed. Results indicated that elaborately encoded central stimuli were better recalled than peripheral details. The CIT was conducted immediately or two weeks after the mock crime and indicated mixed results. Heart rates on peripheral details remained stable over time but increased for central details. In contrast, skin conductance and respiratory data did not differ between the elaborately and shallowly encoded stimuli.

It is apparent that the elaborateness of encoding and its associated recall effect have an impact on the CIT's efficiency. In addition to these studies that systematically manipulate encoding strategies, other studies compare participants who differ in the acquisition of task-relevant information. Those studies compare guilty participants with informed innocents or participants who intend to commit a mock crime. Guilty participants actually commit a mock crime and, hence, gain crime-related information by interacting with the objects (seeing, touching, smelling etc.). Informed innocent participants read or hear basic information about a crime but do not handle or operate with the crime-related objects. Their role is comparable to innocent persons informed by the media. In studies on intended crimes, participants are asked to plan but not to complete a mock crime. Hence, participants in the informed, the intended, and the guilty group have knowledge about the event and, thus, act as different types of deniers in the CIT. Crucially, the three groups deal with the crime-related information differently and thus vary in their levels of processing or elaboration of encoding. Participants enacting the mock crime (i.e., guilty participants) collect the crime-related information incidentally, encode it primarily at a sensory level, and embed it spatiotemporally; semantic encoding plays a minor role. In contrast, participants in informed innocent groups are solely informed about crime-related facts and encode incidentally but also semantically. Participants who intend a mock crime use their long-term memory to mentally complement the event (using scripts or schemata) and generate a mental image of the event. Hence, they encode the information mainly semantically, that is, more elaborately (Craik & Tulving, 1975).

It remains inconsistent whether the CIT can discriminate between participants in the informed, the intended, and the guilty group and, hence, if the CIT is sensitive to different levels of processing (Ben-Shakhar, Gronau, & Elaad, 1999; Bradley, MacLaren, & Carle, 1996; Elaad, 2009; Gamer, 2010; Gamer et al., 2010; Meijer et al., 2010; Meixner & Rosenfeld, 2014; Suchotzki et al., 2013; Winograd & Rosenfeld, 2014). Several authors argue that discrepancies in these results depend to a large extent on the studies' methodological approaches (Bradley, Barefoot, & Arsenault, 2011; Elaad, 2009; Gamer et al., 2010). They primarily criticize that the studies differ in the amount of detail given to informed innocents, in the time for planning the intended mock crime, and – especially – in the instructions for encoding the crime-related information (intentional versus incidental). These considerations again demonstrate the crucial role of encoding for the CIT's accuracy.

Overall, the orienting response theory sufficiently accounts for the fact *that* knowing participants orient towards probe items whereas unknowing (innocent) participants do not. However, it does not explain *why* signal values and, hence, orienting responses vary across different samples of knowing participants (guilty, intended, and informed innocent). Differences in information processing – or, more specifically, in encoding – can complement the theoretical framework for the CIT by supplementing the orienting response theory. Furthermore, this integration of the levels of processing theory in the orienting response theory offers new perspectives for further fields of application – namely for credibility assessment.

Applying a CIT in the Context of Credibility Assessment

As initially described, the theoretical framework for credibility assessment stresses the divergent memory processes of truth tellers and liars. Although truth tellers can rely on sensory input and episodic memory when giving the testimony, liars have to use cognitive scripts and semantic memory to invent a consistent story. The cognitive load theory emphasizes that these contrasting cognitive representations induce differences in cognitive demand and finally in the quality of testimonies. In contrast, the orienting response theory – as the most prominent explanatory model for the CIT – focuses on an attentional rather than a memory approach. However, more recently, information processing and the related memory processes have become topics of interest in CIT research, too. Even though the impact of information processing on the CIT was not embedded in a larger theoretical framework yet, the studies cited above emphasize that depth of encoding has a crucial influence on orienting responses in the CIT. Hence, the relevance of information processing (especially encoding) for both credibility assessment and CIT seems evident.

The classic CIT compares knowing participants who acted as mock crime culprits or read about a crime (guilty participants, i.e., deniers) with a completely unknowing control group (innocents). In contrast, in a CIT for credibility assessment one would need to compare knowing participants who talk about experienced events (truth tellers) with participants who actively invent a story and hence simulate memory (liars). The distinction between truth tellers and liars is far more challenging as both have a certain knowledge about the alleged events. However, the elaboration of encoding of the incident and consequently the cognitive representation of it are likely to differ. Reflecting the relevance of levels of processing for the CIT in the light of credibility assessment leads to the assumption that liars and truth tellers should show different orienting responses in a CIT.

Besides the divergent levels of processing, liars and truth tellers also differ in their emotional and motivational involvement. In contrast to truth tellers, liars engage in strategic self-presentation (Volbert & Steller, 2014). Practically, emotional-motivational factors also influence the CIT by enhancing the signal value and, hence, the orienting towards relevant items (Elaad & Ben-Shakhar, 1989; Verschuere & Ben-Shakhar, 2011). This might induce a stronger orienting towards lie-related stimuli than towards truth-related stimuli.

Lastly, the ADCAT (Walczyk et al., 2014) postulates that the different cognitive demands of truth tellers and liars reflect in RTs. According to ADCAT, a deceitful response takes longer than a truthful one due to the time-consuming activation of the truth, evaluation of costs and benefits of lying (decision component), retrieval of the truth while constructing the lie, and execution of the deceitful response whilst motivational processes also cost cognitive resources. Overall, there is converging evidence that truth tellers and liars should differ in a CIT-like paradigm and that this should be reflected in longer RTs for liars. After all, the levels of processing theory, emotional-motivational factors, and the ADCAT are relevant for both credibility assessment and CIT. This gives reason to assume that liars and truth tellers should show distinct reactions in a CIT.

Methodological Considerations

Despite the theoretical foundation and satisfactory psychometric properties of the CIT, its great disadvantage remains its applicability. The CIT is only suitable when objective evidence is available that can be used for constructing the probe items (e.g., the actual murder weapon "knife"). As already pointed out, these objective proofs are exactly what is lacking when credibility assessment is solicited. Thus, at first sight it seems that the CIT does not seem to be applicable in contexts of credibility assessment in which nothing than a witness's statement is available. However, to the best of my knowledge no previous research has attempted to record psychophysiological or behavioral data while confronting liars and truth tellers with personalized probe items that comprise literal phrases of his or her testimony. Analogous to the classic CIT, participants should deny any knowledge of probe items and irrelevant items but confirm knowledge of target items. The stimulus-response incompatibility for probe items should induce a conflict for truth tellers and liars. However, this conflict is expected to be stronger for liars.

Hypotheses

Memory processes of liars differ crucially from memory processes of truth tellers. Liars should initially engage in an elaborate encoding and rehearsal process; they use the essential facts and broaden them with cognitive scripts to intentionally fabricate and finally memorize a comprehensive story (Volbert, 2010). Giving a false confession in the subsequent interview promotes elaborate encoding as it involves a high cognitive load (inventing a consistent story), motivational factors (self-presentation as trustworthy), and emotional processes (fear of being debunked) (Volbert & Steller, 2014). These emotional-motivational factors get linked with the semantic content of the lie. In sum, these factors should lead to a high signal value of lie-related content. Being confronted with personalized lie-related items in a CIT should induce a strong orienting response and, hence, long RTs as wells as high error rates. Moreover, the ADCAT (Walczyk et al., 2014) postulates that lying involves the activation and further the inhibition of the true response, which additionally explains an increase in RTs. For liars who are confronted with an idiographic probe item of their lie, this truth concerns inventing the probe item. The episode of inventing the lie, its semantic content, the episode of telling it, as well as the corresponding emotional-motivational states are activated and have to be inhibited in a CIT.

In contrast, truth tellers form episodic memory content while experiencing an event. Information is encoded incidentally during the experience, whereas the activation of semantic memory content should play a subordinate role. Due to the low cognitive effort of memorizing and reporting an experience, the absence of any motivation to lie, and the moderate elaboration of encoding, event-related items should have a moderate signal value. Consequently, this should induce moderate RT and error rate effects.

In sum, liars and truth tellers should differ in their elaboration of encoding and storage of the incriminated event as well as in their motivational and affective states. This, in turn, should induce different signal values of items that pertain to fabricated or experienced events. Due to the stronger signal value, liars should show stronger orienting responses and, hence, longer RT effects in a CIT as compared to truth tellers.

Study 1 – Student Sample

Study 1 was conducted to develop a CIT suitable to distinguish truth tellers and liars. Experiencing and fabricating events involve different cognitive processes and resources that induce different signal values for crime-related information. Confronting participants with idiographic items that refer to a personal testimony should induce orienting responses. The strength of this orienting response should be larger for liars than for truth tellers. As initially pointed out, CITs can be assessed using behavioral, neuronal, or physiological paradigms. Since RTs seem to perform best (Verschuere et al., 2010) and are easily and economically measurable, an RT-based Credibility-CIT was developed in this study.

Method

Sample

A sample of 60 psychology students participated for course credit and gave their informed consent. Data from participants with more than 30% incorrect trials on irrelevant items or more than 60% on target items were excluded because those participants either did not understand the task or did not follow the instructions ($n = 4$). One participant was excluded due to software problems. The final sample consisted of $N = 55$ students (82% female) with a mean age of 23.17 years ($SD = 4.84$).

Procedure

The study was conducted using a between-group design with random assignment to the two experimental groups. The participants were informed that some participants would experience and report about certain tasks and others would tell a fabricated story about these events. Immediately afterwards an interviewer and after a one-week interval a “lie detector” would try to find out whether the testimony was true or invented. The participants’ task was to convince the interviewer and the lie detector that their own story was true (for detailed participant’s instructions in German, see Appendices A-F, p. 110-121).

In the truth group, a dyad of two participants (hereafter called truth tellers) participated together but held different roles. One participant (offender) got a list with rough descriptions of nine slightly aversive tasks (e.g., “put feet in slime”, “hold hands in ice water”, “put on a stinky shirt”, see Table 1). He or she was instructed to pick five out of nine tasks and urge the other person (victim) to perform those actions. The victim got a list with the same tasks. Victims were informed that the other participant will ask him or her to complete several unpleasant tasks and that it would be alright to refuse or negotiate with the counterpart. Hence, the aim was not to actually perform each task but to keep the participants busy with the objects. While the dyad interacted and dealt with the tasks, the experimenter waited outside the room and surveilled the participants on a screen to avoid disturbing the dyadic interaction. No video recordings were made.

Liars were randomly assigned to pretend being the victim or the offender during the alleged events. Lying victims and lying offenders received the same list with rough descriptions of the nine tasks that was also given to the truth-telling participants (Table 1). They were instructed pick five out of those nine tasks and fabricate a story on how they performed those. However, liars were not given information about details of the actual tasks at any time. E.g., they did not know the color and texture of the slime for the feet or the design of the stinky shirt. Liars were given 25 minutes to fabricate a detailed story about five of those events and were allowed to take notes.

Table 1

Participant's Instructions and Descriptions of the Actual Tasks.

Instructions given to all Participants	**Details on the Task (Unknown to Liars)**
Eat a cookie with a dry worm lying upon it. [1, 2]	Several butter biscuits with dried mealworms from a pet supply store on them were on a plate. One fresh biscuit was off to the side, not touching a worm at any time.
Explore things in a box you cannot look into. [1, 2]	The box contained oily spaghetti, artificial fur, and jelly.
Hold an object with your arm outstretched. [1, 2]	A cup filled with sugar had to be held as long as possible.
Touch dirty toilet paper. [1, 2]	The piece of toilet paper was smudged with chocolate cream.
Put on a stinky and worn T-shirt. [1, 2]	The shirt was smudged with oil and vinegar.
Put your feet in slime. [1, 2]	The "slime" was water-soluble paste. Cleaning utensils were on hand.
Put your hands in ice water. [1]	A bucket was filled with ice and water. Participants decided for themselves when to finish the task. The "offender" measured the time.
Put your hands in dirt. [2]	The "dirt" was fresh flower soil mixed with water.
Put your hands in a box of slimy condoms. [1]	The condoms were unwrapped, unrolled, and wetted with lubricant.
Put lubricant on a dildo. [1]	The dildo was skin-colored.
Take objects out of a box with your fingers being sticky. [2]	After using hand lotion, participants had to find small objects in a box filled with down feathers.

Note. Truth tellers completed five of the above tasks. Liars had to fabricate a story based on these instructions but did not gain any more details. [1] Task used in Study 1. [2] Task used in Study 2, Study 3, and Study 4.

Immediately after experiencing or fabricating an event, all participants (truth telling victims, truth telling offenders, lying victims, and lying offenders) were interrogated about the occurrences by an interviewer who was blind to the experimental condition. The interviews were recorded on tape because the statements served as a source for the probe items used in the Credibility-CIT (and were analyzed via CBCA, reported in Rönspies, in prep.).

Credibility-CIT

Items. The Credibility-CIT was conducted one week after the first session of experiencing/fabricating, as a study by Seymour and Fraynt (2009) indicated that a one-week interval increases the RT difference between elaborately and shallowly encoded information (for another study manipulating delay, see Carmel et al., 2003). During the RT task, 20 target items (memorized and to-be-recognized items), 80 irrelevant items (distractors), and 20 probe items (critical items) were randomly presented on a computer screen. Each stimulus was a two- or three-word phrase. The 20 probe items were further divided into 10 credibility probes and 10 knowledge probes.

The item construction of the *credibility probes* was an important innovation of this research project. Credibility probes were personalized phrases in the exact words previously used by each participant. The phrases were drawn from the record of each individual's statement and therefore customized for every participant. For example, if a participant told the interviewer "During the experiment, my counterpart asked me to wear a disgusting shirt", the credibility probe item "disgusting shirt" was extracted. All credibility probe items referred to central details of the statement. This procedure was the same for truth tellers and liars. Item extraction was done by an experimenter who was blind to the participants' condition.

The remaining 10 probe items were called *knowledge probes*. Those items were standardized phrases that related to details of the completed tasks. Hence, the knowledge probes were known only to the truth tellers who experienced the events (e.g., the description of slime for the feet: "white slime"). Liars had no access to knowledge about these items. Consequently, truth tellers were expected to react more slowly and less accurately than liars to these items. Knowledge probes were not essential for the credibility assessment via Credibility-CIT. Instead, they served as a manipulation check to test whether knowing participants react systematically to probe items at all.

Credibility-CIT procedure. Participants first had to memorize 20 target items. Each target was presented on the screen for 7,000 ms. The target items' order was randomized to avoid primacy and recency effects. After the targets were shown, participants had to write down all memorized terms. The learning phase and free recall were repeated three times. During the following RT procedure, targets, irrelevant items, and probes were presented in a randomized order on the screen. The question "Do you know this term?" was shown above each item. The participants' task was to press the button for "yes" whenever a memorized term (target item) occurred and the button for "no" for all other stimuli (irrelevant items and both kinds of probe items). This oddball paradigm, in combination with the question "Do you know this term?", induced a stimulus-response incompatibility on probe items for truth tellers and liars. After a maximum response frame of 800 ms, the error message "too slow" occurred. There was a randomly varying inter-stimulus interval of 1,500, 1,750, or 2,000 ms after each item. Every item was presented only once in accordance with findings on the disadvantages of item repetition in the CIT (Ben-Shakhar & Elaad, 2002, 2003). Prior to the test, three practice trials were conducted. In contrast to the actual test phase, an error message appeared when the wrong button was pushed and no response deadline existed. These practice trials were followed by three training trials with the same conditions as in the actual task. Practice and training trials were not included in the data analysis.

Scoring

For the RT scoring, a difference score (hereafter called the RT score) was calculated by subtracting the mean RT of irrelevant items from the mean RT of probe items. High RT scores indicate large RTs on the critical probe items and therefore good discrimination efficiency of the Credibility-CIT. Only responses above 200 ms were included in the calculations. Since latency measures have to deal with the participants' trade-off between speed and accuracy, wrong (incorrect button pressed) or too slow (> 800 ms) responses were replaced with a penalty score. This penalty was calculated by the mean of correct responses plus twice the standard deviation ($M + 2SD$) of the corresponding item category (probe or irrelevant). This error treatment procedure was adopted from a systematic evaluation of scoring algorithms for the Implicit Association Test (IAT) and has proven to be one of the best-performing procedures for the IAT to interpret error trials in the RT-based score (Greenwald, Nosek, & Banaji, 2003).

For the calculation of error rates, pressing the wrong button or reaching the response deadline of 800 ms was counted as an error. Again, a difference score was determined by subtracting the percentage of errors on irrelevant items from the percentage of errors on probe items (the result is hereafter called the error score). High error scores indicate inaccurate responding to probe items. Target items were not included in the calculations.

Reliability Estimations

To estimate the reliability of the RT scores, the item set was divided into two halves of equal size. Since items were presented randomly in the CIT, order effects can be ruled out. A difference score was calculated for each half of the item set (mean RT for half of the probe items – mean RT for half of the irrelevant items). The two RT difference scores served as items for the reliability estimation. The same procedure was used for error scores: the mean error rate on irrelevant items was subtracted from the mean error rate on probe items for both halves of the test. The two error difference scores were used for the reliability estimation.

Results

Truth tellers (n = 27) and liars (n = 28) did not differ in gender distribution ($\chi^2(1)$ = 0.01, p = .95) or age ($t(52)$ = 0.77, p = .45; for one participant, information on age was not available).

RT Scores for Credibility Probes

Since a difference score was used (probes minus irrelevant items, see Method), an RT score around zero indicates similar RTs for probes and irrelevant items. Truth tellers' RT scores (M = 5.93, SD = 42.50) did not differ from zero ($t(26)$ = 0.73, p = .48) but liars' RT scores (M = 31.94, SD = 51.31) were significantly above zero ($t(27)$ = 3.29, $p < .01$). Liars' mean RTs significantly increased for credibility probes but truth tellers did not show a difference. A two-way ANOVA was calculated to test for the influences of credibility (truth teller or liar), role (victim or offender), and their interaction (credibility × role) on the RT score for credibility probes (for descriptive statistics, see Table 2). Participants' RT scores for these customized items differed between truth tellers and liars ($F(1, 51) = 4.12$, $p < .05$, $\eta^2 = .08$) with liars producing larger RTs than truth tellers. Participants' role ($F(1, 51) = 0.09$, $p = .77$, $\eta^2 = .01$) and the interaction term ($F(1, 51) = 0.17$, $p = .68$, $\eta^2 = .01$) did not show significant effects.

Table 2

Descriptive Statistics for Reaction Times and Error Rates by Experimental Groups.

	Truth Tellers n = 27		**Liars** n = 28	
	M	(*SD*)	*M*	(*SD*)
Reaction Times (in milliseconds)				
Credibility probe items	559.06	(58.81)	602.71	(70.95)
Knowledge probe items	583.43	(69.28)	576.77	(85.32)
Irrelevant items	553.13	(49.24)	570.77	(64.37)
Target items	662.11	(57.10)	672.21	(58.10)
RT score for credibility probes	5.93	(42.50)	31.94	(51.31)
RT score for knowledge probes	30.30	(47.03)	6.00	(46.12)
Error Rates (in %)				
Credibility probe items	7.78	(10.50)	13.21	(10.90)
Knowledge probe items	12.96	(11.03)	11.79	(12.19)
Irrelevant items	5.69	(4.44)	7.99	(6.51)
Target items	26.67	(11.60)	29.46	(15.30)
Error score for credibility probes	2.08	(9.40)	5.22	(10.88)
Error score for knowledge probes	7.27	(9.98)	3.79	(9.24)

Note. Wrong or too slow responses were replaced with a penalty score (see Method section). RT score = RT on probe items - RT on irrelevant items. Error score = Errors on probe items - errors on irrelevant items. Response deadline was set to 800 ms.

The main effect of credibility is of special interest, since the difference between liars and truth tellers is the central distinction to be made in credibility assessment. The RT Score yielded a reliability of $r_{tt} = .27$. A *t*-test indicated that truth tellers and liars differed in their RT scores with a medium effect size of $d = 0.55$ ($t(53) = 2.04$, $p < .05$). Moreover, a receiver operating characteristic (ROC) analysis was calculated to test for the Credibility-CIT's efficiency to discriminate truth tellers from liars. In ROC analyses, a test's true positive rate (sensitivity) is plotted against the false positive rate (1 - specificity). An area under the curve (AUC) of .50 indicates a diagnostic value at chance level and an AUC = 1.00 indicates perfect discrimination of a test. For the Credibility-CIT, the ROC analysis yielded a result of AUC = .66 that significantly differed from chance level ($p < .05$, CI [.52; .81]) indicating the Credibility-CIT's ability to significantly discriminate truth tellers from liars.

Error Scores for Credibility Probes

Error scores were also calculated as difference scores, with scores above zero indicating more errors for credibility probe items than for irrelevant items ($r_{tt} = .20$ for error scores). Error scores for liars significantly differed from zero ($M = 5.22$, $SD = 10.88$, $t(27) = 2.54$, $p < .05$), whereas truth tellers' error score equaled zero ($M = 2.08$, $SD = 9.40$, $t(26) = 1.15$, $p = .26$), indicating that liars reacted less accurately to credibility probes than to irrelevant items, but truth tellers did not react systematically. However, ROC analyses (AUC = .61, $p = .14$, CI [.46; .77]) and the two-way ANOVA indicated that error scores for credibility probes did not differ between the two credibility groups ($F(1, 51) = 1.39$, $p = .24$, $\eta^2 = .03$). Additionally, participants' role (victim vs. offender; $F(1, 51) = 0.80$, $p = .38$, $\eta^2 = .02$) and the interaction term (credibility × role; $F(1, 51) = 0.01$, $p = .91$, $\eta^2 = .01$) had no significant influence.

RT Scores for Knowledge Probes

The standardized knowledge probes served as a manipulation check. They comprised terms that were only known to truth-telling participants because those participants actually experienced the events. In support of the hypothesis, truth tellers reacted systematically on knowledge probes; their RT scores differed from zero ($M = 30.30$, $SD = 47.03$, $t(26) = 3.35$, $p < .01$). As expected, liars' RT scores equaled zero ($M = 6.00$, $SD = 46.12$, $t(27) = 0.69$, $p = .50$). Moreover, the ROC curve indicated a significant area under the curve of AUC = .66 ($p < .05$, CI [.51; .80]). That means, with a probability of 66%, the RT-CIT ranks a truth teller as reacting more slowly on knowledge probes than a liar. The RT Score's reliability for knowledge probes was $r_{tt} = .25$.

Error Scores for Knowledge Probes

Error scores for knowledge probes were significantly above zero for both truth tellers ($M = 7.27$, $SD = 9.98$, $t(26) = 3.79$, $p < .01$) and liars ($M = 3.79$, $SD = 9.24$, $t(27) = 2.17$, $p < .05$). Hence, participants in both experimental groups reacted systematically more inaccurately to knowledge probes than to irrelevant items. Means on these error scores did not differ between the groups (AUC = .60, $p = .19$, CI [.45; .75]). Error scores for knowledge probes yielded a reliability of $r_{tt} = .26$.

Discussion

Analyses indicate that truth tellers and liars show different RT patterns in the Credibility-CIT. The slowing of responses caused by testimony-based credibility probes (rather than irrelevant items) was significantly stronger in liars than in truth tellers. I call this the Credibility-CIT effect. In contrast, error scores did not differ between the groups, as has been frequently found in studies on the classic CIT, too (Seymour & Fraynt, 2009; Verschuere et al., 2010).

One could assume that the stronger Credibility-CIT effect in liars is simply due to a better retention rate (Carmel et al., 2003). Truth tellers might have forgotten details of the experience, whereas liars remembered phrases of their lies well. However, this explanation is not sufficient. Since credibility probe items consist of phrases that the participants personally stated during their testimonies, the participants must have consciously encoded these event-related details. However, this does not necessarily mean that they remember the items after the one-week delay before the Credibility-CIT was administered. Since the items referred to the most central details of the statement – items that are usually remembered well (Gamer et al., 2010) – forgetting is not likely. Rather, during item presentation different cognitive processes might be activated, inducing distinct orienting responses. For fabricating the story, liars memorized the given background information, related it to cognitive scripts, took notes on their story, and concentrated on word choice. In doing so, the relevant terms were encoded intentionally, semantically, and therefore more elaborately (Craik & Lockhart, 1972). In contrast, truth tellers experienced events and just recalled them. Hence, their credibility probes were encoded incidentally, at a sensory level, and consequently shallowly.

Considering the knowledge probes, truth tellers reacted systematically but liars did not. This manipulation check indicates that the CIT was generally sensitive to discriminate probes from irrelevant items. Thereby findings on the classic CIT were replicated, showing longer RTs for a knowing than for an unknowing group (Hu, Evans, Wu, Lee, & Fu, 2013; Noordraven & Verschuere, 2013; Seymour, Seifert, Shafto, & Mosmann, 2000; Suchotzki, Verschuere, Peth, Crombez, & Gamer, 2015; Verschuere, Crombez, Degrootte, & Rosseel, 2010; Verschuere, Kleinberg, & Theocharidou, 2015; Visu-Petra, Miclea, & Visu-Petra, 2012).

A limitation to the results are the low reliabilities of the RT scores and error scores. It is a common finding that internal consistencies for mean RTs (here, RT rates) are good to excellent, whereas RT difference scores often produce low internal consistencies (Miller & Ulrich, 2013; Waechter & Stolz, 2015). A study by Miller and Ulrich (2013) used a model based on classical test theory to investigate the psychometric properties of RT rates and RT difference scores. They found that difference scores are much stronger influenced by item number than often assumed. For raw RT rates, ten items tend to produce good internal consistencies, while for RT difference scores hundreds are needed for reliabilities exceeding .80 (Miller & Ulrich, 2013). In the present study, the small number of 10 probe-irrelevant differences is a limitation from a psychometric point of view. However, in the context of CITs, increasing the item number can lead to habituation, strategic responses, and exerting countermeasures (Suchotzki et al., 2017). Therefore, lengthening the Credibility-CIT in future studies is a threat to its validity. Only few papers on the classic CIT report reliability estimations at all. An exception is the study by Noordraven and Verschuere (2013) reporting split-half reliabilities of $r = .38$ and .55 for their probe-irrelevant difference scores in a RT-based CIT. These values exceed the reliabilities found in the Credibility-CIT only slightly.

Study 2 – Replication with a High School Sample

Credibility assessment is mainly commissioned in court when the alleged victim is a child or adolescent. Therefore, Study 2 was conducted to test whether the results of the first study could be replicated in a sample with younger participants. Visu-Petra, Jurje, Ciornei, and Visu-Petra (2016) indicated that the RT-CIT is applicable in young children (7-10 years old). Hence, I expected the Credibility-CIT to discriminate adolescent liars from truth tellers, too. Moreover, a control group was added in order to compare an unknowing group with the two experimental conditions and to rule out the possibility that properties of the idiographic probe items drive the effect. Finally, intelligence was assessed as a potential moderator.

Method

Sample

The sample consisted of $N = 134$ pupils of a German comprehensive school (Gesamtschule). They participated voluntarily during their school lessons and the parents gave informed consent for participation (see Appendix F, p. 121). As in Study 1, participants were excluded if they made more than 30% errors on irrelevant items or more than 60% on target items ($n = 10$). The final sample included $N = 124$ participants (58.1% female) with a mean age of $M = 16.73$ years ($SD = 1.44$, range: 14-19 years). This sample was divided into three groups with $n = 45$ truth tellers (22 offenders, 23 victims), $n = 51$ liars (26 offenders, 25 victims), and $n = 28$ control group participants. Truth tellers, liars, and the control group did not differ in gender distribution ($\chi^2(2) = 3.93$, $p = .14$) or age ($F(2, 121) = 0.68$, $p = .51$).

Procedure

The procedure and scoring algorithm were similar to those used in Study 1 apart from the following modifications: With regard to the younger age of the participants, some of the tasks were replaced (e.g., the tasks with sexual content). The participant's instructions to either fabricate stories or experience the tasks can be found in Appendix A and Appendix B (p. 110 and p. 116, respectively). Moreover, the maximum response frame was extended to 1,200 ms and the number of probe items was increased to 15 credibility and 15 knowledge probes with regard to the low internal consistencies found in Study 1. Finally, a control group was added. The control group did not participate in the first session of the experiment (neither fabricated stories nor experienced the actions and thus were not interviewed). Therefore, they did not gain any knowledge about the actions and accordingly should not react systematically on either knowledge or credibility probes. Each control group participant responded to the item set of one randomly assigned experimental group participant. Intelligence was tested after the RT-CIT procedure was completed.

Intelligence

Cognitive abilities were measured using two subtests of the LPS-2 (Leistungsprüfsystem-2; Kreuzpointner, Lukesch, & Horn, 2013), which is constructed along the intelligence conception by Carroll (1993, 2005). Since data collection was conducted during school lessons, time constraints did not allow for an extensive IQ test. The subtest general knowledge ($\alpha = .89$) was used to measure crystallized intelligence; in this subtest, spelling errors shall be detected in a list of words under time constraints. A second subtest was consulted to assesses cognitive speediness ($\alpha = .81$); here, certain signs must be found in a list and crossed out under time constraints. Since norm data were not available for the subtests, participants' raw scores were transformed into IQ scores by standardization within the final sample. Congruent with the test's norm data, standardization was performed for two age groups separately (14-16 and 17-19 years).

Results

RT Scores for Credibility Probes

As in Study 1, difference scores were calculated with means above zero indicating larger RTs on credibility probe items than on irrelevant items. The RT score's reliability was $r_{tt} = .39$. Participants in all experimental groups reacted systematically to credibility probes (liars: $t(50) = 7.22$, $p < .01$; truth tellers: $t(44) = 6,90$, $p < .01$; control group: $t(27) = 3.85$, $p < .01$; for descriptive statistics, see Table 3). The RT score for credibility probes differed across truth tellers, liars, and the control group ($F(2, 121) = 6.15$, $p < .01$, $\eta^2 = .09$). Post hoc comparisons using REGWQ's test indicated that liars reacted more slowly to their customized credibility probes than truth tellers and the control group did. Truth tellers and the control group did not differ with regards to their RTs.

For credibility assessment, discriminating truth tellers from liars is essential. Focusing on the comparison between these two groups, the two-way ANOVA showed a significant main effect for credibility (truth teller vs. liar, $F(1, 92) = 4.29$, $p < .05$) with liars showing a larger RT score than truth tellers on credibility probes. As expected, the main effect of role (victim vs. offender) was non-significant ($F(1, 92) = 0.27$, $p = .61$), as was the interaction term ($F(1, 92) = 0.72$, $p = .40$). As in Study 1, it was possible to classify truth tellers and liars based on their RTs on the personalized credibility probes ($t(94) = -2.07$, $p < .05$, $d = 0.42$; AUC = .62 (CI [.50; .73], $p < .05$).

Table 3

Descriptive Statistics for Reaction Times and Error Rates by Experimental Group.

	Truth Tellers *n* = 45		**Liars** *n* = 51		**Control Group** *n* = 25	
	M	(*SD*)	*M*	(*SD*)	*M*	(*SD*)
Reaction Times (in milliseconds)						
Credibility probe items	777.52	(117.33)	817.92	(147.11)	735.77	(98.28)
Knowledge probe items	765.20	(126.53)	745.74	(119.82)	711.25	(109.24)
Irrelevant items	723.25	(95.26)	733.66	(100.39)	703.76	(83.75)
Target items	817.98	(111.28)	824.70	(117.73)	786.69	(88.46)
RT score for credibility probes	54.27	(52.77)	84.26	(83.40)	32.01	(44.02)
RT score for knowledge probes	41.94	(69.89)	12.08	(52.02)	7.49	(51.20)
Error Rates (in %)						
Credibility probe items	9.93	(13.53)	12.68	(14.68)	5.00	(7.62)
Knowledge probe items	9.63	(13.53)	4.84	(7.07)	2.86	(5.86)
Irrelevant items	4.14	(5.17)	4.29	(5.30)	2.63	(4.58)
Target items	21.11	(13.69)	21.57	(15.38)	15.18	(10.67)
Error score for credibility probes	5.79	(10.75)	8.39	(11.33)	2.37	(6.15)
Error score for knowledge probes	5.49	(10.03)	0.55	(5.45)	0.22	(3.23)

Note. Wrong or too slow responses were replaced with a penalty score (see Method section). RT score = RT on probe items - RT on irrelevant items. Error score = Errors on probe items - errors on irrelevant items. Response deadline was set to 1,200 ms.

Error Scores for Credibility Probes

Liars and truth tellers showed error scores that were significantly higher than zero for credibility probes (liars: $t(50) = 5.29$, $p < .01$; truth tellers: $t(44) = 3.61$, $p < .01$). The control group did not show systematic error scores ($t(27) = 2.04$, $p = .05$; but note that the significance level was only missed slightly). The ANOVA yielded significant differences among liars, truth tellers, and the control group ($F(2, 121) = 3.20$, $p < .05$, $\eta^2 = .05$). However, REGWQ's post hoc test did not indicate differences in the error scores of the three groups. The ROC analysis showed that error scores did not significantly discriminate truth tellers from liars (AUC = .60, CI [.48; .71], $p = .10$). The error score's reliability was $r_{tt} = .32$.

RT Scores for Knowledge Probes

As in Study 1, knowledge probe items comprised information that was only known to the truth telling participants. The reliability of the RT score for knowledge probes was $r_{tt} = .28$. Truth tellers showed RT scores that were significantly larger than zero when responding to knowledge probes as compared to irrelevant items. In contrast, the liars' and control group's RT did not slow down (truth tellers: $t(44) = 4.03$, $p < .01$; liars: $t(50) = 1.66$, $p = .10$; control group: $t(27) = 0.77$, $p = .45$). The three groups differed in their RT scores ($F(2, 121) = 4.14$, $p < .05$, $\eta^2 = 0.06$). As predicted, planned comparisons revealed that participants possessing knowledge (truth tellers) reacted significantly more slowly on knowledge probes than did the two unknowing groups (liars and the control group; $t(121) = 2.87$, $p < .01$), indicating a successful manipulation check.

Error Scores for Knowledge Probes

Error scores for knowledge probes yielded a reliability of $r_{tt} = .24$. Error scores differed from zero for knowing participants (truth tellers: $t(44) = 3.67$, $p < .01$), but the unknowing participants did not react systematically (liars: $t(50) = 0.72$, $p = .48$; control group: $t(27) = 0.37$, $p = .72$). The ANOVA showed that RT scores differed significantly across the three groups ($F(2, 121) = 7.18$, $p < .01$, $\eta^2 = .11$). Planned comparisons indicated that truth tellers had higher error scores on knowledge probes than the unknowing groups did (liars and control group; $t(53.53) = 3.24$, $p < .01$).

Moderation by Intelligence

Multiple regression analyses were used to test whether the experimental effects were moderated by trait intelligence. Credibility (truth teller or liar), crystallized intelligence, cognitive speediness, and the corresponding interaction terms (credibility × crystallized intelligence; credibility × cognitive speediness; credibility × crystallized intelligence × cognitive speediness) were entered into the regression equation with the RT score or error score for credibility probes as dependent variables (Table 4). All variables were centered. Congruent with the results presented above, credibility predicted the RT score but not the error score for credibility probes. Crystallized intelligence and cognitive speediness did not moderate the Credibility-CIT responses.

Table 4

Moderation by Intelligence.

	RT Score		**Error Score**	
	$R^2 = .07$		$R^2 = .04$	
Predictors	β	*p*	β	*p*
Credibility (Truth Teller vs. Liar)	**.22**	**< .05**	.12	.27
Crystallized Intelligence	.11	.30	-.02	.89
Interaction (Credibility × Cryst. Intelligence)	.01	.99	.05	.62
Cognitive Speediness	.11	.32	.14	.20
Interaction (Credibility × Cogn. Speediness)	.07	.50	.03	.78
3-Way Interaction (Credibility × Cryst. Intell. × Cogn. Speediness)	.02	.87	.04	.73

Note. $N = 96$. All scores z-transformed.

Discussion

Regarding the credibility probes, the results found in Study 1 were replicated in an independent and younger sample. Liars showed slower RTs on credibility probes (as compared to irrelevant items) than truth tellers or the control group did. As compared to Study 1, reliabilities of the RT score were descriptively higher in Study 2 ($r_{\text{tt Study 1}} = .27$ vs. $r_{\text{tt Study 2}} = .39$), while the Credibility-CIT's diagnostic value was slightly lower ($d_{\text{Study 1}} = 0.55$ vs. $d_{\text{Study 2}} = 0.42$). It is reasonable that lengthening the test from 10 to 15 probe items influenced these results. As Suchotzki et al. (2017) considered, a larger number of probe items can induce habituation and the practicing of countermeasures which reduces the test's validity.

RT scores of the control group and the truth tellers were similar, which contradicted the initial hypothesis. However, the most probable explanation is that the control group gained knowledge about the tasks. Since the study was conducted at a school it is not unlikely that the students were curious and chatted about their experience of participation in the study. If this was the case, the control group would represent a shallow encoding group and similarities with truth tellers could be interpreted in the light of the Credibility-CIT's theory. However, a replication with a control group without any knowledge about the critical tasks is necessary.

No influence or interaction effect of crystallized intelligence or cognitive speediness on the Credibility-CIT was found. Therefore, even participants with high intelligence scores were not capable of manipulating their responses in the task. A limiting factor is, however, that due to time constraints only two subtests instead of a whole IQ test were conducted.

Considering the knowledge probes, truth tellers reacted more slowly and more inaccurately on these standardized items than did participants in the unknowing groups (liars and the control group), confirming a successful manipulation check as in Study 1. However, for future studies omission of these knowledge probes is advisable. In the present study design, truth tellers and liars had a different base rate of probe items. Participants with knowledge (truth tellers) see both kinds of probe items (knowledge probes and credibility probes) as critical. In contrast, for liars only the credibility probes induce a cognitive conflict. Studies show that the proportion of probe compared to irrelevant items is crucial for the CIT (Suchotzki, Verschuere, et al., 2015). To equalize this, an exclusive Credibility-CIT without knowledge probe items should be conducted in future studies.

Study 3 – Omission of Knowledge Probes

As discussed in Study 2, there is evidence that the knowledge probes could influence the CIT effect by placing different demands on truth tellers and liars. Hence, Study 3 was conducted to test whether the Credibility-CIT works when knowledge probes are omitted. Moreover, participants were younger than in Study 1 and Study 2. With regard to the application of a CIT for children, I expected the Credibility-CIT to show responses that differentiated between groups in a younger sample, too. As in Study 2, an unknowing control group was used and was expected to show non-specific responses to credibility probe items.

Method

Sample

A sample of 150 school students participated in Study 3. Again, participants with more than 30% errors on irrelevant items or 60% errors on targets were excluded from the calculations (n = 11). The final sample consisted of n = 139 students (40% female; age: M = 13.42 years, SD = 0.98, range 11-16 years). Truth tellers (n = 54), liars (n = 60), and the control group (n = 25) did not differ in gender distribution ($\chi^2(2) = 0.25$, p = .88) or age ($F(2, 138) = 0.37$, p = .69).

Procedure and Scoring

To keep the sample diverse, Study 3 was conducted at two schools (German Gymnasium and Realschule) and a summer academy for mathematically talented high school students. The procedure was similar to that used in Study 1 and Study 2. Stimuli were 15 probes, 15 targets, and 60 irrelevant items. With consideration for the participants' younger age, the response deadline was extended to 1,800 ms. Scoring was the same as in Study 1 and Study 2 (difference score: RT for probes – RT for irrelevant items). Again, only responses that were faster than 200 ms were included in the calculations. Wrong or too slow responses (> 1,800 ms) were replaced with a penalty score (*M* + 2*SD*).

Intelligence

Crystallized intelligence was assessed in more detail than in Study 2. The subtests general knowledge and anagrams from the LPS-2 (Kreuzpointner et al., 2013), as well as the subtest word fluency of the PSB-R 6-13 (Horn, 2003), were averaged to an overall score of crystallized intelligence. Cognitive speediness was measured with the same subtest of the LPS-2 as in Study 2 (Kreuzpointner et al., 2013). As in Study 2, participants' raw scores were transformed into IQ scores by standardization within the final sample for two age groups separately (in this case, ages 11-13 and 14-16 years).

Results

RT Scores for Credibility Probes

Consistent with the hypothesis, truth tellers and liars reacted differentially on credibility probes ($t(53) = 5.20$, $p < .01$ and $t(59) = 5.82$, $p < .01$, respectively), but the control group's RT score did not differ from zero ($t(24) = 1.56$, $p = .13$; for further descriptive statistics, see Table 5; reliability for RT score: $r_{tt} = .55$). The ANOVA showed that the three groups differed in their RT scores ($F(2, 136) = 3.59$, $p < .05$), but post hoc tests indicated that only the control group differed from the two experimental groups; truth tellers ($M = 86.58$, $SD = 122.33$) and liars ($M = 93.59$, $SD = 124.49$) showed similar RT scores ($t(112) = -0.30$, $p = .76$). This contradicted the hypothesis and the results found in the two previous studies.

Regarding the influence of the participant's role (victim vs. offender), the two-way ANOVA showed no significant main effects of either credibility (truth teller vs. liar, $F(1, 110) = 0.06$, $p = .80$) or role (victim vs. offender; $F(1, 110) = 0.08$, $p = .77$), and also no significant interaction effect ($F(1, 110) = 1.18$, $p = .28$).

Table 5

Descriptive Statistics for Reaction Times and Error Rates by Experimental Groups.

	Truth Tellers *n* = 54		**Liars** *n* = 60		**Control Group** *n* = 25	
	M	*(SD)*	*M*	*(SD)*	*M*	*(SD)*
Reaction Times (in milliseconds)						
Credibility probe items	1,009.60	(253.49)	1,028.52	(238.56)	906.72	(205.07)
Irrelevant items	923.02	(180.62)	934.93	(186.63)	884.54	(165.27)
Target items	996.93	(190.62)	1,020.61	(202.22)	1,020.69	(184.64)
RT score	86.58	(122.33)	93.59	(124.49)	22.18	(71.03)
Error Rates (in %)						
Credibility probe items	9.26	(17.63)	11.33	(19.34)	2.67	(5.09)
Irrelevant items	4.75	(7.68)	4.81	(7.08)	2.27	(2.76)
Target items	16.42	(14.40)	18.33	(14.61)	18.13	(12.98)
Error score	4.51	(13.10)	6.53	(14.68)	0.40	(4.06)

Note. Wrong or too slow responses were replaced with a penalty score (see Method section). RT score = RT on probe items - RT on irrelevant items. Error score = Errors on probe items - errors on irrelevant items. Response deadline was set to 1,800 ms.

It was striking that the standard deviations in this sample were twice as large as in Study 2. This might be related to the response deadline of 1,800 ms, which was larger than in Study 1 (800 ms) and Study 2 (1,200 ms). To explore this assumption, the response frame included in the calculations was trimmed post hoc. Table 6 shows that narrowing down the response interval improved the Credibility-CIT's detection efficiency. Trimming the response frame after 1,400 ms or 1,200 ms led to better classification efficiency of the CIT (*d* = 0.25 in both cases). Narrowing down the response interval is accompanied by an exclusion of the valid trials above the limitation criterion. Consequently, a too-strict restriction (up to 1,000 or 800 ms) led to a decline of valid trials (< 75 %) and hence to a poorer classification efficiency. Overall, even with the 1,400 or 1,200 ms deadlines, the RT scores of truth tellers and liars did still not differ significantly.

Table 6

RT scores for trimmed response frames.

Response Frame	**Valid Trials** (in %)	**Truth Tellers**		**Liars**		***t*-test**	***p***
		M	*(SD)*	*M*	*(SD)*		
[200; 1,800 ms]	100	86.58	(122.33)	93.59	(124.49)	*t*(112) = -0.30	.76
[200; 1,600 ms]	98	69.39	(93.23)	82.07	(97.58)	*t*(112) = -0.71	.48
[200; 1,400 ms]	95	56.41	(79.03)	77.84	(88.74)	*t*(112) = -1.36	.18
[200; 1,200 ms]	89	39.74	(63.39)	56.41	(70.30)	*t*(112) = -1.32	.19
[200; 1,000 ms]	73	17.61	(50.13)	29.17	(48.61)	*t*(110) = -1.24	.22
[200; 800 ms]	43	13.65	(37.39)	15.37	(49.61)	*t*(101) = -0.20	.84

Error Scores for Credibility Probes

Error scores (r_{tt} = .60) were significantly higher than zero for truth tellers ($t(53) = 2.53$, $p < .05$) and liars ($t(59) = 3.44$, $p < .01$), while the control group did not show systematic error scores ($t(24) = 0.49$, $p = .63$). The ANOVA indicated that error scores were similar for truth tellers (M = 4.51, SD = 13.10), liars (M = 6.53, SD = 14.68), and the control group (M = 0.40, SD = 4.06; $F(2, 136) = 2.03$, $p = .14$).

Moderation by Intelligence

Table 7 shows the results of the moderator analysis with RT score and error score as dependent variables. For exploratory purposes, the moderation was also calculated with the trimmed RT score [200; 1,400 ms]. However, results between the trimmed and untrimmed RT score differed only marginally. Overall, crystallized intelligence was not significantly related to the Credibility-CIT's classification efficiency ($\beta \leq .09$, $p \geq .35$). However, a high cognitive speediness was related to a lower error score ($\beta = -.24$, $p < .05$). The interaction between credibility (truth teller vs. liar) and cognitive speediness was not meaningfully related to the RT scores or error scores ($\beta = -.03$, $p < .73$). Hence, the CIT's detection efficiency was unaffected by cognitive speediness.

Table 7

Moderation by Intelligence.

	RT Score		Trimmed RT Score [200; 1,400 ms]		Error Score	
	$R^2 = .03$		$R^2 = .03$		$R^2 = .06$	
Predictors	β	*p*	β	*p*	β	*p*
Credibility (Truth Teller vs. Liar)	.02	.87	.14	.18	.08	.44
Crystallized Intelligence	.04	.69	.09	.35	.01	.90
Interaction (Credibility × Cryst. Intelligence)	.02	.85	.06	.53	.03	.74
Cognitive Speediness	-.13	.19	-.06	.57	**-.24**	**< .05**
Interaction (Credibility × Cogn. Speediness)	.06	.56	.01	.96	-.03	.73
3-Way Interaction (Credibility × Cryst. Intell. × Cogn. Speediness)	.07	.47	.03	.80	-.01	.95

Note. *N* = 114. All scores z-transformed.

Discussion

In contrast to the results found in Study 1 and Study 2, truth tellers and liars did not differ in their RT scores in the Credibility-CIT. There are three apparent, possible explanations for this discrepancy. First, the sample might have been too young to deal with the task. However, since error rates in this study were comparable to those found in the two previous studies (compare Table 2, Table 3, and Table 5), task difficulty seems an unlikely explanation. Moreover, a study by Visu-Petra et al. (2016) showed that the classic CIT works even for children aged between seven and ten years. Overall, the sample's age is probably not the cause of the insignificant results.

The second and more likely explanation concerns the extended response deadline of the Credibility-CIT. Seymour et al. (2000) emphasize the necessity of a speeded CIT to avoid strategic responses. They argue that a response deadline around 1,000 ms forces the participant to react automatically – solely based on familiarity or significance – and not consciously. As initially described, this dichotomous classification between familiar and unfamiliar stimuli induces the stimulus-response incompatibility that is essential for a CIT's detection efficiency. Post hoc trimming of the response deadline indicated that a narrower response frame improves the CIT's detection efficiency. Still, RT scores did not differ between truth tellers and liars. It remains unclear to what extend response deadlines can influence CIT effects. To the best of my knowledge, studies systematically evaluating the impact of different response deadlines for the CIT do not exist so far. Not only the classic CIT but also the Credibility-CIT would benefit from further examination of that topic.

A third explanation concerns the omission of knowledge probes. Knowledge probes led to a dissimilar proportion of familiar and unfamiliar items between truth tellers and liars in Study 1 and Study 2. Since the item proportion is crucial for a CIT's detection efficiency (Suchotzki, Verschuere, et al., 2015) this might have triggered the effect. Hence, future studies should again replicate the Credibility-CIT without including knowledge probes.

Regarding the influence of intelligence, neither cognitive speediness nor crystallized intelligence moderated the CIT's detection efficiency. On the one hand, this corresponds to the results found in Study 2. On the other hand, this lack of moderation by intelligence is likely to be related to the invalid overall effect of the experimental group. Hence, substantial conclusions cannot be drawn from this finding.

The control group's unsystematic responses indicate that probe items do not comprise characteristics other than irrelevant items. The study's design and aspects of the participants' anonymity did not allow for separating the control groups for truth tellers and liars. A within-subjects design – with each experimental group and control group participants undergoing a truth telling and lying condition – could solve this issue. This was done in Study 4.

Study 4 – Within-Subjects Design and EEG

Study 4 was conducted to test the Credibility-CIT using a within-participants design with each participant undergoing both a lying and truth-telling condition. With regard to the issues of the control groups in Study 2 and Study 3, a control group that responded to truth-related and lie-related item sets was included. Study 1 and Study 2 did not show differences between victims and offenders in their CIT responses. Hence, in Study 4, all participants were instructed by an experimenter instead of another participant. As in Study 3, knowledge probes were omitted to test whether item proportions drove the effect found in Study 1 and Study 2.

Beyond the already replicated results of the RT-based Credibility-CIT, applying a psychophysiological measure seemed promising. Since studies with the classic CIT show that event-related potentials (ERPs) have a good classification efficiency (Farwell & Donchin, 1991; Meijer, Smulders, Merckelbach, & Wolf, 2007; Rosenfeld, 2011; Suchotzki, Crombez, Smulders, Meijer, & Verschuere, 2015), EEG data should also be useful for the Credibility-CIT.

In addition, cognitive functions and personality traits were assessed to facilitate examination of individual differences in the responsiveness to the Credibility-CIT. Working memory might be related to the Credibility-CIT's detection efficiency, since working memory is involved in information processing and decision making. Given that attentional processes comprise the ability to focus on task-relevant stimuli while ignoring distractors, this capability might also be related to inhibiting deceitful responses in the CIT. Considering personality traits, studies have shown associations between the ability to inhibit task-irrelevant reactions and extraversion (in a Stroop task; Prabhakaran, Kraemer, & Thompson-Schill, 2011). No significant relations between CIT and the propensity towards anxiety occurred (Visu-Petra et al., 2012). Moreover, psychopathy seems to be associated with hyporesponsivity (Verschuere, Crombez, De Clercq, & Koster, 2005; Verschuere, Crombez, Koster, & De Clercq, 2007; Verschuere, Crombez, Koster, & Uzieblo, 2006). Hypo-responsivity has been observed to be accompanied by weaker orienting responses and hence decreased classification accuracy in the CIT for psychopathic participants (for a review see Verschuere et al., 2006).

Method

Sample

A sample of 78 right-handed psychology students participated voluntarily for course credit and gave informed consent. Compared to Study 1 and Study 2, participants had higher error rates on target items. Therefore, data of participants with more than 30% error trials on irrelevant or more than 70% on target items were excluded ($n = 3$). One participant was excluded because she did not follow the experimenter's instructions. The experimental group consisted of $n = 46$ students, and the control group comprised $n = 28$ participants (83% female; age $M = 24.11$ years, $SD = 5.64$). To ensure that the control group did not gain any knowledge about the experiment, data collection was conducted several months after collecting the experimental group's data. The control group only participated in the RT-based Credibility-CIT; EEG data were not recorded.

Procedure

The study was conducted using a within-participants design. Participants were informed that they would have to give a statement about eight to-be-completed tasks; four of them would truly be experienced (truth-telling condition) and the completion of four tasks shall be fabricated (lying condition). An interviewer would try to figure out which part they invented. The three most convincing statements would win vouchers of 50 Euro, 30 Euro, and 20 Euro, respectively (approximately 55, 33, and 22 US Dollars). After a one-week delay, the participants were to complete and attempt to trick an EEG-based “lie detector”.

It was counterbalanced whether the participants experienced the tasks or fabricated the story first. For the truth-telling condition, participants had to deal with four of the tasks as specified in Table 1 (p. 35). For the lying condition, they received a list with rough descriptions of four other tasks listed in Table 1; they had 25 minutes to fabricate a coherent story. Immediately after experiencing and inventing, participants were interrogated about the eight occurrences by an interviewer who was blind to the statements’ condition (and thus credibility). The control group did not participate in this phase of the experiment and was therefore unknowing about the tasks.

Credibility-CIT

Due to the within-participants design, the Credibility-CIT comprised two blocks – a truth block and a lie block: The *truth block* included 15 credibility probe items that were extracted from the participant’s true statement, 15 target items, and a randomly allocated set of 60 irrelevant items. The *lie block* included 15 credibility probes that referred to the participants’ invented statements, the same 15 target items as in the truth block, and another randomly allocated set of 60 irrelevant items. The order of the blocks (truth block or lie block first) was counterbalanced such that every participant was presented with both a truth trials block and a lie trials block, but the order varied. After the two blocks, the whole Credibility-CIT (both blocks) was repeated to obtain a stronger signal-to-noise ratio for the EEG. Following Study 1, which yielded the strongest results thus far, the response deadline was set at 800 ms per item.

In addition to the behavioral data (RTs and error rates), an EEG was recorded from 64 scalp electrodes using electrode caps (ActiCap; Brain Products, Gilching, Germany). Electrodes were placed according to the International 10-10 system. Data were referenced to the mastoid electrodes TP9 and TP10. Impedances were kept below 10kΩ, sampling rate was 500Hz, and the band-pass filter was 0.1-70Hz. Eye movements were corrected for ocular artifacts as proposed by Gratton, Coles, and Donchin (1983). Data were screened for artifacts (amplitudes exceeding ±100µV) and segmented into epochs lasting from 100 ms before to 1,000 ms after stimulus onset.

Cognitive Functions and Personality Traits

After finishing the Credibility-CIT, data for evaluating the relationship between individual differences in response to the CIT were collected. Selective attention was measured using the d2 test (Brickenkamp, 1994). The participant is required to cross out targets ("d"s) within several distractors ("p"s) under time pressure. The total number of targets marked correctly serves as a measure of processing. The total number of errors (distractors crossed out and targets omitted) is a measure of inaccuracy. The subtest digit span of the German Wechsler Adult Intelligence Scale - Revised (HAWIE-R; Tewes, 1991) served as a measure of working memory and attention. Participants must recall digits either in a predetermined order (digit span forward) or in reversed order (digit span backward). Span length served as outcome. The Big Five (extraversion, neuroticism, conscientiousness, agreeableness, and openness to experience) were assessed using the Big Five Inventory (BFI; Lang, Lüdtke, & Asendorpf, 2001). The Dark Triad (psychopathy, narcissism, and machiavellianism) was measured using the Short Dark Triad (SD3; Jones & Paulhus, 2014). For the SD3, Cronbach's alpha was insufficient in the present study (Psychopathy: $\alpha = .49$, Narcissism: $\alpha = .69$, and Machiavellianism: $\alpha = .66$). The low internal consistency of the Dark Triad was probably caused by a restricted score variance or a floor effect.

Scoring

As in the three previous studies, RTs and error rates were scored using difference scores. RT scores and error scores were calculated separately for the truth block and the lie block. Also, the reliability was calculated separately for the truth block and lie block. Data from the blocks that were repeated (to allow adequate data for EEG) were not included in the behavioral scores because studies show that detection efficiency for RTs decreases with repetition (Ben-Shakhar & Elaad, 2002).

Regarding the EEG, the ERP components P200, P300, and LPC (late positive component) at the three midline scalp sites Fz, Cz, and Pz (frontal, central, and parietal, respectively) were analyzed. However, findings were not meaningful (see Results section). Therefore, the results presented here focus only on the P300 component measured at Pz site, which usually produces the most robust effects in a CIT (Farwell & Donchin, 1991; Meixner & Rosenfeld, 2011, 2014; Rosenfeld, 2011). The wave forms revealed that the P300 peaked between 580 and 700 ms after stimulus onset. Hence, the interval was calculated accordingly for probes, irrelevant items, and targets in both the truth block and lie block. Truth and lie blocks were repeated and the EEG data were averaged over the initial and repeated trials.

<u>Results</u>

RT Scores and Error Scores for Credibility Probes

Consistent with the hypotheses, participants in the experimental group reacted systematically on the credibility probe items when they were confronted with lie-related items and truth-related items: The RT scores for the lie bock and for the truth block were significantly higher than zero (lie block: $t(45) = 6.74$, $p < .01$, $r_{tt} = .32$; truth block: $t(45) = 4.29$, $p < .01$, $r_{tt} = .28$; for descriptive statistics, see Table 8). The central distinction in this study is the difference between truth-related and lie-related trials. Responses were stronger slowed down in the lie block than in the truth block ($t(45) = -2.35$, $p < .05$, $d = 0.47$).

Error scores in the lie block and truth block differed from zero, too ($t(45) = 5.77$, $p <.01$, $r_{tt} = .42$ and $t(45) = 3.74$, $p < .01$, $r_{tt} = .09$). The error score in the lie block was higher than in the truth block ($t(45) = -2.22$, $p < .05$, $d = 0.41$). Hence, RT scores and error scores differed when participants either lied or told the truth.

The control group did not react systematically to the credibility probes in either block, as expected. RT scores in lie block and truth block did not differ from zero ($t(27) = 0.04$, $p = .97$ and $t(27) = 0.50$, $p = .62$, respectively). Error scores did not differ from zero, too (lie block: $t(27) = - 0.54$, $p = .59$; truth block: $t(27) = -0.45$, $p = .66$). Consistent with the hypotheses, the control group did not show an Credibility-CIT effect: truth and lie blocks did not differ in RT scores ($t(27) = 0.37$, $p = .71$) and error scores ($t(27) = 0.10$, $p = .92$).

Moreover, the control group had a different reaction pattern than the experimental group. Compared to the control group, the experimental group showed larger RT scores (truth block: $t(72) = 2.19$, $p < .05$, $d = 0.52$; lie block: $t(72) = 4.58$, $p < .01$, $d = 1.14$) and larger error scores (truth block: $t(72) = 2.81$, $p < .01$, $d = 0.70$; lie block: $t(72) = 4.27$, $p < .01$, $d = 1.07$).

Table 8

Descriptive Statistics for Reaction Times and Error Rates by Experimental Condition.

	Experimental Group $n = 46$				**Control Group** $n = 28$			
	Truth Block		**Lie Block**		**Truth Block**		**Lie Block**	
	M	*(SD)*	*M*	*(SD)*	*M*	*(SD)*	*M*	*(SD)*
Reaction Times (in milliseconds)								
Credibility probes	586.01	(56.39)	605.09	(60.85)	567.97	(51.94)	565.57	(55.18)
Irrelevant items	562.62	(40.94)	562.83	(43.98)	564.31	(50.78)	565.37	(42.05)
Target items	696.30	(57.95)	805.87	(66.15)	688.01	(67.27)	810.50	(63.67)
RT score	23.39	(36.98)	42.26	(42.52)	3.67	(38.77)	0.20	(30.15)
Error Rates (in %)								
Credibility probes	10.14	(8.75)	14.64	(11.21)	6.67	(7.26)	5.48	(7.71)
Irrelevant items	5.47	(4.37)	6.16	(4.48)	7.20	(5.54)	6.19	(5.03)
Target items	34.78	(16.74)	31.74	(15.79)	31.43	(14.61)	26.90	(12.76)
Error score	4.67	(8.47)	8.48	(9.97)	-0.54	(6.30)	-0.71	(6.99)

Note. Wrong or too slow responses were replaced with a penalty score (see Method section). RT score = RT on probe items - RT on irrelevant items. Error score = Errors on probe items - errors on irrelevant items. Response deadline was set to 800 ms.

Effects of Trial Repetition

The blocks were repeatedly presented to the participants to obtain a stronger signal-to-noise ratio for the EEG. This offered the opportunity to check for the effects of trial repetition in the behavioral data of the Credibility-CIT. Results showed that the Credibility-CIT effect diminishes with block repetition. The RT score in the truth block (M = 24.19, SD = 32.16) and the lie block (M = 34.94, SD = 33.58) was similar when the RTs were averaged over the initial and repeated block ($t(45) = -1.56$, $p = .13$, $d = 0.33$). The error score also did not differ between lie block and truth block when items were repeated (M = 6.74, SD = 7.46 and M = 5.11, SD = 7.59, respectively; $t(45) = -1.23$, $p = .23$, $d = 0.22$).

Cognitive Functions and Personality Traits

Measures of individual differences were correlated with the Credibility-CIT's detection efficiency (Table 9). According to the previous results, a high detection efficiency of the Credibility-CIT is defined as higher scores in the lie block than the truth block. Therefore, detection efficiency was calculated as the difference between lie and truth block (for RT scores: RT $\text{score}_{\text{lie block}}$ – RT $\text{score}_{\text{truth block}}$; for error scores: error $\text{score}_{\text{lie block}}$ – error $\text{score}_{\text{truth block}}$). High scores indicate strong orienting towards lie-related items.

Working inaccurately in the d2 test was significantly associated with a weaker detection efficiency of the Credibility-CIT ($r = -.29$). Moreover, higher scores on psychopathy were correlated with a weaker orienting towards lie-related items in the Credibility-CIT ($r = -.36$). The other correlations between the Credibility-CIT and cognitive functions or personality traits were not meaningful.

Table 9

Correlations between the Credibility-CIT's Detection Efficiency, Cognitive Functions, and Personality Traits.

	Detection Efficiency using RT Scores	**Detection Efficiency using Error Scores**
Cognitive Functions		
Digit span forward	.10	.10
Digit span backward	-.03	.17
Processing (d2 test)	.13	.00
Inaccuracy (d2 test)	-.29*	-.15
Personality Traits		
Extraversion	.16	.10
Agreeableness	.13	-.05
Conscientiousness	.11	-.03
Neuroticism	.05	.01
Openness	-.03	.04
Narcissism	-.07	.08
Machiavellianism	-.20	-.11
Psychopathy	-.36*	-.15

Note. $N = 46$. * $p < .05$. A high detection efficiency represents a strong orienting towards lie-related items.

To test for individual differences in responsivity, a repeated measures ANCOVA was calculated with cognitive functions and personality trait variables as covariates (Table 10). All covariates were mean-centered. RT scores for the lie block and the truth block served as a repeated measures factor. Overall, the RT effect was independent of any covariate's influences. However, inaccuracy in the d2 test and psychopathy traits nearly reached significance.

Table 10

Detection Efficiency of the Credibility-CIT with Cognitive Functions and Personality Traits as Covariates.

	F **(1, 33)**	***p***	**η^2**
Within-Participants Factor			
Credibility (RT score for lie block vs. RT score for truth block)	5.06	< .05	.13
Covariates			
Cognitive Functions			
Digit span forward	0.19	.67	.01
Digit span backward	0.74	.40	.02
Processing (d2 test)	0.20	.66	.01
Inaccuracy (d2 test)	4.32	.05	.12
Personality Traits			
Extraversion	0.42	.52	.01
Agreeableness	0.67	.42	.02
Conscientiousness	0.03	.85	.00
Neuroticism	0.11	.75	.00
Openness	0.25	.62	.01
Narcissism	0.00	.99	.00
Machiavellianism	0.01	.92	.00
Psychopathy	3.25	.08	.09

Note. $N = 46$. Repeated measures ANCOVA's tests of within-participants effects with RT score for the lie block and RT score for the truth block as repeated measures factor.

Event-Related Potentials

The critical test of the hypothesis was whether credibility probes show different ERPs in lie the blocks and the and truth blocks. A paired samples *t*-test revealed that truth blocks and lie blocks elicited similar P300 amplitudes for credibility probe items ($t(45) = -0.30$, $p = .77$; for descriptive statistics, see Table 11). Hence, ERPs did not discriminate truth from lie in the Credibility-CIT. Further analyses indicated that probe items were processed like irrelevant items, which explains this lack of detection efficiency. A 3 (item category: probe vs. irrelevant vs. target) × 2 (credibility: truth blocks vs. lie blocks) repeated measures ANOVA was calculated with amplitude of the P300 component serving as dependent variable. There was no significant main effect of credibility ($F(1, 45) = 1.70$, $p = .19$, $\eta^2 = .04$) and no interaction effect (credibility × item category; $F(2, 44) = 1.60$, $p = .21$, $\eta^2 = .07$). However, a main effect of item category occurred ($F(2, 44) = 68.68$, $p < .01$, $\eta^2 = .76$). The contrasts indicated that probes and irrelevant items had a similar P300 amplitude ($F(1, 45) = 2.09$, $p = .15$, $\eta^2 = .04$), while targets elicited larger P300 components than probes ($F(1, 45) = 133.72$, $p < .01$, $\eta^2 = .75$).

Table 11

Descriptive Statistics for P300 Amplitudes.

	Truth Block (i.e., truth tellers)		**Lie Block** (i.e., liars)	
P300 Amplitudes (in µV)	*M*	*(SD)*	*M*	*(SD)*
P300 on credibility probe items	3.48	(4.65)	3.60	(4.17)
P300 on irrelevant items	4.16	(4.57)	3.99	(4.50)
P300 on target items	12.38	(7.70)	11.35	(7.78)

Note. $N = 46$. P300 component was recorded at electrode Pz.

Discussion

As in Study 1 and Study 2, lying was accompanied by larger RT scores than truth telling. Hence, results indicating the validity of the Credibility-CIT were replicated for a third time. Moreover, this study was the first that indicated differences in error scores between truth and lie trials in a Credibility-CIT. Lie trials were more inaccurate than truth trials, a finding that corresponds to the orienting response theory.

Different from the null-effect in Study 3, the response deadline in Study 4 was again set to 800 ms (as in Study 1). It seems that narrower response frames are related to the detection efficiency of the Credibility-CIT. As in Study 3, knowledge probes were omitted yet results on credibility probes were still valid. Hence, knowledge probes were probably not the cause of the effects found in Study 1 and Study 2. Rather, the Credibility-CIT effect is indeed triggered by characteristics of the personalized credibility probes. The control group's responses indicated that truth-related and lie-related credibility probes were processed similarly in unknowing participants. These idiographic probe items in cases of lying and telling the truth do not differ in content, language, or syntax – and consequently not in their general signal value. Due to the within-subjects design, systematic sample differences between the conditions (telling the truth versus lying) can be ruled out.

Regarding the measures of individual differences, a strong orienting towards lie-related items in the Credibility-CIT was associated with accurate responses in the d2 test. If accurate responding in general is associated with the Credibility-CIT's detection efficiency, instructions for precise responding in the Credibility-CIT might enhance its validity. This remains a topic for further research. Additionally, psychopathy was associated with a weaker orienting towards lie-related items. This is in line with the hyporesponsivity of persons scoring high on psychopathy described by Verschuere et al. (2006). Overall, most of the cognitive function or personality trait variables were not significantly correlated to the Credibility-CIT. Hence, the essential explanation for differences between the experimental groups is the varying signal value of the individual credibility probe items.

Considering the results of the EEG data, the P300 component peaked rather late (around 620 ms) compared to other studies. This effect may have been caused by the items' complexity (two- or three-word phrases) since the P300 latency is affected by stimulus evaluation (Magliero, Bashore, Coles, & Donchin, 1984; McCarthy & Donchin, 1981). However, probe items did not induce an orienting response that was strong enough to elicit a P300 component at all. Data indicated that on a neuronal level, probes were processed like irrelevant items. However, the strong P300 associated with the target items showed that the EEG was generally sensitive to task-specific and meaningful stimuli. It remains unclear why the basic CIT effect (that probes induce strong P300 components) was not found in this study. Due to the absence of this basic effect, it is not surprising that ERPs did not differ between truth and lie blocks. It can be hypothesized that P300 amplitude's sensitivity to different levels of encoding plays a role. Gamer and Berti (2012) found that in a CIT the P300 is less affected by depth of processing. In their study, P300 amplitudes did not differ between elaborately encoded (central) items and shallowly encoded (peripheral) details. Yet, the similar P300 components of the truth block and the lie block in the current study was probably rather caused by the absence of the basic effect ($P300_{probes} = P300_{irrelevants} < P300_{tagets}$). Overall, it remains unclear why behavioral responses to probe items and irrelevant items differed (RTs and error rates) while they were processed similar on a neuronal level.

Irrespective of the EEG data, the moderate effect sizes of the RT scores repeatedly shown in the Credibility-CIT are a promising result. Hence, it would be informative to determine the overall effect found within this research project and to estimate the expected population effect. A meta-analysis would serve these purposes.

Meta-Analysis

Through Studies 1 through 4, this research project aimed to develop an indirect measure appropriate for contexts of credibility assessment. The CIT – as a theoretically founded and psychometrically sound paradigm – served as a base for these studies. The methodological similarity of the four studies is ideal for meta-analytical purposes. Study 1, Study 2, and Study 4 showed that it is possible to discriminate truth tellers from liars using the RT-based Credibility-CIT: Compared to confronting participants with truth-related stimuli, presentation of lie-related probe items led to a larger RT difference score (RT score = RT on probe items – RT on irrelevant items; see Figure 2). The effect sizes found in Study 1, Study 2 and Study 4 were moderate. In contrast, Study 3 showed non-significant results. Post hoc analyses indicated that this was probably caused by the response deadline of 1,800 ms which was larger than the response frames in the three meaningful studies (800 ms in Study 1 and Study 4; 1,200 ms in Study 2). Hence, the results of Study 3 analyzed and reported in this chapter represent the data trimmed after a response time of 1,400 ms.

In contrast to the strong results of the RT data, error scores significantly discriminated truth tellers from liars only in Study 4 (Figure 3). In studies on the classic CIT, error rates tend to be less valid than RT scores and are typically not consulted as the central differentiating criterion (Seymour & Fraynt, 2009; Verschuere et al., 2010).

Taken together, the four studies build a methodologically homogenous sample appropriate to conduct a meta-analysis. Especially with regard to the promising results of the RT data, it seemed informative to estimate the Credibility-CIT's efficiency to discriminate truth tellers from liars. Hence, a first aim of this meta-analysis was to estimate the overall effect size of this research project. A second goal was to estimate the population effect of the Credibility-CIT's efficiency to discriminate truth tellers from liars.

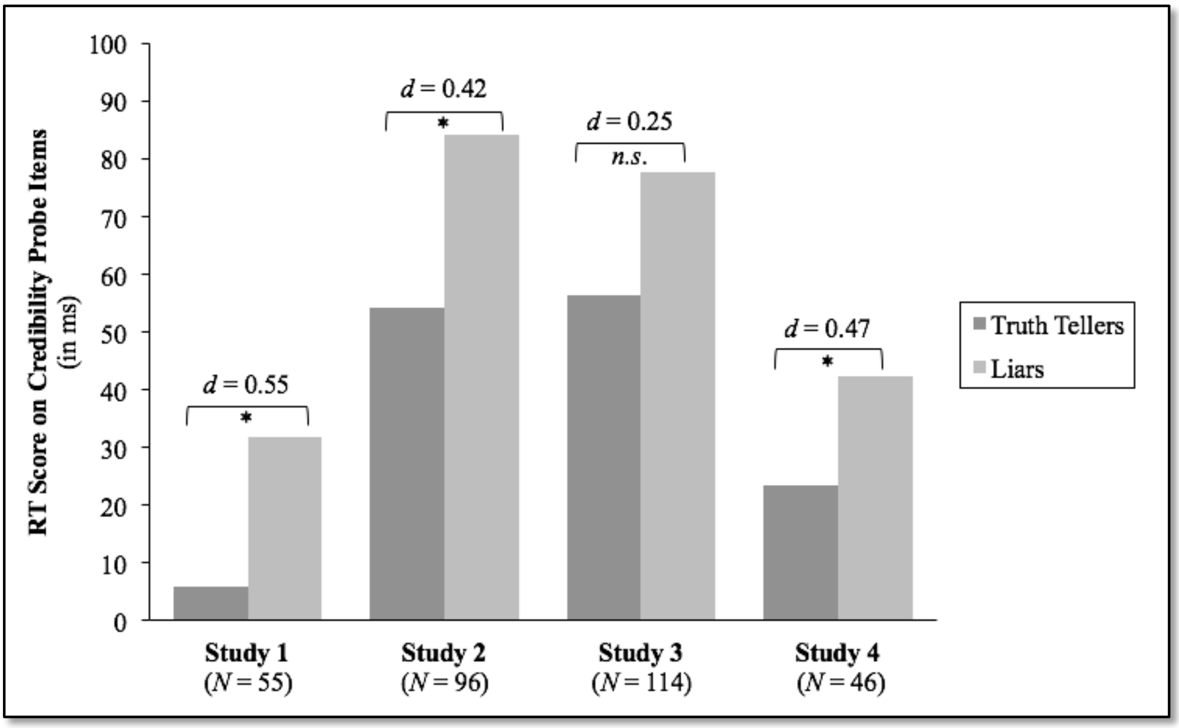

Figure 2. RT scores for truth tellers and liars. RT score = RT on credibility probes - RT on irrelevant items. For the sample sizes, note that Study 4 was a within-subjects design.

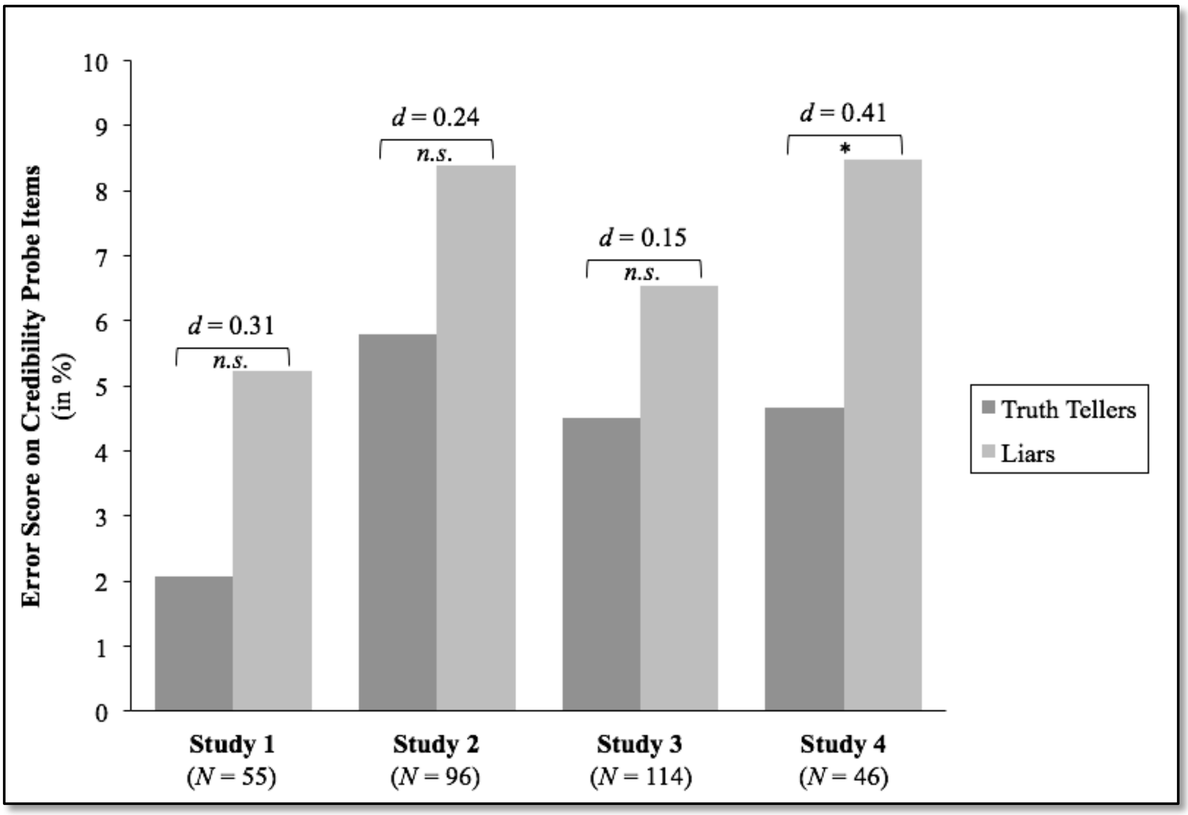

Figure 3. Error scores for truth tellers and liars. Error score = error rates on credibility probes - error rates on irrelevant items. For the sample sizes, note that Study 4 was a within-subjects design.

Method

Inclusion Criteria

This meta-analysis focuses on RTs and error scores as indicators of credibility in a CIT. Since this research project is the first implementation of a Credibility-CIT, only the studies reported above could be included in this meta-analysis (k = 4 studies; n = 311 participants). As this selection makes this sample a comprehensive one, it resolves the need to deal with issues of publication bias (for a recent debate, see Carter, Schönbrodt, Gervais, & Hilgard, 2017). Given the implication of the importance of the response deadline as a moderator of effect, it could be clarifying to examine this variable in a meta-analysis.

However, a sample of k = 4 studies does not allow for that due to a lack of variance.

Effect Size Measures

Cohen's *d* was used as the measure of effect size for the included studies. For the meta-analytical overall effect size, Cohen's *d* tends to overestimate the population effect. Hence, the corrected effect size Hedges' *g* is reported (adapted Hedges' $g = d \times (1 - (3 / (4 \times (N1 + N2) - 9)))$; Hedges, 1981). Calculations were conducted along the IBM SPSS syntax file for meta-analyses provided by Alferes (2003).

Meta-Analytical Procedure

Regarding the meta-analytical method, two main procedures must be distinguished: fixed-effects models and random-effects models. Fixed-effects models assume that there is a fixed underlying effect size in the population. Hence, the studies included in a meta-analysis should represent homogeneous samples of this effect. Consequently, fixed-effect models need to estimate this hypothesized consistency across the included studies. This is usually tested using Cochran's *Q*-test (significant test indicating heterogeneity) or I^2 ($I^2 = 100 \times (Q$ - df) / Q; ranging from 0 = homogenous to 100 = heterogeneous; negative values are set equal to zero; Higgins, Thompson, Deeks, & Altman, 2003). Field and Gillett (2010) argue that human behavior is usually not determined by single causes with fixed effects. Thus, fixed-effects models should not be used to estimate population effects in social sciences. However, fixed-effects models serve well whenever inferences are restricted to the effect size underlying the sampled studies (Hedges & Vevea, 1998). In the present meta-analysis, a fixed-effects model is used for conclusions that are restricted to the reported research project.

In contrast, random-effects models assume that effect sizes vary in the population and, hence, between different studies (i.e., are heterogeneous). They allow for unconditional generalizations beyond the studies included in the meta-analysis (Hedges & Vevea, 1998). A random-effects model was calculated to estimate the overall effect size or population effect of the Credibility-CIT.

Power Analyses

Comparing the four studies meta-analytically yields information about the Credibility-CIT's overall effect size. This, in turn, offers the opportunity to determine each study's statistical power to detect this effect with the given sample sizes. In terms of replicability, the statistical power indicates whether it is reasonable to expect all four studies to provide significant results at all. An underpowered test might be an explanation for the insignificant results found in Study 3. Power analyses were conducted with the software G*Power 3 provided by Faul, Erdfelder, Lang, and Buchner (2007).

Results

RT Score Effect

Estimating the RT effect within this research project, the fixed-effects model yielded a moderate overall effect size of g = 0.40 (95% CI [0.19, 0.61]) for discriminating truth tellers from liars using the Credibility-CIT's RT score. Figure 4 displays the corresponding forest plot. As expected, Cochran's test of heterogeneity was not significant (Q = 1.01, p = .80) and I^2 was negative (I^2 = -197,03), indicating that the included studies are nearly perfectly homogenous.

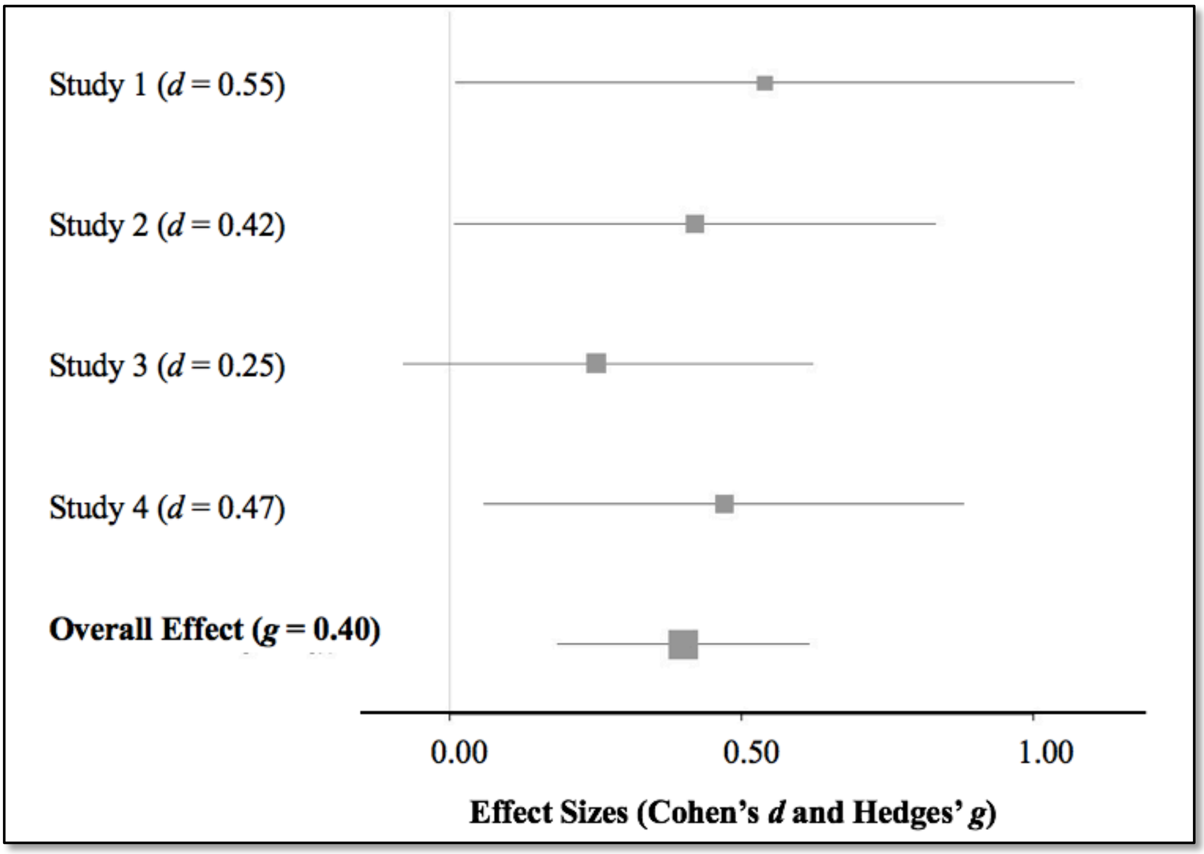

Figure 4. Forest plot of the fixed-effect model's effect sizes for discriminating truth tellers from liars using the RT score. Square size indicates the weighting of each effect size. Line length corresponds to the 95% confidence interval.

To estimate the population effect, a random-effects model was calculated. This indicated an effect size of $g = 0.34$. The 95% confidence interval yielded an effect that ranged from a small to medium effect size [0.25, 0.44].

Error Score Effect

The fixed-effects model indicated that the overall effect size of this research project was small for error scores ($g = 0.26$) and the 95% confidence interval did not include a null effect (95% CI [0.05, 0.47]). Moreover, the four studies built a homogenous subset ($Q = 0.92$, $p = .82$; $I^2 = -226{,}09$). The random-effects model yielded a small effect size for the estimated population effect (*g* = 0.20; 95% CI [0.12, 0.30]).

Power Analyses for RT Scores

Each study's power to detect the Credibility-CIT's overall effect size found in the meta-analysis (*g* = 0.40) was calculated. As evident in Figure 5, detecting an effect of *d* = 0.40 with a typically demanded power of .80 requires a sample size of *N* = 200 participants in a between-group design (i.e., *n* = 100 liars and *n* = 100 truth tellers). As for the studies conducted in this research project, Study 1 had a power of 31% to detect the overall effect of *d* = 0.40 (including *N* = 55 participants). The probability of detecting the effect in Study 2 (*N* = 96) and Study 3 (*N* = 114) were around chance level (1 - β = 49% and 56%, respectively). In contrast, the within-subjects design used in Study 4 had a power of 76% (*N* = 46, within-participants design).

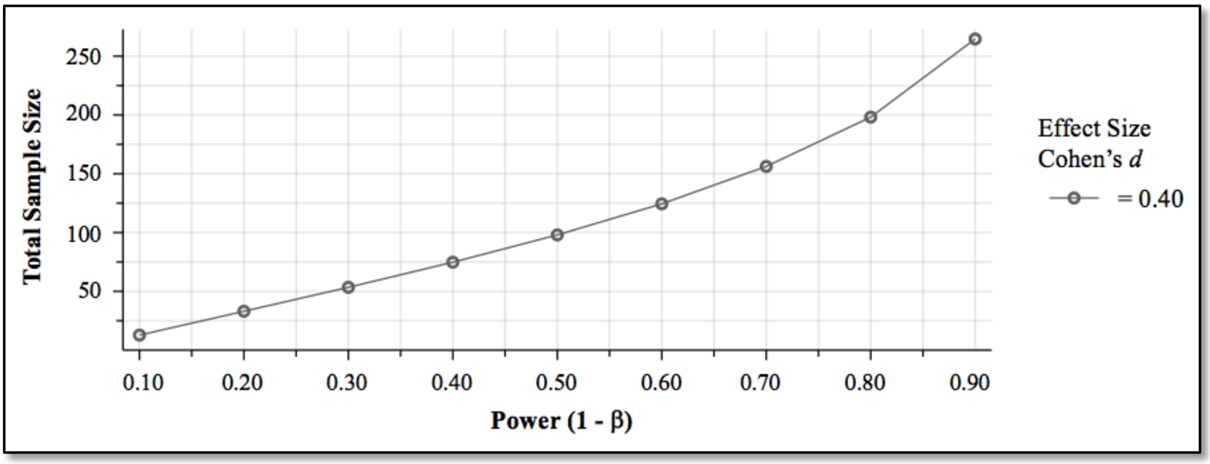

Figure 5. Power analysis for an independent-samples *t*-test using an effect size of *d* = 0.40. α = .05. Graph calculated with and adopted from G*Power (Faul et al., 2007).

Discussion

According to Hedges and Vevea (1998), fixed-effects models only allow for generalizing within the studies included in the meta-analysis. Hence, it can be concluded that the four studies presented within this research project yielded a moderate overall effect in discriminating truth tellers from liars using a RT-based Credibility-CIT (g = .40). Moreover, the tests of heterogeneity indicated that there are no genuine differences underlying the effects found in these four studies. Random-effects models allow for unconditional inferences beyond the studies included (Field & Gillett, 2010; Hedges & Vevea, 1998). The Credibility-CIT's estimated population effect was small (g = 0.34). Regarding error scores, both fixed- and random-effects models yielded small effect sizes (g = 0.26 and g = 0.20, respectively), but the confidence intervals indicated that the effect does not equal zero. Based on the Credibility-CIT's estimated overall effect, the power analyses implied that Study 1, Study 2, and Study 3 were underpowered. This gives a further reason to assume why one out of the four studies conducted was insignificant. Overall, the meta-analysis indicated that the Credibility-CIT has a meaningful potential to discriminate truth-tellers from liars using an RT-based indirect measure.

General Discussion

This study aimed to develop the first RT-based paradigm for credibility assessment. In four studies, participants experienced certain unpleasant actions (truth tellers) or invented a story about these events (liars). During the response time paradigm, called the Credibility-CIT, participants were confronted with personalized items that were phrases from their own statements about the alleged events (credibility probe items). Credibility probes were mixed amongst several irrelevant items and target items (the latter of which were memorized by participants before the test) that served to draw participants' attention. RTs and error rates to these three item categories were recorded. The Credibility-CIT proved to significantly discriminate truth tellers from liars using RTs.

Summary of the Main Findings

RT score discriminated truth tellers from liars. To calculate the RT score, the mean RT to irrelevant items was subtracted from the mean RT to probe items. In three studies, liars had a higher RT score than truth tellers (i.e., liars reacted more slowly than truth tellers to their customized credibility probes than to irrelevant items) with satisfactory effect sizes ($d_{\text{Study 1}} = 0.55$, $d_{\text{Study 2}} = 0.42$, $d_{\text{Study 4}} = 0.47$). In Study 3, the effect was smaller ($d_{\text{Study 3}} = 0.25$) and did not reach significance but yet descriptively resembled the three other studies. Moreover, the power analysis indicated that it is not unlikely to find one insignificant result out of four studies given their small power to find the Credibility-CIT effect at all ($31\% \leq 1 - \beta \leq 76\%$). Subsuming the four studies, the meta-analysis indicated that the RT score of the Credibility-CIT developed in this research project discriminates truth tellers from liars with an overall effect size of $g = 0.40$.

The Credibility-CIT's effect size was therefore smaller than the large effects found in the classic RT-based CIT (*d* = 1.97; Verschuere et al., 2010). The main reason is that the classic CIT compares an experimental group (knowing participants) with an uninformed control group, whereas the Credibility-CIT compares two knowing groups (liars versus truth tellers). In the Credibility-CIT, the comparison between an experimental group and an uninformed control group indicated large effects, too (e.g., *d* = 1.14 between liars and the control group in Study 4). However, this is not the distinction of interest in credibility assessment.

The comparisons with the control group (Study 2, Study 3, and Study 4) indicated that uninformed participants showed an unsystematic RT pattern in the Credibility-CIT. Hence, the probe items selected from the participants' testimonies did not have any characteristics that set them apart from irrelevant items (i.e., in wording or syntax). The only variability is the semantic content – a feature that only knowing participants can identify as task-relevant. This is a fundamental methodological prerequisite for the Credibility-CIT's specificity, i.e., the correct classification of uninformed innocents.

Error scores were less valid. Participants had to deny any knowledge about their personalized credibility probe items. Confirming knowledge by erroneously pressing the "I know it" button and responses reaching the response deadline were counted as errors. Error scores were comparable in all four studies (see Table 2, Table 3, Table 5, and Table 8) ranging between 2-5% for truth tellers and 5-8% for liars. In Study 1, Study 2, and Study 3, error scores did not differ between the experimental groups, which is also a frequent finding in studies on the classic CIT (Seymour & Fraynt, 2009; Verschuere et al., 2010). In Study 4, however, error scores discriminated truth tellers from liars ($d_{\text{Study 4}} = 0.41$). The meta-analytical results indicated an overall effect size of $g = 0.20$ for error scores in the Credibility-CIT.

EEG data did not discriminate truth tellers from liars. Study 4 included assessment of participants' neuronal activity using an EEG-based Credibility-CIT. However, P300 amplitudes did not differ between truth tellers and liars. Probe items did not induce a meaningful P300 component at all, but rather showed the same amplitude as irrelevant items. Based on the absence of this basic CIT effect, a P300 difference between liars and truth tellers could not be expected in this study. Overall, conclusions about the EEG-based Credibility-CIT cannot be drawn from the present study.

Cognitive functions and personality traits did not moderate the Credibility-CIT effect. Regarding the influence of cognitive functions, several previous studies have emphasized the role of executive control (i.e., working memory, inhibitory control, and task switching) for successful deception (Christ et al., 2009; Debey, De Houwer, & Verschuere, 2014; Debey, De Schryver, Logan, Suchotzki, & Verschuere, 2015; Visu-Petra et al., 2012). However, responses in the Credibility-CIT were not related to intelligence or working memory (Study 2, Study 3, and Study 4). The only factor associated with an orienting towards lie-related items was accurate responding in the d2 test (Study 4). Additionally, the Credibility-CIT effect was unrelated to most personality traits (Study 4). An exception is the weaker detection efficiency of the Credibility-CIT for more psychopathic participants. However, regarding the low internal consistency of the psychopathy scale, a further examination of this association is necessary.

Knowledge probes indicated a successful manipulation check. For manipulation checking and in order to replicate findings of the classic RT-CIT, knowledge probes were added as a second probe category in Study 1 and Study 2. Truth tellers reacted more slowly to knowledge probes (as compared to irrelevant items) than did the uninformed groups (liars and the control group). This is consistent with both the hypothesis and the orienting response theory as well as with previous CIT studies (Hu et al., 2013; Noordraven & Verschuere, 2013; Seymour & Kerlin, 2008; Seymour et al., 2000; Suchotzki, Verschuere, et al., 2015; Verschuere et al., 2010, 2015; G. Visu-Petra et al., 2012). However, for the purpose of credibility assessment, knowledge probes were not essential and therefore this item category was omitted in Study 3 and Study 4.

Theoretical Framework for the Credibility-CIT

Based on the results cited above, it is to conclude that liars show stronger orienting responses and are less able to inhibit deceitful responses to credibility probes than truth tellers do. The central question is: Why do lie-related probes have a higher signal value than truth-related stimuli? Explanations for these findings can be drawn from CIT studies on informed innocent participants and on intended mock crimes (e.g., Ben-Shakhar et al., 1999; Bradley et al., 1996; Elaad, 2009; Gamer, 2010; Gamer et al., 2010; Meijer et al., 2010; Meixner & Rosenfeld, 2014; Suchotzki et al., 2013; Winograd & Rosenfeld, 2014).

Gamer (2010) names two main arguments for explaining the different CIT effects in participants who committed a mock crime (guilty participants) as compared to innocent subjects that are solely informed about events (informed innocent participants) or persons who intend to commit a mock crime. Gamer states that differences in encoding of crime-related information and differences in participants' motivation to avoid detection build the core differentiation criteria between the experimental groups. It is reasonable that encoding and emotional-motivational factors also account for the distinction between truth tellers and liars in the Credibility-CIT (Table 12).

Table 12

Differences in Encoding and Motivation for different CIT Participant Groups.

	Encoding	**Motivation to avoid detection**
Classic CIT		
Innocent participants	No encoding of task-relevant information at all	None
Guilty participants	Handling task-related objects, incidental and shallow encoding	High
Informed innocents	Gaining information incidentally, semantic processing, shallow encoding	Moderate
Participants intending to commit a mock crime	Dealing intentionally and mentally with the task, using semantic memory, elaborate encoding	Moderate
Credibility-CIT		
Truth tellers	Handling task-related objects, incidental and shallow encoding	Low
Liars	Dealing intentionally and mentally with the task, using semantic memory, elaborate encoding	High

Elaboration of Encoding during the Interview

According to the cognitive load theory (Volbert & Steller, 2014; Vrij & Mann, 2006), inventing the story and mastering the interview are both highly demanding for liars. Liars must intentionally memorize information on the alleged event and associate the memory content to cognitive scripts about similar occurrences. This utilization of semantic memory content represents an elaborate rehearsal process, which is associated with elaborate encoding (Craik &

Lockhart, 1972). Besides the effort to invent a story, the cognitive load theory states that finally telling the lie during the interview is also more demanding for liars than telling the truth is for truth tellers. During the interrogation, an episodic memory trace is formed which comprises the content of the story and the effort of giving the statement. Hence, the interview serves to consolidate the relevant information again. The fabrication, including the careful selection of central details, and the episodic memory of the challenging interview should lead to an elaborately encoded memory for the person telling the lie. When confronted with the corresponding items in the CIT, the phrases have a high familiarity. This, in turn, induces a high signal value and elicits a strong orienting response. Inhibiting responses to these items is challenging.

According to the enactment effect (Cohen, 1989) one might expect participants who actually performed the tasks – those in the truth telling condition in these studies – to encode these events elaborately and remember them well. However, in Study 1 truth tellers did not react differentially to credibility probes at all. One reason for this might be a lack of memory about the events, which results in poorer retrieval (Bradley et al., 2011). However, it seems unlikely that most participants in the truth-telling group forgot about experiencing these unconventional tasks. Moreover, participants' systematic reactions to the knowledge probes in Study 1 and Study 2 showed that truth tellers remembered the scenario sufficiently. Hence, a lack of memory for truth tellers is an unlikely reason for the differences between the experimental groups. An explanation of this difference might lie in the relevance of the interview. For truth tellers, the interview was neither cognitively demanding (as they just retold previous experiences) nor emotionally challenging because concerns about incredibility were certainly low in this laboratory setting. Hence, the interrogation might play a minor role in consolidation compared to the enactment during which they encoded the stimuli only incidentally. Accordingly, the credibility probes have only a moderate signal value for truth tellers. After all, encoding is less elaborate and consequently credibility probes have a lower significance for truth tellers than for liars.

Another relevant feature related to encoding is the time between encoding and CIT. The interaction of depth of encoding and the interval between encoding and RT-CIT was emphasized in a study by Seymour and Fraynt (2009). They determined that after a one-week interval the RT difference between elaborately and shallowly encoded information was larger than in immediate assessment (see Figure 1, p. 25). In the present Credibility-CIT studies, a one-week interval between the encoding of information (fabricating/experiencing and giving the interview) and the Credibility-CIT was used. Provided that encoding differences drive the varying Credibility-CIT effects between truth tellers and liars, this one-week interval might have enhanced the paradigm's detection accuracy.

Emotional-Motivational Factors

Our "default mode" is to tell the truth; lying is a deviation from this mode and associated with emotional arousal (Elaad & Ben-Shakhar, 1989; Verschuere & Ben-Shakhar, 2011). The effect of emotional-motivational factors can be integrated into the theoretical framework. It is conceivable that liars had a higher motivation to avoid detection than truth tellers had. The liars may have feared being detected twice – in the interview and in the Credibility-CIT. Effort to avoid detection might additionally enhance the noteworthiness of the credibility probe items for liars and, hence, lead to a stronger orienting response. This pattern corresponds to the relevance of emotional arousal for memory processes: Events with emotional relevance lead to higher stress hormone levels and, hence, to stronger memory traces (Cahill, Prins, Weber, & McGaugh, 1994). Both strong orienting and elaborate encoding should make it difficult for participants to quickly classify probes taken from one's previous lies as "unfamiliar" in the Credibility-CIT.

Retrieval of Information during the Credibility-CIT

In addition to the encoding and emotional-motivational explanations put forth by Gamer (2010), the ADCAT (Walczyk et al., 2014) has an explanatory value for understanding patterns of response to the Credibility-CIT. According to ADCAT and other authors (Debey et al., 2014; Verschuere & Shalvi, 2014), every question by default activates the true answer in long-term memory – irrespective of the individual's intention to answer honestly or deceptively. The activated true response automatically enters working memory. In the Credibility-CIT, the to-be answered question is "Do you know this term?". Following this line of reasoning, when confronted with a credibility probe item, for both truth tellers and liars the automatic, true response is to affirm the question, whereas the task requires denying this knowledge. However, the following cognitive processes differ crucially between truth tellers and liars. For liars, the activated truth involves memories about the process of inventing the story and lying in the face of the interviewer during the testimony – the elaborately encoded information. This large amount of information is kept active and has to be processed during all consecutive ADCAT steps; it serves as a retrieval cue in the construction component and has to be inhibited in the action component. Moreover, the evaluation of costs and benefits of lying (decision component) is also more time-consuming for liars as they have a higher motivation to avoid being debunked. These additional cognitive processes add more to RTs for liars than they do for truth tellers.

Limitations and Future Directions

The present project focused on foundational research for the development of the Credibility-CIT. The central aim was the development of the basic paradigm and the replication of the first results. Consequently, there are numerous topics further research should address. First and foremost, the theoretical framework must be specified. The effects of depth of encoding, memory processes, and emotional-motivational factors on the Credibility-CIT are first explanatory approaches. Future studies should systematically manipulate and elucidate these effects.

There is also a need for methodological optimization. Given that this research project was the first attempt to develop an RT-based paradigm for credibility assessment, the effect sizes were satisfying. However, reliabilities were low and an overall effect of $g = 0.40$ is not large enough to use the Credibility-CIT in its present form for single-case diagnostics. Luckily, there are ample opportunities for methodological optimization. For example, results from Study 3 indicated that a systematic evaluation of the impact of response deadlines is a promising avenue to pursue to increase detection efficiency. Not only the Credibility-CIT but also the classic CIT would benefit from further research on the impact of response deadlines.

Varying the number of probe items or manipulating the probe-irrelevant proportion might also have potential to improve the Credibility-CIT's detection efficiency. Consistent with the classical test theory, increasing the number of probe items descriptively enhances the Credibility-CIT's reliability ($r_{tt} = .27$ using 10 probe items in Study 1 versus $.28 \leq r_{tt} \leq .55$ using 15 probes in Studies 2, 3 and 4). However, results on the Credibility-CIT and the classic CIT imply that lengthening the test might be a threat to its validity since habituation and the practicing of countermeasures can occur (Suchotzki, Verschuere, et al., 2015). A systematic evaluation is necessary. Ideally, further strategies than test lengthening should be evaluated to improve the CIT's reliability and consequently its validity. Most central for that purpose is the extraction of meaningful phrases from each participant's testimony that serve as items for the Credibility-CIT.

Testing the Credibility-CIT with further peripheral or central physiological measures could prove informative given that the classic CIT is associated with significant physiological indicators. Combining behavioral (RT) data and physiological indicators would enhance the Credibility-CIT's validity by offering a multi-method approach. In studies on the classic CIT, ERPs (especially P300 amplitudes) and skin conductance yield large effect sizes (Meijer et al., 2014). As discussed above, it remains unclear why the ERP-based Credibility-CIT did not yield meaningful P300 amplitudes for credibility probe items in Study 4.

Regarding studies on further fields of application, a potentially informative research question is whether the Credibility-CIT would yield meaningful results in contexts of suggestion. Suggestions can induce pseudo memories – a circumstance in which persons subjectively believe a fictitious event took place. Interestingly, participants with pseudo memories resemble truth tellers in terms of their motivational and affective states; both groups think they experienced the event and believe their own story. In contrast, the encoding of the event-related information is similar to that of liars as both, liars and people with pseudo-memories, cannot rely on episodic memory. Since recent diagnostic methods such as the SVA cannot reliably discriminate true and suggested statements (Volbert & Steller, 2014), applying the Credibility-CIT for this purpose would represent a major advance for credibility assessment. Admittedly, predicting the direction and size of this effect is challenging with regard to the very basic knowledge the present four studies supply regarding the Credibility-CIT.

However, after refining the theoretical framework and implementing the above-named methodological optimizations, the research potential seems promising.

After optimizing and replicating the results on the Credibility-CIT in further laboratory studies, field studies would give important insight in the practicability of the paradigm. It is conceivable that the elaboration of encoding as well as the emotional and motivational state differ substantially in laboratory versus real-life circumstances. In real-life scenarios, truth tellers should be more emotionally involved and liars should engage in inventing an even more coherent story. Both factors should enhance the signal value of the credibility probe items. It remains unclear whether a more realistic setting would strengthen or reduce the effects found in laboratory work such as that presented in this research report.

Regarding the practical applicability of the Credibility-CIT, the finding that the Credibility-CIT works in the same way for victims and offenders (Study 1 and Study 2) offers an additional area of application, namely the detection of false confessions. Suspects giving false confessions can be considered parallel to those participants who were "offenders" during the initial experience and then lied during their testimony. Suspects who constructed a complex lie about having committed a crime would be expected to undergo similar cognitive and motivational processes as persons who pretend to be the victim of an offense. Hence, the theoretical explanations consulted for the Credibility-CIT also account for a False-Confessions-CIT.

Conclusions

The Credibility-CIT is a promising paradigm to complement classic approaches for credibility assessment, like the SVA. From a practitioner's perspective, the Credibility-CIT is economical and easy to apply subsequent to the SVA interview without influencing the SVA's validity at all. The medium effect size found in Study 1 ($d = 0.55$) indicates the potential of this paradigm.

Even though the current detection efficiency is not sufficient for single-case assessment, with further optimization it might reach or even outperform common witness-related evidence in court, like the CBCA (g = 0.97, Oberlader et al., 2016) or the rather unreliable eyewitnesses (correct classification rate of 50%; Steblay, Dysart, Fulero, & Lindsay, 2001). Therefore, the Credibility-CIT can contribute to the incremental validity of current credibility assessment procedures.

References

Alferes, V. R. (2003). Meta-analysis: Fixed and random effects models [SPSS Syntax File]. Retrieved May 12, 2017, from http://spsstools.net/en/syntax/syntax-index/meta-analysis/meta-analysis-fixed-and-random-effects-models

Allen, J. J. B. (2011). Clinical applications of the concealed information test. In B. Verschuere, G. Ben-Shakhar, & E. Meijer (Eds.), *Memory detection: Theory and application of the concealed information test* (pp. 231–252). Cambridge: University press.

Allen, J. J. B., & Movius, H. L. (2000). The objective assessment of amnesia in dissociative identity disorder using event-related potentials. *International Journal of Psychophysiology*, *38*, 21–41. https://doi.org/10.1016/S0167-8760(00)00128-8

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

Atkinson, R. C., & Shiffrin, R. M. (1968). Human memory: A proposed system and its control processes. In K. W. Spence & J. T. Spence (Eds.), *Psychology of Learning and Motivation* (pp. 89–195). https://doi.org/10.1016/S0079-7421(08)60422-3

Baddeley, A. D. (1992). Working memory. *Science*, *255*, 556–559. https://doi.org/10.1126/science.1736359

Baddeley, A. D. (2000). The episodic buffer: a new component of working memory? *Trends in Cognitive Sciences*, *4*, 417–423.

Baddeley, A. D., & Hitch, G. (1974). Working memory. In G. H. Bower (Ed.), *Psychology of learning and motivation* (pp. 47–89). New York: Academic Press.

Bauer, R. M. (1984). Autonomic recognition of names and faces in prosopagnosia: A neuropsychological application of the guilty knowledge test. *Neuropsychologia*, *22*, 457–469. https://doi.org/10.1016/0028-3932(84)90040-X

Ben-Shakhar, G. (2011). Countermeasures. In B. Verschuere, G. Ben-Shakhar, & E. Meijer (Eds.), *Memory detection: Theory and application of the concealed information test* (pp. 200–214). Cambridge: University press.

Ben-Shakhar, G., & Elaad, E. (2002). Effects of questions' repetition and variation on the efficiency of the guilty knowledge test: A reexamination. *Journal of Applied Psychology*, *87*, 972–977. https://doi.org/10.1037//0021-9010.87.5.972

Ben-Shakhar, G., & Elaad, E. (2003). The validity of psychophysiological detection of information with the guilty knowledge test: A meta-analytic review. *Journal of Applied Psychology*, *88*, 131–151. https://doi.org/10.1037/0021-9010.88.1.131

Ben-Shakhar, G., Gronau, N., & Elaad, E. (1999). Leakage of relevant information to innocent examinees in the GKT: An attempt to reduce false-positive outcomes by introducing target stimuli. *Journal of Applied Psychology*, *84*, 651–660. https://doi.org/10.1037/0021-9010.84.5.651

Bradley, M. T., Barefoot, & Arsenault, A. M. (2011). Leakage of information to innocent suspects. In B. Verschuere, G. Ben-Shakhar, & E. Meijer (Eds.), *Memory detection: Theory and application of the concealed information test* (pp. 187–199). Cambridge: University press.

Bradley, M. T., MacLaren, V. V., & Carle, S. B. (1996). Deception and nondeception in guilty knowledge and guilty actions polygraph tests. *Journal of Applied Psychology*, *81*, 153–160. https://doi.org/10.1037/0021-9010.81.2.153

Brickenkamp, R. (1994). *Test d2: Aufmerksamkeits-Belastungs-Test [The d2 test of attention]*. Göttingen: Hogrefe.

Cahill, L., Prins, B., Weber, M., & McGaugh, J. L. (1994). β-Adrenergic activation and memory for emotional events. *Nature*, *371*, 702–704. https://doi.org/10.1038/371702a0

Carmel, D., Dayan, E., Naveh, A., Raveh, O., & Ben-Shakhar, G. (2003). Estimating the validity of the guilty knowledge test from simulated experiments: The external validity of mock crime studies. *Journal of Experimental Psychology: Applied*, *9*, 261–269. https://doi.org/10.1037/1076-898X.9.4.261

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies*. Cambridge: University Press.

Carroll, J. B. (2005). The three-stratum theory of cognitive abilities. In D. P. Flanagan & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (pp. 69–76). New York: Guilford Press.

Carter, E., Schönbrodt, F., Gervais, W., & Hilgard, J. (2019). Correcting for bias in psychology: A comparison of meta-analytic methods. *Advances in Methods and Practices in Psychological Science*, *2*, 115–144. https://doi.org/10.1177/2515245919847196

Christ, S. E., Van Essen, D. C., Watson, J. M., Brubaker, L. E., & McDermott, K. B. (2009). The contributions of prefrontal cortex and executive control to deception: Evidence from activation likelihood estimate meta-analyses. *Cerebral Cortex*, *19*, 1557–1566. https://doi.org/10.1093/cercor/bhn189

Cohen, R. L. (1989). Memory for action events: The power of enactment. *Educational Psychology Review*, *1*, 57–80. https://doi.org/10.1007/BF01326550

Craik, F. I. M., & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior*, *11*, 671–684. https://doi.org/10.1016/S0022-5371(72)80001-X

Craik, F. I. M., & Tulving, E. (1975). Depth of processing and the retention of words in episodic memory. *Journal of Experimental Psychology: General*, *104*, 268–294. https://doi.org/10.1037/0096-3445.104.3.268

De Houwer, J. (2008). A structural analysis of indirect measures of attitudes. In J. Musch & K. C. Klauer (Eds.), *The psychology of evaluation: Affective processes in cognition and emotion* (pp. 227–252). Mahwah, N.J: Lawrence Erlbaum Associates.

Debey, E., De Houwer, J., & Verschuere, B. (2014). Lying relies on the truth. *Cognition*, *132*, 324–334. https://doi.org/10.1016/j.cognition.2014.04.009

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

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

Elaad, E. (2009). Effects of context and state of guilt on the detection of concealed crime information. *International Journal of Psychophysiology*, *71*, 225–234. https://doi.org/10.1016/j.ijpsycho.2008.10.001

Elaad, E., & Ben-Shakhar, G. (1989). Effects of motivation and verbal response type on psychophysiological detection of information. *Psychophysiology*, *26*, 442–451. https://doi.org/10.1111/j.1469-8986.1989.tb01950.x

Farwell, L. A., & Donchin, E. (1991). The truth will out: Interrogative polygraphy ("lie detection") with event-related brain potentials. *Psychophysiology*, *28*, 531–547. https://doi.org/10.1111/j.1469-8986.1991.tb01990.x

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. https://doi.org/10.3758/BF03193146

Field, A. P., & Gillett, R. (2010). How to do a meta-analysis. *British Journal of Mathematical and Statistical Psychology*, *63*, 665–694. https://doi.org/10.1348/000711010X502733

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

Gamer, M. (2010). Does the guilty actions test allow for differentiating guilty participants from informed innocents? A re-examination. *International Journal of Psychophysiology*, *76*, 19–24. https://doi.org/10.1016/j.ijpsycho.2010.01.009

Gamer, M. (2011). Detecting of deception and concealed information using neuroimaging techniques. In B. Verschuere, G. Ben-Shakhar, & E. Meijer (Eds.), *Memory detection: Theory and application of the concealed information test* (pp. 90–113). Cambridge: University press.

Gamer, M., & Berti, S. (2012). P300 amplitudes in the concealed information test are less affected by depth of processing than electrodermal responses. *Frontiers in Human Neuroscience*, *6*, 1–11. https://doi.org/10.3389/fnhum.2012.00308

Gamer, M., Klimecki, O., Bauermann, T., Stoeter, P., & Vossel, G. (2012). fMRI-activation patterns in the detection of concealed information rely on memory-related effects. *Social Cognitive and Affective Neuroscience*, *7*, 506–515. https://doi.org/10.1093/scan/nsp005

Gamer, M., Kosiol, D., & Vossel, G. (2010). Strength of memory encoding affects physiological responses in the guilty actions test. *Biological Psychology*, *83*, 101–107. https://doi.org/10.1016/j.biopsycho.2009.11.005

Gratton, G., Coles, M. G. H., & Donchin, E. (1983). A new method for off-line removal of ocular artifact. *Electroencephalography and Clinical Neuro-physiology*, *55*, 468–484. https://doi.org/10.1016/0013-4694(83)90135-9

Greenwald, A. G., Nosek, B. A., & Banaji, M. R. (2003). Understanding and using the implicit association test: An improved scoring algorithm. *Journal of Personality and Social Psychology*, *85*, 197–216. https://doi.org/10.1037/0022-3514.85.2.197

Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics*, *6*, 107. https://doi.org/10.2307/1164588

Hedges, L. V., & Vevea, J. L. (1998). Fixed- and random-effects models in meta-analysis. *Psychological Methods*, *3*, 486–504. https://doi.org/10.1037/1082-989X.3.4.486

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

Horn, W. (2003). *Prüfsystem für Schul-und Bildungsberatung für 6. bis 13. Klassen - Revidierte Fassung, PSB-R 6-13 [Testing System for Scholastic and Educational Counseling, Grades 6 to 13 – Revised Version, PSB-R 6-13].* Göttingen: Hogrefe.

Hu, X., Evans, A., Wu, H., Lee, K., & Fu, G. (2013). An interfering dot-probe task facilitates the detection of mock crime memory in a reaction time (RT)-based concealed information test. *Acta Psychologica*, *142*, 278–285. https://doi.org/10.1016/j.actpsy.2012.12.006

Jones, D. N., & Paulhus, D. L. (2014). Introducing the Short Dark Triad (SD3): A brief measure of dark personality traits. *Assessment*, *21*, 28–41. https://doi.org/10.1177/1073191113514105

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

Kreuzpointner, L., Lukesch, H., & Horn, W. (2013). *Leistungsprüfsystem 2 (LPS-2): Manual [Perfomance testing system 2 (LPS-2): Manual]*. Göttingen: Hogrefe.

Lang, F. R., Lüdtke, O., & Asendorpf, J. B. (2001). Testgüte und psychometrische Äquivalenz der deutschen Version des big five inventory (BFI) bei jungen, mittelalten und alten Erwachsenen [Validity and psychometric equivalence of the German version of the Big Five Inventory in young, middle-aged and old adults]. *Diagnostica*, *47*, 111–121. https://doi.org/10.1026//0012-1924.47.3.111

Lubow, R. E., & Fein, O. (1996). Pupillary size in response to a visual guilty knowledge test: New technique for the detection of deception. *Journal of Experimental Psychology: Applied*, *2*, 164–177. https://doi.org/10.1037/1076-898X.2.2.164

Lykken, D. T. (1960). The validity of the guilty knowledge technique: The effects of faking. *Journal of Applied Psychology*, *44*, 258–262. https://doi.org/10.1037/h0044413

Lykken, D. T. (1998). *A tremor in the blood: Uses and abuses of the lie detector*. New York, N.Y.: Plenum Trade.

Magliero, A., Bashore, T. R., Coles, M. G. H., & Donchin, E. (1984). On the dependence of P300 latency on stimulus evaluation processes. *Psychophysiology*, *21*, 171–186. https://doi.org/10.1111/j.1469-8986.1984.tb00201.x

Matsuda, I., Nittono, H., & Allen, J. J. B. (2012). The current and future status of the concealed information test for field use. *Frontiers in Psychology*, *3*, 1–11. https://doi.org/10.3389/fpsyg.2012.00532

McCarthy, G., & Donchin, E. (1981). A metric for thought: A comparison of P300 latency and reaction time. *Science*, *211*, 77–80. https://doi.org/10.1126/science.7444452

Meijer, E. H., Selle, N. K., Elber, L., & Ben-Shakhar, G. (2014). Memory detection with the concealed information test: A meta analysis of skin conductance, respiration, heart rate, and P300 data. *Psychophysiology*, *51*, 879–904. https://doi.org/10.1111/psyp.12239

Meijer, E. H., Smulders, F., Merckelbach, H., & Wolf, A. (2007). The P300 is sensitive to concealed face recognition. *International Journal of Psychophysiology*, *66*, 231–237. https://doi.org/10.1016/j.ijpsycho.2007.08.001

Meijer, E. H., Verschuere, B., & Merckelbach, H. (2010). Detecting criminal intent with the concealed information test. *The Open Criminology Journal*, *3*, 44–47. https://doi.org/10.2174/1874917801003020044

Meixner, J. B., & Rosenfeld, J. P. (2011). A mock terrorism application of the P300-based concealed information test. *Psychophysiology*, *48*, 149–154. https://doi.org/10.1111/j.1469-8986.2010.01050.x

Meixner, J. B., & Rosenfeld, J. P. (2014). Detecting knowledge of incidentally acquired, real-world memories using a P300-based concealed-information test. *Psychological Science*, *25*, 1994–2005. https://doi.org/10.1177/0956797614547278

Miller, J., & Ulrich, R. (2013). Mental chronometry and individual differences: Modeling reliabilities and correlations of reaction time means and effect sizes. *Psychonomic Bulletin & Review*, *20*, 819–858. https://doi.org/10.3758/s13423-013-0404-5

Noordraven, E., & Verschuere, B. (2013). Predicting the sensitivity of the reaction time-based concealed information test. *Applied Cognitive Psychology*, *27*, 328–335. https://doi.org/10.1002/acp.2910

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

Osugi. (2011). Daily application of the concealed information test: Japan. In B. Verschuere, G. Ben-Shakhar, & E. Meijer (Eds.), *Memory detection: Theory and application of the concealed information test* (pp. 253–275). Cambridge: University press.

Prabhakaran, R., Kraemer, D. J. M., & Thompson-Schill, S. L. (2011). Approach, avoidance, and inhibition: Personality traits predict cognitive control abilities. *Personality and Individual Differences*, *51*, 439–444. https://doi.org/10.1016/j.paid.2011.04.009

Rassin, E. (2000). Criteria based content analysis: The less scientific road to truth. *Expert Evidence*, *7*, 265–278. https://doi.org/10.1023/ A:1016627527082

Rönspies, J. (in prep.). *Eine empirische Untersuchung zur Validität ausgewählter Glaubhaftigkeitsmerkmale bei der kriterienorientierten Inhaltsanalyse von Zeugenaussagen [An empirical study on the validity of selected criteria in the criteria-based content analysis of witnesses' testimonies]* (Doctoral Dissertation). University of Bonn, Germany.

Rosenfeld, J. P. (2011). P300 in detecting concealed information. In B. Verschuere, G. Ben-Shakhar, & E. Meijer (Eds.), *Memory detection: Theory and application of the concealed information test* (pp. 63–89). Cambridge: University press.

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

Seymour, T. L., & Fraynt, B. R. (2009). Time and encoding effects in the concealed knowledge test. *Applied Psychophysiology and Biofeedback*, *34*, 177–187. https://doi.org/10.1007/s10484-009-9092-3

Seymour, T. L., & Kerlin, J. R. (2008). Successful detection of verbal and visual concealed knowledge using an RT-based paradigm. *Applied Cognitive Psychology*, *22*, 475–490. https://doi.org/10.1002/acp.1375

Seymour, T. L., Seifert, C. M., Shafto, M. G., & Mosmann, A. L. (2000). Using response time measures to assess “guilty knowledge”. *Journal of Applied Psychology*, *85*, 30–37. https://doi.org/10.1037/0021-9010.85.1.30

Sokolov, E. N. (1963). *Perception and the conditioned reflex*. New York, N.Y.: Macmillian.

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

Sporer, S. L. (2016). Deception and cognitive load: Expanding our horizon with a working memory model. *Frontiers in Psychology*, *7*, 420. https://doi.org/10.3389/fpsyg.2016.00420

Steblay, N., Dysart, J., Fulero, S., & Lindsay, R. C. (2001). Eyewitness accuracy rates in sequential and simultaneous lineup presentations: a meta-analytic comparison. *Law and Human Behavior*, *25*, 459–473.

Suchotzki, K., Crombez, G., Smulders, F. T. Y., Meijer, E., & Verschuere, B. (2015). The cognitive mechanisms underlying deception: An event-related potential study. *International Journal of Psychophysiology*, *95*, 395–405. https://doi.org/10.1016/j.ijpsycho.2015.01.010

Suchotzki, K., Verschuere, B., Crombez, G., & De Houwer, J. (2013). Reaction time measures in deception research: Comparing the effects of irrelevant and relevant stimulus–response compatibility. *Acta Psychologica*, *144*, 224–231. https://doi.org/10.1016/j.actpsy.2013.06.014

Suchotzki, K., Verschuere, B., Peth, J., Crombez, G., & Gamer, M. (2015). Manipulating item proportion and deception reveals crucial dissociation between behavioral, autonomic, and neural indices of concealed information. *Human Brain Mapping*, *36*, 427–439. https://doi.org/10.1002/ hbm.22637

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

Tewes, U. (1991). *Hamburg-Wechsler Intelligenztest für Erwachsene Revision [Wechsler adult intelligence scale - revised]*. Bern: Hans Huber.

Undeutsch, U. (1989). The development of statement reality analysis. In J. C. Yuille (Ed.), *Credibility assessment* (pp. 101–119). https://doi.org/10. 1007/978-94-015-7856-1_6

Verschuere, B., & Ben-Shakhar, G. (2011). Theory of the concealed information test. In B. Verschuere, G. Ben-Shakhar, & E. Meijer (Eds.), *Memory detection: Theory and application of the concealed information test* (pp. 128–148). Cambridge: University press.

Verschuere, B., Crombez, G., De Clercq, A., & Koster, E. H. W. (2004). Autonomic and behavioral responding to concealed information: Differentiating orienting and defensive responses. *Psychophysiology*, *41*, 461–466. https://doi.org/10.1111/j.1469-8986.00167.x

Verschuere, B., Crombez, G., De Clercq, A., & Koster, E. H. W. (2005). Psychopathic traits and autonomic responding to concealed information in a prison sample. *Psychophysiology*, *42*, 239–245. https://doi.org/10.1111/j.1469-8986.2005.00279.x

Verschuere, B., Crombez, G., Degrootte, T., & Rosseel, Y. (2010). Detecting concealed information with reaction times: Validity and comparison with the polygraph. *Applied Cognitive Psychology*, *24*, 991–1002. https://doi.org/10.1002/acp.1601

Verschuere, B., Crombez, G., Koster, E. H. W., & De Clercq, A. (2007). Antisociality, underarousal and the validity of the concealed information polygraph test. *Biological Psychology*, *74*, 309–318. https://doi.org/10.1016/j.biopsycho.2006.08.002

Verschuere, B., Crombez, G., Koster, E. H. W., & Uzieblo, K. (2006). Psychopathy and physiological detection of concealed information: A review. *Psychologica Belgica*, *46*, 99–116. https://doi.org/10.5334/pb-46-1-2-99

Verschuere, B., Kleinberg, B., & Theocharidou, K. (2015). RT-based memory detection: Item saliency effects in the single-probe and the multiple-probe protocol. *Journal of Applied Research in Memory and Cognition*, *4*, 59–65. https://doi.org/10.1016/j.jarmac.2015.01.001

Verschuere, B., & Meijer, E. H. (2014). What's on your mind? Recent advances in memory detection using the concealed information test. *European Psychologist*, *19*, 162–171. https://doi.org/10.1027/1016-9040/a000194

Verschuere, B., & Shalvi, S. (2014). The truth comes naturally! Does it?. *Journal of Language and Social Psychology*, *33*, 417–423. https://doi.org/10.1177/0261927X14535394

Visu-Petra, G., Miclea, M., & Visu-Petra, L. (2012). Reaction time-based detection of concealed information in relation to individual differences in executive functioning. *Applied Cognitive Psychology*, *26*, 342–351. https://doi.org/10.1002/acp.1827

Visu-Petra, L., Jurje, O., Ciornei, O., & Visu-Petra, G. (2016). Can you keep a secret? Introducing the RT-based concealed information test to children. *Psychology, Crime & Law*, *22*, 276–301. https://doi.org/10.1080/1068316X.2015.1109085

Volbert, R. (2010). Aussagepsychologische Begutachtung [Credibility assessment]. In R. Volbert & K.-P. Dahle (Eds.), *Forensisch-psychologische Diagnostik im Strafverfahren [Forensic-psychological assessment in criminmal proceedings]* (pp. 18–66). Göttingen: Hogrefe.

Volbert, R., & Steller, M. (2014). Is this testimony truthful, fabricated, or based on false memory? Credibility assessment 25 years after Steller and Köhnken (1989). *European Psychologist*, *19*, 207–220. 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*, 3–41. https://doi.org/10.1037/1076-8971.11.1.3

Vrij, A., Fisher, R., Mann, S., & Leal, S. (2006). Detecting deception by manipulating cognitive load. *Trends in Cognitive Sciences*, *10*, 141–142. https://doi.org/10.1016/j.tics.2006.02.003

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

Waechter, S., & Stolz, J. A. (2015). Trait Anxiety, State Anxiety, and Attentional Bias to Threat: Assessing the Psychometric Properties of Response Time Measures. *Cognitive Therapy and Research*, *39*(4), 441–458. https://doi.org/10.1007/s10608-015-9670-z

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

Walczyk, J. J., Mahoney, K. T., Doverspike, D., & Griffith-Ross, D. A. (2009). Cognitive lie detection: Response time and consistency of answers as cues to deception. *Journal of Business and Psychology*, *24*, 33–49. https://doi.org/10.1007/s10869-009-9090-8

Walczyk, J. J., Roper, K. S., Seemann, E., & Humphrey, A. M. (2003). Cognitive mechanisms underlying lying to questions: Response time as a cue to deception. *Applied Cognitive Psychology*, *17*, 755–774. https://doi.org/10.1002/acp.914

Walczyk, J. J., Schwartz, J. P., Clifton, R., Adams, B., Wei, M., & Zha, P. (2005). Lying person-to-person about life events: A cognitive framework for lie detection. *Personnel Psychology*, *58*, 141–170. https://doi.org/10.1111/j.1744-6570.2005.00484.x

Winograd, M. R., & Rosenfeld, J. P. (2014). The impact of prior knowledge from participant instructions in a mock crime P300 concealed information test. *International Journal of Psychophysiology*, *94*, 473–481. https://doi.org/10.1016/j.ijpsycho.2014.08.002

Appendix

Appendix A: Teilnehmerinformation für lügende Teilnehmer, Rolle „Opfer / Akteur“ [Participant information for liars, role: victim]

Appendix B: Teilnehmerinformation für lügende Teilnehmer, Rolle „Täter / Versuchsleiter“ [Participant information for liars, role: offender]

Appendix C: Teilnehmerinformation für wahraussagende Teilnehmer, Rolle „Opfer / Akteur“ [Participant information for truth tellers, role: victim]

Appendix D: Teilnehmerinformation für wahraussagende Teilnehmer, Rolle „Täter / Versuchsleiter“ [Participant information for truth tellers, role: offender]

Appendix E: Einverständniserklärung der Schüler [Informed consent]

Appendix F: Elterninformation und Elterneinverständnis [Informed consent for participant’s parents]

Appendix A

Teilnehmerinformation für lügende Teilnehmer, Rolle „Opfer / Akteur“

Liebe(r) Interessent(in),

universitätbonn

vielen Dank für dein Interesse an unserer Studie!

In der Studie geht es um Zeugenaussagen. Manchmal kommt es vor, dass Zeugen bei der Polizei oder vor einem Richter lügen. Wir wollen deshalb erforschen, wie gut Personen Aussagen erfinden können. Wir untersuchen, ob Personen, die sich ein Erlebnis ausdenken eine genau so gute Zeugenaussage machen können, wie Personen, die das Erlebnis wirklich hatten. Wenn du mit der Teilnahme einverstanden bist, werden wir dich deshalb bitten, eine gewisse Geschichte zu erfinden. Danach sollst du zu diesem angeblichen Erlebnis eine möglichst überzeugende „Zeugenaussage“ machen. Du hast 30 Minuten Zeit, dir die Geschichte auszudenken. Das Erlebnis, das du erfinden sollst, hat folgende Rahmenhandlung:

Stell dir vor, du würdest mit einer anderen Person hier in die Schule kommen, um an der Studie teilzunehmen. Der andere würde die Aufgabe bekommen, dich dazu bringen, mehrere sehr unangenehme oder eklige Aktionen durchzuführen. Er dürfte aus der folgenden Liste fünf Dinge auswählen, die du danach machen solltest:

- *Einen Keks essen, auf dem ein getrockneter Wurm liegt*
- *Verschiedene Dinge in einer Box ertasten, in die man nicht hineinschauen kann*

- *Möglichst lange den Arm nach vorne ausstrecken und dabei einen Gegenstand festhalten*
- *Schmutziges Klopapier anfassen*
- *Ein benutztes, stinkendes T-Shirt überziehen*
- *Füße in Schleim tauchen*
- *Hände in Dreck halten*
- *Mit klebrigen Händen Sachen aus einer Kiste holen*

Der andere Teilnehmer würde fünf der Aktionen auswählen und die Utensilien dafür würden in einen Raum hier in der Schule gebracht. Er würde versuchen dich dazu zu bringen, möglichst alle der fünf ausgesuchten Aufgaben zu erledigen. Dir gefällt das aber gar nicht gut, denn die Aktionen sind wirklich unangenehm. Vielleicht würdest du dich sogar gegen manche Aufgaben wehren.

Bitte versuche, dich gut in diese Situation hineinzuversetzen, denn du wirst in einer halben Stunde zu diesem Erlebnis gefragt. Die Interviewerin weiß nicht, ob du wirklich etwas erlebt hast und sie darf auf keinen Fall merken, dass du dir alles nur ausgedacht hast. Bitte denke dir daher möglichst detailliert aus, was du angeblich erlebt hast. Du kannst auch viel dazu erfinden, aber es müssen fünf der aufgelisteten Aktionen in deiner Aussage vorkommen. Erzähle ihr, was dir angeblich passiert ist und was der andere mit dir gemacht hat.

Bitte versuche, möglichst überzeugend deine angeblichen Erlebnisse zu schildern. Du hast nachher für deine Aussage so lange Zeit, wie du brauchst.

Zum Abschluss weisen wir dich noch darauf hin, dass für etwaige Unfälle während der Untersuchung eine verschuldensabhängige Haftung bzw. Versicherung besteht.

Wenn du nicht mehr an der Studie teilnehmen möchtest, kannst du das jederzeit sagen und solltest du Fragen haben, so kannst du sie gerne stellen.

Mit einer Teilnahme würdest du uns sehr helfen und einen wichtigen Beitrag zur psychologischen Forschung leisten. Wenn du bereit bist, teilzunehmen, fülle bitte auf der folgenden Seite die schriftliche Einverständniserklärung aus und unterschreibe sie.

Wir danken dir ganz herzlich für dein Interesse an unserer Studie!

Prof. Dr. Rainer Banse, Kathrin Eickmeier und Jelena Rönspies

Appendix B

Teilnehmerinformation für lügende Teilnehmer, Rolle „Täter / Versuchsleiter“

Liebe(r) Interessent(in),

vielen Dank für dein Interesse an unserer Studie!
In der Studie geht es um Zeugenaussagen. Manchmal kommt es vor, dass Zeugen bei der Polizei oder vor einem Richter lügen. Wir wollen deshalb erforschen, wie gut Personen Aussagen erfinden können. Wir untersuchen, ob Personen, die sich ein Erlebnis ausdenken eine genau so gute Zeugenaussage machen können, wie Personen, die das Erlebnis wirklich hatten. Wenn du mit der Teilnahme einverstanden bist, werden wir dich deshalb bitten, eine gewisse Geschichte zu erfinden. Danach sollst du zu diesem angeblichen Erlebnis eine möglichst überzeugende „Zeugenaussage“ machen. Du hast 30 Minuten Zeit, dir die Geschichte auszudenken. Das Erlebnis, das du erfinden sollst, hat folgende Rahmenhandlung:

Stell dir vor, du würdest mit einer anderen Person in die Schule kommen, um an der Studie teilzunehmen. Du würdest die Rolle eines Versuchsleiters bekommen und solltest den anderen Teilnehmer dazu bringen, einige sehr unangenehme oder eklige Aktionen durchzuführen. Du dürftest aus der folgenden Liste fünf Dinge auswählen, die der andere danach machen soll:

- *Einen Keks essen, auf dem ein getrockneter Wurm liegt*
- *Verschiedene Dinge in einer Box ertasten, in die man nicht hineinschauen kann*
- *Möglichst lange den Arm nach vorne ausstrecken und dabei einen Gegenstand festhalten*
- *Schmutziges Klopapier anfassen*
- *Ein benutztes, stinkendes T-Shirt überziehen*
- *Füße in Schleim tauchen*
- *Hände in Dreck halten*
- *Mit klebrigen Händen Sachen aus einer Kiste holen*

Du würdest fünf der Aktionen auswählen und die Utensilien dafür würden in einen Raum hier in der Schule gebracht. Du würdest versuchen die andere Person dazu zu bringen, möglichst alle der fünf ausgesuchten Aufgaben zu erledigen. Dem anderen würde das aber gar nicht gut gefallen, denn die Aktionen sind wirklich unangenehm. Vielleicht würde er sich sogar gegen die Durchführung wehren.

Bitte versuche, dich gut in diese Situation hineinzuversetzen, denn du wirst in einer halben Stunde zu diesem Erlebnis gefragt. Die Interviewerin weiß nicht, ob du wirklich etwas erlebt hast und sie darf auf keinen Fall merken, dass du dir alles nur ausgedacht hast. Bitte denke dir daher möglichst detailliert aus, was du angeblich erlebt hast. Du kannst auch viel dazu erfinden, aber es müssen fünf der aufgelisteten Aktionen in deiner Aussage vorkommen. Erzähle ihr, was dir angeblich passiert ist und was du mit dem anderen gemacht hast.
Bitte versuche, möglichst überzeugend deine angeblichen Erlebnisse zu schildern. Du hast nachher für deine Aussage so lange Zeit, wie du brauchst.

Zum Abschluss weisen wir dich noch darauf hin, dass für etwaige Unfälle während der Untersuchung eine verschuldensabhängige Haftung bzw. Versicherung besteht.

Wenn du nicht mehr an der Studie teilnehmen möchtest, kannst du das jederzeit sagen und solltest du Fragen haben, so kannst du sie gerne stellen.

Mit einer Teilnahme würdest du uns sehr helfen und einen wichtigen Beitrag zur psychologischen Forschung leisten. Wenn du bereit bist, teilzunehmen, fülle bitte auf der folgenden Seite die schriftliche Einverständniserklärung aus und unterschreibe sie.

Wir danken dir ganz herzlich für dein Interesse an unserer Studie!

Prof. Dr. Rainer Banse, Kathrin Eickmeier und Jelena Rönspies

Appendix C

Teilnehmerinformation für wahraussagende Teilnehmer, Rolle „Opfer / Akteur"

universitätbonn

Liebe(r) Interessent(in),

vielen Dank für dein Interesse an unserer Studie! Du hilfst uns damit sehr und leistest damit einen wichtigen Beitrag zur psychologischen Forschung.
In der Studie geht es um Zeugenaussagen. Manchmal kommt es vor, dass Zeugen bei der Polizei oder vor einem Richter lügen. Wir wollen deshalb erforschen, wie gut Personen Aussagen erfinden können. Wir untersuchen, ob Personen, die sich ein Erlebnis ausdenken eine genau so gute Zeugenaussage machen können, wie Personen, die das Erlebnis wirklich hatten. Wenn du mit der Teilnahme einverstanden bist, wirst du daher heute eine Situation erleben und danach zum Geschehenen eine „Zeugenaussage" machen.

<u>Das erwartet dich heute:</u>
Ein anderer Teilnehmer wird dich dazu bringen wollen, einige unangenehme oder eklige Aktionen durchzuführen. Manche Aktionen empfindest du aber vielleicht als spannend oder als Herausforderung.
Der andere sucht aus der folgenden Liste <u>fünf</u> Aktionen aus, die du gleich machen sollst:

- Einen Keks essen, auf dem ein getrockneter Wurm lag

- Verschiedene Dinge in einer Box ertasten, in die man nicht hineinschauen kann
- Möglichst lange den Arm nach vorne ausstrecken und dabei einen Gegenstand festhalten
- Schmutziges Klopapier anfassen
- Ein benutztes, stinkendes T-Shirt überziehen
- Füße in Schleim tauchen
- Hände in Dreck halten
- Mit klebrigen Händen Sachen aus einer Kiste holen

Niemand zwingt dich zu der Teilnahme und es ist immer möglich, einfach aufzuhören. Du musst nichts tun, was du nicht möchtest und du kannst mit dem anderen darüber diskutieren. Wir als Studienleiter befinden uns im Nebenraum und sehen euch auf einem Monitor.

Zum Abschluss weisen wir dich noch darauf hin, dass für etwaige Unfälle während der Untersuchung eine verschuldensabhängige Haftung bzw. Versicherung besteht.
Wenn du nicht mehr an der Studie teilnehmen möchtest, kannst du das jederzeit sagen und solltest du Fragen haben, so kannst du sie gerne stellen.

Mit einer Teilnahme würdest du uns sehr helfen und einen wichtigen Beitrag zur psychologischen Forschung leisten. Wenn du bereit bist, teilzunehmen, fülle bitte auf der folgenden Seite die schriftliche Einverständniserklärung aus und unterschreibe sie.
Wir danken dir ganz herzlich für dein Interesse an unserer Studie!

Prof. Dr. Rainer Banse, Kathrin Eickmeier und Jelena Rönspies

Appendix D

Teilnehmerinformation für wahraussagende Teilnehmer, Rolle „Täter / Versuchsleiter"

universitätbonn

Liebe(r) Interessent(in),

vielen Dank für dein Interesse an unserer Studie!
In der Studie geht es um Zeugenaussagen. Manchmal kommt es vor, dass Zeugen bei der Polizei oder vor einem Richter lügen. Wir wollen deshalb erforschen, wie gut Personen Aussagen erfinden können. Wir untersuchen, ob Personen, die sich ein Erlebnis ausdenken eine genau so gute Zeugenaussage machen können, wie Personen, die das Erlebnis wirklich hatten. Wenn du mit der Teilnahme einverstanden bist, wirst du daher heute eine Situation erleben und danach zum Geschehenen eine „Zeugenaussage" machen.

Das erwartet dich heute:
Deine Aufgabe wird ist es, einen anderen Teilnehmer dazu zu bringen, einige unangenehme oder eklige Aktionen durchzuführen. Bitte suche aus der folgenden Liste fünf Aktionen aus, die die andere Person machen soll:

- Einen Keks essen, auf dem ein getrockneter Wurm lag
- Verschiedene Dinge in einer Box ertasten, in die man nicht hineinschauen kann
- Möglichst lange den Arm nach vorne ausstrecken und dabei einen Gegenstand festhalten

- Schmutziges Klopapier anfassen
- Ein benutztes, stinkendes T-Shirt überziehen
- Füße in Schleim tauchen
- Hände in Dreck halten
- Mit klebrigen Händen Sachen aus einer Kiste holen

Der andere wird diese Dinge sicherlich nicht ganz freiwillig machen, aber du sollst es schaffen, dass er möglichst viele der Dinge erledigt. Versuche möglichst gut, den anderen zu überzeugen. Wir als Versuchsleiter befinden uns im Nebenraum und sehen euch auf einem Monitor.

Zum Abschluss weisen wir dich noch darauf hin, dass für etwaige Unfälle während der Untersuchung eine verschuldensabhängige Haftung bzw. Versicherung besteht.

Wenn du nicht mehr an der Studie teilnehmen möchtest, kannst du das jederzeit sagen und solltest du Fragen haben, so kannst du sie gerne stellen.

Mit einer Teilnahme würdest du uns sehr helfen und einen wichtigen Beitrag zur psychologischen Forschung leisten. Wenn du bereit bist, teilzunehmen, fülle bitte auf der folgenden Seite die schriftliche Einverständniserklärung aus und unterschreibe sie.

Wir danken dir ganz herzlich für dein Interesse an unserer Studie!

Prof. Dr. Rainer Banse, Jelena Rönspies und Kathrin Eickmeier

Appendix E

Einverständniserklärung der Schüler

Einverständniserklärung

Ich ____(Name) bin mit der Teilnahme an der Studie einverstanden.

- Ich bin darüber informiert worden, dass die Teilnahme an der Studie freiwillig erfolgt und von mir jederzeit ohne Angabe von Gründen und ohne negative Konsequenzen beendet werden kann.
- Ich bin damit einverstanden, dass meine Aussage auf Tonband aufgenommen wird. Die Aufnahme wird von der Studienleitung nur bis zum Abschluss der Auswertung der Studie gespeichert und dann sofort gelöscht. Die übrigen Daten werden anonym gespeichert und erlauben keine Rückschlüsse auf meine Person.
- Ich bin darüber aufgeklärt worden, dass für etwaige Unfälle während der Untersuchung eine verschuldensabhängige Haftung bzw. Versicherung besteht.
- Mir ist bewusst, dass ich nicht mit anderen Schülern über den Verlauf der Studie sprechen darf.

Datum, Unterschrift____

Appendix F

Elterninformation und Elterneinverständnis

universitätbonn

Prof. Dr. Rainer Banse
Jelena Rönspies, M.Sc. Psych.
Kathrin Eickmeier, M.Sc. Psych.

Institut für Psychologie
Abteilung Sozial- und Rechtspsychologie

Kaiser-Karl-Ring 9
53111 Bonn

Tel. +49 228 / 734190
Mail: k.eickmeier@uni-bonn.de
www.psychologie.uni-bonn.de

Wie gut können Kinder und Jugendliche Zeugenaussagen machen?

Liebe Eltern,

im Rahmen unserer Studie am Institut für Psychologie der Universität Bonn erforschen wir die Eigenschaften von Zeugenaussagen von Kindern und Jugendlichen. Dabei interessieren uns sowohl Aussagen über tatsächlich erlebte Ereignisse als auch erfundene Schilderungen.

Nicht selten muss in Strafprozessen anhand der Zeugenaussagen von Kindern und Jugendlichen entschieden werden, ob eine Straftat tatsächlich stattgefunden hat. Daher ist es sehr wichtig Kriterien zu bestimmen, die eine Unterscheidung von erfundenen oder erlebten Ereignissen erlauben. Ziel unserer Studie ist es, die psychologischen Methoden in diesem Bereich noch zuverlässiger zu machen.

In unserer Studie werden Kinder und Jugendliche zu zweit einige Aufgaben entweder tatsächlich erledigen oder sich den Ablauf der Aufgaben nur ausdenken. Einige dieser Aufgaben sind so gestaltet, dass sie – immer auf spielerischer Ebene – ein bisschen eklig sind oder Mut erfordern (z.B. schleimige

Lebensmittel anfassen, blind Gegenstände ertasten). Unabhängig davon, ob die Aufgabe erlebt

oder erfunden wurde, werden die Kinder anschließend zu der Situation befragt. Erfundene und erlebte Aussagen werden dann von uns anhand verschiedener Kriterien verglichen. Zusätzlich wird von den Kindern ein kurzes, neu entwickeltes Computerprogramm bearbeitet, das zwischen den beiden Gruppen unterscheiden soll. Dabei soll eine Reihe von Begriffen auswendig gelernt und diese in einer Menge anderer Wörter wiedererkannt werden.

Wenn Ihr Kind Interesse an der Teilnahme hat, müsste im Rahmen des Psychologie-Unterrichts an der IGS Beuel an zwei Terminen teilnehmen. Der erste Termin dauert 45 bis 60 Minuten. In dieser Zeit wird das Geschehen, je nach Gruppenzugehörigkeit, erlebt bzw. erfunden und danach eine Aussage dazu gemacht. Das Erlebnis wird auf Video und die Aussage auf Tonband aufgezeichnet, damit die Aussage im Nachhinein von den Studienleitern ausgewertet werden kann. Beim zweiten Termin wird die Aufgabe am Computer erledigt, welche ca. 30 Minuten in Anspruch nimmt. Als kleines Dankeschön bekommen die Teilnehmer die Möglichkeit an der Verlosung mehrerer Gutscheine teilzunehmen. Alle erhobenen Daten werden vollständig anonym erhoben, so dass keine Rückschlüsse von den erhobenen Daten auf die Teilnehmer möglich sind. Die Tonband- und Videoaufzeichnungen werden nach Abschluss der Auswertungen unverzüglich und unwiderruflich gelöscht. Es wird sichergestellt, dass zu keinem Zeitpunkt Rückschlüsse auf die Ergebnisse einzelner Kinder gezogen werden können. Sämtliche Aufgaben, die die Teilnehmer absolvieren, werden permanent von der Studienleitung (die von geschulten Psychologinnen übernommen wird) überwacht. Die Kinder können zudem die Teilnahme an der Studie ohne Weiteres und zu jedem Zeitpunkt abbrechen.

Zum Abschluss weisen wir Sie noch darauf hin, dass für etwaige Unfälle während der Untersuchung eine verschuldensabhängige Haftung bzw. Versicherung besteht.

Wir würden uns freuen, wenn Ihr Kind an unserer Studie teilnehmen möchte und Sie ihr Einverständnis dazu geben. Durch eine Teilnahme leisten Sie und Ihr Kind einen wichtigen Beitrag zur Forschung im Bereich kindlicher Zeugenaussagen.

Wenn Sie weitere Fragen haben sollten, können Sie sich gerne über die oben genannten Kontaktdaten mit uns in Verbindung setzen.

Wir danken Ihnen für Ihr Interesse und verbleiben mit freundlichen Grüßen,

Prof. Dr. Rainer Banse, Jelena Rönspies und Kathrin Eickmeier

(Diese Seite ist vom Teilnehmer zum ersten Studientermin mitzubringen)

universitätbonn

Einverständniserklärung Studie zu Zeugenaussagen von Kindern und Jugendlichen

Name des Kindes: ________________________________

Mein Kind hat bekannte Allergien oder Unverträglichkeiten

☐ Ja ☐ Nein

Wenn ja, folgende: ________________________________

Hiermit erlaube ich meiner Tochter/meinem Sohn an der Studie zu Zeugenaussagen der Universität Bonn teilzunehmen. Ich bin darüber aufgeklärt worden, dass die Universität Bonn für etwaige Unfälle meines Kindes während der Teilnahme an der Studie verschuldensabhängig haftet.

Datum und Unterschrift des Erziehungsberechtigten:

____