Fachbereich Erziehungswissenschaft und Psychologie

der Freien Universität Berlin

Differential Bias and Suggestiveness

Cognitive Patterns and Emotional Reactivity as Predictors of Bias and Suggestion in Child Sexual Abuse Investigations

Dissertation

zur Erlangung des akademischen Grades

Doktorin der Philosophie (Dr. phil.)

vorgelegt von

Elsa Kristin Billie Gewehr (M.Sc., M.Sc.)

Berlin, Oktober 2024

Erstgutachterin:

Prof. Dr. Renate Volbert

Zweitgutachter:

Prof. Dr. Stefan Krumm

Datum der Disputation: 31.01.2025

Abstract

This thesis examines how cognitive and emotional individual differences predispose biased mindsets and suggestiveness in formal and informal questioning of children about child sexual abuse (CSA) suspicions. Across five empirical studies, presented in three articles, the Cognitions and Emotions about Child Sexual Abuse (CECSA) scales were developed, validated, and tested for their relationship to biased judgments and suggestive questioning, and for their responsiveness to training. Article 1 outlines the development and initial validation of three CECSA scales – Naive Confidence (NC), Emotional Reactivity (ER), and Justice System Distrust (JSD) – in a sample of 801 human sciences students. The scales demonstrated good model fit, acceptable to good reliability and, importantly, predicted participants' bias toward the abuse hypothesis when judging vague CSA suspicions. Article 2 presents three mock case studies for predicting bias and suggestive questioning using varying formats for question posing: a single-choice format, a free-writing format, and a natural language format in a virtual reality (VR) simulation. Results for a total of 674 students from diverse disciplines (human sciences, teaching, police studies) showed that NC and ER, but not JSD, robustly predicted biased mindsets and a suggestive questioning style across the three studies and a meta-analytic integration. Article 3 evaluates a training program aimed at improving questioning techniques and related constructs. A secondary analysis of the data showed that a two-day seminar-style training significantly reduced NC and ER scores, while the results for JSD were inconclusive. These findings establish the CECSA scales NC and ER as reliable predictors of bias and a suggestive questioning style and show that both are modifiable through training. The scales are of diagnostic and evaluative value for training development or personnel selection, and can be used and extended to further investigate differential aspects of child sexual abuse investigations.

Table of Contents

Prologue

Questioning children about their autobiographical experiences is a fundamental element of child sexual abuse investigations. Many children face multiple rounds of questioning, beginning with informal conversations in childcare, health, or education settings, followed by child protection services, and potential formal interviews with law enforcement personal or forensic psychologists. Given that child sexual abuse (CSA) often leaves no physical traces or other corroborative evidence, these interactions, whether formal or informal, are often central in determining whether a child has experienced abuse (Cirlugea & O'Donohue, 2016; Talwar et al., 2024). However, they are also prone to errors with potentially far-reaching consequences (Lilienfeld, 2016) and are widely recognized as a complex task that requires specialized conversational skills (J. Johnson et al., 2016; Korkman et al., 2024). Professionals need to navigate a balance between providing sufficient emotional and cognitive support to encourage truthful disclosures from a child, while maintaining an open and neutral stance to avoid conveying their own assumptions about the events in question, as those might distort the child's account. Unfortunately, many professionals from a variety of sectors feel severely unprepared for this task and do not receive specialized training (Baginsky, 2003; Goldman, 2007; Greytak, 2009; Lemaigre et al., 2017).

Decades of legal psychology research on the influence of different question types on children's accounts have resulted in consensual recommendations on how to interview children in forensic settings (Korkman et al., 2024): In order to obtain truthful disclosures and comprehensive and accurate reports, interviewers must establish rapport with the child, provide socio-emotionally support throughout the entire interaction, and focus on open-ended questions such as invitations to tell what happened. Interviewers further need to acknowledge their own biases, work against them by actively maintaining an open mindset, and avoid any type of

suggestive questioning or behaviors. Various initiatives have developed empirically-based half-structured protocols (e.g., the NICHD protocol; La Rooy et al., 2015) to guide and standardize forensic child interviewing. In contrast, informal conversations with children in education, health, or child protection settings have received less scholarly attention, and often lack official guidelines or protocols (Talwar et al., 2024). From a legal psychology perspective, best-practice recommendations for forensic interviews – providing rapport and socio-emotional support, maintaining an open mindset, and using open, non-suggestive questions – should equally apply to informal conversations with children, albeit with less emphasis on obtaining detailed information. However, research has shown that in practice, both conversations and interviews are often not conducted in line with these evidence-based recommendations. Professionals seem to have particular difficulty maintaining an open mindset and avoiding suggestive questions (Andrews & Lamb, 2021; Brubacher et al., 2016; Fessinger & McAuliff, 2020; M. Johnson et al., 2015; Korkman et al., 2014; Marchant et al., 2020).

Bias and Suggestion

Planned communication with children about possible abuse experiences is usually based on an adult's a priori hypothesis about what may have happened to the child. While in forensic interviews, these hypotheses are based on prior investigations, in more informal settings, they often arise based on potentially concerning, but nonspecific behaviors of a child (e.g., moodiness, social isolation, aggression, wetting, touching genitals, or using sexual language; Volbert & Kuhle, 2019). Because such behaviors can result from a variety of experiences or developmental issues and are empirically not or only weakly associated with sexual abuse (Kendall-Tucket et al., 1993; Lewis et al., 2016), they are not valid diagnostic indicators of CSA. Given the widespread prevalence of such behaviors and the comparatively low population base

rate of CSA (Stoltenborgh et al., 2015), there are many more non-abused than abused children who exhibit these behaviors (Talwar et al., 2024). As a result, false a priori hypotheses are an inevitable part of CSA investigations. When working on or confronted with such cases, the possibility of being mistaken must be actively considered, as well as alternative explanations for the origin of observed behavior (Korkman et al., 2024).

One of the biggest pitfalls in interviewing and talking to children is to become cognitively or emotionally attached to an a priori assumption, and then question children with the aim to confirm this hypothesis. This process has often been observed in laboratory and field studies and has been termed 'interviewer bias' (Brown & Lamb, 2015; Ceci et al., 2016; Ceci & Bruck, 1995; Duke et al., 2016; Korkman et al., 2024; Powell et al., 2012; Rohrabaugh et al., 2016). It corresponds to the general human tendency to gather confirmatory evidence for one's beliefs through belief-consistent information processing, instead of considering alternative explanations or falsifying beliefs, which is known as 'confirmation bias' (Neal et al., 2022; Nickerson, 1998; Oeberst & Imhoff, 2023).[1] Although the issue of interviewer bias has been well-known in the legal psychology literature since at least the 1990's (Bruck & Ceci, 1997; Ceci & Bruck, 1995), and is based on the older literature on experimenter expectancy effects (Rosenthal, 1976), it has scientifically been somewhat neglected in recent decades. More effort has been devoted to studying the behavioral manifestation of interviewer bias, that is, suggestive questions and other suggestive behaviors (Ceci & Bruck, 2006; O'Donohue & Cirlugea, 2021; Zhang et al., 2022).

---

[1] Because interviewer bias is prevalent not only in forensic interviews, but also in informal conversations with children that aim to clarify abuse suspicions (Brubacher et al., 2016; Marchant et al., 2020; O'Donohue & Cirlugea, 2021), broader terms such as a 'biased mindset' or simply 'bias' are used in this dissertation to refer to confirmatory processes when questioning children.

Suggestive questioning refers to utterances in which adults – intentionally or not – communicate their own assumptions about a child's potential experience to the child (Ceci et al., 2016). This includes providing the child with information about one's own beliefs (e.g., by weaving belief-consistent information into a question) or implying that certain (belief-consistent) responses from the child will be more valued than others. Efforts have been made to define suggestive questions in linguistic terms for research and training purposes. Typical suggestive utterances are closed-ended questions with new information that ask for monosyllabic responses (e.g., "Did he hurt you?"; "Did your uncle do this or was it your ant?"). Emphasizing the expected response (e.g., "Your ant did this, *didn't she?*"; "*I'm sure* that you didn't like that") can further increase suggestive pressure. Often, suggestion occurs in more subtle ways, such as social or peer pressure (e.g., "Sarah already told me that something bad happened, don't you want to tell me too?"), selective reinforcement of belief-consistent responses (e.g., "It's very brave of you to tell me about these bad things"; "Now that's is a bit much, don't you think?"), emotional appraisal ("That's terrible"), or invitations to speculate or imagine (e.g., "What do you think would have happened if you had stayed longer?"; "How would you feel if he did that to you?"). Such utterances do not openly invite the children to report from their own memories, but instead have the potential to influence children's subsequent statements and reports through socio-emotional (e.g., peer pressure, social desirability) or cognitive (e.g., memory alteration or retrieval interference) mechanisms (Ceci et al., 2016; Ceci & Bruck, 1993).

Research has shown that children's responses to suggestive questions are not only less detailed and less accurate than to open questions, they also often falsely confirm the interviewer's false beliefs (Brown & Lamb, 2015; Ceci et al., 2016; Ceci & Friedman, 2000). The long-term consequences of suggestive interviewing can be detrimental. In particular,

repeated or intense exposure to suggestion can foster false beliefs about autobiographical experiences, distort memory or even lead to the development of pseudo-memories: Vivid recollections of events that are subjectively believed in but not actually based on real experiences (Howe & Knott, 2015; Scoboria et al., 2017). Extensive empirical research and many real-life cases have shown that false memories can develop even for strongly adverse, and personally and emotionally relevant events, such as having experienced sexual abuse (Brewin & Andrews, 2017; Oeberst et al., 2021; Patihis & Pendergrast, 2019). This poses a serious threat to children's well-being. For example, false memories about CSA can lead to similar psychopathological symptoms as actual CSA experiences (Baldwin et al., 2024a; Porter et al., 2007).

Suggestive processes can emerge in all types of allegations, but they are a particular issue in alleged cases of CSA, where the absence of physical or corroborative evidence often leaves the child's testimony as the primary, and sometimes only, piece of evidence (Cirlugea & O'Donohue, 2016; Talwar et al., 2024). Indeed, field research has repeatedly shown that suggestive questioning frequently occurs in CSA investigations in both early informal conversations (Brubacher et al., 2016; Korkman et al., 2014; Marchant et al., 2020) and formal forensic interviews (Andrews & Lamb, 2021; Cirlugea & O'Donohue, 2016; Fessinger & McAuliff, 2020; M. Johnson et al., 2015; Lamb et al., 2007). Because reports from false memories are often indistinguishable from true reports (Korkman et al., 2024; Wachendörfer & Oeberst, 2023), suggestive processes pose a serious burden to the effective juridical prosecution of CSA allegations.

With the development of structured interview protocols and extensive training programs, researchers have found a way to considerably reduce suggestive and otherwise undesired questioning (Akca et al., 2021; La Rooy et al., 2015; Lamb et al., 2011; Zajac & Brown, 2018).

However, these approaches are not universally applied and not accessible to all professionals, particularly not to those outside of forensic settings. Unfortunately, many shorter and simpler training endeavors have proven unsuccessful, and research suggests that behavioral training needs to be repeated regularly to maintain it's positive effect on questioning styles (Brubacher et al., 2022; Powell, 2008; Powell et al., 2022). On the other hand, research also shows variability in the degree of suggestiveness between interviewers, indicating that interviewers' individual differences can contribute to the degree of suggestive pressure placed on children (Brubacher et al., 2014; Finnilä-Tuohimaa et al., 2008; M. Johnson et al., 2015; Kask et al., 2022; Pompedda et al., 2022). This highlights the need to develop a deeper understanding of the factors underlying biased and suggestive questioning – such as individual differences – and to find ways to mitigate it at levels other than time-consuming and costly behavioral training.

Differential Bias and Suggestiveness

Despite consensus on the highly individual nature of child interviewing (Korkman et al., 2024; Lilienfeld, 2016; Talwar et al., 2024), and the relevance of the dyadic adult-child relationship for successful interviewing (Lavoie et al., 2021; Saywitz et al., 2015), individual differences in interviewing performance, particularly bias and suggestiveness, have scarcely been studied. Equally little is known about how informal conversations with children are influenced by individual differences in adults, although the lack of guidelines likely corresponds to even greater individualization. Instead, bias and suggestive questioning are usually discussed as an innate aspect of human nature, that is aggravated by situational factors and can be defeated by behavioral training. Exemplary situational factors that have been shown to bias forensic decision-making and interviewing are the position of the commissioning party (e.g., prosecution or defense, "adversarial allegiance") or information about the case or the child that are irrelevant

to the diagnostic question (e.g., race or gender; Huang & Bull, 2021; Neal et al., 2022; O'Donohue & Cirlugea, 2021). Research on how to defeat bias and suggestion in CSA investigations typically advises behavioral training of some form. This can be learning how to actively consider alternative hypotheses (Gumpert & Lindblad, 2000; Korkman et al., 2024; Oberlader et al., 2024; O'Donohue et al., 2013; Otgaar et al., 2017; Zapf & Dror, 2017), but most often involves direct training of non-suggestive questioning, e.g., by learning to formulate open questions and practice to suppress suggestive questions in simulated child interviews (Akca et al., 2021; Powell et al., 2022).

Notably, while a differential perspective is largely absent in applied research on interviewer bias and suggestion, the broader scientific discussion on mechanisms of human biases in general and in the forensic domain is increasingly taking individual differences into account. For example, Neal et al. (2022), who conducted a systematic review on cognitive biases in the forensic sector, named the identification of individual differences that "function as risk and protective factors against bias" as one of the most important future research endeavors to advance our understanding of cognitive biases. Oeberst & Imhoff (2023), who proposed a general framework for cognitive biases across various subtypes, more specifically conceptualize individual beliefs, developed from autobiographical experiences, as the starting point of any biased information processing. Zapf and Dror (2017) introduced a hierarchical taxonomy on sources of bias in the forensic domain, which reaches from "basic human nature and the cognitive architecture of the brain" at the bottom level, across influences from "environment, culture, and experiences", up to more situational "case-specific influences". In the middle range of their taxonomy, the authors identified individual level sources of bias, such as motivations, preferences, attitudes, and thinking styles. They argue that identifying individual characteristics

that influence forensic decision-making will be a crucial first step to study ways to mitigate differential bias, such as the development of individual profiles for personnel selection or skill development. To sum, different models and taxonomies name individual differences as an important potential source of bias, but little effort has been made to identify the specific characteristics that can bias forensic decisions and behaviors.

Returning to child interviewing research, despite its general focus on situational sources of bias, some researchers have recently discussed the influence of individual differences on interviewing performance. For example, Akca et al. (2021), who conducted a systematic review on interview trainings, discussed police officers' personality traits as a source of variations in interview success and differential training effects. In a recent work by an international consortium on "urgent issues and prospects on investigative interviews with children and adolescents", Talwar et al. (2024) highlight the need to study individual-level predictors of false CSA allegations (e.g., beliefs about diagnosing CSA from behavior). In a similar effort, a European consortium (Korkman et al., 2024) who presented a "white paper on forensic child interviewing", identified interviewer bias, influenced by professionals' cognitive styles, attitudes, or beliefs, as one of the main risk factors for child-interviewing.

Nonetheless, empirical research on differential aspects of bias and suggestiveness is limited: Studying Big Five personality traits, Acka & Eastwood (2021) found higher extraversion and lower agreeableness to relate to more inappropriate (i.e., leading, long or complex) questions in 130 mock interviews with adult witnesses. The authors argued that these traits tempt interviewers to be more talkative themselves instead of listening to the perspectives of the children. However, Melinder et al. (2020) found no convincing pattern connecting Big Five personality traits to self-reported interviewing style among 46 police officers.

Cognitive Sources

Some studies assessed whether cognitive patterns, such as attitudes, beliefs, or information processing styles, relate to bias or suggestiveness: Everson and Sandoval (2011) found that child-protection professionals' ($N > 1000$) general propensity to prioritize sensitivity (i.e., avoid false denials) or specificity (i.e., avoid false allegations), as well as their general skepticism toward children's abuse reports biased their judgements when evaluating children's credibility in mock CSA cases (also see Fessinger & McAuliff, 2020 for similar findings). Finnilä-Tuohimaa et al. (2008) developed four scales to measure more specific attitudes and beliefs when dealing with CSA allegations: Unconditionally trusting children's abuse reports (Pro-Child Scale), intuitively evaluating CSA allegations (Intuition Scale), believing children are unable to disclose abuse by themselves (Disclosure Scale), and pessimistic views about the willingness and ability of the justice system to prosecute CSA (Anti Criminal Justice System Scale). All four scales related to biased mindsets toward the abuse hypothesis when students and mental health professionals evaluated cases with unspecific and vague CSA allegations (Finnilä-Tuohimaa et al., 2008). For the first two scales, this was true even if the case information included strong suggestive questioning of the children (Finnilä-Tuohimaa et al., 2009).

Emotional Sources

Next to cognitive patterns, emotional reactions are a well-known source of bias (Lerner et al., 2015), in particular confirmation bias (Jonas et al., 2006). This is grounded in valence-based theories (e.g., Feeling-as-information theory; Schwarz, 2012), according to which the valence (positive vs. negative) of an emotional reaction directly informs judgement and choice. Valence also activates emotionally congruent memories, which are used as further informational indicators (Schwarz, 2012). Because CSA is an emotionally charged crime not only for lay

people but also for many professionals (Magnusson et al., 2021; McCartan et al., 2015; Olaguez et al., 2023; Segal et al., 2022; Segal, Kaniušonytė, et al., 2023), individual levels of emotionality regarding CSA may influence professionals' propensity for bias and suggestiveness when interacting with the children involved. Studying the role of emotions in child interviewing, Segal et al. (2023) reported that interviewers' facially observed emotions of anger in simulated child interviews went along with posing more closed (versus open) questions. Further support that negative emotions can induce bias comes from research on suspect interviews or verdicts, where negative emotions such as anger and disgust were linked to more confrontational interviewing styles (Magnusson et al., 2021), and a greater inclination toward guilty verdicts and harsher punishments (Olaguez et al., 2023; Salerno, 2021).

Overall, research on differential bias and suggestiveness is scarce, but some studies indicate that cognitive patterns, such as specific attitudes, beliefs, or information processing styles, as well as emotional reactions, may predispose biased mindsets and suggestive questioning in conversations and interviews with children about CSA. Increasing knowledge about these predispositions may help to select adequate personal for questioning children and identify their individual training needs. If the individual differences that predispose bias and suggestion are receptive to intervention, interview training curricula might be enriched by modules that target these characteristics directly and thus increase their impact on interviewer performance.

Research Objectives

The overarching objective of this thesis is to investigate individual differences in cognitive patterns and emotional reactivity as sources of biased mindsets and suggestive questioning in child sexual abuse investigations. The focus is not only on forensic interviews of

children but also on informal conversations with children that aim to clarify suspicions of CSA in child-protection, health or educational settings. The thesis consists of five empirical studies that are summarized in in three articles (i.e., published articles or manuscripts submitted for publication). The first article reports the development of a self-report instrument on Cognitions and Emotions about Child Sexual Abuse (CECSA), including its validation as a tool to predicting biased evaluations in CSA allegations. The second article investigates how the CECSA scales predict suggestive questioning across a series of three studies and a meta-analytical integration. The third article describes a randomized controlled trial to evaluate a conversational training program for school settings, in which the CECSA scales were used as one of the instruments to measure training success. For the scope of this thesis, the focus is on how participants' CECSA scores can be influenced by the seminar intervention of the training program, which teaches evidence-based handling of CSA suspicions. Because Article 3 only reported results for one of the CECSA scales, results for the remaining CECSA scales are reported in a secondary analysis in this thesis.

Together, the studies of this thesis contribute to understanding differential components of bias and suggestiveness – a strongly understudied subject in the field of child interview research. They also offer practical insights for improving CSA interviews and conversations trough informing personnel selection, efficient allocation of resources, and the improvement of training curricula.

Article 1:

Cognitions and Emotions about Child Sexual Abuse (CECSA): Development of a Self-Report Measure to Predict Interviewer Bias

Status: Published in *Psychology, Crime, and Law* (01.01.2025)

Elsa Gewehr[1,2], Renate Volbert[1], Marie Merschhemke[4], Pekka Santtila[3], and Simone Pülschen[4]

[1] Psychologische Hochschule Berlin, [2] Universität Kassel, [3] New York University Shanghai, [4] Europa Universität Flensburg

This is an original manuscript of an article published by Taylor & Francis in PSYCHOLOGY, CRIME & LAW on 01.01.2025, available online: **https://doi.org/10.1080/1068316X.2024.2443448**

Gewehr, E., Volbert, R., Merschhemke, M., Santtila, P., & Pülschen, S. (2025). Cognitions and emotions about child sexual abuse (CECSA): development of a self-report measure to predict bias in child sexual abuse investigations. *Psychology, Crime & Law*, 1–21. https://doi.org/10.1080/1068316X.2024.2443448

Abstract

A biased mindset can foster confirmatory reasoning and suggestive questioning when adults talk to children about abuse suspicions in child-protection, healthcare, educational or investigative settings. We developed a self-report instrument on Cognitions and Emotions about Child Sexual Abuse (CECSA) that may predict individual propensity for a bias toward the abuse hypothesis. Three subscales, 23 items in total, were created in a sample of 801 students of human sciences via exploratory factor analysis and Ant Colony Optimization. The “Naïve Confidence” subscale reflects overestimating one's ability to recognize abused children and overestimating the accuracy of children’s abuse reports, the "Emotional Reactivity" subscale measures the intensity of one's emotional reactions towards the topic of child sexual abuse (CSA), and the "Justice System Distrust" subscale covers distrusting the justice system’s ability to prosecute CSA. The CECSA showed adequate model fit and good internal consistencies. Correlations with other self-report measures demonstrated convergent validity. All subscales predicted biased evaluations towards the abuse hypothesis in scenarios of children displaying unspecific behavioral problems. Prospectively, the CECSA may be used to evaluate training programs or to assess training needs of professionals who talk to children about CSA suspicions.

Keywords: Child Sexual Abuse, Interviewer Bias, Cognitive Styles, Attitudes, Emotions

Cognitions and Emotions about Child Sexual Abuse (CECSA): Development of a Self-Report Measure to Predict Bias in Child Sexual Abuse Investigations

C*onfirmation bias* is a well-known psychological phenomenon, describing the human tendency to generate a single hypothesis to explain a new observation and then pursue its confirmation even in the light of disconfirming evidence. It seems to be driven by the difficulty to take different optional causes into account simultaneously. Instead, people often develop one single hypothesis based on prior beliefs or experiences and strive to consolidate it by searching or prioritizing confirmatory evidence. Contradictory information is often ignored, and ambiguous or non-diagnostic information often interpreted in line with the initial hypothesis (for overviews, see Neal et al., 2022; Nickerson, 1998; Oeberst & Imhoff, 2023). As a task- and domain-specific subtype, *interviewer bias* has gained attention in the field of legal psychology (Brown & Lamb, 2015; Ceci & Bruck, 1995, 2006; Duke et al., 2016; Huang & Bull, 2021; Korkman et al., 2024; Powell et al., 2012; Rohrabaugh et al., 2016). It describes the often-observed tendency of interviewers to strive for a confirmation of their a priori hypothesis about a criminal case by pursuing a confirmatory statement from the interviewee. Interviewer bias has been found to lead to suggestive questioning, that is, the integration of the interviewer's presumptions into their questions (Ceci & Bruck, 2006; O'Donohue & Cirlugea, 2021; Zhang et al., 2022). It also promotes sensemaking, which describes the interpretation of senseless or ambiguous responses in line with one's a priori hypothesis (Dana et al., 2013).

Both processes – suggestive questioning and sensemaking – are especially prominent in interviews with children, compared to adult interviews (Bruck & Ceci, 1997; Quas et al., 2007). That is because, for one thing, children more often than adults give answers that are ambiguous,

do not quite fit the question, or seem senseless from an adult perspective and thus pave the way for sensemaking (Korkman et al., 2008; Perez et al., 2022). Second, while children are generally capable of making comprehensive and accurate autobiographical statements, the younger they are, the more prompts and memory cues they need to do so (Fivush, 2011). This can lead to suggestive questioning on the adults' behalf. The often-observed adults' habit to ask children for confirmations ("You like strawberry ice cream, right?") instead of information or opinions ("Tell me about your favorite ice creams!") can further increase this suggestive tendency.

Suggestive questioning, in turn, systematically results in children's answers being less detailed and less accurate (Ceci et al., 2016; Ceci & Friedman, 2000). Repeated or intensive suggestive questioning can taint children's memories or even foster the development of entirely false memories: Mental representations of events that feel like and are believed to be autobiographical memories but are not based on actual experiences (Howe & Knott, 2015; Scoboria et al., 2017).

The existence and detrimental consequences of a biased mindset, suggestive questioning and sensemaking have been recognized not only for formal forensic interviews (e.g., police or expert witness interviews; hereafter referred to as "interviews"), but also for informal conversations between adults and children that aim to clarify abuse suspicions in child-protection, healthcare, or educational settings (hereafter referred to as "conversations"; Brubacher et al., 2016; Korkman, Juusola, et al., 2014; Marchant & Turner, 2017; O'Donohue & Cirlugea, 2021).

To counter bias and suggestive questioning, forensic interview techniques and extensive interview protocols have been developed and evaluated (e.g., the NICHD protocol; La Rooy et al., 2015; Lamb et al., 2007). Corresponding programs train open and non-suggestive

questioning as a skill. While programs that focus on increasing knowledge only often fail to result in behavioral changes, more complex trainings that include multiple sessions of practice and detailed feedback have shown to improve interviewing performance (Akca et al., 2021; Kask et al., 2022; Powell, 2008). However, such trainings are cost- and time-intensive and most professionals do not have the opportunity for regular training and feedback. That is especially true if talking to children about forensically relevant issues is not a regular task, as is the case for most child-protection, healthcare, or educational personnel. Thus, hoping to get beyond behavioral trainings and develop measures to tackle biased questioning at its roots, some scholars have called for a stronger research focus on the bias itself to understand the mechanisms behind suggestive questioning and sensemaking in interviews and conversations with children (Ceci & Bruck, 2006; Huang & Bull, 2021; O'Donohue & Cirlugea, 2021).

Understanding a Biased Mindset as a Differential Construct

Biased questioning of children is usually discussed as a situational phenomenon: Wherever interviews or clarifying conversations are initiated based on an abuse hypothesis – which is the case for most CSA interviews and conversations – above-described confirmatory processes may come into play. Correspondingly, most research has focused on situational aspects that aggravate this bias, such as case-irrelevant information or the position of the commissioning party in forensic interviews (e.g., Huang & Bull, 2021; Neal et al., 2022; O'Donohue & Cirlugea, 2021). Although interviews and conversations are highly individual in nature and studies report considerable between-person variance in interviewing performance (e.g., Finnilä-Tuohimaa et al., 2008; Pompedda et al., 2022) and case evaluations (Everson & Sandoval, 2011), differential aspects of bias and suggestiveness are rarely examined: Are some people more prone than others to fall for a bias toward the abuse hypothesis and into using suggestive questioning and

sensemaking? If so, what distinguishes these individuals from those who stay more open-minded? Besides some efforts investigating the role of Big Five personality traits in interviewing performance (Akca & Eastwood, 2021; Melinder et al., 2020), scholars have mainly suggested cognitive and emotional factors as individual level sources of bias in CSA investigations.

Cognitive Sources of Bias

Some researchers have pointed out that practitioners handling suspicions of CSA often seem to be guided more by lay convictions, erroneous beliefs, or personal attitudes than by empirical research findings (Herman, 2005; Horner et al., 1993; McGuire & London, 2017; Melinder et al., 2004; Patihis et al., 2014; Pelisoli et al., 2015). Survey studies show that many teachers (Márquez-Flores et al., 2016), investigative interviewers (Davey & Hill, 1995), judges and jurors (Goodman-Delahunty et al., 2017; Korkman, Svanbäck, et al., 2014), child-protection workers (Erens et al., 2020), social workers, psychiatrists or psychotherapists (Finnilä-Tuohimaa et al., 2005, 2009; Patihis et al., 2014; Pelisoli et al., 2015; Schemmel et al., 2024) hold misconceptions about sexual abuse, children's memories or disclosure patterns that contradict empirical evidence. Beyond surveying faulty knowledge, some efforts have been made to define and measure specific attitudes, thinking styles or information processing strategies (hereafter summarized as "cognitive styles") that may enhance a propensity for bias when dealing with CSA suspicions. For example, Everson and Sandoval (2011) defined diagnostic foci that may differ between professional groups and influence decision making in CSA cases: The fear of undercalling abuse (emphasizing sensitivity), the fear of overcalling abuse (emphasizing specifity), and a general skepticism towards child abuse reports. All three constructs influenced the evaluations of mock CSA suspicion cases of diverse professionals (e.g., child-protection, law enforcement or mental health professionals).

A Finnish research group developed a preliminary self-report instrument on specific attitudes and beliefs that may facilitate bias when dealing with CSA suspicions ("Child Sexual Abuse Attitude and Belief Scales" [CSAABS]; Finnilä-Tuohimaa et al., 2008). Four subscales each describe a different cognitive stance: (1) The *Pro-Child Scale* reflects unconditionally trusting children's abuse reports and dismissing the possibility of false or distorted allegations. (2) The *Disclosure Scale* measures the conviction that children rarely disclose abuse by themselves, and that disclosure must be facilitated at any cost, even through suggestive questioning. (3) The *Intuition Scale* describes an intuitive thinking style when evaluating CSA suspicions. (4) Finally, the *Anti Criminal Justice System Scale* summarizes pessimistic views about the justice system's competence and willingness to prosecute sexual delinquency. Most surveyed students, mental health professionals, investigative interviewers and judges held biased attitudes and beliefs only to small or moderate extent, but a non-negligible minority showed quite extreme values on the CSAABS. This variance was largely not explainable by the level of professional experience, and, among healthcare professionals, former training experience was surprisingly associated with more biased attitudes and beliefs (Finnilä-Tuohimaa et al., 2008, 2009; Korkman, Svanbäck, et al., 2014; Lahtinen et al., 2017).

Finnilä-Tuohimaa et al. (2008) also investigated the effect of their attitudes and beliefs scales on biased decision-making in CSA evaluations: Students and mental health professionals who possessed stronger attitudes and beliefs were more prone to conclude that CSA had taken place based on children's unspecific behavioral problems (e.g., wetting or moodiness) or sexualized behavior or interest in sexual topics (e.g., touching their own genitals or asking where babies come from), none of which are valid diagnostic indicators for sexual abuse (Kendall-Tucket et al., 1993; Lewis et al., 2016). Participants with more strongly held attitudes and beliefs

were also more likely to doubt a court's decision for acquittal that was based on a lack of evidence (i.e., the presumption of innocence) in such cases. Previous versions of two of the subscales (Pro Child and Anti Criminal Justice System; Finnilä-Tuohimaa et al., 2009) were also associated with lower sensitivity to suggestive questioning: Healthcare professionals with strong compared to moderate attitudes rated the probability of CSA higher and more strongly voted for conviction, even when they had read the highly suggestive child interviews conducted in those cases. Overall, these studies provide evidence for the assumption that cognitive constructs, such as specific attitudes, beliefs, thinking- or information processing styles can foster biased decision-making in CSA evaluations. The mechanisms of this process are not yet well understood. Possibly, these cognitive tendencies relate to underestimating the shortcomings of the evidence (Finnilä-Tuohimaa et al., 2009) or the base rates of false and distorted allegations, to ignoring alternative hypotheses, or to lowering subjective standards for burden of proof – all of which may foster confirmatory information processing and biased judgements and decision-making. Especially intuitive decision-making has been associated with systematic cognitive biases (Gilovich et al., 2002; Neal et al., 2022), in particular for domains with uncertain outcomes and no possibility for valid feedback (Hogarth, 2010), which is true for most CSA cases. Perceiving the justice system as incompetent may additionally trigger a need for compensation at the level of individual cases, leading to a further biased stance in favor of the abuse hypothesis (Finnilä-Tuohimaa et al., 2009).

Emotional Sources of Bias

The first reaction to a stimulus is often an emotional one (Zajonc, 1980), influencing attention and decision-making before more deliberate cognitive processes come into play. As such, emotions are a well-known source of cognitive biases (Lerner et al., 2015), such as the

confirmation bias (Jonas et al., 2006). In experimental studies, negative emotions like anger and disgust have been associated with more confrontational suspect interviews (Magnusson et al., 2021), and stronger preferences for guilty verdicts and harsh punishments (see Salerno, 2021, for a review). This may be due to a heightened desire to blame and punish a perpetrator and a lowered standard for burden of proof, which can induce confirmatory information processing (Salerno, 2021). Because child sexual abuse is an emotionally charged crime (e.g., Cheung & Boutte-Queen, 2000; Magnusson et al., 2021), emotions may influence confirmation- and interviewer bias especially strongly in this domain. However, to our knowledge, the influence of individual emotions on the level of bias in CSA investigations has not yet been empirically investigated.

Study Aims

We aimed to develop a self-report instrument measuring cognitive and emotional individual differences in handling CSA suspicions that predict vulnerability for a biased stance toward the abuse hypothesis. The target group is professionals who are not specifically trained in forensic interviewing or assessment, but who (regularly or irregularly) conduct informal initial interviews with children to clarify suspicions of sexual abuse in child protection, health care, education, or similar settings. Accordingly, we selected a sample of human sciences students who will become such professionals in the future. Regarding the measurement of cognitive sources of bias, a foundation has been built by the work surrounding the Child Sexual Abuse Attitudes and Beliefs Scale (CSAABS; Finnilä-Tuohimaa et al., 2008), which measures four cognitive constructs, each defined as a set of attitudes and beliefs towards CSA suspicions and investigations (Intuition, Pro Child, Disclosure, and Anti Criminal Justice System). However, several psychometric and conceptual issues leave the CSAABS in need for further development.

For example, the empirical base to distinguish attitude items (i.e., subjective opinions) from false belief items (i.e., convictions countering empirical evidence) has not been clearly reported and appears debatable for several items. Also, some items seem to represent information processing styles rather than attitudes or beliefs. Some item-factor allocations are questionable (low factor loadings and low content congruence), the translation procedure between English and Finnish has not been described, and the published article reports only 36 items of a 40-item-instrument. We therefore aimed to develop psychometrically sound scales to assess cognition related individual differences that influence bias in CSA investigations, by taking the constructs and item pool of the CSAABS as a starting point.

Adding to the cognitive style scales, we also aimed to construct one or more scales to assess emotional influences on bias. In particular, we intended to measure the self-reported intensity of negative emotional reactions when faced with the topic of CSA. As the influence of different emotions on bias has scarcely been researched, we took basic negative emotions such as anger, sadness, disgust, shame (Ekman, 1999) and hatred, when faced with the topic of CSA, as a starting point for the development of the emotion items. We assumed that an overall heightened emotional reactivity towards the topic of CSA would increase bias in individual cases of CSA. This reasoning was based on the notion that, in general, decision-making is not only influenced by one's emotions in the decision situation, but also by one's general emotional reaction toward the decision topic, and by one's emotional experiences from similar prevailing situations (Schwarz, 2012).

Methods

Ethical approval for this study was granted by the ethics committees of the FernUniversität in Hagen (EA_79_2019) and the Psychologische Hochschule Berlin (decision 09/19/2018).

Participants and Procedure

A total of $N = 1{,}153$ undergraduate and graduate students of psychology, pedagogy, educational studies, and social work across three German universities gave informed consent and took part in anonymous, 30- to 60- minute-long surveys, in either paper-pencil or online format. Depending on the universities' regulations, participants were compensated with study credits or could win 10 or 20 € vouchers. All participants worked on a pool of 66 items that was set up for developing the target questionnaire CECSA (see item pool description below). For subsamples of $n = 259$ to 391 participants, respectively, different validation instruments were added (aiming for samplesizes > 250 to obtain stable correlations; Schönbrodt & Perugini, 2013) as well as further questionnaires for the purpose of other research aims (see Table S1 in the online supplement for a list of instruments per subsample). All materials were presented in German.

For the scale construction, we removed participants who showed careless responding within the CECSA item pool (more than 10% missing values or more than 10 identical values in a row [max LongString; Meade & Craig, 2012]), leaving 801 participants for the scale construction. The same careless responding analysis was conducted for each of the validation instruments, which led to the exclusion of one further participant for one scale (Negative Emotionality). The final subsamples for the validation analyses consisted of 256 to 390 participants.

In the scale construction sample of 801 participants, 80.8% were women, 19.0% men, 0.3% ($n = 2$) identified as other genders, and 0.1% ($n = 1$) provided no gender information. Age

ranged from 18 to 61 years (M = 26.9, SD = 7.1, Mdn = 25; five participants did not report their age) and 12.4% reported having children on their own. Also, 12.7% had participated in other training programs about handling CSA suspicions (1.4% did not answer this item) and 21.1% had discussed the topic during university lectures (1.3% did not answer this item). Finally, 17.7% reported having been subjected to sexual assault (as children or adults), 78.2% negated this question while 4.1% did not answer it.

Materials and Measures

CECSA Item Pool Development

Aiming to develop scales assessing cognitive and emotional individual differences in handling CSA suspicions, five initial constructs were used as a starting point to set up an item pool: Four cognitive style constructs, based on the four subscales of the "Child Sexual Abuse Attitude and Belief Scale" (CSAABS; Finnilä-Tuohimaa et al., 2008), and one newly developed emotional construct. Across the five constructs, the initial item pool consisted of 66 items.

Cognitive style constructs: The four cognitive style constructs, adopted from the subscales of the CSAABS, were originally labelled as attitudes and beliefs regarding (1) *Disclosure,* (2) *Pro-Child*, (3) *Intuition*, and (4) *Anti Criminal Justice System.* We abandoned the original authors' distinction between attitudes (i.e., subjective opinions) and false beliefs (i.e., convictions countering scientific evidence) because of the lacking empirical base for this distinction and a strong conceptual overlap, and because some items seemed to rather present information processing styles or other cognitive tendencies. Instead, we used the term "cognitive styles" to summarize items on attitudes, beliefs, thinking styles and other cognitive patterns. To set up the item pool for the cognitive style constructs, we used 32 of the original CSAABS items (see

online supplement 2.1 for details on the item selection). These were supplemented with 21 newly developed items, intended to assess the core concepts in more depth (4 to 8 items per construct).

Emotional construct: To incorporate the emotional construct, 13 newly developed items on *Emotional Reactivity* towards CSA were included, either describing general emotionality (e.g., "The issue of child sexual abuse is more emotionally charged to me than it is to most other people", 4 items), or specific basic negative emotions such as anger, sadness, disgust, shame (Ekman, 1999) and hatred (9 items). To cover a broad scope of emotion-eliciting situations, encounters with CSA cases in real life (e.g., "When sexual abuse of children is discussed, I often feel sadness") and through media (e.g., "When the media reports about child sexual abuse, I often feel anger") were included.

All questionnaires were administered in German. CSAABS items were translated from the original Finnish and the published English into a common German version. Small adjustments were made to simplify complex or ambiguous wordings. Items describing previous encounters with CSA cases were adjusted to describe hypothetical or future situations. New items were developed in German. To obtain a matching English version of the final questionnaire for international usage, all items were independently back- and forth-translated between German and English (following recommendations for item translation by Schmitt & Eid, 2007). Agreement to all items were indicated on a 6-point scale (1 = *fully disagree*, 2 = *mostly disagree,* 3 = *somewhat disagree,* 4 = *somewhat agree,* 5 = *mostly agree,* 6 = *fully agree*). See supplement 5 for the initial item pool and Table 1 for the final item selection (German Version in supplement 4).

Scenario Ratings

Participants were presented with three scenarios describing suspected cases of CSA based on unspecific behavioral problems (e.g., mood problems, wetting) and mild age-appropriate sexual interest or behaviors (e.g., touching one's genitals; scenarios adopted from Finnilä-Tuohimaa et al., 2008; see supplement 2.2). For each scenario, participants had to indicate their belief that CSA had taken place on a 4-point scale (1 = *no*, 2 = *rather not*, 3 = *rather yes,* 4 = *yes*). For the third scenario, participants additionally read a follow-up story describing how the staff of a psychiatric clinic concluded that the child had been sexually abused, although the child had not made such a statement, how the police found no further evidence and how a court ultimately acquitted the suspect due to a lack of evidence. Participants were then asked to subjectively evaluate the court's decision on a 4-point scale (1 = *correct*, 2 = *rather correct*, 3 = *rather false,* 4 = *false*). Higher ratings on both measures (belief that CSA had taken place and evaluating the acquittal as false) were interpreted as more biased judgments, because unspecific behavioral problems and mild sexual interests and behaviors are empirically either not or only weakly associated to sexual abuse experiences (Kendall-Tucket et al., 1993; Lewis et al., 2016), especially given that many more non-abused than abused children exhibit such problems, it is not valid to infer abuse experiences from any single or a combination of these behavioral observations without an incriminating statement or hint from the child or a third person or other external evidence.

Self-Report Validation Measures

Instruments to test convergent validity were selected based on the cognitive and emotional constructs of the initial item pool, because we expected the final scales to cover similar concepts. All validation instruments were included in the main data collection. The

specific validation hypotheses were formulated after constructing the CECSA scales and are listed in Table 2.

For the cognitive style constructs, we included validation instruments on intuitive thinking- and decision-styles (German versions of the Faith in Intuition subscale from the Rational-Experiential-Inventory [REI; Epstein et al., 1996; Keller et al., 2000], and the questionnaire on Preference for Intuition and Deliberation [PID; Betsch, 2004]) and different measures on attitudes towards (in)justice (General Belief in a Just World Scale [Dalbert et al., 1987]; four items on punitive attitudes towards sexual offenders from the Scale on Punitive Attitudes [Armborst, 2014, 2017]; and the Observers Perspective subscale from the Justice Sensitivity Shortscales [Baumert et al., 2014; Beierlein et al., 2013], focusing on injustice done to others)

For the emotional construct, we included validation instruments on general empathy and negative emotions (Scales for the Assessment of Empathic Abilities [E-Scale; Leibetseder et al., 2001]; Negative Emotionality scale from the Big-Five Inventory 2 [BFI-2; Danner et al., 2016; Soto & John, 2017]), as well as one more specific item on the frequency of anger about sexual assault (developed to accompany the Scale on Punitive Attitudes; Armborst, 2014) and above mentioned scale on Justice Sensitivity (Baumert et al., 2014; Beierlein et al., 2013).

Statistical Analyses

We analyzed the data in three steps to (1) explore the factor structure of the initial 66 items pool, (2) create short scales with optimized psychometric properties through an Ant Colony Optimization (ACO) algorithm, and (3) test hypotheses for convergent validity. The analyses were conducted with R (v4.1.1.; R Core Team, 2021), mainly using the R packages stuart (Schultze, 2019) and Mplus (Muthén & Muthén, 2017).

Explorative Scale Construction

We conducted an exploratory factor analysis (EFA) with the initial pool of 66 items to explore the dimensional structure of the CECSA by applying oblimin rotation and maximum likelihood estimation. Parallel analysis (based on principal components, Timmerman & Lorenzo-Seva, 2011) and a screeplot were used as statistical retention criteria and complemented by theoretical considerations. We aimed to include at least 4 items per factor, with thresholds for item inclusion set to >.4 for factor loadings (i.e., convergent validity), and to > .2 for differences between primary and secondary loadings (i.e., discriminant validity).

Item Selection via Ant Colony Optimization

Ant Colony Optimization (ACO) is an automated item selection strategy that can be used to create short scales based on a predefined latent structure and customizable selection criteria (Schroeders et al., 2016; Schultze, 2017). In an iterative process, combinations of items are repeatedly selected and evaluated in order to optimize the predefined selection criteria. Throughout the process, items from more advantageous combinations have a higher chance to get selected in subsequent iterations, which successively leads to an optimized item set. This procedure is arguably superior to conventional strategies of manual item selection (e.g., sequentially selecting items based on their loadings; Leite et al., 2008; Olaru et al., 2019).

We aimed to compile a short questionnaire based on the initial item pool and the scale structure explored via EFA. We combined three selection criteria to simultaneously optimize (a) model fit, (b) reliability, and (c) predictive validity. Model fit was assessed via the comparative fit index (CFI; RMSEA was not used as an optimization criterion, because even randomly selected item sets often achieved values close to the benchmark of .06 [Hu & Bentler, 1999]). Reliability was assessed via internal consistencies of the scales (Cronbach's α and McDonald's

ω) and factor loadings of the items (with equal weights). Predictive validity was defined as the extent to which the scales predicted the four scenario tasks in multiple regression analyses (average R^2; subsample of 348 participants). The ACO models were estimated using the WLSMV estimator for ordinal variables. We estimated different models with varying numbers of items per scale (4 to 12) and selected the best solution regarding our selection criteria.

Convergent Validation

To assess convergent validity, we calculated Pearson correlations of the resulting CECSA scales with other established self-report measures, according to theoretically expected overlaps between the respective constructs. All validation measures, hypotheses, and subsample sizes are listed in Table 2.

Results

Explorative Scale Construction

Considering different retention criteria (screeplot, parallel analysis, our item inclusion criteria, and theoretical considerations), we extracted a three-factor solution from the initial pool of 66 items. The screeplot had suggested to extract either 2 or 4 factors and a parallel analysis a maximum of 8 factors (see Figure S3 in the online supplement), but inspecting solutions with 1 to 8 factors, only 3 factors met our item inclusion criteria and were theoretically coherent. The first factor ("Naïve Confidence" [NC]), included 14 items from the cognitive style constructs "Intuition" and "Pro Child" and describes overreliance in one's CSA recognition abilities (e.g., "I would trust my first impression when assessing whether a child was sexually abused or not") and in child abuse reports (e.g., "It is very unlikely that sexually abused children exaggerate when they tell about an abusive experience"). The second factor ("Emotional Reactivity" [ER]) consisted of 10 items from the emotional construct (e.g., "When it comes to the topic of child

sexual abuse, I react very emotionally"). The third factor, ("Justice System Distrust" [JSD]) included 8 items from the cognitive style construct "Anti Criminal Justice System" that reflected a distrust in the justice system's competence and willingness to handle CSA cases (e.g., "When it comes to child sexual abuse, courts are not taking children seriously enough"). The factors correlated positively, with the highest correlation found between NC and JSD (Pearson's $r = .45$) and lower correlations between ER and JSD ($r = .32$), and between NC and ER ($r = .20$). Item factor loadings are depicted in supplement 5.

Ant Colony Optimization (ACO)

We set the ACO algorithm to find an optimal set of items for the three scales, based on the three-factor solution from the EFA. However, for each scale, all initial items from the constructs that were represented by the corresponding EFA factor were included for the ACO procedure (i.e., all items from the "Intuition" and "Pro Child" constructs for the scale on Naïve Confidence, all items from the "Emotional Reactivity" construct for the corresponding scale, and items from the "Anti Criminal Justice System" construct for the scale on Justice System Distrust), resulting in 51 items. The items on the "Disclosure" cognitions were not included, because this construct was not empirically supported by the EFA.

After running ACO procedures with varying numbers of items per scale (4 to 12), we selected a solution with 11 items for the Naïve Confidence scale, 6 items for the Emotional Reactivity scale, and 6 items for the Justice System Distrust scale, as this solution reached the highest values on the optimization criteria, while also depicting the respective constructs in sufficient breadth. The final set of 23 items including descriptive statistics, latent factor loadings and each item's origin is depicted in Table 1. The final solution achieved acceptable model fit ($N = 801$, χ^2 (227) = 591.746, $p < .001$, CFI = .94; RMSEA = .045 [CI90% = .040, .049], SRMR =

0.045) and all three scales showed good internal consistencies (Emotional Reactivity: Cronbach's α = .88, McDonald's ω = .88; Naïve Confidence: α = .82, ω = 82; Justice System Distrust: α = .83, ω = 83). Item factor loadings varied between λ = .39 – .86 (see Table 1).

All three scales were positively intercorrelated, with the highest factor correlation found between NC and JSD (Pearson's *r* = .44) and lower correlations between ER and JSD (*r* = .32), and between NC and ER (*r* = .17). The scales NC and JSD were normally distributed with mean values around the center of the scale (NC: *M* = 3.06, *SD* = 0.69, *Mdn* = 3.09; JSD: *M* = 3.25, *SD* = 0.92, *Mdn* = 3.17), while the scale ER was slightly left-skewed (*M* = 4.55, *SD* = 1.03, *Mdn* = 4.67).

For the selection criterion of predictive validity, we assessed predictions for each scenario task separately, because their intercorrelations were rather low (r = |.05| - |-.29|; see Table S3 in the online supplement for correlations and descriptive values of the scenario tasks). Multiple regression analyses showed that the three CECSA scales explained between 6.2% and 15.3% of the variance for each of the four scenario tasks (see Table 3). The highest variance explanation was found for rating the correctness of the suspect's acquittal. Each scale uniquely contributed to predicting at least one rating task. Whereas the scale Naïve Confidence uniquely predicted each rating task (β = .17 – .28), the scale Emotional Reactivity contributed to predicting one of the abuse probability ratings (β = .12) and the scale Justice System Distrust contributed to predicting the correctness of acquittal rating (β = .26).

Convergent Validation

Most of the self-report validation instruments had acceptable to good internal consistencies within our data (Cronbach's α = .76 - .87; see Table 2). Only the internal consistencies of the E-Scales were somewhat unsatisfactory (α = .60 and .68). The scale Naïve

Confidence, as expected, showed small to moderate positive associations with faith in intuition (r = .21) and preference for intuition (r = .14), but not the expected negative association with preference for deliberation. As expected, the scale Emotional Reactivity correlated positively and quite strongly with measures of empathy (r = .38 - .42), and anger about sexual assault (r = .28), and small to moderately with negative emotionality (r = .12) and sensitivity to injustice done to others (r = .19). For the scale Justice System Distrust, as expected, stronger distrust was quite highly associated with stronger punitive attitudes (r = -.32), but we did not find the predicted associations with a belief in a just world or justice sensitivity (see Table 2 for more information).

Discussion

Following the idea that specific cognitive styles and emotional reactivity may be associated with a biased mindset when handling CSA suspicions, we present the development and validation of a self-report instrument on Cognitions and Emotions about Child Sexual Abuse (CECSA). The scale construction was based on a former questionnaire on attitudes and beliefs (CSAABS; Finnilä-Tuohimaa et al., 2008) and on theoretical considerations about emotional sources of bias. Item selection was conducted by means of an Ant Colony Optimization procedure (ACO; Schultze, 2017).

The resulting instrument consisted of 23 self-descriptive statements, grouped to three scales: “Naïve Confidence” (NC) reflects an overestimation of one’s ability to (intuitively) recognize abused children and an overestimation of the reliability of child abuse reports, “Emotional Reactivity” (ER) covers the intensity of one’s emotional reactions towards the topic of CSA, and “Justice System Distrust” (JSD) measures a distrust in the justice system’s competence and willingness to handle CSA cases. All three scales showed good internal consistencies and moderate intercorrelations, supporting the distinction into three scales and their

compilation into one instrument. Speaking in favor of predictive validity, all three scales predicted biased evaluations in CSA scenarios, indicating an association between the scales and a biased stance of prematurely confirming the abuse hypothesis when allegations are based on unspecific behavioral observations that are not diagnostic for CSA.

Convergent validity was demonstrated by correlations between the scales and theoretically overlapping self-report instruments. The Naïve Confidence scale overlapped with a general preference for intuitive decision-making, reflecting the many items on intuitively evaluating CSA suspicions. Against our assumption, NC did not correlate negatively with a preference for deliberate decision-making, suggesting that intuition and deliberation are not necessarily opposing constructs, as already discussed by Betsch (2004). The Emotional Reactivity scale showed overlaps with general negative emotionality, empathy, and sensitivity to injustice done to others, underlining a general emotional-empathic component of the scale, but it also correlated with more specific anger about sexual assault. The Justice System Distrust scale overlapped with punitive attitudes towards sexual offenders, but not with a general belief in a just world or justice sensitivity, suggesting that a distrust in the justice system to handle sex crimes may be independent of more general justice attitudes.

Because the CECSA scales were developed among human sciences students, who make up future child-protection, health, or education personnel, we recommend the practical use of the instrument for these professions. It is also these professionals, who usually hold conversations with children about abuse suspicions without specialized training (Brubacher et al., 2016; Cerezo & Pons-Salvador, 2004; Schols et al., 2013). The utility of the CECSA scales may extend to police officers, who occasionally conduct, but are not specifically trained for child interviews, although empirical validation and inquiry into specific challenges of such samples is warranted

(e.g., low variance or faking-good may be an issue for the Justice System Distrust scale in police samples). Generalizability to professionally trained forensic child interviewers (e.g., specialized police officers or forensic psychologists) cannot not be assumed without empirical validation, especially if their training included evidence-based recommendations for child interviewing (Korkman et al., 2024), active debiasing, or emotional coping strategies. From a psychometric perspective, important next steps for the CECSA scales are to confirm their structure and predictive validity in independent samples, including samples of native English speakers and working professionals, and to assess retest-reliability. Encouragingly, a small study on similar attitudinal scales from the CSAABS reported good retest-reliability (r = .82 to .91) for 26 students after three weeks (Finnilä-Tuohimaa et al., 2008). Regarding predictive validity, we provided first evidence that the CECSA scales predict bias in CSA evaluations. Most strongly, this seems to be the case for the Naïve Confidence scale. Future studies need to expand on testing convergent and predictive validity, specifically assess whether the CECSA scales also predict the behavioral manifestations of a biased mindset when talking to children about abuse suspicions: Suggestive questioning and sensemaking. Assessing changes of CECSA scores as a function of intervention will be another future endeavor to evaluate the practical utility of the scale.

For scientific purposes, the CECSA may be used to evaluate interviewer or conversational training programs, or to study the role of bias in forensic decision-making (e.g., in mock jury research; Goodman-Delahunty et al., 2021). From a practical perspective, the CECSA may be utilized to assess the individual training needs of prospective or practicing professionals that talk to children about CSA suspicions.

When designing training programs for such professionals, it seems advisable to integrate education on children's memory, disclosure, and statement patterns, the fallacies of human judgment and the need to evaluate alternative hypotheses, as well as emotional coping strategies to complement practical interview or conversational training (Finnilä-Tuohimaa et al., 2008; O'Donohue & Cirlugea, 2021). Influencing cognitive and emotional patterns in such a way may be more sustainable than mere behavioral training of non-suggestive questioning, which are cost-intensive and often yield only small long-term effects if not repeated regularly (Johnson et al., 2015; Poole, 2016). However, changing attitudes or similar individual traits can be a challenging process too (Albarracin & Shavitt, 2017), and research still needs to show that cognitions and emotions about CSA can be influenced through training endeavors. Promisingly, Lahtinen et al. (2017) reported that a one-year training program, including above mentioned aspects, decreased problematic attitudes and beliefs in a group of 27 investigative interviewers, and the effect largely sustained at a one-year follow-up.

It is important to stress that the CECSA is not an exhaustive collection of cognitive styles, emotions, or individual differences that can foster a biases mindset when handling CSA suspicions. Rather, it is a first compilation that future research may expand on. It is also worth mentioning that this compilation emphasizes only certain possible biases: While the CECSA focusses on differential constructs that predict a biased stance towards the abuse hypotheses (thus risking false positives), other research has focused on a biased stance towards the non-abuse hypothesis (risking false negatives, e.g., Cromer & Goldsmith, 2010). The latter bias, possibly expressed in a general disbelief of child abuse accounts, may be driven by attitudes such as trivializing abuse, or blaming the victims or by holding incorrect stereotypical concepts about CSA (Collings, 1997). However, while CSA stereotypes (e.g., that perpetrators are mostly

strangers or that only emotional reports of abuse are credible) seem to somewhat persist among laypeople (Márquez-Flores et al., 2016), but not necessarily among professionals (Korkman, Svanbäck, et al., 2014), trivialization or victim-blaming do not seem to be widespread (Cromer & Goldsmith, 2010; Goodman-Delahunty et al., 2017), in particular among childcare professionals in the United States (Rohrabaugh et al., 2016).

Results of the CECSA scales should not be evaluated normatively or as depicting incorrect knowledge. That is because while some items are countered by scientific evidence (e.g., "suggestive interview techniques only influence children's memories of details and banal things"; see Scoboria et al., 2017), others currently lack the empirical base for a valid evaluation (e.g., "when it comes to child sexual abuse, courts are not taking children seriously enough), are non-normative in nature (e.g., items on subjective emotions or impressions), or would be evaluated differently depending on country and region (e.g., item on juridical fairness; see Cross et al., 2003; Ernberg et al., 2018).

Conclusion

We developed and validated a self-report instrument with three scales on Cognitions and Emotions about Child Sexual Abuse (CECSA) that predict a biased stance toward the abuse hypothesis in CSA investigations. Relevant in particular for child-protection, healthcare or educational personnel, who professionally talk to children about CSA suspicions, the CECSA scales may be used to assess individual training needs or to evaluate existing training programs for handling CSA suspicions or for conducting conversations or interviews with potentially abused children.

Tables

Table 1

CECSA Subscales, Items and Descriptive Statistics

Item no.	Subscale / Item	*M*	*SD*	λ
	Naive Confidence	**3.06**	**0.69**	
1	I would trust my first impression when assessing whether a child was sexually abused or not.*	3.34	1.29	.63
3	You generally already know whether a child has been sexually abused or not before talking with him/her.*	4.21	1.07	.65
7	I cannot imagine that I would be fooled by a child when it comes to sexual abuse.	3.60	1.09	.57
8	False allegations of child sexual abuse are very rare.*	4.21	0.98	.42
11	Suggestive interview techniques only influence children's memories of details and banal things.*	3.99	1.20	.40
12	Even if children do not yet dare to tell about sexual abuse, I would very probably be able to recognize if something like this had happened to them.	3.50	1.09	.68
14	Adults who work a lot with children professionally, probably recognize intuitively whether a child is telling the truth about sexual abuse or not.	4.47	0.90	.59
17	Children have no reason to say that they have been sexually abused, if something like this has not actually happened to them.	3.01	1.15	.48
19	You can recognize whether a child was suggestively influenced.*	4.14	0.95	.43
20	It is very unlikely that sexually abused children exaggerate when they tell about an abusive experience.	4.58	1.00	.40
23	I can tell if a child is telling the truth about a sexual abuse.*	4.48	0.98	.70
	Emotional Reactivity	**4.55**	**1.03**	
2	When it comes to the topic of child sexual abuse, I react very emotionally.	4.36	1.27	.62
5	When the media reports about child sexual abuse, I often feel a lot of anger.	3.28	1.37	.86
10	When the media reports about child sexual abuse, I often feel a lot of disgust.	4.72	1.19	.66
15	When the media reports about child sexual abuse, I often feel strong hatred towards the offender.	4.75	1.21	.79
18	When the topic of child sexual abuse is discussed, I often feel sadness.	2.77	1.14	.69
22	When the topic of child sexual abuse is discussed, I often feel anger.	2.35	1.06	.79

	Justice System Distrust	**3.25**	**0.92**	
4	When it comes to child sexual abuse, courts are not taking children seriously enough.	4.12	1.10	.70
6	In cases of child sexual abuse, it is easy for a good lawyer to get an acquittal for the suspect.	3.41	1.16	.67
9	As long as there is no clear evidence, it is hopeless to report child sexual abuse to the police.	3.28	1.25	0.39
13	In cases of child sexual abuse, courts usually hesitate to convict the suspect.*	3.68	1.31	0.75
16	I don't have faith in the potential of the justice system to prosecute perpetrators of sexual abuse.*	2.76	1.15	0.74
21	Reports of child sexual abuse are not taken seriously enough by the police.	5.30	1.09	0.79

Note. N = 801. The numbering refers to the order of the items as presented to the participants. * = Item originated from the CSAABS scales. λ = standardized factor loading.

Table 2

Overview of Validation Measures, Descriptive Values, Validation Hypotheses and Results

Validation Construct (Instrument, Subscale)	Descriptive Values				Validation Hypothesis	Validation Results	
	n	*M*	*SD*	α		*r*	*p*
Faith in Intuition (REI, Faith in Intuition)	303	4.37	0.84	.87	NC (+)	**.21**	< .001
Preference for Intuition (PID, Intuition)	303	3.52	0.57	.76	NC (+)	**.14**	.016
Preference for Deliberation (PID, Deliberation)	303	3.88	0.60	.79	NC (-)	-.07	.252
Cognitive Empathy (E-Scale, Cognitive Concern)	256	2.70	0.53	.60	ER (+)	**.42**	< .001
Emotional Empathy (E-Scale, Emotional Concern)	256	2.99	0.46	.68	ER (+)	**.38**	< .001
Anger about sexual assault (Item „Frequency of anger")	259	3.18	1.44	-	ER (+)	**.28**	< .001
Negative Emotionality (BFI-2, Negative Emotionality)	390	2.57	0.62	.87	ER (+)	**.12**	.017
Justice Sensitivity (JSS, Observer's Perspective)	259	4.17	1.13	.80	ER (+)	**.19**	.002
					JSD (+)	.05	.456
Punitive attitudes about sexual offenders (PAS, Items on sexual delinquency)	259	2.69	0.71	.76	JSD (-)	**-.32***	< .001
Belief in a Just World (GBJW)	259	4.44	0.95	.84	JSD (-)	-.08	.196

Note. α = Cronbach's α; Validation Hypothesis = expected association between validation instrument and CECSA subscale; expected direction given in parenthesis. *r* = Pearson's correlations; significant correlations ($p < .05$) are depicted in bold. * = A negative value on the PA scale indicates a strong punitive attitude. NC = Naïve Confidence, ER = Emotional Reactivity, JSD = Justice System Distrust, REI = Rational Experience Inventory (Keller et al., 2000), PID = Preference for Intuition and Deliberation (Betsch, 2004), E-Scale = Scale for the Assessment of Empathic Abilities (Leibetseder et al., 2001), the Item „Frequency of anger about sexual assault" was introduced by Armborst et al. (2014), BFI-2 = Big-Five Inventory 2 (Soto & John, 2017), JSS = Justice Sensitivity Shortscales (Baumert et al., 2012), PAS = Punitive Attitudes Scale (Armborst, 2017), GBJW = General Belief in a Just World Scale (Dalbert et al., 1987).

Table 3

Results of four Multiple Regression Analyses, each Predicting one Scenario Rating from the three CECSA Subscales

	CECSA Subscales	β	p	R^2
Scenario 1: Probability of CSA				.062
	Naïve Confidence	**.17**	.013	
	Emotional Reactivity	**.12**	.047	
	Justice System Distrust	.04	.612	
Scenario 2: Probability of CSA				.058
	Naïve Confidence	**.26**	<.001	
	Emotional Reactivity	-.05	.427	
	Justice System Distrust	.01	.427	
Scenario 3: Probability of CSA				.091
	Naïve Confidence	**.28**	<.001	
	Emotional Reactivity	.09	.135	
	Justice System Distrust	.01	.936	
Scenario 3: Correctness of Acquittal				.153
	Naïve Confidence	**.27**	<.001	
	Emotional Reactivity	-.01	.890	
	Justice System Distrust	**.23**	.001	

Note. N = 384. The regression analyses were run with structural equation modeling. Thus, the CECSA subscales were modelled as latent factors. β = standardized regression coefficient. Significant coefficients (p <.05) are depicted in bold.

Data Availability Statement

The data that support the findings of this study are openly available in the Open Science Framework (OSF) at https://osf.io/uqhp4/.

Supplemental Material

Supplemental Material for this Study can be found on OSF: https://osf.io/uqhp4/.

CRediT Authorship Contribution Statement:

Elsa Gewehr: Conceptualization (lead), Methodology, Resources (lead), Investigation (lead), Data Curation, Formal analysis, Writing - Original Draft, Project administration. **Renate Volbert:** Conceptualization (supporting), Resources (supporting), Writing – Review & Editing, Supervision (equal). **Marie Merschhemke:** Resources (supporting), Writing - Review & Editing. **Pekka Santtila:** Resources (supporting), Writing - Review & Editing. **Simone Pülschen:** Conceptualization (supporting), Resources (supporting), Investigation (supporting), Writing – Review & Editing, Supervision (equal).

Acknowledgments

We would like to thank Kristin Jankowsky, Ulrich Schroeders and Johannes Zimmermann for their methodological advice. We would also like to thank Dahlnym Yoon for providing the opportunity for joint data collection and Larissa Jotzeit, Elisabeth Abel, Jenni Marie Klier, and Claudia Wenzel for carrying out parts of the data collection.

Funding Information:

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Declaration of Interest Statement

We have no conflicts of interest to disclose.

References

Akca, D., & Eastwood, J. (2021). The impact of individual differences on investigative interviewing performance: A test of the police interviewing competencies inventory and the five factor model. *Police Practice and Research*, *22*(1), 1027–1045. https://doi.org/10.1080/15614263.2019.1644177

Akca, D., Larivière, C. D., & Eastwood, J. (2021). Assessing the efficacy of investigative interviewing training courses: A systematic review. *International Journal of Police Science & Management, 23*(1), 73–84. https://doi.org/10.1177/14613557211008470

Albarracin, D., & Shavitt, S. (2017). Attitudes and Attitude Change. *Annual review of psychology, 69,* 299-327.

Armborst, A. (2014). Kriminalitätsfurcht und punitive Einstellungen: Indikatoren, Skalen und Interaktionen. *Soziale Probleme, 25*(1), 105–142.

Armborst, A. (2017). How fear of crime affects punitive attitudes. *European Journal on Criminal Policy and Research, 23*(3), 461–481. https://doi.org/10.1007/s10610-017-9342-5

Baumert, A., Beierlein, C., Schmitt, M., Kemper, C. J., Kovaleva, A., Liebig, S., & Rammstedt, B. (2014). Measuring Four Perspectives of Justice Sensitivity With Two Items Each. *Journal of Personality Assessment, 96*(3), 380–390. https://doi.org/10.1080/00223891.2013.836526

Beierlein, C., Baumert, A., Schmitt, M., Kemper, C. J., & Rammstedt, B. (2013). Four Short Scales for Measuring the Personality Trait of “Justice Sensitivity.” *methoden, daten, analysen, 7*(2), 279–310. https://doi.org/10.12758/mda.2013.015

Betsch, C. (2004). Präferenz für Intuition und Deliberation (PID). Inventar zur Erfassung von affekt- und kognitionsbasiertem Entscheiden. *Zeitschrift Für Differentielle Und Diagnostische Psychologie, 25*(4), 179–197.

Brown, D. A., & Lamb, M. E. (2015). Can Children Be Useful Witnesses? It Depends How They Are Questioned. *Child Development Perspectives, 9*(4), 250–255. https://doi.org/10.1111/cdep.12142

Brubacher, S. P., Powell, M. B., Snow, P. C., Skouteris, H., & Manger, B. (2016). Guidelines for teachers to elicit detailed and accurate narrative accounts from children. *Children and Youth Services Review, 63*, 83–92. https://doi.org/10.1016/j.childyouth.2016.02.018

Bruck, M., & Ceci, S. J. (1997). The suggestibility of young children. *Current Directions in Psychological Science, 6(*3), 75–79.

Ceci, S. J., & Bruck, M. (1995). *Jeopardy in the courtroom: A scientific analysis of children's testimony.* American Psychological Association.

Ceci, S. J., & Bruck, M. (2006). Children's suggestibility: Characteristics and mechanisms. In *Advances in Child Development and Behavior* (Vol. 34, pp. 247–281). Elsevier. https://doi.org/10.1016/S0065-2407(06)80009-1

Ceci, S. J., & Friedman. (2000). The suggestibility of children: Scientific research and legal implications. *Cornell Law Review, 86*(1), 34–108.

Ceci, S. J., Hritz, A., & Royer, C. (2016). Understanding Suggestibility. In W. O'Donohue & M. Fanetti (Eds.), *Forensic Interviews Regarding Child Sexual Abuse: A Guide to Evidence-Based Practice.* (pp. 141–153). Springer.

Cerezo, M. A., & Pons-Salvador, G. (2004). Improving child maltreatment detection systems: A large-scale case study involving health, social services, and school professionals. *Child Abuse & Neglect, 28*(11), 1153–1169. https://doi.org/10.1016/j.chiabu.2004.06.007

Cheung, M., & Boutte-Queen, N. M. (2000). Emotional responses to child sexual abuse: A comparison between police and social workers in Hong Kong. *Child Abuse & Neglect, 24*(12), 1613–1621. https://doi.org/10.1016/S0145-2134(00)00203-9

Collings, S. (1997). Development, Reliability, and Validity of the Child Sexual Abuse Myth Scale. *Journal of Interpersonal Violence, 12*(5), 665–674.

Cromer, L. D., & Goldsmith, R. E. (2010). Child Sexual Abuse Myths: Attitudes, Beliefs, and Individual Differences. *Journal of Child Sexual Abuse, 19*(6), 618–647. https://doi.org/10.1080/10538712.2010.522493

Cross, T. P., Walsh, W. A., Simone, M., & Jones, L. M. (2003). Prosecution of Child Abuse: A Meta-Analysis of Rates of Criminal Justice Decisions. *Trauma, Violence, & Abuse*, *4*(4), 323–340. https://doi.org/10.1177/1524838003256561

Dalbert, C., Montada, L., & Schmitt, M. (1987). Glaube an eine gerechte Welt als Motiv: Validierungskorrelate zweier Skalen. *Psychologische Beiträge, 29,* 596–615.

Dana, J., Dawes, R., & Peterson, N. (2013). Belief in the unstructured interview: The persistence of an illusion. *Judgment and Decision Making, 8*(5), 9.

Danner, D., Rammstedt, B., Bluemke, M., Treiber, L., Berres, S., Soto, C., & John, O. (2016). Die deutsche Version des Big Five Inventory 2 (BFl-2). Zusammenstellung Sozialwissenschaftlicher Ltems Und Skalen. https://doi.org/10.6102/zis247

Davey, R. I., & Hill, J. (1995). A study of the variability of training and beliefs among professionals who interview children to investigate suspected sexual abuse. *Child Abuse & Neglect, 19*(8), 933–942. https://doi.org/10.1016/0145-2134(95)00055-D

Duke, M., Elisabeth, R., & Price, H. (2016). Avoiding problems in child abuse interviews and investigations, in Forensic Interviews Regarding Child Sexual Abuse. In W. O'Donohue & M. Fanetti (Eds.), *Forensic Interviews Regarding Child Sexual Abuse: A Guide to Evidence-Based Practice.* (pp. 179–195). Springer.

Ekman, P. (1999). Basic Emotions. In T. Dalgeish & M. Power (Eds.), *Handbook of Cognition and Emotion.* Wiley.

Epstein, S., Pacici, R., & Denes-Raj, V. (1996). Individual differences in intuitive-experiential and analytical-rational thinking styles*. Journal of Personality and Social Psychology, 71*(2), 390–405. https://doi.org/doi.org/10.1037//0022- 3514.71.2.390

Erens, B., Otgaar, H., Patihis, L., & de Ruiter, C. (2020). Beliefs About Children's Memory and Child Investigative Interviewing Practices: A Survey in Dutch Child Protection Professionals from 'Safe Home.' *Frontiers in Psychology, 11.* https://doi.org/10.3389/fpsyg.2020.546187

Ernberg, E., Magnusson, M., & Landström, S. (2018). Prosecution of Child Sexual Abuse Cases Involving Preschool-Aged Children: A Study of Swedish Cases from 2010 to 2014. *Journal of Child Sexual Abuse*, *27*(7), 832–851. https://doi.org/10.1080/10538712.2018.1501786

Everson, M. D., & Sandoval, J. M. (2011). Forensic child sexual abuse evaluations: Assessing subjectivity and bias in professional judgements. *Child Abuse & Neglect, 35*(4), 287–298. https://doi.org/10.1016/j.chiabu.2011.01.001

Finnilä-Tuohimaa, K., Santtila, P., Björnberg, L., Hakala, N., Niemi, P., & Sandnabba, K. (2008). Attitudes related to child sexual abuse: Scale construction and explorative study among psychologists. *Scandinavian Journal of Psychology, 49*(4), 311–323. https://doi.org/10.1111/j.1467-9450.2008.00635.x

Finnilä-Tuohimaa, K., Santtila, P., Sainio, M., Niemi, P., & Sandnabba, K. (2005). Connections between experience, beliefs, scientific knowledge, and self-evaluated expertise among investigators of child sexual abuse in Finland. *Scandinavian Journal of Psychology, 46*(1), 1–10. https://doi.org/10.1111/j.1467-9450.2005.00429.x

Finnilä-Tuohimaa, K., Santtila, P., Sainio, M., Niemi, P., & Sandnabba, K. (2009). Expert judgment in cases of alleged child sexual abuse: Clinicians' sensitivity to suggestive influences, pre-existing beliefs and base rate estimates. *Scandinavian Journal of Psychology, 50*(2), 129–142. https://doi.org/10.1111/j.1467-9450.2008.00687.x

Fivush, R. (2011). The Development of Autobiographical Memory. *Annual Review of Psychology*, *62*(1), 559–582. https://doi.org/10.1146/annurev.psych.121208.131702

Gilovich, T., Griffin, D., & Kahneman, D. (Eds.). (2002). *Heuristics and biases: The psychology of intuitive judgment.* Cambridge university press.

Goodman-Delahunty, J., Martschuk, N., & Cossins, A. (2017). What Australian Jurors Know and Do Not Know about Evidence of Child Sexual Abuse. 86-103. *Criminal Law Journal*, *41*(2), 86-103.

Goodman-Delahunty, J., Martschuk, N., Lee, E., & Cossins, A. (2021). Greater Knowledge Enhances Complainant Credibility and Increases Jury Convictions for Child Sexual Assault. *Frontiers in Psychology*, *12*, 3362. https://doi.org/10.3389/fpsyg.2021.624331

Herman, S. (2005). Improving Decision Making in Forensic Child Sexual Abuse Evaluations. *Law and Human Behavior, 29*(1), 87–120. https://doi.org/10.1007/s10979-005-1400-8

Hogarth, R. M. (2010). *Educating Intuition.* University of Chicago Press. https://press.uchicago.edu/ucp/books/book/chicago/E/bo3624460.html

Horner, T., Guyer, M., & Kalter, Neil. (1993). The biases of child sexual abuse experts: Believing is seeing. *Bulletin of the American Academy of Psychiatry & the Law, 21*(3), 281–292.

Howe, M. L., & Knott, L. M. (2015). The fallibility of memory in judicial processes: Lessons from the past and their modern consequences. *Memory, 23*(5), 633–656. https://doi.org/10.1080/09658211.2015.1010709

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal, 6*(1), 1–55. https://doi.org/10.1080/10705519909540118

Huang, C.-Y., & Bull, R. (2021). Applying Hierarchy of Expert Performance (HEP) to investigative interview evaluation: Strengths, challenges and future directions. *Psychiatry, Psychology and Law, 28*(2), 255–273. https://doi.org/10.1080/13218719.2020.1770634

Johnson, M., Magnussen, S., Thoresen, C., Lønnum, K., Burrell, L. V., & Melinder, A. (2015). Best Practice Recommendations Still Fail to Result in Action: A National 10-Year Follow-up Study of Investigative Interviews in CSA Cases: Follow-up study of investigative interviews. *Applied Cognitive Psychology, 29*(5), 661–668. https://doi.org/10.1002/acp.3147

Jonas, E., Graupmann, V., & Frey, D. (2006). The Influence of Mood on the Search for Supporting Versus Conflicting Information: Dissonance Reduction as a Means of Mood Regulation? *Personality and Social Psychology Bulletin, 32*(1), 3–15. https://doi.org/10.1177/0146167205276118

Kask, K., Pompedda, F., Palu, A., Schiff, K., Mägi, M.-L., & Santtila, P. (2022). Transfer of Avatar Training Effects to Investigative Field Interviews of Children Conducted by Police Officers. *Frontiers in Psychology, 13.* https://www.frontiersin.org/articles/10.3389/fpsyg.2022.753111

Keller, J., Bohner, G., & Erb, H.-P. (2000). Intuitive und heuristische Urteilsbildung—Verschiedene Prozesse? *Zeitschrift für Sozialpsychologie, 31*(2), 87–101. https://doi.org/10.1024//0044-3514.31.2.87

Kendall-Tucket, K., Williams, L. M., & Finkelhor, D. (1993). Impact of sexual abuse on children: A review and synthesis of recent empirical studies. *Psychological Bulletin, 113*(1), 164–180. https://doi.org/10.1037/0033-2909.113.1.164

Korkman, J., Juusola, A., & Santtila, P. (2014). Who made the disclosure? Recorded discussions between children and caretakers suspecting child abuse. *Psychology, Crime & Law, 20*(10), 994–1004. https://doi.org/10.1080/1068316X.2014.902455

Korkman, J., Otgaar, H., Geven, L. M., Bull, R., Cyr, M., Hershkowitz, I., Mäkelä, J.-M., Mattison, M., Milne, R., Santtila, P., van Koppen, P., Memon, A., Danby, M., Filipovic, L., Garcia, F. J., Gewehr, E., Gomes Bell, O., Järvilehto, L., Kask, K., … Volbert, R. (2024). White paper on forensic child interviewing: Research-based recommendations by the European Association of Psychology and Law. *Psychology, Crime & Law*, *0*(0), 1–44. https://doi.org/10.1080/1068316X.2024.2324098

Korkman, J., Santtila, P., Drzewiecki, T., & Kenneth Sandnabba, N. (2008). Failing to keep it simple: Language use in child sexual abuse interviews with 3–8-year-old children. *Psychology, Crime & Law, 14*(1), 41–60. https://doi.org/10.1080/10683160701368438

Korkman, J., Svanbäck, J., Finnilä, K., & Santtila, P. (2014). Judges' views of child sexual abuse: Evaluating beliefs against research findings in a Finnish sample. *Scandinavian Journal of Psychology, 55*(5), 497–504. https://doi.org/10.1111/sjop.12147

La Rooy, D., Brubacher, S. P., Aromäki-Stratos, A., Cyr, M., Hershkowitz, I., Korkman, J., Myklebust, T., Naka, M., Peixoto, C. E., Roberts, K. P., Stewart, H., & Lamb, M. E. (2015). The NICHD protocol: A review of an internationally-used evidence-based tool for training child forensic interviewers. *Journal of Criminological Research, Policy and Practice, 1*(2), 76–89. https://doi.org/10.1108/JCRPP-01-2015-0001

Lahtinen, H.-M., Korkman, J., Laitila, A., & Mehtätalo, L. (2017). The effect of training on investigative interviewers' attitudes and beliefs related to child sexual abuse. *Investigative Interviewing: Research and Practice, 8*(1), 16–30.

Lamb, M. E., Orbach, Y., Hershkowitz, I., Esplin, P. W., & Horowitz, D. (2007). A structured forensic interview protocol improves the quality and informativeness of investigative interviews with children: A review of research using the NICHD Investigative Interview Protocol. *Child abuse & neglect*, *31*(11-12), 1201-1231. 10.1016/j.chiabu.2007.03.021

Leibetseder, M., Laireiter, A.-R., Riepler, A., & Köller, T. (2001). E-Skala: Fragebogen zur Erfassung von Empathie - Beschreibung und psychometrische Eigenschaften. *Zeitschrift Für Differentielle Und Diagnostische Psychologie, 22*(1), 70–85. https://doi.org/10.1024//0170-1789.22.1.70

Leite, W. L., Huang, I.-C., & Marcoulides, G. A. (2008). Item Selection for the Development of Short Forms of Scales Using an Ant Colony Optimization Algorithm. *Multivariate Behavioral Research, 43*(3), 411–431. https://doi.org/10.1080/00273170802285743

Lerner, J. S., Li, Y., Valdesolo, P., & Kassam, K. S. (2015). Emotion and Decision Making. *Annual Review of Psychology, 66*(1), 799–823. https://doi.org/10.1146/annurev-psych-010213-115043

Lewis, T., McElroy, E., Harlaar, N., & Runyan, D. (2016). Does the impact of child sexual abuse differ from maltreated but non-sexually abused children? A prospective examination of the impact of child sexual abuse on internalizing and externalizing behavior problems. *Child Abuse & Neglect*, *51*, 31–40. https://doi.org/10.1016/j.chiabu.2015.11.016

Magnusson, M., Joleby, M., Luke, T. J., Ask, K., & Lefsaker Sakrisvold, M. (2021). Swedish and Norwegian Police Interviewers' Goals, Tactics, and Emotions When Interviewing Suspects of Child Sexual Abuse. *Frontiers in Psychology, 12.* https://doi.org/10.3389/fpsyg.2021.606774

Marchant, R., & Turner, L. (2017). 'Opening Doors': Best practice when a young child might be showing or telling you that they are at risk. *Early Years Educator, 19*(6), 54–60. https://doi.org/10.12968/eyed.2017.19.6.54

Márquez-Flores, M. M., Márquez-Hernández, V. V., & Granados-Gámez, G. (2016). Teachers' Knowledge and Beliefs About Child Sexual Abuse. *Journal of Child Sexual Abuse, 25*(5), 538–555. https://doi.org/10.1080/10538712.2016.1189474

McGuire, K., & London, K. (2017). Common Beliefs About Child Sexual Abuse and Disclosure: A College Sample. *Journal of Child Sexual Abuse, 26*(2), 175–194. https://doi.org/10.1080/10538712.2017.1281368

Melinder, A., Brennen, T., Husby, M. F., & Vassend, O. (2020). Personality, confirmation bias, and forensic interviewing performance. *Applied Cognitive Psychology, 34*(5), 961–971. https://doi.org/10.1002/acp.3674

Melinder, A., Goodman, G. S., Eilertsen, D. E., & Magnussen, S. (2004). Beliefs about child witnesses: A survey of professionals. *Psychology, Crime & Law, 10*(4), 347–365. https://doi.org/10.1080/10683160310001618717

Muthén, L., & Muthén, B. (2017). Mplus. Statistical analysis with latent variables.

Neal, T. M. S., Lienert, P., Denne, E., & Singh, J. P. (2022). A general model of cognitive bias in human judgment and systematic review specific to forensic mental health. *Law and Human Behavior, 46*(2), 99. https://doi.org/10.1037/lhb0000482

Nickerson, R. S. (1998). Confirmation Bias: A Ubiquitous Phenomenon in Many Guises. *Review of General Psychology, 2*(2), 175–220. https://doi.org/10.1037/1089-2680.2.2.175

O'Donohue, W., & Cirlugea, O. (2021). Controlling for Confirmation Bias in Child Sexual Abuse Interviews. *The Journal of the American Academy of Psychiatry and the Law, 49*(3), 371–380. https://doi.org/10.29158/JAAPL.200109-20

Oeberst, A., & Imhoff, R. (2023). Toward Parsimony in Bias Research: A Proposed Common Framework of Belief-Consistent Information Processing for a Set of Biases. *Perspectives on Psychological Science*, 17456916221148147.

Olaru, G., Schroeders, U., Hartung, J., & Wilhelm, O. (2019). Ant Colony Optimization and Local Weighted Structural Equation Modeling. A Tutorial on Novel Item and Person Sampling Procedures for Personality Research. *European Journal of Personality, 33*(3), 400–419. https://doi.org/10.1002/per.2195

Patihis, L., Ho, L. Y., Tingen, I. W., Lilienfeld, S. O., & Loftus, E. F. (2014). Are the "Memory Wars" Over? A Scientist-Practitioner Gap in Beliefs About Repressed Memory. *Psychological Science, 25*(2), 519–530. https://doi.org/10.1177/0956797613510718

Pelisoli, C., Herman, S., & Dell'Aglio, D. D. (2015). Child sexual abuse research knowledge among child abuse professionals and laypersons. *Child Abuse & Neglect, 40*, 36–47. https://doi.org/10.1016/j.chiabu.2014.08.010

Perez, C. O., London, K., & Otgaar, H. (2022). A review of the differential contributions of language abilities to children's eyewitness memory and suggestibility. *Developmental Review, 63*, 101009. https://doi.org/10.1016/j.dr.2021.101009

Pompedda, F., Zhang, Y., Haginoya, S., & Santtila, P. (2022). A Mega-Analysis of the Effects of Feedback on the Quality of Simulated Child Sexual Abuse Interviews with Avatars. *Journal of Police and Criminal Psychology*. https://doi.org/10.1007/s11896-022-09509-7

Poole, D. A. (2016). *Interviewing children: The science of conversation in forensic contexts.* American Psychological Association.

Powell, M. B. (2008). Designing Effective Training Programs for Investigative Interviewers of Children. *Current Issues in Criminal Justice, 20*(2), 189–208. https://doi.org/10.1080/10345329.2008.12035804

Powell, M. B., Hughes-Scholes, C. H., & Sharman, S. J. (2012). Skill in Interviewing Reduces Confirmation Bias: Confirmation bias and interviews. *Journal of Investigative Psychology and Offender Profiling, 9*(2), 126–134. https://doi.org/10.1002/jip.1357

Quas, J. A., Malloy, L. C., Melinder, A., Goodman, G. S., D'Mello, M., & Schaaf, J. (2007). Developmental differences in the effects of repeated interviews and interviewer bias on

young children's event memory and false reports. *Developmental Psychology, 43*(4), 823–837. https://doi.org/10.1037/0012-1649.43.4.823

R Core Team. (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing. https://www.R-project.org/.

Rohrabaugh, M., London, K., & Hall, A. K. (2016). Planning the Forensic Interview. In W. O'Donohue & M. Fanetti (Eds.), *Forensic Interviews Regarding Child Sexual Abuse: A Guide to Evidence-Based Practice* (pp. 197–218). Springer International Publishing. https://doi.org/10.1007/978-3-319-21097-1_11

Salerno, J. M. (2021). The Impact of Experienced and Expressed Emotion on Legal Factfinding. *Annual Review of Law and Social Science, 17*, 181-203. https://doi.org/10.1146/annurev-lawsocsci-021721-072326

Schemmel, J., Datschewski-Verch, L., & Volbert, R. (2024). Recovered memories in psychotherapy: a survey of practicing psychotherapists in Germany. *Memory*, 1-21. https://doi.org/10.1080/09658211.2024.2305870

Schmitt, M., & Eid, M. (2007). Richtlinien für die Übersetzung fremdsprachlicher Messinstrumente. *Diagnostica, 53*(1), 1–2. https://doi.org/10.1026/0012-1924.53.1.1

Schols, M. W., de Ruiter, C., & Öry, F. G. (2013). How do public child healthcare professionals and primary school teachers identify and handle child abuse cases? A qualitative study. *BMC Public Health*, *13*, 807–807.

Schönbrodt, F. D., & Perugini, M. (2013). At what sample size do correlations stabilize? *Journal of Research in Personality, 47*(5), 609–612. https://doi.org/10.1016/j.jrp.2013.05.009

Schroeders, U., Wilhelm, O., & Olaru, G. (2016). Meta-Heuristics in Short Scale Construction: Ant Colony Optimization and Genetic Algorithm. *PLOS ONE, 11*(11), e0167110. https://doi.org/10.1371/journal.pone.0167110

Schultze, M. (2017). *Constructing Subtests Using Ant Colony Optimization.* https://doi.org/10.13140/RG.2.2.25738.52167

Schultze, M. (2019). *STUART: Subtests Using Algorithmic Rummaging Techniques.* https://cran.microsoft.com/snapshot/2022-03-13/web/packages/stuart/index.html

Schwarz, N. (2012). Feelings-as-information theory. In P. A. M. Van Lange, A. W. Kruglanski, & E. T. Higgins (Eds.), *Handbook of theories of social psychology, Vol. 1* (pp. 289–308). Sage Publications Ltd. https://doi.org/10.4135/9781446249215.n15

Scoboria, A., Wade, K. A., Lindsay, D. S., Azad, T., Strange, D., Ost, J., & Hyman, I. E. (2017). A mega-analysis of memory reports from eight peer-reviewed false memory implantation studies. *Memory, 25*(2), 146–163. https://doi.org/10.1080/09658211.2016.1260747

Soto, C. J., & John, O. P. (2017). The next Big Five Inventory (BFI-2): Developing and assessing a hierarchical model with 15 facets to enhance bandwidth, fidelity, and predictive power. *Journal of Personality and Social Psychology, 113*(1), 117–143. https://doi.org/10.1037/pspp0000096

Timmerman, M. E., & Lorenzo-Seva, U. (2011). Dimensionality assessment of ordered polytomous items with parallel analysis. *Psychological Methods*, *16*(2), 209–220. https://doi.org/10.1037/a0023353

Zajonc, R. B. (1980). Feeling and thinking: Preferences need no inferences. *American Psychologist, 35*(2), 151–175.

Zhang, Y., Segal, A., Pompedda, F., Haginoya, S., & Santtila, P. (2022). Confirmation bias in simulated CSA interviews: How abuse assumption influences interviewing and decision-making processes? Legal and Criminological Psychology, *27*(2), 314-328. https://doi.org/10.1111/lcrp.1221

Article 2:

Predicting Suggestive Questioning from Cognitions and Emotions about Child Sexual Abuse across Three Study Paradigms

Status: Invited to resubmit after revision at *Law and Human Behavior* (01.12.2024)

Elsa Gewehr[1,2], Marie Merschhemke[3], Simone Pülschen[3], Dietrich Pülschen[4], and Renate Volbert[1]

[1] Psychologische Hochschule Berlin, [2] Universität Kassel, [3] Europa Universität Flensburg, [4]Fachhochschule für Verwaltung und Dienstleistung, Fachbereich Polizei, Altenholz

Gewehr, E., Merschhemke, M., Pülschen, S., Pülschen, D., & Volbert, R. (2024). *Predicting Suggestive Questioning from Cognitions and Emotions about Child Sexual Abuse across Three Study Paradigms*. PsyArXiv. https://doi.org/10.31234/osf.io/qyfd9

Abstract

Objective: Although interviews and conversations are highly individual in nature, and suggestiveness is a major pitfall when questioning children, individual differences in interviewer bias and suggestiveness remain understudied. We assessed the influence of Cognitions and Emotions about Child Sexual Abuse (CECSA) on suggestive questioning and a biased mindset toward the abuse hypothesis across a series of three studies with varying mock paradigms and a meta-analytical integration.

Hypotheses: For all studies, we expect the scores on the three CECSA scales (Naive Confidence, Emotional Reactivity, and Justice System Distrust) to be associated with the number of suggestive questions and indicators of a biased mindset.

Method: In all studies, participants read mock cases about children displaying mild behavioral symptoms that were unspecific but gave rise to suspecting sexual abuse. In Study 1, 285 human sciences students further read interview transcripts and selected questions suitable to pose to the child. In Study 2, 241 police students read interview transcripts and freely formulated questions to pose to the children. In Study 3, 148 teaching students interviewed virtual children using natural language in a Virtual Reality setting.

Results: Across three studies and their meta-analytical integration, we found robust evidence that Naive Confidence and Emotional Reactivity, but not Justice System Distrust, significantly predict bias and suggestive questioning, with effect sizes of $b = .14–.37$. The newly developed measures to assess suggestive questioning validly captured a unidimensional trait of suggestive questioning but showed unsatisfying reliability.

Conclusions: The findings enhance our understanding of individual differences in suggestive questioning and bias and can inform the development, customization, and evaluation

of interviewer training programs, as well as the selection of interview personnel. We also provide recommendations to increase the reliability of the suggestiveness measures in future research.

Keywords: Child Sexual Abuse, Suggestion, Bias, Interviewing, Emotions

Predicting Suggestive Questioning from Cognitions and Emotions about Child Sexual Abuse Across Three Study Paradigms

Interviewing or talking to a child is a highly individual process prone to human error and at risk for false positive or false negative conclusions (Korkman et al., 2024; Lilienfeld, 2016). A common side finding of forensic interview research is that interviewers vary not only in the degree to which they strive to avoid either false positives or false negatives (Everson & Sandoval, 2011; Fessinger & McAuliff, 2020) but also, on a more behavioral level, in their degree of suggestiveness when talking to children (e.g., Brubacher et al., 2014; Finnilä-Tuohimaa et al., 2008; Johnson et al., 2015; Kask et al., 2022; Pompedda et al., 2022). While it has been acknowledged that children differ in their susceptibility to suggestion (Bruck & Melnyk, 2004; Klemfuss & Olaguez, 2020), variation in adults' suggestiveness has essentially gone unexamined (for a notable exception, see the work of Melinder et al., 2020, on Big Five personality traits and intervierwer performance).

The most acknowledged mechanism behind suggestive questioning is interviewer bias, a subtype of confirmation bias (Brown & Lamb, 2015; Ceci & Bruck, 2006; Powell et al., 2012; for details on confirmation bias, see Oeberst & Imhoff, 2023) of highly individual nature. It describes the a priori belief of an adult that a specific event (e.g., sexual abuse) has taken place and the following pursuit of its confirmation through mechanisms such as belief-consistent information processing and suggestive questioning to evoke confirmatory responses from the child (Ceci et al., 2016; Ceci & Bruck, 2006; Melinder et al., 2020; O'Donohue & Cirlugea, 2021; Zhang et al., 2022). Thus, suggestive questioning can be seen as the behavioral enactment of a biased interviewer's mindset (O'Donohue & Cirlugea, 2021).

The scientific community currently recommends mitigating interviewer suggestiveness on a behavioral level through the usage of structured interview protocols, repeated training of open questioning, and active consideration of alternative hypotheses—all of which have been shown to reduce suggestive questioning and improve child sexual abuse (CSA) investigations (Korkman et al., 2024; Lamb et al., 2011; Zajac & Brown, 2018). Identifying individual characteristics that are associated with adults' suggestiveness could further assist in promoting non-suggestive questioning, especially for earlier phases of CSA investigations. Knowledge about such individual differences could, for example, be used to select adequate personnel for conducting conversations or interviews with children, identify individual training needs of current personnel, or evaluate interviewer training programs, including their differential effects.

Questionnaire on Cognitions and Emotions about Child Sexual Abuse (CECSA)

Based on findings about attitudes, information processing styles or emotions that correlate with biased forensic decision-making (Finnilä-Tuohimaa et al., 2008, 2009; Magnusson et al., 2021; Neal et al., 2022; Salerno, 2021), Gewehr et al. (2023) developed the self-report instrument Cognitions and Emotions about Child Sexual Abuse (CECSA; see Table 1). It consists of three subscales, each of which is associated with bias when handling CSA allegations: Naive Confidence (NC; overestimation of one's innate ability to recognize abused children and of the accuracy of children's abuse reports), Emotional Reactivity (ER; intensity of emotional reactions to the topic of CSA), and Justice System Distrust (JSD; distrust in the justice system to adequately prosecute CSA). Convergent validity was demonstrated by theoretically derived correlations with other measures (NC: preference for intuitive decisions; ER: negative emotionality, empathy, sensitivity to injustice, anger about sexual assault; JSD: punitive attitudes toward sexual offenders). Further validity of the ER scale was demonstrated by Segal et al.

(2022, 2023), who showed that higher ER scores are associated with stronger self-reported and facially expressed emotions, specifically anger, during mock CSA interviews with virtual children. For the NC scale, Krause et al. (2024) showed that teaching students' NC scores can be reduced through lectures on evidence-based handling of CSA allegations.

The potential of the CECSA scales to predict a biased mindset was shown by Gewehr et al. (2023): Participants read scenario cases of children exhibiting unspecific behavioral problems (e.g., mood problems, bed wetting) or mild sexual behaviors (e.g., touching their own genitals) and were asked to rate the likelihood of CSA having taken place and their perception of fairness if a suspect in such a case were acquitted. Because the described behaviors are widespread and empirically not necessarily indicative of CSA (Kendall-Tucket et al., 1993; Lewis et al., 2016), they will, given the population base rate of CSA (Stoltenborgh et al., 2015), be exhibited much more often by non-abused than by abused children (for details, see commentary by London in Talwar et al., 2024). Thus, rating the likelihood of CSA as high and perceiving an acquittal as unfair were considered to be indicators of bias toward the abuse hypothesis. As expected, all three CECSA subscales were positively associated with a biased mindset (β = .12 - .28; for details on the CECSA scales, see the *Measures* section of Study 1).

Present Studies

So far, no study has examined whether the CECSA scales also predict behavioral enactment of a biased mindset, i.e., the suggestiveness of adults' questions when talking to children about CSA allegations. We aimed to close this gap with a series of three studies on the association between the CECSA scales and the tendency to pose suggestive questions in mock conversations about CSA suspicions. Using newly developed measures of suggestiveness, we assessed the tendency to pose suggestive questions in three different modalities with increasing

ecological validity: a single-choice format where participants selected their preferred question out of options with varying suggestiveness (Study 1), a free writing format where participants freely wrote down questions which were then coded for suggestiveness (Study 2), and a natural language format where participants freely posed verbal questions to a virtual child in an interactive virtual reality interview (Study 3). In Studies 1 and 2, we also assessed bias by asking participants to rate the likelihood of CSA in the mock cases. We varied the populations by sampling human science students, prospective police officers, and teaching students. Finally, we integrated the findings meta-analytically. With the multi-study approach, we aimed to evaluate the robustness of effects from individual studies, reduce the influence of methodological and sample artifacts, and explore the strengths and weaknesses of the different suggestiveness measures.

Study One: Single-Choice Format

The first study examined associations between the CECSA scales and suggestive questioning via a single-choice selection of questions in written mock conversations. A possible presence of bias was measured by asking participants how strongly the information from the mock cases and conversations indicate CSA. The sample was drawn from human sciences students. We expected higher values on each of the three CECSA scales to be associated with a higher number of selected suggestive questions (H1.1.1–H1.1.3) and higher ratings of CSA indicativity (H1.2.1–H1.2.3).

Method

All data, code, and materials of this study are publicly available in the open science framework (OSF; https://osf.io/hjkt2/?view_only=2b18866d0bf24a448280641f0e69aead). Study 1 was not preregistered, but its hypotheses, data exclusions, and statistical analyses were aligned

with the preregistrations for Studies 2 and 3 unless stated otherwise. All measures and data exclusions are described in the following. Ethical approval for this study was granted by the ethics committees of the FernUniversität in Hagen (EA_79_2019) and the Psychologische Hochschule Berlin (granted on 09/19/2018). All participants provided written informed consent.

Participants and Procedure

We recruited undergraduate and graduate university students of the human sciences (psychology, teaching, social work, pedagogy) for a voluntary online survey that took 35–45 minutes and was optionally rewarded with study credits and/or participation in a raffle for €10 vouchers. The survey consisted of a) questions about demographic characteristics and prior experiences with the topic of CSA, b) four mock cases about CSA suspicions, including case descriptions, interview transcripts, and tasks to select interview questions and rate the indicativity for CSA, and c) the CECSA questionnaire (see *Materials* for details). Further items fulfilled purposes outside of the present study (Negative Emotionality scale from the Big Five Inventory 2 [BFI-2; Soto & John, 2017], CECSA item development pool, evaluations of case vignettes [all described in Gewehr et al., 2023]).

The sample size was determined by an a priori power analysis for bivariate linear regressions (G*power, Faul et al., 2009). We aimed for a sample size of $N = 150$ to be able to detect effect sizes of $b \geq .2$ with a power of .8 and a type 1 error probability of .05. From 332 participants who took part in the study, we excluded (based on exclusion criteria preregistered for Studies 2 and 3) 45 participants due to missing data on the dependent variable (selected questions) and two participants due to careless responding (maximal longstring > 10) in the original CECSA item pool, leaving a final sample of 285 participants. We further excluded participants from individual analyses if they showed > 20% missing data on the respective

CECSA scale (pairwise completion), which left 257–266 participants per analysis (see Table 5). For participants with ≤ 20% missing data on an individual CECSA scale, we applied mean ipsative imputation. The participants' demographic characteristics and experiences regarding CSA are summarized in Table 2. Note that the sample of Study 1 was also part of the CECSA development sample (Gewehr et al., 2023).

Measures

Mock Conversations about CSA Suspicions with Single-Choice Question Selection. Participants read four mock cases about CSA suspicions. Each case includes a case description with information about the child, their environment, and recent unspecific symptoms, or problematic or unusual behaviors that are not necessarily indicative of sexual abuse (e.g., poor performance at school, social isolation, wetting their pants, playing naked). To provoke suspicion about a possible perpetrator, each case includes information about an adult who regularly spends alone time with the child and about whom the child talks negatively. For example, there is eight-year-old Paula, who enjoys ballet and attends tutoring sessions due to school struggles. Recently, she's been avoiding the restroom and often wet herself. Her tutor suggests doubling her sessions to help with school overload. However, Paula tells her teacher that she is afraid of the tutor, that he is unpleasant and often gets angry (see online Supplement 4 for all mock case descriptions). Each case description is followed by an excerpt from a conversation between the child and their teacher, who worries about the child having experienced sexual abuse. Over the course of the conversation, the child emphasizes their dislike of the suspect but does not mention sexual abuse or any other type of maltreatment. On three occasions in the excerpt, participants are asked to imagine themselves as the teacher and select the most suitable next question to pose to the child to further clarify the suspicion. They are presented with a single-choice item with four possible

(randomized) questions. Unknown to the participants, the four question options systematically vary, being a) open vs. closed and b) suggestive vs. non-suggestive (see Table 3 for examples). While open questions are broad in nature and encourage a free account, closed questions limit responses by proposing specific options. Specifically, the open questions are formulated as either invitations or directive questions, while the closed questions are formulated as option-posing questions (see Table 7 for the question type definitions, which are in line with common forensic taxonomies, e.g., Korkman et al., 2024). Suggestion is defined as the inclusion of a new piece of information regarding a possible adverse experience that has not been named by the child. After each mock case, participants are asked to rate how strongly they perceive the information from the case and conversation to be indicative of CSA on a scale from 0 (no indication of CSA) to 100 (clear indications of CSA) and how certain they were of this judgment (0 = very uncertain, 100 = very certain). Overall, participants are presented with four mock cases and conversations. They select three questions for each, resulting in 12 questions being selected.

The comparability of the four mock cases in terms of CSA indicativity was tested in a *pilot study*, which also helped to develop the questions for the single-choice items. Here, $N = 26$ participants read the above-described mock cases and conversation excerpts, but instead of selecting questions, they were asked to make up their own suitable next questions to ask the child and write them down. As in the main study, they also rated how strongly each of the cases indicated CSA. In the pilot study, we found no significant differences between the mean ratings of CSA indicativity of the four mock cases ($M = 60.31–70.9$; repeated measures ANOVA: $F(3) = 1.57$, $p = .061$, $\eta_G^2 = .024$), suggesting appropriateness to integrate them into a common measure. To develop the single-choice items for Study 1, we qualitatively derived typical questions that

participants from the pilot study often came up with and partially adjusted them to fit the predefined question categories. For detailed information on the pilot study, see Supplement 1.

Self-Report Questionnaire on Cognitions and Emotions about Child Sexual Abuse (CECSA). The Cognitions and Emotions about Child Sexual Abuse (CECSA; Gewehr et al., 2023) questionnaire is a self-report instrument that measures various cognitive and emotional patterns that can predict an individual's bias toward the abuse hypothesis when handling CSA allegations. It comprises 23 self-descriptive statements (see Table 1) organized into three subscales. The Naive Confidence (NC) scale assesses overestimation of one's (intuitive) capability to recognize whether a child has been sexually abused (e.g., "I would trust my first impression when assessing whether a child was sexually abused or not") and of the reliability of children's abuse reports (e.g., "It is very unlikely that sexually abused children exaggerate when they tell about an abusive experience"). The Emotional Reactivity (ER) scale measures the intensity of emotional responses to the subject of child sexual abuse (e.g., "When it comes to the topic of child sexual abuse, I react very emotionally"). The Justice System Distrust (JSD) scale evaluates skepticism regarding the justice system's competence and commitment to prosecuting CSA cases (e.g., "When it comes to child sexual abuse, courts are not taking children seriously enough"). For all items, agreement is indicated on a 6-point scale (1 = *fully disagree*, 2 = *mostly disagree,* 3 = *somewhat disagree,* 4 = *somewhat agree,* 5 = *mostly agree,* 6 = *fully agree*). The three subscales showed good internal consistencies (Cronbach's α = .82-.88, McDonald's ω = .82-.88) and moderate intercorrelations (Pearson's r = .17-.44) in the development sample of humanities students. As described in the introduction, higher ratings on each of the CECSA scales were associated with increased ratings of CSA likelihood in the mock cases of children

with unspecific behavioral issues. This indicates that the scales predict a biased mindset toward the abuse hypothesis.

Data Analysis

All statistical analyses were conducted with R (v1.4.1717; R Core Team, 2021). To derive a variable for suggestiveness, the number of suggestive questions from the 12 selected questions was summed. To assess whether these 12 indicators of suggestiveness measure a common unidimensional latent construct, we conducted a confirmatory factor analysis (CFA) using the R package lavaan (Rosseel, 2012) as a preliminary non-preregistered step. Details of the CFA are described in Supplement 2.

To test the hypotheses that the CECSA scales relate to suggestive questioning (H1.1.1–H1.1.3), we ran three bivariate regression analyses with the three CECSA scale scores as the independent variables and the total number of suggestive questions as the dependent variable. Exploratorily, we also ran a multiple regression analysis to explore the unique contributions of each CECSA scale in predicting the number of suggestive questions. Because the dependent variable strongly deviated from a normal distribution (skewness = 1.27), we opted for ordinal regression models for all four analyses (adhering to the preregistered criterion from Studies 2 and 3 to approximate normality when skewness is below 1). Following recommendations by Bürkner and Vuorre (2019), all ordinal regression models were fitted using the R package *brms* (Bürkner et al., 2023), applying Bayesian modeling with uninformative prior distributions for cumulative probit models. All resulting parameters are reported on a latent normally distributed variable that is estimated to underlie the ordinally distributed dependent variable. To evaluate the hypotheses, we report standardized regression coefficients (*b*; *M* of the posterior distribution) and credible intervals (CI) that include 95% of the posterior distribution. Credible intervals that do not include

zero allow rejection of the null hypothesis with a 5% type I error probability (similar to "significance" in frequentist statistics).

To test the hypotheses that the CECSA scales relate to the CSA indicativity ratings (H1.2.1–H1.2.3), we ran three bivariate linear regressions (skewness parameters indicated approximate normality, see Table 4) with the three CECSA scale scores as the independent variables and the mean CSA indicativity score (across four mock cases) as the dependent variable. Exploratorily, we further ran a multiple regression analysis to explore the unique contributions of each of the CECSA scales in predicting suggestive questioning.

Results

Descriptive and Preliminary Analyses

The three CECSA scales showed good internal consistencies (α = .79–.90) and moderate intercorrelations (r = .31–.40). Table 4 includes further descriptive results for the CECSA scales, and supplementary Table S5 includes further descriptive results for the items. Across the 12 indicators of suggestiveness (i.e., 12 occasions to select a question across four mock conversations), suggestive (vs. non-suggestive) questions were chosen rather rarely (M = 2.07, SD = 1.98, 17.25%). Closed (vs. open) questions were selected more often but still rarely (M = 4.88, SD = 2.57; see Table 4 for details). The CFA of the 12 indicators of suggestiveness showed a good fit for a unidimensional model (χ^2 [54.0] = 52.02, p = .551, CFI = 1.0, RMSEA = 0.0, SRMR = 0.09). Reliability, as measured through McDonald's Omega, was ω = .66, and all but one of the factor loadings were $\lambda > .4$ (for details, see supplementary Table S3). The four mock cases and conversational transcripts were, on average, rated as being rather indicative of CSA (M = 66.55, SD = 15.06), and participants felt rather certain (M = 63.16, SD = 18.1) about these

judgments (see Table 4 for details). Intercorrelations between all relevant study variables and demographic variables are shown in supplementary Table S6.

Regression Analyses

All ordinal regression models to predict suggestive questioning from the CECSA scales showed a good fit (Rhat ≤ 1.1, effective sample size > 1000; Bürkner & Vuorre, 2019). As predicted (H1.1.1–H1.1.3), each of the individual CECSA scales positively related to the number of suggestive questions. Detailed results are shown in Table 5 (estimates are plotted and reported on the latent normally distributed probit scale). Standardized regression coefficients indicate that a one-level increase on a CECSA scale is associated with an increase in suggestive questioning of b = .22 standard deviations for the NC scale, b = .26 for the ER scale, and b = .13 for the JSD scale. All 95% credible intervals excluded zero, indicating strong evidence. Exploratory multiple regression analysis showed that only NC (b = .15) and ER (b = .21; both CI_{95} excluding zero) were uniquely associated with suggestive questioning (JSD: b = .01, CI_{95} [-.13, .16]).

The results of the linear regressions to predict CSA indicativity ratings are summarized in Table 6. As predicted (H1.2.1–H1.2.3), all three CECSA scales positively related to the rating of the mock cases as indicative of CSA. Standardized regression coefficients were b = .36 (NC), b = .24 (ER), and b = .22 (JSD; all p < .001). In the exploratory multiple regression analysis, only NC (b = .29, p < .001) and ER (b = .13, p < .05) showed unique significant associations with CSA indicativity (JSD: b = .07, p = .292).

Discussion

In line with our hypotheses, all three CECSA scales significantly predicted suggestive questioning and a biased mindset toward the abuse hypothesis, with Naive Confidence (NC) having the largest effect sizes. The results of NC and Emotional Reactivity (ER) remained robust

in multiple regression analyses. All three scales also proved good reliability. The newly developed mock cases and conversation excerpts succeeded in creating slight to moderate suspicions about CSA, which still remained somewhat uncertain, as shown by the CSA indicativity and certainty ratings of 66% and 63%, respectively. The suggestiveness of the 12 chosen questions validly measured a common latent construct—a suggestive questioning style—as shown by the good CFA fit for a unidimensional model. The reliability of measuring suggestiveness ($\omega = .66$) was close to what is considered acceptable for early stages of research ($> .7$; Lance et al., 2006; Nunnally & Bernstein, 1994). Because participants selected few suggestive questions on average ($M = 2.07$; 17.25%), the limited variance in the number of suggestive questions may have impaired reliability and further potential for prediction through the CECSA scales. It may be that some participants tend to pose suggestive questions when they have to formulate them themselves but recognize the superiority of open, non-suggestive questions when they are offered as options (similar results were found for suspect interviews; Lidén et al., 2018; May et al., 2021).

Study Two: Free Writing Format

The second study examined associations between the CECSA scales and suggestive questioning in a free writing format, whereby participants generated their own questions to pose to a child in written mock conversations. Thus, by measuring suggestive questioning closer to how it is applied in conversational practice, we increase the ecological validity compared to Study 1. As in Study 1, bias was measured through CSA indicativity ratings. The sample consisted of prospective police inspectors at a university of applied sciences. As in the previous study, we expected higher values on each of the three CECSA scales to relate to a higher number of suggestive questions (H1.1.1–H1.1.3) and higher ratings of CSA indicativity (H1.2.1–H1.2.3).

Method

All hypotheses, data cleaning procedures, and analyses for this study were preregistered (https://aspredicted.org/GGN_GKW) unless stated otherwise. Open data and code for this study can be found via OSF (https://osf.io/hjkt2/?view_only=2b18866d0bf24a448280641f0e69aead). All measures and data exclusions are described in the following. We did not renew the ethics vote from Study 1 because there were only small changes in the study material. All participants provided written informed consent.

Participants and Procedure

We collected data from 241 undergraduate students who were in their first semester of studying to become police inspectors at a German university of applied sciences for public administration and whose curriculum had not yet covered CSA or child interviewing. Participation was voluntary, took place during lectures in a paper-pencil format, lasted 30 to 45 minutes, and was not compensated. The survey consisted of a) questions about demographic characteristics and prior experiences with the topic of CSA, b) four mock cases and conversations about CSA suspicions, with instructions to freely write suitable questions and rate CSA indicativity, and c) the CECSA questionnaire (see *Materials* for details). For other research aims, the survey also assessed CSA knowledge and punitive attitudes.

Based on an a priori power analysis for bivariate linear regressions (G*Power; Faul et al., 2009), we aimed for a sample size of $N = 150$ to be able to detect effect sizes of $b \geq .2$ with a power of .8 and a type 1 error probability of .05. We collected data from 241 participants to account for expected exclusions of 20% due to missing data because we did not control for completion of the paper-pencil questionnaires. Data exclusion was carried out as preregistered. After excluding 26 participants due to missing data for the dependent variable (formulated

questions), a final sample of 215 participants remained. For the CECSA items, careless responders (maximal longstring > 10) were not identified. Participants with > 20% missing data on an individual CECSA scale were excluded from the respective analysis (pairwise completion), which left between 212 and 214 participants per analysis (see Table 5). For participants with ≤ 20% missing data on an individual CECSA scale, we applied mean ipsative imputation. Participants' demographic characteristics and experiences regarding CSA are summarized in Table 2.

Measures

Mock Conversations about Child Sexual Abuse Suspicions with Free Writing of Questions. Participants read four mock cases and excerpts from conversations about CSA suspicions as described in Study 1 but framed as a police interview with children. On three occasions in each interview, participants are asked to freely come up with a suitable next question to pose to the child to clarify the abuse suspicion. Overall, each participant writes down 12 questions. As in Study 1, participants are asked to rate the indicativity of CSA and the certainty of their judgements on scales from 0 to 100.

Coding. To code the questions generated by participants, two independent coders were trained by the first author. Both coders coded all questions regarding a) the formal question category (seven options) and b) the presence and type of suggestion (five options). The coding scheme, depicted in detail in Table 7, was developed in line with common taxonomies from forensic interviewing literature (Korkman et al., 2024; Pompedda et al., 2015). Disagreements between the coders were resolved through discussion between one coder and the first author until a consensus was reached. For the present study, the more comprehensive codings were collapsed into two binary variables: suggestive (vs. non-suggestive) questions and closed (vs. open)

questions (other questions that did not suit this taxonomy were not included; see Table 7 for allocations of question categories to binary variables). As in Study 1, the dependent variable was derived by summing the suggestive questions of each participant across the 12 indicators. Interrater reliability for the sum of suggestive questions was excellent with ICC(2.2) = .90 (CI_{95} [.87, .93]) (Koo & Li, 2016).

Self-Report Questionnaire on Cognitions and Emotions about Child Sexual Abuse (CECSA). See Study 1 for a description of the questionnaire on Cognitions and Emotions about Child Sexual Abuse (CECSA; Gewehr et al., 2023).

Data Analysis

All statistical analyses were conducted with R (v1.4.1717; R Core Team, 2021). In a preliminary non-preregistered step, we conducted a CFA to assess whether the 12 indicators of suggestiveness measure a common unidimensional latent construct (i.e., a suggestive questioning style) using the R package lavaan (Rosseel, 2012; see Supplement 2 for details on the CFA). To test the hypotheses that the CECSA scales relate to the number of suggestive questions (H2.1.1–H2.1.3), we ran three bivariate regression analyses with each of the CECSA scale scores, respectively, as the independent variables and the number of suggestive questions as the dependent variable. Exploratorily, we also ran a multiple regression analysis to explore the unique contributions of each of the CECSA scales in predicting the number of suggestive questions. We conducted the same analyses to test the hypotheses that the CECSA scales relate to the CSA indicativity ratings (H2.1.1–H2.1.3), where the mean CSA indicativity score (across four mock cases) served as the dependent variable. All regression analyses were conducted using linear models, as the distributions of the dependent variables adhered to the preregistered criteria for approximate normality (skewness < 1).

Results

Descriptive and Preliminary Analyses

Of the CECSA scales, ER showed good internal consistency (α = .83), and values for NC and JSD were just below acceptable (α = .69, .68) for early research stages (Lance et al., 2006). Scale intercorrelations were low to moderate (*r* = .05–.19). Further descriptive scale statistics can be found in Table 4 and item statistics in supplementary Table S5. Across the 12 occasions, participants generated *M* = 3.84 (*SD* = 2.12, 32%) suggestive questions and *M* = 5.18 (*SD* = 2.39, 43%) closed questions (see Table 4 for details). The CFA of the 12 indicators of suggestiveness showed a good fit for a unidimensional model (χ^2 [54.0] = 52.77, *p* = .522, CFI = 1.0, RMSEA = 0.0, SRMR = 0.08). Reliability was low (ω = .52), and the factor loadings, although all positive, included seven loadings below λ = .4 (for details, see supplementary Table S3). The four mock cases and conversations were rated as being rather indicative of CSA (*M* = 69.97, *SD* = 13.1), and participants felt rather certain (*M* = 70.54, *SD* = 14.44) about their judgments (see Table 4 for details). Intercorrelations between all relevant variables are shown in the supplementary Table S7.

Regression Analyses

The results of the linear regression models to predict suggestive questioning are summarized in Table 5. Only ER significantly related to the number of suggestive questions (*b* = .14, *p* < .05). The exploratory multiple regression analysis showed no significant effects. The results of the linear regression models to predict CSA indicativity ratings are summarized in Table 6. Only NC showed a significant effect in the bivariate (*b* = .37, *p* < .001) and multiple (*b* = .36, *p* < .001) regression analyses.

Discussion

In Study 2, only the Emotional Reactivity scale predicted suggestive questioning, and only the Naive Confidence scale predicted bias (NC was robust in a multiple regression), providing support for 1/3 of our hypotheses. As in Study 1, the mock cases were perceived as somewhat indicative of CSA with moderate certainty. Compared to the numbers from Study 1, the participants posed a substantial number of suggestive questions (M = 3.84; 32%) in the free-writing format of Study 2. This is in line with suspect interview research reporting stronger belief-consistent questioning in free-writing formats than in single-choice formats, which has been attributed to stronger ecological validity regarding cognitive load (Lidén et al., 2018; May et al., 2021). Therefore, free writing seems to be the more suitable approach to capture indicators of confirmatory processes, such as a suggestive questioning style.

The good model fit of the CFA supported the notion that the suggestiveness of the 12 questions is explained by a latent factor (i.e., a suggestive questioning style). However, despite the increase in suggestive questions compared to Study 1, the 12 indicators measured latent suggestiveness with only questionable reliability (ω = .52; various low factor loadings), which was lower than in Study 1. Surprisingly, the CECSA scales NC and JSD also showed lower reliability (α = .69, .68) than in Study 1 or the scale development study (Gewehr et al., 2023), while ER again showed good reliability (α = .83). Thus, the lack of support for our hypotheses linking suggestive questioning to NC and JSD may be attributed to the insufficient reliability of the constructs on both ends. To compensate for unsystematic measurement error of the indicators and thereby increase reliability, further studies should not only use a format that evokes many suggestive questions but also increase the overall number of questions participants have to come up with in the mock conversations. Furthermore, the sample size should be increased because

although the sample of Study 2 ($N = 215$) exceeded the planned sample size ($N = 150$, powered to detect effects of $b \geq .2$), it may be fruitful to be able to detect smaller effects as well. In addition, simulations by Schönbrodt and Perugini (2013) suggest that a sample size of 250 is needed to obtain stable estimates of bivariate associations.

Study Three: Natural Language in an Interactive Virtual Reality Paradigm

Study 3 further increased ecological validity compared to Studies 1 and 2 by testing whether the CECSA scales relate to verbal suggestive questioning in interactive virtual reality mock conversations with virtual children about suspicions of child endangerment (e.g., CSA). The number of possible questions was also increased compared to Studies 1 and 2. Data was drawn from a study evaluating a training program on interview skills among teaching students ("ViContact", Krause et al., 2024). We expected higher values on each of the three CECSA scales to relate to a higher number of suggestive questions (H3.1–H3.3).

Method

All hypotheses, data cleaning procedures, and analyses for this study were preregistered (https://aspredicted.org/blind.php?x=7yc7b9) unless stated otherwise. Open data and code are available via OSF (https://osf.io/hjkt2/?view_only=2b18866d0bf24a448280641f0e69aead). All measures and data exclusions are described in the following. The study was approved by the ethics committee of the Psychologische Hochschule Berlin (granted on 09/19/2018). All participants provided written informed consent.

Participants

In total, 148 students training to become teachers participated in a Virtual Reality (VR) study with multiple sessions designed to evaluate interventions of a training program for child interviewing (Krause et al., 2024). For the analyses presented here, we used data from the

participants' first VR session, which was conducted prior to any training intervention. Adhering to preregistered criteria, we excluded one careless responder (maximal longstring > 10 in the CECSA scales) as well as six participants due to technical issues during VR sessions, leaving 141 participants for our analyses. While the original sample size was determined by the resources available for the evaluation study (Krause et al., 2024), we deemed the final sample of 141 participants sufficient for the present objectives, as, according to a sensitivity analysis for bivariate linear regressions (G*Power; Faul et al., 2009), it allowed the detection of effect sizes of $b \geq .2$ with a power of .8 and a type 1 error probability of .05. Participants' demographic characteristics and experiences regarding CSA are summarized in Table 2.

Procedure

Participation in the VR study was voluntary. The baseline session lasted 90 minutes and was compensated with €25. Participants provided demographic data and information about experiences with the topic of CSA and filled in the CECSA questionnaire (and a questionnaire on self-efficacy by Mensing et al. [2024], which is not relevant to the present study). They went through a familiarization phase in the virtual environment and then conducted two 10-minute VR interviews with two different virtual children, one male and one female (see *Measures* for details on the virtual children and interviews). In one interview, the child approached the participant and started the conversation, saying that they had something to tell (child-initiated interview). In the other interview, the participant had to initiate the conversation themself (teacher-initiated interview). Prior to each interview, participants read a case vignette about the child, including information about age, family, friends, housing, hobbies, school behavior, and, for teacher-initiated interviews, a concerning observation of the child's behavior (see Supplement 5 for a vignette example). The participant's task was to find out what had happened to the virtual child

and who was involved. Unknown to participants, each child had stored background information about having experienced one out of three possible critical events: a) sexual abuse, b) another protection issue (e.g., physical abuse), or c) a stressful event not relevant for child protection (e.g., an argument with another child). The balancing procedure assigning the four conditions (virtual child [male and female], interview initiator, and critical event) to each interview are detailed in Krause et al. (2024). After each interview, participants were asked to briefly write down what they had found out about the child's critical experience and were presented with multiple-choice questions about the event-type (a, b, or c) and the persons involved. Participants also filled in VR experience questionnaires before and after the interviews (not relevant here; see Krause et al. [2024] for details). At all times in the laboratory, a research assistant was present with the participant, guiding them through the study procedure and coding their questions during the interviews (see *Coding scheme*).

Materials & Measures

Virtual Reality Interviews and Children. In a three-dimensional virtual reality setting, entered via a headset, the participant sits in a classroom behind a teacher's desk and faces a virtual child. Key aspects of the case vignette are presented on a virtual notepad on the desk. Communication with the virtual children occurs via natural verbal language, with unlimited questioning in a ten-minute timeframe. For technical details, see Krause et al. (2024) and Barbe et al. (2023).

The participant meets one of eight virtual children designed to simulate conversational behavior typical of ten-year-old children. Each is equipped with an individual memory covering everyday topics (e.g., family, friends, school) and one critical event, which can be a) sexual abuse, b) another child protection issue, or c) a less harmful negative event. The selection of the

virtual child's answers occurs through a combination of a) automatic identification of keywords from participants' questions, b) human coding of question categories, and c) a probabilistic algorithm based on empirical findings on children's response patterns (see Supplement 6 for details). The coding of questions is performed simultaneously during the interview by a human operator (see *Coding Scheme* for details).

Coding Scheme and Procedure. Ten operators were trained to code the VR interviews. During each interview, one operator simultaneously coded each utterance in terms of a) the formal question category (six options) and b) the presence and type of suggestion (three options). Further codings regarding rapport were conducted but are not analyzed here. The operators were blind to the present hypotheses. Interrater agreement was assessed by coding interview transcripts and indicated good performance for formal question categories (Fleiss' $\kappa = .81$) and the presence and type of suggestion or rapport ($\kappa = .75$). For the present study, the codings were collapsed into the two binary variables suggestive (vs. non-suggestive) question and closed (vs. open) question (utterances of greeting and saying goodbye to the child were excluded; see Table 7 for allocations of question categories to binary variables). The dependent variables for our analyses were derived by calculating the mean number of suggestive questions per interview across the participants' two interviews.

Self-Report Questionnaire on Cognitions and Emotions about Child Sexual Abuse (CECSA). See Study 1 for a description of the questionnaire on Cognitions and Emotions about Child Sexual Abuse (CECSA; Gewehr et al., 2023).

Data Analysis

Statistical analyses were conducted with R (v1.4.1717; R Core Team, 2021). In a preliminary non-preregistered step, we estimated the reliability of measuring a suggestive

questioning style via correlations and split-half reliabilities (Spearman-Brown prediction formula) of the number and percentage of suggestive questions between the participants' two VR interviews. A CFA was not feasible here because conversational turns (i.e., virtual children's utterances and opportunities for participants to pose questions) were neither standardized nor limited in number, so they could not be aggregated into a fixed set of items across participants.

To test the hypotheses that the CECSA scales relate to suggestive questioning (H3.1–H3.3), we ran three bivariate regression analyses with each of the CECSA scale scores, respectively, as the independent variables and the mean number of suggestive questions per interview as the dependent variable. In three further multiple regression analyses, we exploratorily added the total number of questions posed as a control variable for each CECSA scale's prediction of suggestiveness. A last exploratory multiple regression analysis assessed the unique contributions of each of the three CECSA scales in predicting the mean number of suggestive questions.

Results

Descriptive and Preliminary Analyses

Internal consistencies of the CECSA scales were close to acceptable for NC ($\alpha = .69$), acceptable for JSD ($\alpha = .76$), and good for ER ($\alpha = .82$). Scale intercorrelations were low to moderate ($r = .02–.28$). Further descriptive scale statistics are shown in Table 4 and item statistics in supplementary Table S5. Across the two VR interviews, participants posed $M = 26.35$ ($SD = 5.48$) questions on average per conversation, of which $M = 12.18$ ($SD = 3.78$; 46.22%) were closed questions. Only a few questions per conversation ($M = 2.69$, $SD = 2.01$; 10.21%) were suggestive (see Table 4 for details). The correlation between the numbers of suggestive questions in the participants' two interviews was $r = .41$, and split-half reliability

was .58. Similarly, the correlation between the percentages of suggestive questions in the two interviews was $r = .45$, and split-half reliability was .61, which overall indicates questionable reliability.

Regression Analyses

Table 5 shows the results of the linear regression models to predict suggestive questioning (H3.1–H3.3). Only NC significantly related to the number of suggestive questions in the bivariate analysis ($b = .17$, $p < .05$) and the multivariate analysis that included all CECSA scales ($b = .18$, $p < .05$). Results of the multiple regression analyses controlling for the overall number of questions were non-significant (see supplementary Table S4).

Discussion

In Study 3, suggestive questioning was predicted only by the Naive Confidence (NC) scale (robustly in a multiple regression), providing support for only one of our three hypotheses. Participants posed few suggestive questions ($M = 2.69$) per conversation. Given the large total number of questions on average ($M = 26.35$), this reveals a much lower rate of suggestiveness (10.21%) relative to the two previous studies. At first, this may seem to counter the assumption that paradigms that are more ecologically valid, especially in terms of cognitive load, are more suitable for capturing confirmatory processes (Lidén et al., 2018; May et al., 2021). However, the paradigms vary in other ways as well, making them difficult to compare. For example, the paper-pencil designs (Studies 1 and 2) provide only four question opportunities per case, which renders each individual question relevant for clarifying the suspicion. This may prompt stronger confirmatory processes for each question compared to the mock conversations of Study 3, where the dialogues unfold more slowly and utterances also follow other goals, such as establishing rapport. Thus, from an assessment perspective, the VR paradigm, despite its improved ecological

validity and larger number of questions, is not more suitable for capturing confirmatory processes and suggestive questioning.

Perhaps due to its rare occurrence, suggestiveness was measured with only questionable reliability, as indicated by the split-half reliabilities and intercorrelations between participants' conversations. As in the previous studies, the limited presence and reliability of suggestiveness may have impeded further prediction through the CECSA scales, and an increased sample size would provide the power to detect smaller effects and provide higher stability of estimates (Schönbrodt & Perugini, 2013). This may be particularly relevant for reassessing the non-significant effect of Emotional Reactivity (ER; b = .07). The effect found for Justice System Distrust (JSD), however, was not even in the positive range (b = -.01) and therefore provides no support for our hypothesis.

Meta-analytical integration

To evaluate the predictive validity of the CECSA scales across studies, we meta-analytically integrated the results of the bivariate regression analyses from Studies 1, 2, and 3. Specifically, we ran six (non-preregistered) fixed effects meta-analyses using the R package metafor (Viechtbauer, 2010): Three meta-analyses focused on the bivariate regression coefficients for predicting suggestive questioning from each of the three CECSA scales across Studies 1, 2, and 3 (N = 640 participants, k = 3 studies), and three meta-analyses focused on the bivariate regression coefficients for predicting CSA indicativity from each of the three CECSA scales across Studies 1 and 2 (N = 400 participants, k = 2 studies). The decision for the fixed-effects variant was based on the small number of included studies (k = 2; k = 3), which would result in imprecise estimations of between-study variance in random-effects meta-analyses (Hedges & Vevea, 1998).

Results

All meta-analyses yielded significant results. Suggestive questioning was significantly predicted by the three CECSA scales Naive Confidence (b = .16, SE = .04, p < .001, CI_{95} [.08, .24]), Emotional Reactivity (b = .17, SE = .04, p < .001, CI_{95} [.09, .26]), and Justice System Distrust (b = .08, SE = .04, p = .040, CI_{95} [.004, .16]). Similarly, CSA indicativity was meta-analytically predicted by the CECSA scales Naive Confidence (b = .37, SE = .04, p < .001, CI_{95} [.29, .45]), Emotional Reactivity (b = .19, SE = .05, p < .001, CI_{95} [.10, .28]), and Justice System Distrust (b = .11, SE = .05, p = .016, CI_{95} [.02, .20]). Notably, the JSD scale showed the smallest effect sizes in both analyses on a significance level of $\alpha = .05$, while all other effects were significant on the level of $\alpha = .001$.

General Discussion

We tested the impact of Cognitions and Emotions about Child Sexual Abuse (CECSA) on suggestive questioning and bias with a series of three studies and a meta-analytical integration. Overall, we found robust evidence that the subscales Naive Confidence and Emotional Reactivity, but not Justice System Distrust, can predict a suggestive questioning style and a biased mindset toward the abuse hypothesis.

Prediction of Suggestive Questioning and Bias through Cognitive and Emotional Patterns

Across all studies, we found good evidence that the Naive Confidence (NC) scale predicts suggestiveness, with significant effects in the meta-analytical integration (b = .16) and two of the individual studies (b = .17, .22), including the most ecologically valid natural language paradigm. Similarly, we found good evidence that Emotional Reactivity (ER) predicts suggestiveness, with significant effects in the meta-analytical integration (b = .17) and two individual studies (b = .26, .14; not in the natural language paradigm). For Justice System

Distrust (JSD), we found a significant effect only in the least ecologically valid single-choice paradigm (b = .15) and a tiny, barely significant effect in the meta-analysis (b = .08, α = .04).

When predicting a biased mindset based on the CSA indicativity ratings of Studies 1 and 2, we found results similar to those for suggestiveness. NC showed the largest predictive effects in both individual studies (b = .36, .37) and in the meta-analysis (b = .37). ER showed a significant effect in Study 1 (b = .24) and in the meta-analysis (b = .19), and JSD did as well, but to a smaller degree (Study 1: b = .22; meta-analysis: b = .11).

Overall, Naive Confidence and Emotional Reactivity predicted both bias and suggestive questioning rather robustly and largely in line with our hypotheses, and they showed significant (non-preregistered) meta-analytical effects. The effect sizes (b = .14–.37) are in line with those commonly found in personality, social, and applied psychology (r ≈ 20; Bosco et al., 2015; Gignac & Szodorai, 2016; Richard et al., 2003). While the well-known heuristic by Cohen (1988) would classify them as low- to medium-sized, more contemporary, empirically derived guidelines would consider them medium to large effects (Bosco et al., 2015; Funder & Ozer, 2019; Gignac & Szodorai, 2016). Notably, when predicting behavior based on individual differences, the real-life impact of the often small effects (Bosco et al., 2015) lies in their accumulation across repeated behaviors (Funder & Ozer, 2018; Thielmann et al., 2020). As such, even small influences of the interviewer's cognitive and emotional characteristics on their degree of suggestiveness can accumulate to create strong suggestive pressure when children are interviewed at length or repeatedly.

This underlines the validity of the two CECSA scales NC and ER and their practical utility to, for example, assist in the selection of suitable interview personnel or the identification of interviewers' individual training needs. In addition, training curricula on evidence-based

interviewing could be enriched or even individualized by modules on emotional coping strategies (targeting high Emotional Reactivity) and on abuse indicators, children's disclosure patterns, and human judgment fallacies (targeting Naive Confidence). While many participants would benefit from such content, individually customized curricula could allow more efficient allocation of resources. Furthermore, training programs that target such bias-associated characteristics can be evaluated using the CECSA scales NC and ER. Regarding the Justice System Distrust scale, we do not deem the (especially meta-analytical) evidence from our analyses strong enough to recommend practical or scientific application when it comes to predicting suggestiveness or bias.

From an empirical knowledge perspective, the present results help to shed light on the understudied issue of differential suggestiveness: Individuals who naively believe that they can innately or intuitively identify sexually abused children and that children will only provide accurate information have a higher risk of falling into suggestive questioning, as do people who tend to have strong emotional reactions to the issue of CSA. However, individuals' skepticism regarding the judicial system's ability to adequately prosecute CSA does not necessarily relate to their degree of suggestiveness when questioning children.

Performance of the Mock Case Material and Measures of Suggestive Questioning

The newly developed mock case material used in Studies 1 and 2 successfully triggered some degree of suspicion in participants regarding potential sexual abuse experiences of the children, as indicated by the CSA indicativity and certainty ratings between 63–71%.

The 12 indicators of suggestive questioning in both Study 1 (single-choice selected questions) and Study 2 (coded written questions) showed good fit for a unidimensional model in confirmatory factor analyses, indicating appropriateness to summarize them in a common construct measuring suggestive questioning style. However, the reliability of the measure was

not satisfying (Study 1: ω = .66, Study 2: ω = .52), possibly due to the dichotomous nature of the items and the rather low absolute number of suggestive questions (Study 1: *M* = 2.07; 17%; Study 2: *M* = 3.84, 32%), which limits the potential to explain latent variance. This is in line with or even increased compared to previous research, which has typically found suggestive questions to be rare (but consequential) with 8–15% suggestive questions in field studies (Cederborg et al., 2000; Johnson et al., 2015; Korkman et al., 2008; Peixoto et al., 2017) and even rarer (5–9%) in laboratory mock interviews (Cyr et al., 2021; Kask et al., 2022; Pompedda et al., 2015; Sternberg et al., 2001) across varying samples and professions (see Brubacher et al., 2014, for an example of 33% among teachers).

Although non-suggestive questions remained more frequent, participants posed almost double the number of suggestive questions in the ecologically more valid free writing format (32%) than in the single-choice paradigm (17%). This is in line with findings from suspect-interview research, where free writing also prompted more belief-consistent questioning than single-choice formats (Lidén et al., 2018; May et al., 2021). It seems that people generally do appreciate the superiority of non-suggestive questions, at least if they are made aware of the option. However, they less often formulate non-suggestive questions themselves, either because such questions don't come to mind if not explicitly mentioned or because it is cognitively more demanding to generate a good question than to detect one (Lidén et al., 2018). Indeed, confirmation bias is thought to flare up under high working memory demands (Evans & Stanovich, 2013; Neal et al., 2022), and the same can be assumed for suggestive questioning. Therefore, we suggest that future studies aiming to capture suggestiveness in a paper-pencil design use a free writing format because it is more ecologically valid and prompts more suggestive questions than the single-choice paradigm. However, we recommend increasing the

number of suggestiveness indicators to compensate for unsystematic measurement error, thus improving reliability (e.g., by increasing the number of mock cases or the number of question opportunities within the existing, or possibly prolonged, mock interviews). Measuring suggestiveness in a more fine-grained way (i.e., with more levels than our dichotomous variable) could further increase reliability.

The same recommendations apply for measuring suggestive questioning through natural language in dynamic virtual interviews, as participants in Study 3 posed similarly few suggestive questions (M = 2.69 per conversation; 10%) and reliability was similarly unsatisfying. Although the dynamic virtual conversations increase ecological validity and question opportunities compared to Studies 1 and 2, they are not necessarily more suitable for prompting suggestive questions. The low reliability can also be explained by the varying difficulty of indicators within and between participants, as, in contrast to Studies 1 and 2, each virtual interview unfolded individually, meaning that the context in which a question had to be posed (i.e., the previous response of the child) was different for each question. Note that the limited reliability of the suggestiveness measurement in Studies 1-3 does not necessarily limit the reported prediction findings, as the model fit indicated high validity of a unidimensional construct. Rather, it suggests that future studies with higher reliability may be able to explain an even larger share of variance in suggestive questioning.

Limitations and Future Directions

A general limitation of our studies is the unknowable generalizability of the suggestiveness and bias measures and the CECSA predictions to real-world settings. We partially addressed this issue by increasing the ecological validity across the study designs, but future studies ought to assess associations between the three measures and real-life interview

performances of working professionals. In addition, testing the long-term stability of the CECSA scales and a suggestive questioning style has yet to be tested, which is a prerequisite for assuming stable individual differences. Future studies should close this gap by reassessing individuals after varying timeframes.

Over the three studies, we not only varied the suggestiveness measure but also used different samples (human science vs. police vs. teaching students), so we cannot with certainty attribute the differences between the study results to either of the two factors. As the human science sample (Study 1) consisted of almost 50% teaching students, a large overlap with the teaching students sample (Study 3) seems likely, but sample differences could account for the results of the police students (Study 2). In particular, the increased number of suggestive questions in Study 2, which we attributed to the free writing format, may have been due to characteristics of the police student sample. In contrast, the CECSA scores were remarkably similar for the three samples, with the exception of the police students' descriptively lower mean JSD score. Given that these students decided to work for and are educated by law enforcement, this is unsurprising. Regarding the prediction findings, in particular the robust findings for NC and ER, we have no theoretical reason to assume that associations between these scales and suggestiveness would vary among professions. Nevertheless, future studies should reassess the predictions using the free writing format with different samples.

There were some exceptions to the overall pattern of support for our hypotheses regarding NC and ER: In Study 2, NC did not predict suggestion, and ER did not predict bias, and in Study 3, ER did not predict suggestion. We attribute these inconsistencies to methodological limitations, such as the limited reliability of the suggestiveness measures, the partially only acceptable reliability of individual CECSA scales, and the limited sample sizes.

Notably, Study 1, which had the largest sample size and showed the highest reliability values for suggestive questioning and the CECSA scales, also provided support for all our hypotheses.

The meta-analyses provided a way to quantitatively integrate the findings of the three studies and compensate for their moderate sample sizes. However, the fixed-effects variant restrictively assumes the existence of only one true effect, while there may be heterogeneity between the true effects of our studies. A random-effects meta-analysis was not feasible because the small number of studies would have led to imprecise estimations of between-study variance (Hedges & Vevea, 1998), but future meta-analyses that are able to integrate more studies should opt for a random-effects approach.

Opening the perspective, future studies could investigate other individual differences that may predict bias and suggestive questioning. A bottom-up approach involving exploratory correlation of various personality traits with suggestiveness in a large sample could be fruitful given the current lack of empirical knowledge about the relation of suggestiveness to specific traits. As previously discussed, studies aiming to measure suggestiveness behaviorally should include a large enough number of suggestiveness indicators (i.e., occasions to pose questions) to account for the rare occurrence of suggestive questions.

Conclusion

Across a series of three studies and a meta-analytical integration, we found a biased mindset and a suggestive questioning style to be robustly predicted by two scales of the questionnaire on Cognitions and Emotions about Child Sexual Abuse. Individuals high in Naive Confidence, who overestimate both their ability to recognize abused children and the accuracy of children's abuse reports, are at risk of posing more suggestive questions and drawing biased conclusions. Similarly, individuals high in Emotional Reactivity regarding child sexual abuse

more strongly tend toward suggestive questioning and biased evaluations. We did not find sufficient support for an association between Justice System Distrust and suggestive questioning or bias. The results further our understanding of differential aspects of suggestive questioning and can be of practical use when developing or evaluating interviewer training programs or when selecting suitable personnel to talk to children about potential experiences of sexual abuse.

Tables

Table 1

CECSA Items and Allocations to Scales

Nr.	Scale	Item
1	NC	I would trust my first impression when assessing whether a child was sexually abused or not.
2	ER	When it comes to the topic of child sexual abuse, I react very emotionally.
3	NC	You generally already know whether a child has been sexually abused or not before talking with him/her
4	JSD	When it comes to child sexual abuse, courts are not taking children seriously enough.
5	ER	When the media reports about child sexual abuse, I often feel a lot of anger.
6	JSD	In cases of child sexual abuse, it is easy for a good lawyer to get an acquittal for the suspect.
7	NC	I cannot imagine that I would be fooled by a child when it comes to sexual abuse.
8	NC	False allegations of child sexual abuse are very rare.
9	JSD	As long as there is no clear evidence, it is hopeless to report child sexual abuse to the police.
10	ER	When the media reports about child sexual abuse, I often feel a lot of disgust.
11	NC	Suggestive interview techniques only influence children's memories of details and banal things.
12	NC	Even if children do not yet dare to tell about sexual abuse, I could very probably look at them if something like this had happened to them.
13	JSD	In cases of child sexual abuse, courts usually hesitate to convict the suspect.
14	NC	Adults who work a lot with children professionally, probably recognize intuitively whether a child is telling the truth about sexual abuse or not.
15	ER	When the media reports about child sexual abuse, I often feel strong hatred towards the offender.
16	JSD	I don't have faith in the potential of the justice system to prosecute perpetrators of sexual abuse.
17	NC	Children have no reason to say that they have been sexually abused, if something like this has not actually happened to them.
18	ER	When the topic of child sexual abuse is discussed, I often feel sadness.
19	NC	You can recognize whether a child was suggestively influenced.
20	NC	It is very unlikely that sexually abused children exaggerate when they tell about an abusive experience.
21	JSD	Reports of child sexual abuse are not taken seriously enough by the police.
22	ER	When the topic of child sexual abuse is discussed, I often feel anger.
23	NC	I can tell if a child is telling the truth about a sexual abuse.

Note. NC = Naive Confidence; ER = Emotional Reactivity; JSD = Justice System Distrust.

Table 2

Demographic Characteristics and Prior Experiences of Participants in Studies 1 - 3

	Study 1 N = 285		Study 2 N = 215		Study 3 N =141	
	n	%	n	%	n	%
Gender						
Female	223	78.2	72	33.5	119	84.4
Male	60	21.1	140	65.1	21	14.9
Other	0	0	0	0	0	0
No response	2	0.7	3	1.4	1	0.7
University subject						
Psychology	79	27.7	-	-	-	-
Social Work	31	10.9	-	-	-	-
Teaching	132	46.3	-	-	141	100
Educational Science	1	0.4	-	-	-	-
Pedagogy	23	8.1	-	-	-	-
Other	0	0	-	-	-	-
No response	19	6.7	-	-	-	-
Academic experiences with study topic						
Yes	51	17.9	9	4.2	1	0.7
No	234	82.1	128	59.5	139	98.6
No response	0	0	78	36.3		
Other educational experiences with study topic						
Yes	39	13.7	13	6.0	6	4.3
No	246	86.3	199	92.6	134	95.0
No response	0	0	3	1.4	1	0.7
Victimization to sexual assault						
Yes	53	18.6	-	-	7	5.0
No	215	75.4	-	-	131	92.9
No response	17	6.0	-	-	3	2.1
Parenthood						
Yes	16	5.6	-	-	-	-
No	269	94.4	-	-	-	-

Note. Participants' mean age was 24.86 years (SD = 5.48) in Study 1, 22.17 years (SD = 4.41; three participants provided no information) in Study 2, and 23.63 years (SD = 3.51; one participant provided no information) in Study 3.
Academic experiences and other educational experiences with the study topic refer to prior discussions about how to handle CSA allegations in university or other educational contexts. Victimization by sexual assault refers to having been victimized by sexual assault during childhood, adolescence, or adulthood. Parenthood refers to having children of one's own. The dash ("-") indicates that the variable was not assessed, or the response option was not given.

Table 3

Scheme of Four Question Types for the Single-Choice Items and Example Questions

	Non-Suggestive	Suggestive
Open Question	"What do you do when you two play together?"	"What is *uncomfortable* for you when you two play together?"
Closed Question	"Do you play board games, or do you play a different kind of game?"	"Is it the games that are *uncomfortable* to you or is he doing something else that is making you feel *uncomfortable?"*

Note. In the examples, the child has so far *not* mentioned feeling *uncomfortable* when playing games.

Table 4

Descriptive Results for the Mock Case- and Conversation Variables and Descriptive Results and Intercorrelations for the CECSA Scales in Studies 1 - 3

									Pearson's r	
Study	Subscale	n	M	SD	Mdn	skew	kurtosis	α	NC	ER
Study 1										
	Sum of Suggestive Questions	285	2.07	1.98	1	1.24	1.8	.63		
	Sum of Closed Questions	285	4.88	2.57	5	.21	-.69	.64		
	CSA Indicativity	285	66.55	15.06	67	-.24	-.38	.77		
	Certainty	285	63.16	18.1	64.75	-.22	-.19	.80		
	Naive Confidence (NC)	260	3.12	0.65	3.09	-.07	-.23	.79		
	Emotional Reactivity (ER)	257	4.68	1.05	4.83	-.87	.61	.90	**.31****	
	Justice System Distrust (JSD)	266	3.39	0.86	3.5	0	-.22	.80	**.40****	**.32****
Study 2										
	Sum of Suggestive Questions	215	3.84	2.12	4	.5	-.23	.54		
	Sum of Closed Questions	215	5.18	2.39	5	.04	-.55	.46		
	CSA Indicativity	215	69.97	13.1	70	-.59	.48	.6		
	Certainty	215	70.54	14.44	71.25	-.27	-.24	.75		
	Naive Confidence (NC)	212	3.35	0.58	3.36	-.18	-.21	.69		
	Emotional Reactivity (ER)	214	4.67	0.85	4.67	-.35	-.52	.83	**.19****	
	Justice System Distrust (JSD)	213	2.72	0.71	2.67	.39	-.15	.68	.12	.05
Study 3										
	Overall number of questions	141	26.35	5.48	26	-.09	-.14	-		
	Sum of Suggestive Questions	141	2.69	2.01	2.5	.89	.3	-		

Study	Subscale	n	M	SD	Mdn	skew	kurtosis	α	Pearson's r NC	Pearson's r ER
	Sum of Closed Questions	141	12.18	3.78	12	.31	.02	-		
	Naive Confidence (NC)	141	3.18	0.54	3.18	-.1	-.17	.69		
	Emotional Reactivity (ER)	141	4.58	0.8	4.67	-.63	.91	.82	.14	
	Justice System Distrust (JSD)	141	3.44	0.76	3.5	-.02	.22	.76	**.28****	.02

Note. * indicates p < .05. ** indicates p < .01.

Table 5

Results of Individual, Multiple, and Meta-Analytically Integrated Ordinal and Linear Regression Analyses to Predict Suggestive Questioning in Studies 1 – 3

Study	Analysis	Subscale	*n*	*b*	*SE*	95% CI *LL*	95% CI *UL*	*p*
Study 1								
	Ordinal Regression	Naive Confidence	260	**.22**	.07	.09	.35	-
	Ordinal Regression	Emotional Reactivity	257	**.26**	.07	.13	.29	-
	Ordinal Regression	Justice System Distrust	266	**.15**	.06	.02	.27	-
	Multiple Ordinal Regression	Naive Confidence	256	**.15**	.07	.00	.30	-
		Emotional Reactivity		**.21**	.07	.07	.35	-
		Justice System Distrust		.01	.07	-.13	.16	-
Study 2								
	Linear Regression	Naive Confidence	212	.09	.07	-.05	.23	.188
	Linear Regression	Emotional Reactivity	214	**.14**	.07	.01	.28	.035
	Linear Regression	Justice System Distrust	213	.05	.07	-.09	.18	.501
	Multiple Linear Regression	Naive Confidence	208	.08	.07	-.06	.22	.286
		Emotional Reactivity		.13	.07	-.004	.27	.057
		Justice System Distrust		.04	.07	-.09	.18	.522
Study 3								
	Linear Regression	Naive Confidence	141	**.17**	.08	.00	.33	.047
	Linear Regression	Emotional Reactivity	141	.07	.09	-.10	.23	.440
	Linear Regression	Justice System Distrust	141	-.01	.09	-.18	.16	.924
	Multiple Linear Regression	Naive Confidence	141	**.18**	.09	.004	.35	.045
		Emotional Reactivity		.04	.09	-.13	.21	.625
		Justice System Distrust		-.06	.09	-.23	.11	.498
Meta-Analyses (k = 3)								
		Naive Confidence	613	**.16**	.04	.08	.24	< .001
		Emotional Reactivity	612	**.17**	.04	.09	.26	< .001
		Justice System Distrust	620	**.08**	.04	.004	.16	.040

Note. CI = Credible interval for Study 1, confidence interval for Studies 2 and 3. Bold font indicates a significance level of at least α = .05. For Study 1, *p* values are not indicated because

of Bayes parameter estimations. Instead, 95% CIs that do not include zero indicate strong evidence. Parameters of Study 1 are reported on a latent normally distributed variable underlying the ordinal items.

Table 6

Results of Individual, Multiple, and Meta-Analytically Integrated Linear Regression Analyses to Predict CSA Indicativity Ratings in Studies 1 and 2

Study	Analysis	Subscale	*n*	*b*	*SE*	95% CI		*p*
						LL	*UL*	
Study 1								
	Linear Regression	Naive Confidence	260	**.37**	.06	.25	.48	< .001
	Linear Regression	Emotional Reactivity	257	**.24**	.06	.12	.36	< .001
	Linear Regression	Justice System Distrust	266	**.22**	.06	.10	.34	< .001
	Multiple Linear Regression	Naive Confidence	256	**.29**	.07	.17	.42	< .001
		Emotional Reactivity		**.13**	.06	.01	.26	.037
		Justice System Distrust		.07	.07	-.06	.20	.292
Study 2								
	Linear Regression	Naive Confidence	212	**.37**	.06	.24	.50	< .001
	Linear Regression	Emotional Reactivity	214	.13	.07	.00	.27	.052
	Linear Regression	Justice System Distrust	213	-.04	.07	-.18	.09	.524
	Multiple Linear Regression	Naive Confidence	208	**.36**	.07	.23	.49	< .001
		Emotional Reactivity		.07	.07	-.06	.20	.299
		Justice System Distrust		-.09	.07	-.22	.04	.165
Meta-Analyses (k = 2)								
		Naive Confidence	472	**.37**	.04	.29	.45	< .001
		Emotional Reactivity	471	**.19**	.05	.10	.28	< .001
		Justice System Distrust	479	**.11**	.05	.02	.20	.016

*Note.*CI = Confidence Interval. Bold font indicates a significance level of at least α = .05.

Table 7

Binary and Comprehensive Question Categories for Study 2 and 3

Formal Question Categories		
Closed Questions	Choice question	Questions containing two or more options for the child to choose from. ("Were you at home or at school when that happened?")
	Yes-no question	Questions containing new information provided by the interviewer for confirmation or negation. ("Did you have a nice time with your friend?")
Open Questions	Invitation to tell	Utterances inviting a free narrative, allowing the child to present their recollections without introducing any information. (For example, "Tell me what happened", "Tell me more about that", "Tell me about your handball training".)
	Facilitator	Repeating, summarizing or paraphrasing what the child has said without adding new or false information; Utterances signaling understanding for the child or short utterances indicating active listening ("U-hu", "Okay", "Yes?").
	Directive question	"Wh"-questions asking for specific details of a situation or broader topic without introducing new information.
Other Question types	Incomprehensible question	Questions that are difficult to understand for children because of their length, complexity, or ambiguity.
	Repetition Request	Request to repeat the last answer (applies only to study 3)
	Unspecified	Questions that do not fit in any of the categories above. (applies only to study 2)

Suggestiveness		
Suggestive	Specific suggestion (assuming maltreatment)	Yes-no questions that contain information conforming to a scheme of child abuse or in which the interviewer communicates what answer is expected, and paraphrases or comments that contain new information (not just schema-conforming) that the child has not previously expressed.
	Specific suggestion (assuming no maltreatment)	(only applies to study 2; did not occur)

	Unspecific suggestion	Utterances that do not contain specific new information about the event in question, but encourage speculation or communicate an expectation through strong evaluations, negative feedback on the child's statement, or claims about already knowing what happened.
	Pressure	Pressure or manipulation to evoke a narrative response, without suggesting specific information. (Applies as separate category only to study 2; was integrated into "Unspecific suggestion" in study 3)
Not suggestive	No suggestion	-

Note. To collapse the formal question categories into the bivariate variable "open vs. closed questions," questions labeled as "other question types" were ignored. In Study 2, the category "specific suggestion (assuming no maltreatment)" was included in the coding scheme but not used, as no such question was produced by the participants.

Data Availability Statement

The data, code, material, and preregistered design and analyses plans used for this study are openly available via the Open Science Framework (OSF; https://osf.io/hjkt2/?view_only=2b18866d0bf24a448280641f0e69aead).

Supplemental Material

Supplemental Material for this study can be found on OSF: https://osf.io/hjkt2/?view_only=2b18866d0bf24a448280641f0e69aead.

Funding Information and Declaration of Interest

Study 1 and 2 did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors. The data used for Study 3 was collected with funding of the German Federal Ministry of Education and Research (01SR1703, 01SR2111). We have no conflicts of interest to disclose.

CRediT Authorship Contribution Statement

Elsa Gewehr: Conceptualization (lead), Data Curation, Formal analysis, Investigation (Study 1 & 2: lead, Study 3: equal), Methodology, Project administration (Study 1: lead, Study 2: lead, Study 3: supporting), Resources (Study 1: lead, Study 2: lead, Study 3: equal), Visualization, Writing - Original Draft, Project administration (Study 1: lead, Study 2: lead, Study 3: equal). **Marie Merschhemke:** Investigation (Study 1: supporting, Study 3: equal), Project administration (Study 3: supporting), Resources (Study 1: supporting, Study 3: equal), Writing - Review & Editing. **Simone Pülschen:** Funding acquisition (Study 3: equal), Project administration (Study 3: lead, equal), Investigation (supporting), Resources (Study 1 & Study 2: supporting, Study 3: equal), Writing – Review & Editing, Supervision (support). **Dietrich Pülschen:** Project administration (Study 2: supporting), Resources (Study 2: supporting),

Investigation (Study 2: supporting), Writing - Review & Editing. **Renate Volbert:** Conceptualization (supporting), Funding acquisition (Study 3: equal), Project administration (Study 3: lead, equal), Resources (Study 1 & Study 2: supporting, Study 3: equal), Writing – Review & Editing, Supervision (lead).

Acknowledgements

We thank Larissa Jotzeit, Elisabeth Abel, Lennart Bayer, Jenni Marie Klier, Claudia Wenzel, Harriet Sewald and Svenja Seuffert for carrying out parts of the data collection and coding many questions. We thank our collaborators in the ViContact research project – Anett Tamm, Niels Krause, Hermann Barbe, Bruno Siegel, Peter Fromberger, and Jürgen Müller – for co-developing the Virtual Reality system and jointly collecting the data used for Study 3 of this article. We also thank Kristin Jankowksy for methodological comments on an earlier version of this article.

References

Barbe, H., Müller, J. L., Siegel, B., & Fromberger, P. (2023). An Open Source Virtual Reality Training Framework for the Criminal Justice System. *Criminal Justice and Behavior*, *50*(2), 294–303. https://doi.org/10.1177/00938548221124128

Bosco, F. A., Aguinis, H., Singh, K., Field, J. G., & Pierce, C. A. (2015). Correlational effect size benchmarks. *Journal of Applied Psychology*, *100*(2), 431–449. https://doi.org/10.1037/a0038047

Brown, D. A., & Lamb, M. E. (2015). Can Children Be Useful Witnesses? It Depends How They Are Questioned. *Child Development Perspectives*, *9*(4), 250–255. https://doi.org/10.1111/cdep.12142

Brubacher, S. P., Powell, M., Skouteris, H., & Guadagno, B. (2014). An Investigation of the Question-Types Teachers Use to Elicit Information From Children. *The Australian Educational and Developmental Psychologist*, *31*(2), 125–140. https://doi.org/10.1017/edp.2014.5

Bruck, M., & Melnyk, L. (2004). Individual differences in children's suggestibility: A review and synthesis. *Applied Cognitive Psychology*, *18*(8), 947–996. https://doi.org/10.1002/acp.1070

Bürkner, P.-C., Gabry, J., Weber, S., Johnson, A., Modrak, M., Badr, H. S., Weber, F., Ben-Shachar, M. S., Rabel, H., Mills, S. C., & Wild, S. (2023). *brms: Bayesian Regression Models using "Stan"* (Version 2.20.4) [Computer software]. https://cran.r-project.org/web/packages/brms/index.html

Bürkner, P.-C., & Vuorre, M. (2019). Ordinal Regression Models in Psychology: A Tutorial. *Advances in Methods and Practices in Psychological Science*, *2*(1), 77–101. https://doi.org/10.1177/2515245918823199

Ceci, S. J., & Bruck, M. (2006). Children's suggestibility: Characteristics and mechanisms. In *Advances in Child Development and Behavior* (Vol. 34, pp. 247–281). Elsevier. https://doi.org/10.1016/S0065-2407(06)80009-1

Ceci, S. J., Hritz, A., & Royer, C. (2016). Understanding Suggestibility. In W. O'Donohue & M. Fanetti (Eds.), *Forensic Interviews Regarding Child Sexual Abuse: A Guide to Evidence-Based Practice.* (pp. 141–153). Springer.

Cederborg, A.-C., Orbach, Y., Sternberg, K. J., & Lamb, M. E. (2000). Investigative interviews of child witnesses in Sweden. *Child Abuse & Neglect*, *24*(10), 1355–1361. https://doi.org/10.1016/S0145-2134(00)00183-6

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

Cyr, M., Dion, J., Gendron, A., Powell, M., & Brubacher, S. (2021). A test of three refresher modalities on child forensic interviewers' posttraining performance. *Psychology, Public Policy, and Law*. https://doi.org/10.1037/law0000300

Evans, J. St. B. T., & Stanovich, K. E. (2013). Dual-Process Theories of Higher Cognition: Advancing the Debate. *Perspectives on Psychological Science*, *8*(3), 223–241. https://doi.org/10.1177/1745691612460685

Everson, M. D., & Sandoval, J. M. (2011). Forensic child sexual abuse evaluations: Assessing subjectivity and bias in professional judgements. *Child Abuse & Neglect*, *35*(4), 287–298. https://doi.org/10.1016/j.chiabu.2011.01.001

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, *41*(4), 1149–1160. https://doi.org/10.3758/BRM.41.4.1149

Fessinger, M. B., & McAuliff, B. D. (2020). A national survey of child forensic interviewers: Implications for research, practice, and law. *Law and Human Behavior*, *44*(2), 113–127. https://doi.org/10.1037/lhb0000368

Finnilä-Tuohimaa, K., Santtila, P., Björnberg, L., Hakala, N., Niemi, P., & Sandnabba, K. (2008). Attitudes related to child sexual abuse: Scale construction and explorative study among psychologists. *Scandinavian Journal of Psychology*, *49*(4), 311–323. https://doi.org/10.1111/j.1467-9450.2008.00635.x

Finnilä-Tuohimaa, K., Santtila, P., Sainio, M., Niemi, P., & Sandnabba, K. (2009). Expert judgment in cases of alleged child sexual abuse: Clinicians' sensitivity to suggestive influences, pre-existing beliefs and base rate estimates. *Scandinavian Journal of Psychology*, *50*(2), 129–142. https://doi.org/10.1111/j.1467-9450.2008.00687.x

Funder, D. C., & Ozer, D. J. (2019). Evaluating Effect Size in Psychological Research: Sense and Nonsense. *Advances in Methods and Practices in Psychological Science*, *2*(2), 156–168. https://doi.org/10.1177/2515245919847202

Gewehr, E., Volbert, R., Merschhemke, M., Santtila, P. O., & Pülschen, S. (2023). *Cognitions and Emotions about Child Sexual Abuse (CECSA): Development of a Self-Report Measure to Predict Interviewer Bias* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/qcfvb

Gignac, G. E., & Szodorai, E. T. (2016). Effect size guidelines for individual differences researchers. *Personality and Individual Differences*, *102*, 74–78. https://doi.org/10.1016/j.paid.2016.06.069

Hedges, L. V., & Vevea, J. L. (1998). Fixed- and random-effects models in meta-analysis. *Psychological Methods*, *3*(4), 486–504. https://doi.org/10.1037/1082-989X.3.4.486

Johnson, M., Magnussen, S., Thoresen, C., Lønnum, K., Burrell, L. V., & Melinder, A. (2015). Best Practice Recommendations Still Fail to Result in Action: A National 10-Year Follow-up Study of Investigative Interviews in CSA Cases: Follow-up study of investigative interviews. *Applied Cognitive Psychology*, *29*(5), 661–668. https://doi.org/10.1002/acp.3147

Kask, K., Pompedda, F., Palu, A., Schiff, K., Mägi, M.-L., & Santtila, P. (2022). Transfer of Avatar Training Effects to Investigative Field Interviews of Children Conducted by Police Officers. *Frontiers in Psychology*, *13*. https://www.frontiersin.org/articles/10.3389/fpsyg.2022.753111

Kendall-Tucket, K., Williams, L. M., & Finkelhor, D. (1993). Impact of sexual abuse on children: A review and synthesis of recent empirical studies. *Psychological Bulletin*, *113*(1), 164–180. https://doi.org/10.1037/0033-2909.113.1.164

Klemfuss, J. Z., & Olaguez, A. P. (2020). Individual Differences in Children's Suggestibility: An Updated Review. *Journal of Child Sexual Abuse*, *29*(2), 158–182. https://doi.org/10.1080/10538712.2018.1508108

Koo, T. K., & Li, M. Y. (2016). A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *Journal of Chiropractic Medicine*, *15*(2), 155–163. https://doi.org/10.1016/j.jcm.2016.02.012

Korkman, J., Otgaar, H., Geven, L. M., Bull, R., Cyr, M., Hershkowitz, I., Mäkelä, J.-M., Mattison, M., Milne, R., Santtila, P., van Koppen, P., Memon, A., Danby, M., Filipovic, L., Garcia, F. J., Gewehr, E., Gomes Bell, O., Järvilehto, L., Kask, K., … Volbert, R. (2024). White paper on forensic child interviewing: Research-based recommendations by the European Association of Psychology and Law. *Psychology, Crime & Law*, *0*(0), 1–44. https://doi.org/10.1080/1068316X.2024.2324098

Korkman, J., Santtila, P., Westeråker, M., & Sandnabba, N. K. (2008). Interviewing techniques and follow-up questions in child sexual abuse interviews. *European Journal of Developmental Psychology*, *5*(1), 108–128. https://doi.org/10.1080/17405620701210460

Krause, N., Gewehr, E., Barbe, H., Merschhemke, M., Mensing, F., Siegel, B., Müller, J. L., Volbert, R., Fromberger, P., Tamm, A., & Pülschen, S. (2024). How to prepare for conversations with children about suspicions of sexual abuse? Evaluation of an interactive virtual reality training for student teachers. *Child Abuse & Neglect*, *149*, 106677. https://doi.org/10.1016/j.chiabu.2024.106677

Lamb, M. E., La Rooy, D., Malloy, L., & Katz, C. (Eds.). (2011). *Children's testimony: A handbook of psychological research and forensic practice* (2. ed.). Wiley-Blackwell.

Lance, C. E., Butts, M. M., & Michels, L. C. (2006). The Sources of Four Commonly Reported Cutoff Criteria: What Did They Really Say? *Organizational Research Methods*, *9*(2), 202–220. https://doi.org/10.1177/1094428105284919

Lewis, T., McElroy, E., Harlaar, N., & Runyan, D. (2016). Does the impact of child sexual abuse differ from maltreated but non-sexually abused children? A prospective examination of the impact of child sexual abuse on internalizing and externalizing behavior problems. *Child Abuse & Neglect*, *51*, 31–40. https://doi.org/10.1016/j.chiabu.2015.11.016

Lidén, M., Gräns, M., & Juslin, P. (2018). The presumption of guilt in suspect interrogations: Apprehension as a trigger of confirmation bias and debiasing techniques. *Law and Human Behavior*, *42*(4), 336–354. https://doi.org/10.1037/lhb0000287

Lilienfeld, S. O. (2016). Forensic Interviewing for Child Sexual Abuse: Why Psychometrics Matters. In W. T. O'Donohue & M. Fanetti (Eds.), *Forensic Interviews Regarding Child Sexual Abuse: A Guide to Evidence-Based Practice* (pp. 155–178). Springer International Publishing. https://doi.org/10.1007/978-3-319-21097-1_9

Magnusson, M., Joleby, M., Luke, T. J., Ask, K., & Lefsaker Sakrisvold, M. (2021). Swedish and Norwegian Police Interviewers' Goals, Tactics, and Emotions When Interviewing Suspects of Child Sexual Abuse. *Frontiers in Psychology*, *12*. https://doi.org/10.3389/fpsyg.2021.606774

May, L., Stein, A., Gundlach, T. E., & Volbert, R. (2021). Schuldig bei Verdacht? Prüfstrategien in Beschuldigtenvernehmungen. *Monatsschrift für Kriminologie und Strafrechtsreform*, *104*(2), 81–91. https://doi.org/10.1515/mks-2021-0102

Melinder, A., Brennen, T., Husby, M. F., & Vassend, O. (2020). Personality, confirmation bias, and forensic interviewing performance. *Applied Cognitive Psychology*, *34*(5), 961–971. https://doi.org/10.1002/acp.3674

Mensing, F., Gewehr, E., Merschhemke, M., & Pülschen, S. (2024). Measuring teacher's capabilities: Development of the CSA-SE scale for assessing teachers' self-efficacy in addressing suspected cases of child sexual abuse. *Child Protection and Practice*, *2*, 100049. https://doi.org/10.1016/j.chipro.2024.100049

Neal, T. M. S., Lienert, P., Denne, E., & Singh, J. P. (2022). A general model of cognitive bias in human judgment and systematic review specific to forensic mental health. *Law and Human Behavior*, *46*(2), 99. https://doi.org/10.1037/lhb0000482

Nunnally, J., & Bernstein, I. (1994). The assessment of reliability. In *Psychometric Theory* (3nd ed., pp. 248–292). McGraw-Hill.

O'Donohue, W., & Cirlugea, O. (2021). Controlling for Confirmation Bias in Child Sexual Abuse Interviews. *The Journal of the American Academy of Psychiatry and the Law*, *49*(3), 371–380. https://doi.org/10.29158/JAAPL.200109-20

Oeberst, A., & Imhoff, R. (2023). Toward Parsimony in Bias Research: A Proposed Common Framework of Belief-Consistent Information Processing for a Set of Biases. *Perspectives on Psychological Science*, *18*(6), 1–24. https://doi.org/10.1177/17456916221148147

Peixoto, C. E., Fernandes, R. V., Almeida, T. S., Silva, J. M., La Rooy, D., Ribeiro, C., Magalhães, T., & Lamb, M. E. (2017). Interviews of Children in a Portuguese Special Judicial Procedure. *Behavioral Sciences & the Law*, *35*(3), 189–203. https://doi.org/10.1002/bsl.2284

Pompedda, F., Zappalà, A., & Santtila, P. (2015). Simulations of child sexual abuse interviews using avatars paired with feedback improves interview quality. *Psychology, Crime & Law*, *21*(1), 28–52. https://doi.org/10.1080/1068316X.2014.915323

Pompedda, F., Zhang, Y., Haginoya, S., & Santtila, P. (2022). A Mega-Analysis of the Effects of Feedback on the Quality of Simulated Child Sexual Abuse Interviews with Avatars. *Journal of Police and Criminal Psychology*. https://doi.org/10.1007/s11896-022-09509-7

Powell, M. B., Hughes-Scholes, C. H., & Sharman, S. J. (2012). Skill in Interviewing Reduces Confirmation Bias: Confirmation bias and interviews. *Journal of Investigative Psychology and Offender Profiling*, *9*(2), 126–134. https://doi.org/10.1002/jip.1357

R Core Team. (2021). *R: A language and environment for statistical computing.* [Computer software]. R Foundation for Statistical Computing. https://www.R-project.org/.

Richard, F. D., Bond, C. F., & Stokes-Zoota, J. J. (2003). One Hundred Years of Social Psychology Quantitatively Described. *Review of General Psychology*, *7*(4), 331–363. https://doi.org/10.1037/1089-2680.7.4.331

Rosseel, Y. (2012). lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software*, *48*(2). https://doi.org/10.18637/jss.v048.i02

Salerno, J. M. (2021). The Impact of Experienced and Expressed Emotion on Legal Factfinding. *Annual Review of Law and Social Science*, *17*, 181-203. https://doi.org/10.1146/annurev-lawsocsci-021721-072326

Schönbrodt, F. D., & Perugini, M. (2013). At what sample size do correlations stabilize? *Journal of Research in Personality*, *47*(5), 609–612. https://doi.org/10.1016/j.jrp.2013.05.009

Segal, A., Bakaitytė, A., Kaniušonytė, G., Ustinavičiūtė-Klenauskė, L., Haginoya, S., Zhang, Y., Pompedda, F., Žukauskienė, R., & Santtila, P. (2023). Associations between emotions and psychophysiological states and confirmation bias in question formulation in ongoing simulated investigative interviews of child sexual abuse. *Frontiers in Psychology*, *14*. https://www.frontiersin.org/articles/10.3389/fpsyg.2023.1085567

Segal, A., Pompedda, F., Haginoya, S., Kaniušonytė, G., & Santtila, P. (2022). Avatars with child sexual abuse (vs. No abuse) scenarios elicit different emotional reactions. *Psychology, Crime & Law*, 1–21. https://doi.org/10.1080/1068316X.2022.2082422

Sternberg, K. J., Lamb, M. E., Davies, G. M., & Westcott, H. L. (2001). The Memorandum of Good Practice: Theory versus application. *Child Abuse & Neglect*, *25*(5), 669–681. https://doi.org/10.1016/S0145-2134(01)00232-0

Stoltenborgh, M., Bakermans-Kranenburg, M. J., Alink, L. R. A., & van IJzendoorn, M. H. (2015). The Prevalence of Child Maltreatment across the Globe: Review of a Series of Meta-Analyses: Prevalence of Child Maltreatment across the Globe. *Child Abuse Review*, *24*(1), 37–50. https://doi.org/10.1002/car.2353

Talwar, V., Crossman, A., Block, S., Brubacher, S., Dianiska, R., Espinosa Becerra, A. K., Goodman, G. S., Lamb, M., London, K., La Rooy, D., Lyon, T. D., Malloy, L., Maltby, L., Nguyen Greco, V. P., Powell, M., Quas, J., Rood, C. J., Spyksma, S., Szojka, Z., … Wylie, B. (2024). Urgent Issues and Prospects on Investigative Interviews with Children and Adolescents. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4916643

Viechtbauer, W. (2010). Conducting Meta-Analyses in *R* with the metafor Package. *Journal of Statistical Software*, *36*(3). https://doi.org/10.18637/jss.v036.i03

Zajac, R., & Brown, D. A. (2018). Conducting Successful Memory Interviews with Children. *Child and Adolescent Social Work Journal*, *35*(3), 297–308. https://doi.org/10.1007/s10560-017-0527-z

Zhang, Y., Segal, A., Pompedda, F., Haginoya, S., & Santtila, P. (2022). Confirmation bias in simulated CSA interviews: How abuse assumption influences interviewing and decision-making processes? *Legal and Criminological Psychology*, *27*(2), 314–328. https://doi.org/10.1111/lcrp.12213

Article 3:

How to Prepare for Conversations with Children about Suspicions of Sexual Abuse? Evaluation of an Interactive Virtual Reality Training for Student Teachers

Status: published in *Child Abuse and Neglect* (08.02.2024)

Niels Krause[1]*, Elsa Gewehr[1,2]*, Hermann Barbe[3], Marie Merschhemke[4], Frieda Schifner[4], Bruno Siegel[3], Jürgen L. Müller[3], Renate Volbert[1], Peter Fromberger[3]†, Anett Tamm[1]†, Simone Pülschen[4]†

[1]Psychologische Hochschule Berlin, [2]Universität Kassel, [3]Klinik für Psychiatrie und Psychotherapie, Forensische Psychiatrie, Universitätsmedizin Göttingen, [4]Europa-Universität Flensburg

*The first two authors share first authorship; †The last three authors share last authorship

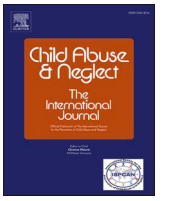
Krause, N., Gewehr, E., Barbe, H., Merschhemke, M., Schifner, F., Siegel, B., Müller, J., Volbert, R, Fromberger, P., Tamm, A., & Pülschen, S. (2024). How to prepare for conversations with children about suspicions of sexual abuse? Evaluation of an interactive virtual reality training for student teachers. *Child Abuse & Neglect*, *149*, 106677. https://doi.org/10.1016/j.chiabu.2024.106677

Child Abuse & Neglect 149 (2024) 106677

Contents lists available at ScienceDirect

Child Abuse & Neglect

journal homepage: www.elsevier.com/locate/chiabuneg



How to prepare for conversations with children about suspicions of sexual abuse? Evaluation of an interactive virtual reality training for student teachers☆

Niels Krause [a,*,1], Elsa Gewehr [a,b,1], Hermann Barbe [c], Marie Merschhemke [d], Frieda Mensing [d], Bruno Siegel [c], Jürgen L. Müller [c], Renate Volbert [a], Peter Fromberger [c,2], Anett Tamm [a,2], Simone Pülschen [d,2]

[a] *Psychologische Hochschule Berlin, Germany*
[b] *Universität Kassel, Germany*
[c] *Klinik für Psychiatrie und Psychotherapie, Forensische Psychiatrie Universitätsmedizin Göttingen, Germany*
[d] *Europa-Universität Flensburg, Germany*

ARTICLE INFO

Keywords:
Child interviewing
Child sexual abuse
Virtual reality
Training
Teacher professionalization

ABSTRACT

Background: Training for child interviewing in case of suspected (sexual) abuse must include ongoing practice, expert feedback and performance evaluation. Computer-based interview simulations including these components have shown efficacy in promoting open-ended questioning skills.
Objective: We evaluated ViContact, a training program for childcare professionals on conversations with children in case of suspected abuse.
Participants and setting: 110 student teachers were divided into four groups and took part either in a two-hour virtual reality training through verbal interaction with virtual children, followed by automated, personalized feedback (VR), two days of online seminar training on conversation skills, related knowledge and action strategies (ST), a combination of both (ST + VR), or no training (control group, CG).
Methods: We conducted a pre-registered, randomized-controlled evaluation study. Pre-post changes on three behavioral outcomes in the VR conversations and two questionnaire scores (self-efficacy and – undesirable – naïve confidence in one's own judgment of an abuse suspicion) were analyzed via mixed ANOVA interaction effects.
Results: Combined training vs. CG led to improvements in the proportion of recommended questions ($\eta_p^2 = 0.75$), supportive utterances ($\eta_p^2 = 0.36$), and self-efficacy ($\eta_p^2 = 0.77$; all $ps < .001$). Both interventions alone improved the proportion of recommended questions (VR: $\eta_p^2 = 0.67$, ST: $\eta_p^2 = 0.68$, $ps < .001$) and self-efficacy (VR: $\eta_p^2 = 0.24$, ST: $\eta_p^2 = 0.65$, $ps < .001$), but not supportive utterances (VR: $\eta_p^2 = 0.10$, ST: $\eta_p^2 = 0.13$, both n. s.).
Conclusions: The combination of VR and ST proved most beneficial. Thus, VR exercises should not replace, but rather complement classical training approaches.

☆ We have no conflict of interest to disclose. This study was funded by the German Federal Ministry of Education and Research (01SR1703, 01SR2111).

* Corresponding author.
E-mail address: n.krause@phb.de (N. Krause).

[1] The first two authors share first authorship.
[2] The last three authors share last authorship.

https://doi.org/10.1016/j.chiabu.2024.106677
Received 15 September 2023; Received in revised form 15 January 2024; Accepted 24 January 2024
Available online 8 February 2024

1. Introduction

Research has repeatedly shown that interviewers questioning children about suspected sexual abuse often do not apply the interview techniques recommended by best practice guidelines (see Faller, 2015, for an overview), but rather ask many specific closed questions. In reaction to this, a number of training programs on best-practice interviewing have been developed. Early attempts were based on classical didactics (classroom-based or online seminar settings, theoretical lectures, discussions and written and partnered roleplay exercises) and often did not improve the interview quality to a satisfactory extent (see Benson & Powell, 2015, for an overview). Based on former research findings, Powell (2008) identified six key features of effective interviewer trainings: (1) establish an understanding of the core underlying principles of effective interviewing, (2) provide an interview framework that maximizes narrative detail, (3) give clear instructions about its application, and enable (4) effective ongoing practice, (5) expert feedback and (6) regular performance evaluations. The latter three aspects – practice, feedback, and performance evaluation – imply that a program must either follow up on participants' field interviews or develop mock interview scenarios that provide realistic training opportunities. Feedback on field interviews has been shown to lead to increased use of open questions when combined with the use of the NICHD protocol (e.g., Cyr & Lamb, 2009). However, following up on field interviews is often legally or organizationally challenging (Powell & Barnett, 2015) and using real cases for training of otherwise untrained personnel may be unethical, as mistakes can be highly consequential. In addition, feedback on field interviews is only possible regarding the types of questions asked (so called process feedback), but not regarding the accuracy of an interviewer's conclusion about a case (so called outcome feedback), as the ground truth is usually unknown (Pompedda et al., 2017).

Mock interviews allow for process as well as outcome feedback. However, children themselves cannot serve as interview partners due to ethical considerations. Adult actors, in turn, must be trained to simulate the conversational behavior of children realistically (e.g., different levels of detail and accuracy in response to question types) in order to create a valid learning environment (Nicol et al., 2023; Powell et al., 2022). For example, Powell and colleagues trained post-graduate students, who were already experienced with child interviewing, for 25 h across 12 weeks to act the role of a child (Powell et al., 2008a, 2008b). Interview roleplays with such extensively trained actors led to an increased usage of recommended questions for interviewers applying NICHD or other highly structured protocols (Lawrie et al., 2020, 2021; Powell et al., 2008a, 2008b, 2014), which was not the case for roleplays with untrained actors (Powell et al., 2008b). However, the degree to which even specifically trained actors' responses mirror empirical knowledge about children's response patterns has not yet been evaluated, questioning the validity of such exercises for real-world settings.

An alternative to roleplays are computer-based interview simulations in which participants question fictitious children that are represented through videoclips or virtual characters. The children's responses are selected out of a predefined spectrum of possible answers in order to match the interviewer's questions. Two such approaches are known to date:

An Australian research team (Guadagno & Powell, 2012) developed the “Unreal Interviewing” simulation (later called “Live-Simulation”; Røed, Powell, et al., 2023), where participants choose one out of four presented questions on a computer screen to ask a video-taped five-year old child about an alleged sexual abuse. The child's answers are predefined for every question, mirror research findings on response patterns of five-year-olds and are played as videoclips. Written feedback is provided for every question. The simulation has been embedded in comprehensive e-learning programs that span across various sessions and also include theoretical lectures, other exercises (Powell et al., 2014), a structured interview protocol or roleplays with actors (Benson & Powell, 2015; Lawrie et al., 2021). Overall, these programs have led to considerable increases in the proportion of open-ended questions between 24 % and 36 % in subsequent roleplays (Benson & Powell, 2015; Powell et al., 2014), with the improvements remaining reasonably stable at follow-ups between 3 and 12 months later. The isolated effect of the computer-based interview simulation was only measured in a small sample of 36 teachers, where a similarly large increase of open questions was found. Note that none of these studies compared the effects with a control group.

A Finnish-Italian research group developed the “Empowering Interviewer Training” (EIT; Pompedda, 2018), a program where participants interview a virtual child that consists of animated morphed images of real children and is presented on a two-dimensional computer screen. Here, participants can interact with the child by freely posing verbal questions, while the child's responses consist of videoclips with predefined answers. The latter are either manually selected by an operator (Pompedda et al., 2015), or, in a revised version of the program, automatically selected by a probabilistic response algorithm, after an operator manually categorized the question (Krause et al., 2017). The response algorithm is based on experimental evidence on children's responses to different question types (e.g., high probability of a narrative response after open-ended questions; see Pompedda et al. (2015) for a list of references). EIT now consists of sixteen 4 and 6-year-old children, each of which possesses an individual “memory”, i.e., a set of predefined sentences, that can be launched as responses. Those include neutral and event-relevant information about either a sexual abuse or an event that led to a false assumption of abuse. Based on a serious gaming approach, it is the participants' task to find out whether the virtual child has been sexually abused or not within a ten-minute interview. Through the response algorithm, the chance of narrative responses and relevant correct information increases if participants adhere to the recommendations of best-practice interviewing. After each interview, participants document their conclusion about what happened to the child and receive personal feedback from a researcher about their questioning style (process feedback) and the accuracy of their conclusion (outcome feedback; see Pompedda, 2018, for a comprehensive description).

Studies evaluating the revised EIT with feedback found remarkable increases in the proportion of recommended questions between 20 % and 53 % across four to eight interviews among students and psychologists of different countries (Haginoya et al., 2020, 2021; Krause et al., 2017; Pompedda, 2018). The combination of process and outcome feedback outperformed groups with only one type of feedback (Pompedda et al., 2017, 2022) and the increased open questioning partially transferred to interviews with real children in

mock- and field interviews (Kask et al., 2022; Pompedda et al., 2020).

In recent efforts, an international collaboration (Norwegians, Australians, and others) has been reporting the ongoing development of different components for a conversational chatbot for child interviewing, including an artificial intelligence (AI) based conversational engine and different versions of text-based, two-dimensional video and three-dimensional VR children (Salehi et al., 2022). First technical proof-of-concept and user experience studies report promising results for AI components (categorization of questions, virtual child responses, and feedback) and the users' preference for VR over other visual technologies (Hassan et al., 2022). While technologically promising, a comprehensive training concept and the proof of an educational effect on the various skills involved in child interviewing are still pending. In a preliminary evaluation, repeated conversations with one text- and audio-based virtual 6-year-old girl with physical abuse experiences, supplemented by feedback, led to a slight increase of open-ended questions, but not to a decrease of undesired (e.g., closed or leading) questions and had no advantage regarding self-efficacy and perceived learning usefulness compared to conversations without feedback (Røed, Baugerud, et al., 2023).

To date, no study has compared participants of a computer-based interview training to participants that received no or only classical classroom-based or online seminar trainings. Further, the focus has been on improving open and non-suggestive questioning, while no study has investigated the effect of computer-based training on the provision of socio-emotional support.

While existing training programs largely address forensic interviewers, it has increasingly been acknowledged that there is also a need to train childcare professionals on how to talk to children about suspected sexual abuse (Brubacher et al., 2016; Cerezo & Pons-Salvador, 2004; Glammeier, 2019; Rheingold et al., 2014; Schols et al., 2013; Tener & Sigad, 2019; Volbert, 2015). Teachers and other educators that see children on a regular basis are in a unique position to identify behavioral changes, talk with children about difficulties in their lives, serve as recipients of disclosure and thus detect possible cases of CSA. Although such cases are often initially recognized in school (Sedlak et al., 2010), and forwarded to child protection services by teachers (Goebbels et al., 2008; Walsh et al., 2012), teachers also largely report feeling unprepared for these situations (Goebbels et al., 2008; Greytak, 2009; Tener & Sigad, 2019) and display a lack of confidence (Goldman, 2007) and knowledge (Márquez-Flores et al., 2016; McKee & Dillenburger, 2009) about how to respond appropriately to suspicions of CSA. Indeed, there has been evidence for underreporting (Goebbels et al., 2008) but also overreporting (King & Scott, 2014) of CSA by teachers, which emphasizes the need for further training. CSA prevention programs for teachers exist (Rheingold et al., 2014; Topping & Barron, 2009), but they rarely inform about how to conduct conversations with children. In general, best-practice recommendations for forensic interviews (rapport, simple language, open-ended questioning, avoidance of suggestion) also apply to conversations with children in school (Brubacher et al., 2016). However, there is no need for teachers to obtain a detailed account from children. Instead, they only need the information necessary to decide upon possible next steps for child protection (e.g., did something aversive happen that is relevant for child protection, what happened broadly, who is the possible perpetrator, is the child currently in danger and is help already in place; Volbert, 2015). In addition, whereas forensic interviews are usually carefully planned in advance, conversations about child abuse in school can, from a teacher's perspective, happen both in a planned way (teacher-initiated conversations) and in unplanned, perhaps surprising situation when a child decides to approach the teacher with a serious topic (child-initiated conversations), and specialized trainings for school professionals should address both types of situations (Volbert, 2015).

Samples of teachers studied by Brubacher et al. (2014, 2015) mainly used specific or leading questions in mock interviews about CSA and benefited considerably from a simple virtual training, where they read about best-practice interviewing, chose appropriate questions and received children's responses as videoclips as well as feedback on their choices.

So far, all known computer-based interview simulations have been conducted with a two-dimensional representation of children on a computer screen. Instead, three-dimensional viewing (e.g., in Virtual Reality applications) can enhance the feeling of presence and thus increase the ecological validity of stimuli (Schultheis & Rizzo, 2001). VR has successfully been applied for diagnostic and psychotherapeutic interventions (Emmelkamp & Meyerbröker, 2021) and to practice medical patient interviews (Talbot & Rizzo, 2019). Similarly, VR has the potential to improve the ecological validity of training scenarios for forensic-psychological or psychiatric staff (for overviews see Barbe et al., 2020, 2023; Fromberger et al., 2014, 2018) or child interviews in pedagogic and child protection settings.

We developed “ViContact”, a program combining VR-based simulated conversations with automated feedback and classical seminar training delivered online in order to qualify teacher students for talking to children in cases of suspected sexual abuse.

To evaluate the program, we conducted a randomized controlled, four-group, pre-post evaluation study comparing the effects of each intervention alone and both combined with a control group that received no training. Participants' interviewing behavior within simulated conversations, their self-reported attitudes and cognitions about CSA and self-efficacy for handling abuse suspicions served as main outcomes. We hypothesized that the VR training would improve conversational skills, the online seminar training would decrease problematic attitudes and cognitions, and both interventions would increase self-efficacy. The study was pre-registered on AsPredicted.org (available at aspredicted.org/8eg5n.pdf).

2. Methods

2.1. Participants

Student teachers ($N = 148$) took part in two waves of data collection between October 2020 to March 2021 and between March to May 2022. The first data collection had to be aborted due to governmental restrictions concerning the COVID pandemic. In order to reach the planned number of complete cases, a second data collection was organized. Fourteen participants were excluded due to violations of the study protocol ($n = 7$) or technical issues with the VR environment ($n = 7$) and further 24 participants did not

complete all parts of the study (because of illness, $n = 5$; cancellation of one cohort due to pandemic restrictions, $n = 11$; or missing out on a study session without specified reason, $n = 8$) resulting in a final sample size of $N = 110$ complete cases (92 women, 17 men; for one participant from control group, demographical data were not available). The number of dropouts did not differ significantly between groups (ST + VR: $n = 13$, VR: $n = 9$, ST: $n = 6$, CG: $n = 10$; $\chi^2 = 1.44$, $p = .70$).

Participants in the final sample ranged in age from 19 to 48 years ($M = 23.8$, $SD = 3.9$). Six persons (5 %) reported having experienced sexual abuse as children themselves. Professional experience in dealing with (suspected) cases of sexual abuse was indicated by one subject (1 %), and six (5 %) had previously attended some form of training on the topic. Seventy-three participants (66 %) had no experience with virtual reality, 29 (26 %) reported one-time use, eight (7 %) stated that they had been in a VR environment more than once, and one participant (1 %) indicated regular use. Fourteen participants (13 %) reported playing computer games more frequently than once a month, 55 (50 %) played once in a while, and 30 (27 %) had no experience with computer games at all.

All participants provided written informed consent before participating in the study. In accordance with the Bonn Ethics Declaration (Poelchau et al., 2015), for potential cases of psychological distress related to the topic of CSA, psychological support was available close to the study site. Nevertheless, no such situation occurred. The ethics committee of Psychologische Hochschule Berlin approved the study.

2.2. Design

Upon subscription, participants were randomly assigned to one out of four intervention groups by blocked randomization: Combined seminar and VR training (ST + VR, $n = 29$), VR training only (VR, $n = 27$), seminar training only (ST, $n = 26$) and a control group receiving no training (CG, $n = 28$). Group assignments were determined by throwing a dice. If all slots for one experimental group in a cohort were already taken, the dice was thrown again until a participant could be assigned to a free slot.

All participants filled in self-report questionnaires (demographical data, *self-efficacy*, *Cognitions and Emotions about Child Sexual Abuse* (CECSA; Gewehr et al., 2023)), and conducted two simulated conversations for a baseline measurement. In order to evaluate the participants' experience with the VR environment, they filled in the *Simulator Sickness Questionnaire* (SSQ; Kennedy et al., 1993) immediately before each conversation and after the last conversation. Additionally, participants completed the *Ingroup Presence Questionnaire* (IPQ; Schubert et al., 2001), the *Social Presence Questionnaire* (SPQ; Bailenson et al., 2005) and the *VR Simulation Realism Scale* (VSRS; Poeschl & Doering, 2013) after the end of the second simulated conversation. All questionnaires were presented via a web interface on a laptop (except the paper-pen demographical data form, for data protection reasons). Within the following two weeks, participants received training or no training according to their experimental group and finally completed a post-test session that included two additional simulated conversations (accompanied by the VR experience questionnaires mentioned above) and filling in the self-efficacy and CECSA questionnaires for a second time.

Five measures served as main outcomes, three of which were behavioral measures from the simulated conversations (*proportion of recommended questions, proportion of supportive utterances* and *tasks concerning a supportive opening and closure of the conversation*) and two of which were self-report questionnaires ((1) a newly developed instrument aimed at capturing *self-efficacy* beliefs of pedagogical professionals about handling suspected cases of child sexual abuse and (2) *Cognitions and Emotions about Child Sexual Abuse*, CECSA; Gewehr et al., 2023).

Fig. 1. Screenshot from the ViContact VR environment (left) and a participant wearing head mounted display and controller.

2.3. Materials

2.3.1. Virtual reality environment

The ViContact VR environment (Fig. 1) has been developed as a practice opportunity for conversations about a possible abuse suspicion with computer-generated virtual children. It has been primarily designed to provide a training environment where participants can exercise their conversational skills and receive personalized feedback. In the current study, the VR conversations also served to measure baseline and post-test performance.

2.3.1.1. Virtual children: memories and interaction. Eight virtual characters (four male, four female) have been developed to simulate the conversational behavior of ten-year-old children. These *virtual children* dispose of pre-defined memories about their everyday life as well as about one *critical event* that involves an experience of either (a) sexual abuse, (b) another child protection issue (e.g., physical abuse), or (c) a stressful event with no need to intervene (e.g., an argument with another child). We decided to implement three types of critical events in order to teach open-mindedness: In early conversations with children, teachers should openly aim to find out whether there is anything currently bothering the child and not only whether or not the child has been sexually abused.

The virtual children can be communicated with via natural language processing, involving speech-to-text and text-to-speech components and a text-based dialog management system (ChatScript, Wilcox & Wilcox, 2013). As in the Empowering Interviewer Training (EIT) approach (Krause et al., 2017; Pompedda, 2018; Pompedda et al., 2015, 2020, 2022), the memory content for each virtual child is stored in a set of narrative responses that the virtual child can reveal according to probabilistic algorithms that depend on the participant's questioning style. The use of recommended (open, non-suggestive and simply phrased) questions leads to narrative responses with a high probability, while the use of non-recommended (closed, suggestive or too complex) questions leads to generic responses like “yes”, “no”, or “I don't know”, that are selected randomly and thus do not contain informational value. Closed, and even more so suggestive questions more often lead to confirming (e.g., “yes”) than disconfirming (e.g., “no”) answers, which mimics the confirmatory effect elicited through suggestive questioning (see Supplement S2 for an overview of the probabilistic relations between interviewer utterance categories and answer categories). In ViContact, these algorithms have been designed adaptively to account for socio-emotional support: The probability of narrative responses after recommended questions increases with the use of socio-emotionally supportive and rapport-building utterances.

In order to automatically select appropriate answers from the child's memory, content-based processing of participants' utterances has been implemented within ChatScript. Information on the child's everyday life is organized in six *neutral topics* with four narrative responses each concerning a specific domain, such as family, a best friend, or school. For example, “Tell me about your brother” is coded as invitation. Correspondingly, in 20 % of the cases, a generic answer like “I don't know” or “Hm” is selected, while in 80 % of the cases, a narrative response is selected. ChatScript recognizes the word “brother” and thus selects an answer from the neutral topic “brother”, for example: “When Leon plays FIFA he allows me to watch.”

The *critical topic* contains ten narrative responses: Six responses with information on the context of the critical event and the person (s) involved, a seventh response which discloses information on the event and three responses with further elaborations on the event. Each virtual child has been developed in three versions that differ in the type of critical event (i.e., in the last four narrative responses of the critical topic).

Half of the virtual conversations have been designed as *child-initiated* (i.e., the child approaches the teacher in order to tell him or her something) and the other half as *teacher-initiated* (i.e., the teacher initiates the conversation because of a suspicion that something might have happened to the child). Virtual children in the child-initiated condition start the conversation by themselves, are already willing to talk at the beginning (i.e., likely to answer open-ended questions with a narrative response) and disclose the critical event by their own initiative. In the teacher-initiated condition, the participant has to start the conversation. The virtual children are less talkative at the beginning (i.e., less likely to give a narrative response) and not yet ready to disclose the critical event. They only do so, if the participant has engaged in rapport-building before addressing a potentially unpleasant issue.

To sum up, only when participants interview in a supportive manner and use recommended questions, do the virtual children reveal their memory about the critical event via ten narrative responses from the critical topic. When questioned differently, they tend to give uninformative (generic, randomly selected) answers which may nevertheless lead the interviewer to construe an event they believe to have learned about from the child (e.g., when the interviewer asks many option-posing questions that the virtual child acquiesces to).

2.3.1.2. Conversations. Participants are asked to engage in ten-minute-long conversations with each virtual child within the virtual environment of a classroom, in an assumed ten-minutes school break. They are instructed to gather enough information during the conversation to assess what the child has experienced: (a) sexual abuse, (b) another event that requires intervention or (c) another stressful event that does not require intervention; and which person(s) from the child's environment were involved in the event in question. Before the start of each conversation, participants read a case vignette (for an example, see supplement S1) with information on the child's age, living conditions, family, friends, hobbies, and behavior at school. Additionally, for teacher-initiated conversations, a worrying observation in the child's behavior is mentioned as the immediate reason for the conversation.

Each virtual conversation consists of three phases: At the beginning (the *opening phase*), participants' task is to build rapport with the child. In the child-initiated condition, where the virtual child is already willing to talk and disclose, this is limited to employing open-ended questions and kindly refusing the request for confidentiality made by the virtual child. In the teacher-initiated condition, participants' task is to clearly state that they want to talk to the child, transparently communicate their reasons for it and ask questions

Table 1
Question categories.

Formal categories		
Recommended	Invitations	Utterances inviting a free narrative, allowing the child to present their recollections without introducing any information. (For example, "Tell me what happened", "Tell me more about that", "Tell me about your handball training".)
	Facilitators	Repeating, summarizing or paraphrasing what the child has said without adding new or false information; Utterances signaling understanding for the child or short utterances indicating active listening ("U-hu", "Okay", "Yes?").
	Directive questions	"Wh"-questions asking for specific details of a situation or broader topic without introducing new information.
Not recommended	Choice questions	Questions containing two or more options for the child to choose from. ("Were you at home or at school when that happened?")
	Yes-no questions	Questions containing new information provided by the interviewer for confirmation or negation. ("Did you have a nice time with your friend?")
	Incomprehensible questions	Questions that are difficult to understand for children because of their length, complexity, or ambiguity.
Suggestion vs. support		
Recommended	Supportive utterances	Utterances that help establishing rapport between teacher and child by communicating understanding, acceptance and a genuine interest in the child's well-being and in their opinions and emotions.
Not recommended	Specific suggestive	Yes-no questions that contain information conforming to a scheme of child abuse or in which the interviewer communicates what answer is expected, and paraphrases or comments that contain new information (not just schema-conforming) that the child has not previously expressed.
	Unspecific suggestive	Utterances that do not contain specific new information about the event in question, but encourage speculation or communicate an expectation through strong evaluations, negative feedback on the child's statement, or claims about already knowing what happened.

about neutrally or positively connotated topics to build rapport with the child before addressing potentially problematic issues. When participants adhere to these tasks and employ supportive utterances, the virtual child becomes gradually more talkative (i.e., the probability of a narrative response rises). Asking the virtual child about something stressful or problematic right at the beginning without building rapport first results in the child becoming less talkative. After asking a few recommended questions about a neutral or positive topic, the child becomes ready to disclose (i.e., will talk about the critical topic if asked) and eventually gives a hint towards the critical topic by her- or himself. If a participant asks only closed-ended questions, the virtual child does not talk about the critical topic at all and the conversation remains in the opening phase.

The *main phase* begins when the child starts to talk about the critical event. Participants' task is to elicit accurate information about the critical event by asking open-ended questions. Additionally, they are supposed to continue employing supportive utterances in order to maintain rapport, e.g., by telling they care for the child, showing understanding, asking for the child's opinion or validating emotional utterances of the child.

Eight minutes after the start of the conversation, a bell rings to announce the near end of the school break and with it the conversation. Even if the child has not yet disclosed the critical event, the *closing phase* begins. In the closing phase, the participants are supposed to end the conversation in a supportive manner, providing a feeling of safety to the child. Their task is to explain their next steps (e.g., talk to a colleague about how to help the child), ask the child whether he or she needs any immediate support and show their availability for further conversations. Then, the participants can either end the conversation themselves or it is automatically ended after 10 min.

After each conversation, participants are asked to report their conclusion about the type of critical event and the persons involved through choice questions and write a short paragraph about the course of the event within the web interface used for the questionnaires.

2.3.1.3. Technical setup. The VRCT framework (Barbe et al., 2023) is used in an adapted version to perform the virtual reality trainings, to provide questionnaires, to provide the feedback to the participant and to save all data in a database. In short, the VRCT framework consists of 5 different program modules: virtual environment, speech-to-text engine, operator application, conversation engine and web server.

The virtual environment is presented via a Head Mounted Display (HMD; HTC Vive Pro 2). It includes the representation of the virtual children as well as a typical (virtual) classroom where the conversation takes place (see Fig. 1). In it, the participant sits at a teacher's desk facing the virtual child which sits next to the desk. A summary of the vignette is written on a notebook that is placed on the desk. For visual representation within the VR environment, eight virtual characters have been chosen from a set of 16 three-dimensional child characters, designed and evaluated at the Human Medical Center Göttingen (Bonnet et al., 2018).

The interaction between the participant and the virtual child happens via natural (verbal) language, using microphone and speakers of the HMD. In order to ask a question, participants have to press and hold a button on a controller which belongs to the HMD. When the button is released, an audio file is generated and sent to a speech-to-text module for transcription. The transcribed question is

then forwarded to the operator application on a separate laptop. It serves a human operator, among other things, to categorize the transcribed question according to the scheme explained below. Once the question has been categorized, it is forwarded to the ChatScript-based conversation engine. Here, based on the response behavior described above, the most suitable answer possible is searched for and sent back to the operator application. Finally, the answer from the conversation engine is forwarded to the virtual environment, where the matching audio file with the spoken answer (created in advance using German Wavenet voices from Google Cloud text-to-speech, https://cloud.google.com/text-to-speech) is played through the headphones of the HMD.

2.3.1.4. Coding scheme. The operators coded question types, suggestiveness, supportiveness and conversational tasks as described in Table 1. The coding scheme was adapted from Pompedda et al. (2015) who used question categories from the empirical literature on interviewing children (e.g., Korkman et al., 2006, 2008; Lamb et al., 1996; Sternberg et al., 1996; for a full list of references, see Pompedda et al., 2015). We extended the coding scheme with respect to supportive and rapport-building utterances, building on empirical findings (Hershkowitz et al., 2015; Lamb et al., 2018; Tamm et al., 2021).

2.3.1.5. Performance measures. To assess the quality of participants' questioning style, three measures were calculated: (1) The *proportion of recommended questions* (non-suggestive invitations, directive questions and facilitators) among all questions posed in a conversation, (2) the *proportion of supportive utterances* among all utterances in a conversation, and (3) a score indicating how many of the conversational *tasks concerning a supportive opening and closure* of the conversation were completed (see section "Conversations" for details). "Utterances" were defined as all verbal actions by a participant towards a virtual child. The term "questions" refers to all utterances except comments from the introduction or closure tasks and requests for repeating an answer that a participant had not understood. The *proportion of recommended questions* was calculated dividing the number of recommended questions by the total number of questions (i.e., the total number of utterances minus introductory and closure comments and requests for repeating an answer). The *proportion of supportive utterances* was calculated dividing the number of supportive utterances by the total number of utterances in a conversation. There was one task in the opening phase of the child-initiated conversation, three tasks in the opening phase of the teacher-initiated conversation and three tasks in the closure phase of both conversations, resulting in a total score between zero and ten points for two conversations (one teacher-initiated and one child-initiated).

The written answers of the participants concerning their *conclusion about the critical event* were coded as correct if they corresponded with the virtual child's critical event memory (specifically, when the critical event according to the seventh answer from the critical topic was included), mentioned the person involved and did not contain any information related to child abuse that was not present in the child's memory.

2.3.1.6. Automated feedback. After each VR training conversation, the VRCT framework provided automated digital feedback on the correctness of the participant's conclusion concerning the critical event, on their questioning style and on their performance concerning rapport-building and support including examples from their past conversation. Specifically, concerning the conclusions, participants' answers were compared with the correct solutions. Concerning questioning style, a percentage distribution regarding the participants' usage of recommended and non-recommended question categories in the last conversation was displayed along with positive and negative examples, short paragraphs describing each question category and recommendations for subsequent conversations. For supportive utterances, positive examples were displayed together with an explanatory paragraph. Conversational tasks concerning a supportive opening and closure of the conversation were fed back as either completed or not completed. In the latter case, a model solution was given. An illustrative example for the feedback that was shown to VR training participants is displayed in Supplement S3.

2.3.2. Online seminar training

The online seminar training (ST) consisted of seven modules addressing relevant knowledge as well as best practice recommendations based on pedagogical as well as forensic and developmental psychological research evidence: (1) definition and phenomenology of CSA, (2) guidelines and legal requirements for handling CSA suspicions in school, (3) children's disclosure patterns and ways for teachers to facilitate disclosure, (4) judgment, decision making, and bias, (5) children's memory and ways of communication, (6) conversational methods for talking to children about abuse suspicions and (7) documenting a conversation. The main focus of the seminar was on conversational skills, with module 6 taking up roughly half of the course time. Here, topics such as how to start a conversation, how to respond to a disclosure, which questions to use in order to elicit free narrative recall from a child (open-ended questions, avoiding to introduce information, avoiding suggestive questioning and otherwise suggestive behavior), socio-emotional support, child-appropriate language and appropriate closure of a conversation were addressed.

The online seminar consisted of input sections, moderated discussions, individual and small group exercises, videos, and recap quizzes. For example, participants practiced developing alternative hypotheses about ambiguous cases based on vignettes, recognizing different types of questions and supportive interviewing techniques and making up own examples. In one exercise, participants watched three videos with adult actors that simulated teacher-child conversations, in which the teacher talks to the child (1) in an open-ended fashion, (2) in a suggestive manner, holding an abuse hypothesis and (3) in a suggestive manner, holding the hypothesis that the child is exaggerating and nothing serious has happened. For all videos, participants were asked to classify the teacher's questions and discuss their observations about what attitudes and assumptions the teachers revealed. At the end of the seminar, participants were familiarized with a practical guideline sheet on preparing, conducting and documenting conversations with children about abuse suspicions and planning subsequent actions.

The seminar training did not include any roleplay exercises in order to prevent conceptual overlap with the VR Training. A training manual with course material and additional information for participants and instructors has been prepared in German language and is available from the last author.

2.3.3. Self-report questionnaires

To measure *Self-Efficacy*, we developed and used a redacted and shortened version of a compilation of items by König et al. (2015) which aims at capturing self-efficacy beliefs of pedagogical professionals about handling suspected cases of child sexual abuse (supplement S4). Its fourteen items are answered on a 4-point Likert scale ranging from 1 (incorrect) to 4 (fully correct), for example: "When confronted with a potential case of child sexual abuse, I know how to behave." A sum score is calculated ranging between 14 and 56.

The questionnaire *Cognitions and Emotions about Child Sexual Abuse* (CECSA; Gewehr et al., 2023) subsumes three scales on different cognitions and emotions that are related to interviewer confirmation bias in cases of alleged child sexual abuse. It has been developed based on the CSA Attitudes and Beliefs Scale (CSAABS; Finnilä-Tuohimaa et al., 2008). Here, we used only the scale *Naïve Confidence* (NC), which includes 11 items describing an uncritical confidence in one's own ability to recognize abuse experiences in children (e.g., "I would trust my first impression when assessing whether a child was sexually abused or not") and an uncritical acceptance of all statements made by children about sexual abuse (e.g., "Children have no reason to say that they have been sexually abused, if something like this has not actually happened to them"). All items are answered on a 6-point Likert scale ranging from 1 (fully disagree) to 6 (fully agree), resulting in a sum score between 11 and 66.

2.4. Procedure

Ten operators were trained in a two-day online course to code the participants' utterances in the VR conversations. After the operator training, the operators coded three sets of four conversation transcripts (one set at the end of the course, two days, and four days later, respectively) and received personal feedback on their coding performance. For the last set of transcripts coded, interrater agreements of Fleiss' $\kappa = 0.75$ (for coding an utterance as neutral vs. supportive vs. unspecific suggestive vs. specific suggestive), $\kappa = 0.81$ (for coding of formal question categories) and $\kappa = 0.86$ (for coding of conversational tasks) were achieved, indicating good interrater agreement (Fleiss, 1971). Participants' experimental conditions were not blinded to the operators.

The study was carried out in eight cohorts with a maximum capacity of 20 participants each, divided equally between experimental groups. All participants attended a baseline session of 90 min. Upon arrival at the laboratory they signed informed consent and confidentiality agreements, filled in a demographics sheet and baseline questionnaires. Participants then went through a familiarization phase within the virtual environment including a speech recognition tutorial. Subsequently, they had one child-initiated and one teacher-initiated conversation (or vice versa; both without feedback; for details see supplement S5, "Counterbalancing of VR Conversations") as a baseline performance measurement. Before and after each conversation, participants filled in the VR experience questionnaires. One to three days after baseline, participants of the ST and ST + VR group attended the seminar training, which was held online via a video conferencing tool (WebEx Training, version WBS33; Cisco Systems, Inc., San Jose, CA, USA). Each course was conducted in a co-training manner by two members of the research team for up to twelve participants and lasted 13 h across two days.

Four to ten days after baseline, participants of the VR and ST + VR group attended their VR training sessions. At the beginning of a training session, participants read an instruction sheet that contained instructions according to the main training goals, asking them to (1) adopt an open-minded attitude, consider different possible outcomes of the conversation, (2) employ open-ended, child-appropriate and non-suggestive questions, and (3) use rapport building and supportive interviewing techniques at the beginning, during the course and at the end of the conversations. Subsequently, they conducted two child-initiated and two teacher-initiated conversations with feedback.

Fourteen days after baseline, participants were supposed to attend the post-test session which had the same procedure as the baseline measurement. Due to organizational challenges, only 60 % of the participants were tested exactly 14 days later. Eighty percent of the post-test sessions took place 13 to 15 days after baseline and the remaining 20 % between 10 and 18 days after baseline. Pre-posttest intervals did not significantly differ between groups ($F = 0.76$, $p = .52$). For each testing or training session, participants received a compensation of 25€, with total compensations between 50€ (control group) and 125€ (combined training group).

2.5. Hypotheses

For all behavioral outcomes (*recommended questions, supportive utterances, opening and closure tasks*), we hypothesized that VR and combined training participants would show a stronger increase between baseline and post-test than the online seminar training and control group participants, mirroring findings on the effectivity of computer-based, practical and feedbacked training versus classical seminar-style approaches in changing interviewer behavior (for an overview, see Benson & Powell, 2015).

We hypothesized that the participants of both single and the combined training groups would experience a stronger pre-post increase in *self-efficacy* than control group. Following the model of self-efficacy determinants by Gist and Mitchell (1992), we assumed that both training interventions affect self-efficacy in different ways: While the VR training is supposed to improve the participants' actual conversational skills and enable experiences of mastery when successfully applying the newly acquired skills, the seminar training in turn provides relevant context knowledge and best practice recommendations on conversations with children about possible abuse. Thus, we expected the effects of both interventions on self-efficacy to be complementary and to add up to a stronger improvement in the combined training group compared to both single-intervention groups.

Table 2
Group means and standard deviations of numbers of utterances and questions per conversation and main outcome variables at baseline and post-test.

		% Recommended questions		% Supportive utterances		Opening and closure tasks		Naïve confidence		Self-efficacy		Total number of utterances		Number of questions	
		M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Control group (n = 28)	Baseline	42.78	10.52	18.40	7.02	3.04	1.60	33.50	5.69	34.61	5.96	30.46	5.35	28.54	5.54
	Post-test	41.31	8.96	14.75	8.68	3.07	1.27	32.71	6.32	33.68	5.91	33.39	4.73	31.59	4.96
Seminar training (n = 26)	Baseline	40.62	10.65	18.57	9.34	3.23	1.42	33.46	6.20	33.96	5.88	29.10	4.23	27.23	4.55
	Post-test	71.46	13.11	24.64	12.65	4.46	1.79	28.00	6.59	44.38	4.84	28.98	4.68	26.83	4.91
VR training (n = 27)	Baseline	42.33	10.84	22.03	9.85	2.78	1.28	35.26	6.67	36.93	6.86	27.54	5.08	27.34	4.81
	Post-test	76.50	14.04	25.45	10.32	7.04	1.74	33.93	7.91	39.78	6.65	28.98	5.33	26.60	3.63
Combined training (n = 29)	Baseline	41.01	9.91	18.70	9.21	2.28	1.03	34.66	5.00	30.76	6.16	28.69	4.45	25.91	5.05
	Post-test	84.43	10.93	28.97	9.44	6.41	1.64	30.24	6.33	47.17	4.39	29.31	3.72	25.94	4.93

Note. "Utterances" are defined as all verbal actions by a participant towards a virtual child. The term "questions" refers to all utterances except comments from to the introduction or closure tasks and requests for repeating an answer.

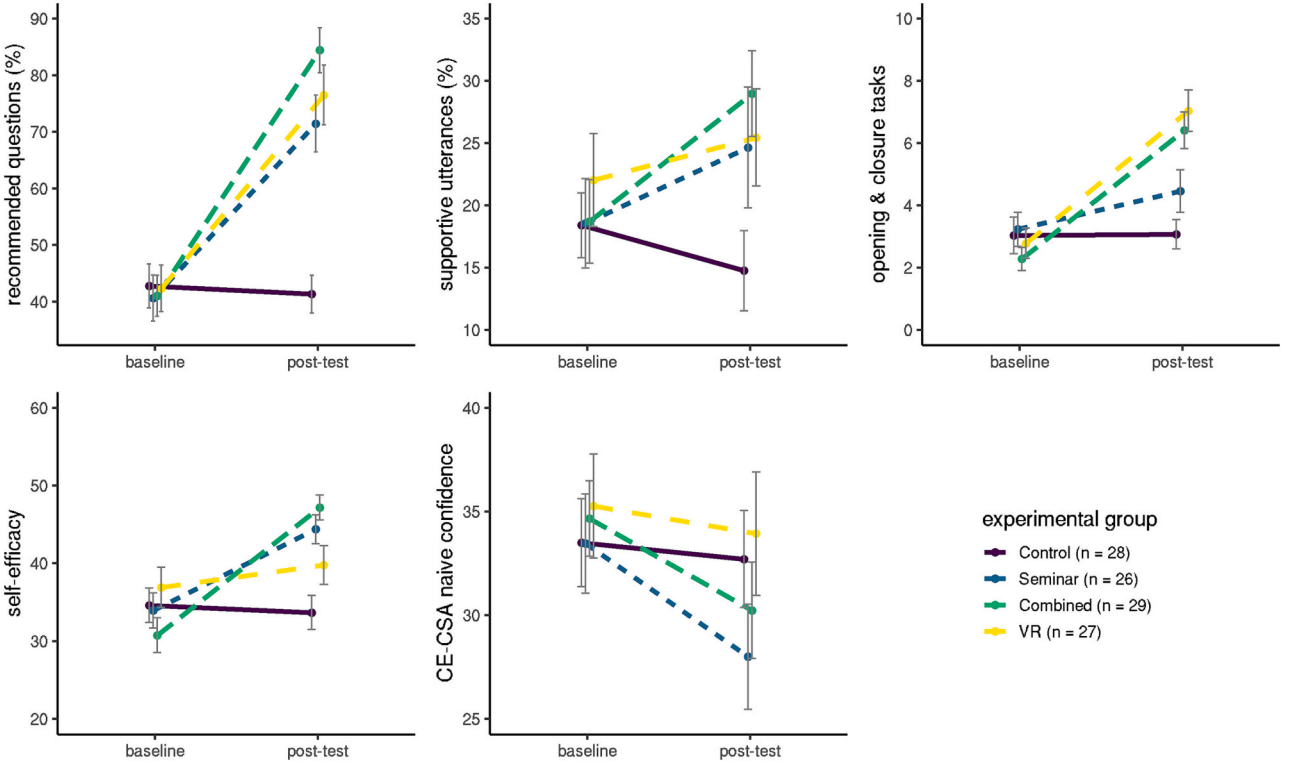


Fig. 2. Means of main outcomes at baseline and post-test by training group. *Note*. Error bars display 95 % confidence intervals for between-subjects effects.

As bias and overconfidence were among the central topics of the seminar training and not directly addressed in the VR training, we supposed that the seminar training and combination groups would show a stronger reduction in *Naïve Confidence* scores than VR and control participants.

2.6. Statistical analyses

Analysis scripts and data files can be found at osf.io/v8g4m. We performed all calculations in R (version 3.6.3; R Core Team, 2020) using the RStudio environment (RStudio Team, 2022). Descriptive statistics of the main outcomes were calculated using the psych package (version 2.2.5; Revelle, 2022). As preregistered, we conducted global 4 (group, between) by 2 (pre-post, within) mixed ANOVAs on each of the five main outcomes using the rstatix package (version 0.7.0; Kassambara, 2021). In case of a statistically significant interaction term, 2 (group) by 2 (pre-post) mixed ANOVAs were calculated for pairwise comparisons of training effects between groups according to the hypotheses specified. Study-wide Bonferroni correction was used to conservatively control for multiple testing. Thus, we set significance levels to $\alpha = 0.01$ for global ANOVAs and $\alpha = 0.05/21 = 0.0024$ for planned 2 by 2 ANOVAs. Further exploratory 2 by 2 ANOVAs were then conducted in order to compare the pre-post differences of groups for which we had not predicted any significant difference with $\alpha = 0.05/30 = 0.0017$, correcting for the total number of tests including both planned and exploratory comparisons.

Our resources allowed for a planned sample size of 100 participants (25 per group), with an additional 20 % of time slots planned to account for anticipated dropout. Power analysis with G*Power (version 3.1.9.6; Faul et al., 2009) for 2 × 2 ANOVAs presupposing the achievable sample size, $\alpha = 0.0024$, r (pre-post) = 0.60, resulted in a planned power of $1 - \beta = 0.81$ for an intermediate effect ($\eta_p^2 = 0.06$) and $1 - \beta = 0.99$ for a large effect ($\eta_p^2 = 0.14$) sensu Cohen (1988).

As there was a substantial number of dropouts in our study, as well as a previously unplanned second data collection wave in order to reach our planned sample size of complete cases, we additionally calculated intent-to-treat analyses using all available data via a structural equation modeling (SEM) approach and checked for possible differences between the two cohorts via multigroup SEM. Method and results are explained in detail in the supplement S7.

For exploratory purposes, we also calculated the proportion of *correct conclusions about the critical event* in each experimental group at baseline and test. As the sample size was not sufficient for a logistic regression of reasonable power, only descriptive results are reported.

3. Results

3.1. Descriptives and preliminary analyses

In the two baseline VR conversations, participants across all conditions used mainly yes-no-questions (Mean total number per conversation: $M = 11.2$, $SD = 4.3$) and directive questions ($M = 6.7$, $SD = 3.6$); followed by invitations ($M = 2.4$, $SD = 2.0$) and facilitators ($M = 2.2$, $SD = 2.0$). Even fewer specific suggestive ($M = 1.4$, $SD = 1.5$), unspecific suggestive ($M = 1.2$, $SD = 1.5$), choice questions ($M = 1.1$, $SD = 1.4$) and incomprehensible questions ($M = 1.0$, $SD = 1.3$) were used. In child-initiated baseline conversations, participants elicited $M = 4.7$ ($SD = 3.2$) critical details (narrative answers from the critical topic), $M = 4.1$ ($SD = 2.8$) neutral details

Table 3
Interaction terms of 2 (group) × 2 (pre-coutcomes.

Outcome	Hypothesis	F	p	$p < \alpha$	η_p^2
% Recommended questions	ST + VR > CG	160.303	<.001	*[a]	0.745
	ST + VR > ST	10.313	.002	*	0.163
	VR > CG	105.876	<.001	*	0.666
	VR > ST	0.748	.391		0.014
% Supportive utterances	ST + VR > CG	31.308	<.001	*	0.363
	ST + VR > ST	1.565	.216		0.029
	VR > CG	5.904	.019		0.100
	VR > ST	0.500	.483		0.010
Opening and closure tasks	ST + VR > CG	61.494	<.001	*	0.528
	ST + VR > ST	23.432	<.001	*	0.307
	VR > CG	56.960	<.001	*	0.518
	VR > ST	22.464	<.001	*	0.306
Self-efficacy	ST + VR > CG	186.003	<.001	*	0.772
	ST > CG	96.698	<.001	*	0.650
	VR > CG	16.813	<.001	*	0.241
	ST + VR > ST	14.540	<.001	*	0.215
	ST + VR > VR	93.931	<.001	*	0.635
Naïve confidence (NC)	ST + VR > CG	5.969	.018		0.098
	ST > CG	10.576	.002	*	0.169
	ST + VR > VR	3.947	.052		0.068
	ST > VR	7.493	.009		0.128

[a] Bonferroni-corrected significance threshold was set to $\alpha = 0.0024$.

Table 4
Interaction terms of 2 (group) × 2 (pre-post) exploratory ANOVAs on main outcomes.

Outcome	Groups	*F*	*p*	$p < \alpha$	η_p^2	Observed effect
% Recommended questions	ST + VR vs. VR	4.896	.031		0.083	
	ST vs. CG	109.016	<.001	*[a]	0.677	ST > CG
% Supportive utterances	ST + VR vs. VR	6.393	.014		0.106	
	ST vs. CG	7.526	.008		0.126	
Opening and Closure Tasks	ST + VR vs. VR	0.045	.833		0.001	
	ST vs. CG	4.144	.047		0.074	
Self-efficacy	ST vs. VR	33.797	<.001	*	0.399	ST > VR
Naïve confidence	VR vs. CG	0.165	.686		0.003	
	ST + VR vs. CG	0.411	.524		0.008	

[a] Bonferroni-corrected significance threshold was set to $\alpha = 0.0017$.

(narrative answers from neutral topics and specific answers), and $M = 10.9$ ($SD = 3.5$) option responses (affirmation, negation or options). In teacher-initiated baseline conversations, they elicited $M = 2.6$ ($SD = 1.9$) critical details, $M = 4.5$ ($SD = 2.3$) neutral details and $M = 12.2$ ($SD = 4.1$) option responses. Correlations between question and response categories are to be found in supplement S6.

Descriptive statistics for total number of utterances and number of actual questions per conversation can be found in Table 2. We checked for differences in the total number of utterances in the VR conversations between groups and timepoints: The mean number of utterances differed between groups ($F = 4.23$, $p = .007$, $\eta_p^2 = 0.11$), as well as between pre- and post-test ($F = 8.13$, $p = .005$, $\eta_p^2 = 0.07$). Post-hoc *t*-tests with Bonferroni correction revealed a significant increase between pre- and post-test in control group ($p = .002$, $d = 0.66$) and significant differences between control group vs. all three training groups at post-test (all $p \leq .001$, $d = 0.87 \ldots 0.94$). For number of actual questions asked per interview, a similar pattern was observed, with post-hoc t-tests showing a significant increase in control group ($p = .002$, d = 0.65) and significant differences between control group and all training groups at post-test (all $p \leq .001$, $d = 0.96 \ldots 1.15$). Descriptives for the main outcomes are shown in Table 2 and Fig. 2.

3.2. Hypothesis testing

3.2.1. Complete case analyses

Significant group by pre-post global interaction terms were observed for all main outcomes: Proportion of recommended questions ($F = 57.34$, $p < .001$, $\eta_p^2 = 0.62$), proportion of supportive utterances ($F = 7.14$, $p < .001$, $\eta_p^2 = 0.17$), *self-efficacy* ($F = 74.58$, $p < .001$, $\eta_p^2 = 0.68$), naïve confidence ($F = 4.63$, $p = .004$, $\eta_p^2 = 0.12$) as well as opening and closure tasks ($F = 26.87$, $p < .001$, $\eta_p^2 = 0.43$).

Table 3 shows the interaction terms of all planned 2 (group) by 2 (pre-post) ANOVAs on the main outcomes and Table 4 shows the results of additional exploratory 2 by 2 ANOVAs. On *proportion of recommended questions*, the results conformed with our hypotheses concerning higher pre-post improvements for combined and VR training vs. no training as well as the advantage of combined over seminar training. The expected advantage of VR training over seminar training, however, was not found. Instead, exploratory comparisons showed an advantage of seminar training over control group.

On *proportion of supportive utterances*, we found only the predicted advantage of combined training over control group to be statistically significant, whereas an intermediate effect in the predicted direction between VR and CG remained below the significance threshold. Only small and non-significant differences emerged between combined training vs. ST and VR vs. ST. Exploratory comparisons showed medium-sized, but non-significant advantages of ST + VR over VR and of ST over CG.

For *opening and closure tasks*, the results conformed our hypotheses: Both the combined and the VR training led to stronger improvements than seminar training or no training. The exploratory analyses did not show any unexpected interactions.

Likewise, the results concerning *self-efficacy* were in line with our expectations: Both single interventions showed an advantage over control group and combined training led to the highest improvement among all experimental groups, with a significant advantage over VR, ST and CG, respectively. Unexpectedly, also the self-efficacy improvement in the seminar training group was found to be stronger than in the VR training group.

Concerning *naïve confidence*, only the predicted advantage of seminar training over control group was found to be statistically significant. Neither the expected advantages of combined training over control group and VR training group, nor an advantage of seminar training over VR training could be corroborated. Exploratory comparisons did not show any unexpected group differences in pre-post-change.

3.2.2. Intent-to-treat analyses

The intent-to-treat analyses resulted in an identical pattern of statistically significant and non-significant differences in pre-post improvement between the experimental groups (see supplement S7). Concerning possible differences between the two data collection cohorts, none of the model comparisons indicated an improved fit when allowing the model parameters to vary between cohorts.

3.3. Exploratory analyses

We observed a decrease in the proportion of correct conclusions in the control group from 23 % at baseline to 12 % at post-test, while that proportion increased in all training groups, with the smallest increase in the ST group (12 % to 27 %), a somewhat

stronger increase in the VR group (17 % to 35 %) and the strongest increase in the combined training group (21 % to 59 %). The proportions were calculated by dividing the number of correct conclusions per group at a given measurement occasion by the total number of conversations in the respective group at that measurement occasion (two conversations per participant).

4. Discussion

We conducted a pre-registered, randomized-controlled study to evaluate ViContact, a two-fold training program on conversations with children about possible (sexual) abuse consisting of simulated conversations in a VR environment followed by individual feedback and an online seminar training. We aimed to assess the impact of both training modes separately and in combination on participant's conversational behavior, self-efficacy as well as their attitudes regarding suspected child abuse cases.

4.1. VR conversation performance

All three training groups achieved substantial improvements on the proportion of *recommended questions* (open-ended, non-suggestive and child-appropriate), a measure that has been widely used as an outcome in former child interviewer training studies. The largest effects were found in the combined training (from 41 % at baseline to 84 % at post-test) and the VR training group (from 42 % to 77 %); the effect in the online seminar training group (from 41 % to 71 %) was statistically smaller than in the combined, but not smaller than in the VR training group. These effect sizes were similar to those found for the Empowering Interviewer Training (EIT; different studies found increases between 20 % and 53 % across four to eight interviews; Haginoya et al., 2020, 2021; Krause et al., 2017; Pompedda, 2018). While we had expected the advantage of both groups that received VR training over our control group, the effect of the online seminar training alone compared to the control group was unexpected. Apparently, although no interview roleplays were included, the broad range of other exercises on appropriate questions in the seminar (writing tasks, discussions, watching videos) made it possible for the participants to apply the newly acquired skills in the VR conversations.

ViContact is the first computer-based program to also include the provision of rapport and socio-emotional support as a training goal, as this has long been recognized as a central technique to support children's disclosures (Saywitz et al., 2015). On the *percentage of supportive utterances*, only the combined training group showed a significant increase (19 % to 29 %); VR or seminar training alone did not lead to a significant increase. Apparently, providing socio-emotional support requires a deeper understanding of the construct than was provided in the brief instructions for the VR training alone, but also more practice and feedback than was provided in the seminar training alone. Another reason for the good results of the seminar training regarding recommended questions and the superiority of the combined training regarding supportive utterances could be the inclusion of videos with actors exemplifying good and bad mock interviewing in the seminar training, because similar interventions (so called “modeling”) have increased interviewer performance in other studies (Haginoya et al., 2021).

In line with our hypotheses, both the combined training (2.3 to 6.4 points) and the VR training group (2.8 to 7.0) showed an improvement on *supportive opening and closure tasks*, while the online seminar training did not lead to a significant improvement. This may have been due to the word-by-word instruction in advance of the VR training, while in the seminar training, the issue was addressed in a more general way.

On a descriptive level, we explored the correctness of participants' conclusions about the virtual children's critical memories regarding CSA, a child protection issue or another event with no need to intervene. While in the control group, the percentage of correct conclusions decreased from baseline (23 %) to post-test (12 %), all training groups saw an increase of correct conclusions, with the largest increase (12 % to 59 %) in the combined training group and smaller increases in the single intervention groups (VR: 17 % to 35 %, ST: 12 % to 27 %). The improvements in the combined training group resemble to findings by Pompedda et al. (2022), who saw an average number of 11 % correct conclusions in their no-feedback condition and an increase from 5 % (first conversation) to 50 % correct conclusions (eighth conversation) in the feedback condition.

Our results are in line with the established notion that practical training combined with feedback improves interviewing skills beyond classical seminar-style approaches (Benson & Powell, 2015). It allows for repeated practice in a safe learning environment and enables gradual improvement through process- and outcome-feedback. While some computer-based programs work with written dialog (e.g., “Unreal Interviewing”; Guadagno & Powell, 2012), ViContact allows for a verbal interaction with a virtual child, similar to the “Empowering Interviewer Training” (EIT; Pompedda, 2018).

4.2. Self-report questionnaires

Teachers have reported a lack of confidence in their own abilities to handle abuse suspicions (Goebbels et al., 2008; Goldman, 2007; Greytak, 2009; Tener & Sigad, 2019). Raising our trainees' level of self-efficacy for real-life situations thus has been a central goal of ViContact. The training effect on self-efficacy score was largest after combined training (from a score of 31 at baseline to 47 at post-test), somewhat smaller in the ST group (34 to 44), and smallest, although still significantly different from control group, after VR training only (37 to 40). This mostly conforms with our hypothesis that each training mode would have an individual impact on self-efficacy, with both effects adding up when the trainings are combined. We did not expect the advantage of the online seminar training over VR training, but this difference can possibly be explained with the broader range of topics and the longer duration of the seminar training.

Besides facilitating concrete behavior change, we also deemed it necessary to address broader underlying attitudes and cognitions regarding child abuse suspicions in order to prevent expectancy-driven, confirmatory, and suggestive questioning. We expected the

seminar training to lead to a decrease in the CECSA Naïve Confidence score, with no such effect in the VR or control group. Only the decrease in the ST group (33.5 to 28.0) differed statistically from that in the control group (33.5 to 32.7). All other group differences remained insignificant, although the change in the combined training group (34.7 to 30.2) was numerically not much smaller than that in the ST group.

4.3. General discussion

The present study is the first to compare a computer-based training on child interviewing with a classical seminar training (delivered online) and with a control group receiving no training. It is also the first to assess the combined effect of VR and online seminar training beyond the individual interventions. Indeed, the combined training was the only one improving both outcome measures of socio-emotional support, it saw the highest increase in self-efficacy and, on a descriptive level, it led to the highest increase of correct conclusions. Interestingly, the online seminar training alone was already effective in altering conversational behavior, but the combined training exceeded this effect. Thus, rather than fully replacing classical seminar trainings with virtual interview simulations, combining both interventions seems to be the most promising approach.

Most training programs to date have been designed for forensic interviewers. ViContact, instead, targets teachers and student teachers and could similarly be used by other childcare professionals. Existing educational programs for those professional groups almost never teach conversational skills for child protection matters (Rheingold et al., 2014; Topping & Barron, 2009), although a large training need has been identified (Brubacher et al., 2014; Cerezo & Pons-Salvador, 2004). Correspondingly, our virtual children are designed to resemble children at an average elementary school age (eight to ten years old), while other computer-based trainings have simulated younger children (four to six years old; Guadagno & Powell, 2012; Pompedda, 2018). In addition, previous interview simulations have instructed participants to find out whether a child was sexually abused or not. However, especially early conversations in school or child protection settings require a more open-minded approach considering the broad range of (adverse) events that a child may have experienced. We therefore programmed various critical adverse memories for our virtual children and had participants distinguish between sexual abuse, other child protection issues, and smaller-scaled stressful events, in order to teach open-mindedness.

4.4. Limitations and future directions

One limitation to our study design is that the practical training completed through VR consisted of the same task that was used for the behavioral outcome measures, which may have led to an advantage for the groups with VR training. However, such an effect would have been mitigated by the fact that participants of all groups got the opportunity to familiarize with the VR conversations during the baseline measurement.

Moreover, the coding system and conversational algorithm might encourage an undesired gamification effect. That is, some participants might learn to stereotypically repeat certain recommended questions or supportive utterances, because the coding system and algorithm do not detect repeated questions. While this can pose methodological issues in an evaluation study, trainees who seek to acquire skills for real-life contexts can be expected to try and lead the virtual conversations in the way they would also talk to a real child.

Concerning statistical power, our results showed that the assumption of a pre-post correlation of $r = 0.60$ only held for the *naïve confidence* score. The other measures' pre-post correlation ranged between $r = 0.10$ for *proportion of recommended questions* and $r = 0.31$ for *self-efficacy*. Thus, as a lower boundary, achieved power for *recommended questions* was satisfying only for large effects ($\eta_p^2 = 0.14$, $1 - \beta = 0.86$), but nor for intermediate effects ($\eta_p^2 = 0.06$, $1 - \beta = 0.33$). Considering the magnitude of the training effects we found, which were similar to those reported in previous studies (e.g., Brubacher et al., 2015 reported $\eta_p^2 = 0.30 \ldots 0.60$ for pre-post-change in several desired and undesired question types; Krause et al., 2017 reported $\eta_p^2 = 0.39$ for a group by time interaction between feedback and control group after eight conversations), we still consider our analyses sufficiently powered to answer our research questions. One might also argue that for a training aimed at remarkable behavior change, intermediate effects would not be practically significant.

Compared to mock interviews with trained actors (e.g., Lawrie et al., 2021; Powell et al., 2014), computer-based interview simulations can provide more scalable and standardized training opportunities and control the imitation of children's response patterns. However, as we intentionally did not include role play conversations in the online seminar training, it cannot be directly inferred from our results, whether VR conversations lead to better results than role-play exercises. Comparing the training effects of role-plays vs. computer-based simulations, but also of three-dimensional vs. two-dimensional simulations will be an important research endeavor for the future.

In order to transfer the ViContact training program from laboratory to practical settings, the current VR software is being further developed to enable usage without specific IT competences. In addition, a follow-up project addresses further analysis and optimization of the virtual children's conversation engine. Currently, ViContact runs with human operators, who manually code the questions and, for the present study, were not blinded to the experimental condition, which potentially limits reliability (interrater reliability was assessed via paper-pencil, but not within the virtual environment). This human component also limits a larger scaled distribution of the training. An important future step is to exchange the human operator by an automated coding of questions, for example, through a machine learning model, to simplify its application (similar first steps have been done for the EIT, but with only moderate classification accuracy; see Haginoya et al., 2023) or by fine-tuning pre-existing large language models (LLM's; as done by Røed, Baugerud, et al., 2023). Transforming parts of the online seminar training (e.g., theoretical input and individual exercises) into an automated e-learning system could further increase scalability and standardization of the combined training approach. Instead of full automation, however,

a blended learning approach may be advisable, where advantages of asynchronous e-learning (e.g., individual timing and pace) are combined with those of traditional (online or face-to-face) student-teacher group-encounters (e.g., allowing for questions, discussions, and cooperative learning; Davis et al., 2018; Means et al., 2013), with the latter also providing a space to share and discuss experiences of dealing with CSA suspicions. Lastly, future studies need to investigate the longevity of ViContact's training effects as well as the transfer of its effects on real life interviews.

4.5. Conclusion

Evaluating "ViContact", a training program on how to conduct conversations with children about suspected (sexual) abuse, we contrasted a practical Virtual Reality training component with an online seminar training and tested the effect of their combined application. Both interventions were shown to be effective in improving participant's conversational skills but were most helpful when applied in combination.

CRediT authorship contribution statement

Niels Krause: Data curation, Formal analysis, Investigation, Methodology, Resources, Validation, Visualization, Writing – original draft. **Elsa Gewehr:** Data curation, Investigation, Methodology, Resources, Writing – original draft. **Hermann Barbe:** Methodology, Software, Writing – review & editing. **Marie Merschhemke:** Data curation, Investigation, Resources, Writing – review & editing. **Frieda Mensing:** Data curation, Investigation, Writing – review & editing. **Bruno Siegel:** Methodology, Software, Writing – review & editing. **Jürgen L. Müller:** Conceptualization, Funding acquisition, Methodology, Project administration, Resources, Supervision, Writing – review & editing. **Renate Volbert:** Conceptualization, Funding acquisition, Methodology, Project administration, Supervision, Writing – review & editing. **Peter Fromberger:** Conceptualization, Funding acquisition, Methodology, Project administration, Resources, Software, Supervision, Writing – review & editing. **Anett Tamm:** Methodology, Project administration, Resources, Supervision, Writing – review & editing. **Simone Pülschen:** Conceptualization, Data curation, Funding acquisition, Investigation, Methodology, Project administration, Resources, Supervision, Writing – review & editing.

Data availability

Data and analysis scripts available at the Open Science Framework: osf.io/v8g4m

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.chiabu.2024.106677.

References

Bailenson, J. N., Swinth, K., Hoyt, C., Persky, S., Dimov, A., & Blascovich, J. (2005). The independent and interactive effects of embodied-agent appearance and behavior on self-report, cognitive, and behavioral markers of copresence in immersive virtual environments. *Presence: Teleoperators and Virtual Environments, 14*(4), 379–393. https://doi.org/10.1162/105474605774785235

Barbe, H., Müller, J. L., Siegel, B., & Fromberger, P. (2023). An open source virtual reality training framework for the criminal justice system. *Criminal Justice and Behavior*, 294–303. https://doi.org/10.1177/00938548221124128

Barbe, H., Siegel, B., Müller, J. L., & Fromberger, P. (2020). Welches Potenzial haben virtuelle Realitäten in der klinischen und forensischen Psychiatrie? Ein Überblick über aktuelle Verfahren und Einsatzmöglichkeiten. *Forensische Psychiatrie, Psychologie, Kriminologie, 14*(3), 270–277. https://doi.org/10.1007/s11757-020-00611-2

Benson, M. S., & Powell, M. B. (2015). Evaluation of a comprehensive interactive training system for investigative interviewers of children. *Psychology, Public Policy, and Law, 21*(3), 309–322. https://doi.org/10.1037/law0000052

Bonnet, I., Riese, D., Jordan, K., Müller, J. L., & Fromberger, P. (2018). Evaluation eines Sets virtueller Charaktere zur VR-Forschung im forensischen Kontext. In *EFPPP Jahrbuch 2018: Empirische Forschung in der forensischen Psychiatrie, Psychologie und Psychotherapie* (pp. 11–21). MWV Medizinisch Wissenschaftliche Verlagsgesellschaft, 1., Bd. 7, S.

Brubacher, S. P., Powell, M., Skouteris, H., & Guadagno, B. (2014). An investigation of the question-types teachers use to elicit information from children. *Australian Educational and Developmental Psychologist, 31*(2), 125–140. https://doi.org/10.1017/edp.2014.5

Brubacher, S. P., Powell, M., Skouteris, H., & Guadagno, B. (2015). The effects of e-simulation interview training on teachers' use of open-ended questions. *Child Abuse & Neglect, 43*, 95–103. https://doi.org/10.1016/j.chiabu.2015.02.004

Brubacher, S. P., Powell, M. B., Snow, P. C., Skouteris, H., & Manger, B. (2016). Guidelines for teachers to elicit detailed and accurate narrative accounts from children. *Children and Youth Services Review, 63*, 83–92. https://doi.org/10.1016/j.childyouth.2016.02.018

Cerezo, M. A., & Pons-Salvador, G. (2004). Improving child maltreatment detection systems: A large-scale case study involving health, social services, and school professionals. *Child Abuse & Neglect, 28*(11), 1153–1169. https://doi.org/10.1016/j.chiabu.2004.06.007

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2. Aufl.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Cyr, M., & Lamb, M. E. (2009). Assessing the effectiveness of the NICHD investigative interview protocol when interviewing French-speaking alleged victims of child sexual abuse in Quebec. *Child Abuse & Neglect, 33*(5), 257–268. https://doi.org/10.1016/j.chiabu.2008.04.002

Davis, D., Chen, G., Hauff, C., & Houben, G.-J. (2018). Activating learning at scale: A review of innovations in online learning strategies. *Computers & Education, 125*, 327–344. https://doi.org/10.1016/j.compedu.2018.05.019

Emmelkamp, P. M., & Meyerbröker, K. (2021). Virtual reality therapy in mental health. *Annual Review of Clinical Psychology, 17*, 495–519. https://doi.org/10.1146/annurev-clinpsy-081219-115923

Faller, K. (2015). Forty years of forensic interviewing of children suspected of sexual abuse, 1974–2014: Historical benchmarks. *Social Sciences, 4*(1), 34–65. https://doi.org/10.3390/socsci4010034
Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods, 41*(4), 1149–1160. https://doi.org/10.3758/BRM.41.4.1149
Finnilä-Tuohimaa, K., Santtila, P., Björnberg, L., Hakala, N., Niemi, P., & Sandnabba, K. (2008). Attitudes related to child sexual abuse: Scale construction and explorative study among psychologists. *Scandinavian Journal of Psychology, 49*(4), 311–323. https://doi.org/10.1111/j.1467-9450.2008.00635.x
Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin, 76*(5), 378.
Fromberger, P., Jordan, K., & Müller, J. L. (2014). Anwendung virtueller Realitäten in der forensischen Psychiatrie: Ein neues Paradigma? *Der Nervenarzt, 85*(3), 298–303. https://doi.org/10.1007/s00115-013-3904-7
Fromberger, P., Jordan, K., & Müller, J. L. (2018). Virtual reality applications for diagnosis, risk assessment and therapy of child abusers. *Behavioral Sciences & the Law, 36*(2), 235–244. https://doi.org/10.1002/bsl.2332
Gewehr, E., Volbert, R., Merschhemke, M., Santtila, P. O., & Pülschen, S. (2023). Cognitions and Emotions about Child Sexual Abuse (CECSA): Development of a self-report measure to predict interviewer bias [preprint]. *PsyArXiv*. https://doi.org/10.31234/osf.io/qcfvb
Gist, M. E., & Mitchell, T. R. (1992). Self-efficacy: A theoretical analysis of its determinants and malleability. *Academy of Management Review, 17*(2), 183–211.
Glammeier, S. (2019). Sexuelle Gewalt und Schule. Hrsg.. In M. Wazlawik, A. Henningsen, A. Dekker, H.-J. Voß, & A. Retkowski (Eds.), *Sexuelle Gewalt in pädagogischen Kontexten* (pp. S. 197–209). Springer Fachmedien Wiesbaden GmbH
Goebbels, A. F. G., Nicholson, J. M., Walsh, K., & De Vries, H. (2008). Teachers' reporting of suspected child abuse and neglect: Behaviour and determinants. *Health Education Research, 23*(6), 941–951. https://doi.org/10.1093/her/cyn030
Goldman, J. D. G. (2007). Primary school student-teachers' knowledge and understandings of child sexual abuse and its mandatory reporting. *International Journal of Educational Research, 46*(6), 368–381. https://doi.org/10.1016/j.ijer.2007.09.002
Greytak, E. A. (2009). Are teachers prepared? Predictors of teachers' readiness to serve as mandated reporters of child abuse. In *ProQuest dissertations and theses* (p. S. 380). http://repository.upenn.edu/edissertations/57.
Guadagno, B., & Powell, M. B. (2012). E-simulations for the purpose of training forensic (investigative) interviewers. Hrsg.. In D. Holt, S. Segrave, & J. L. Cybulski (Eds.), *Professional education using e-simulations: Benefits of blended learning design* (pp. S. 71–87). Business Science Reference
Haginoya, S., Ibe, T., Yamamoto, S., Yoshimoto, N., Mizushi, H., & Santtila, P. (2023). AI avatar tells you what happened: The first test of using AI-operated children in simulated interviews to train investigative interviewers. *Frontiers in Psychology, 14*, 579. https://doi.org/10.3389/fpsyg.2023.1133621
Haginoya, S., Yamamoto, S., Pompedda, F., Naka, M., Antfolk, J., & Santtila, P. (2020). Online simulation training of child sexual abuse interviews with feedback improves interview quality in Japanese university students. *Frontiers in Psychology, 11*(May). https://doi.org/10.3389/fpsyg.2020.00998
Haginoya, S., Yamamoto, S., & Santtila, P. (2021). The combination of feedback and modeling in online simulation training of child sexual abuse interviews improves interview quality in clinical psychologists. *Child Abuse & Neglect, 115*, Article 105013. https://doi.org/10.1016/j.chiabu.2021.105013
Hassan, S. Z., Salehi, P., Røed, R. K., Halvorsen, P., Baugerud, G. A., Johnson, M. S., … Sabet, S. S. (2022). Towards an AI-driven talking avatar in virtual reality for investigative interviews of children. In *Proceedings of the 2nd workshop on games systems* (pp. 9–15). https://doi.org/10.1145/3534085.3534340
Hershkowitz, I., Lamb, M. E., Katz, C., & Malloy, L. C. (2015). Does enhanced rapport-building alter the dynamics of investigative interviews with suspected victims of intra-familial abuse? *Journal of Police and Criminal Psychology, 30*(1), 6–14. https://doi.org/10.1007/s11896-013-9136-8
Kask, K., Pompedda, F., Palu, A., Schiff, K., Mägi, M.-L., & Santtila, P. (2022). Transfer of avatar training effects to investigative field interviews of children conducted by police officers. *Frontiers in Psychology, 13*, 75311. https://doi.org/10.3389/fpsyg.2022.753111
Kassambara, A. (2021). rstatix: Pipe-friendly framework for basic statistical tests (0.7.0) [software]. https://CRAN.R-project.org/package=rstatix.
Kennedy, R. S., Lane, N. E., Berbaum, K. S., & Lilienthal, M. G. (1993). Simulator sickness questionnaire: An enhanced method for quantifying simulator sickness. *The International Journal of Aviation Psychology, 3*(3), 203–220. https://doi.org/10.1207/s15327108ijap0303_3
King, C. B., & Scott, K. L. (2014). Why are suspected cases of child maltreatment referred by educators so often unsubstantiated? *Child Abuse and Neglect, 38*(1), 1–10. https://doi.org/10.1016/j.chiabu.2013.06.002
König, E., Hoffmann, U., Liebhardt, H., Michi, E., Niehues, J., & Fegert, J. M. (2015). Ergebnisse der Evaluation des Online-Kurses "Prävention von sexuellem Kindesmissbrauch". In *Sexueller Missbrauch von Kindern und Jugendlichen* (pp. S. 15–24). Springer.
Korkman, J., Santtila, P., Drzewiecki, T., & Sandnabba, N. K. (2008). Failing to keep it simple: Language use in child sexual abuse interviews with 3–8-year-old children. *Psychology, Crime & Law, 14*(1), 41–60. https://doi.org/10.1080/10683160701368438
Korkman, J., Santtila, P., & Sandnabba, N. K. (2006). Dynamics of verbal interaction between interviewer and child in interviews with alleged victims of child sexual abuse. *Scandinavian Journal of Psychology, 47*(2), 109–119. https://doi.org/10.1111/j.1467-9450.2006.00498.x
Krause, N., Pompedda, F., Antfolk, J., Zappalá, A., & Santtila, P. (2017). The effects of feedback and reflection on the questioning style of untrained interviewers in simulated child sexual abuse interviews. *Applied Cognitive Psychology, 31*(2), 187–198. https://doi.org/10.1002/acp.3316
Lamb, M. E., Brown, D. A., Hershkowitz, I., Orbach, Y., & Esplin, P. W. (2018). *Tell me what happened: Questioning children about abuse* (2nd ed.). Wiley-Blackwell.
Lamb, M. E., Hershkowitz, I., Sternberg, K. J., Esplin, P. W., Hovav, M., Manor, T., & Yudilevitch, L. (1996). Effects of investigative utterance types on Israeli children's responses. *International Journal of Behavioral Development, 19*(3), 627–638. https://doi.org/10.1080/016502596385721
Lawrie, M., Brubacher, S. P., Earhart, B., Powell, M. B., Steele, L. C., & Boud, D. (2021). Testing the effectiveness of a blended vulnerable witness training for forensic interviewers. *Journal of Family Trauma, Child Custody & Child Development, 18*(3), 279–297. https://doi.org/10.1080/26904586.2021.1894303
Lawrie, M., Brubacher, S. P., Powell, M. B., & Boud, D. (2020). Forensic interviewers' perceptions of the utility of mock interviews with trained actors as a training tool for child interviewing. *Child Abuse and Neglect, 106*, Article 104553. https://doi.org/10.1016/j.chiabu.2020.104553
Márquez-Flores, M. M., Márquez-Hernández, V. V., & Granados-Gámez, G. (2016). Teachers' knowledge and beliefs about child sexual abuse. *Journal of Child Sexual Abuse, 25*(5), 538–555. https://doi.org/10.1080/10538712.2016.1189474
McKee, B. E., & Dillenburger, K. (2009). Child abuse and neglect: Training needs of student teachers. *International Journal of Educational Research, 48*(5), 320–330. https://doi.org/10.1016/j.ijer.2010.03.002
Means, B., Toyama, Y., Murphy, R., & Baki, M. (2013). The effectiveness of online and blended learning: A meta-analysis of the empirical literature. *Teachers College Record: The Voice of Scholarship in Education, 115*(3), 1–47. https://doi.org/10.1177/016146811311500307
Nicol, A., Szojka, Z. A., Watkins, C. D., Gabbert, F., & La Rooy, D. (2023). A systematic examination of actor and trainee interviewer behaviour during joint investigative interviewing training. *Journal of Police and Criminal Psychology, 38*(3), 593–606. https://doi.org/10.1007/s11896-023-09577-3
Poelchau, H.-W., Briken, P., Wazlawik, M., Bauer, U., Fegert, J., & Kavemann, B. (2015). Bonner Ethik-Erklärung. *Zeitschrift für Sexualforschung, 28*(02), 153–160. https://doi.org/10.1055/s-0035-1553220
Poeschl, S., & Doering, N. (2013). The German VR Simulation Realism Scale—Psychometric construction for virtual reality applications with virtual humans. *Studies in Health Technology and Informatics, 191*, 33–37.
Pompedda, F. (2018). Training in investigative interviews of children: Serious gaming paired with feedback improves interview quality (Doctoral dissertation) https://www.doria.fi/handle/10024/152565.
Pompedda, F., Antfolk, J., Zappalà, A., & Santtila, P. (2017). A combination of outcome and process feedback enhances performance in simulations of child sexual abuse interviews using avatars. *Frontiers in Psychology, 8*, 1474. https://doi.org/10.3389/fpsyg.2017.01474
Pompedda, F., Palu, A., Kask, K., Schiff, K., Soveri, A., Antfolk, J., & Santtila, P. (2020). Transfer of simulated interview training effects into interviews with children exposed to a mock event. *Nordic Psychology, 73*(1), 43–67. https://doi.org/10.1080/19012276.2020.1788417
Pompedda, F., Zappalà, A., & Santtila, P. (2015). Simulations of child sexual abuse interviews using avatars paired with feedback improves interview quality. *Psychology, Crime & Law, 21*(1), 28–52. https://doi.org/10.1080/1068316X.2014.915323
Pompedda, F., Zhang, Y., Haginoya, S., & Santtila, P. (2022). A mega-analysis of the effects of feedback on the quality of simulated child sexual abuse interviews with avatars. *Journal of Police and Criminal Psychology, 37*, 485–497. https://doi.org/10.1007/s11896-022-09509-7

Powell, M. B. (2008). Designing effective training programs for investigative interviewers of children. *Current Issues in Criminal Justice, 20*(2), 189–208. https://doi.org/10.1080/10345329.2008.12035804

Powell, M. B., & Barnett, M. (2015). Elements underpinning successful implementation of a national best-practice child investigative interviewing framework. *Psychiatry, Psychology and Law, 22*(3), 368–377. https://doi.org/10.1080/13218719.2014.951112

Powell, M. B., Brubacher, S. P., & Baugerud, G. A. (2022). An overview of mock interviews as a training tool for interviewers of children. *Child Abuse & Neglect, 129*, Article 105685. https://doi.org/10.1016/j.chiabu.2022.105685

Powell, M. B., Fisher, R. P., & Hughes-Scholes, C. H. (2008a). The effect of intra- versus post-interview feedback during simulated practice interviews about child abuse. *Child Abuse and Neglect, 32*(2), 213–227. https://doi.org/10.1016/j.chiabu.2007.08.002

Powell, M. B., Fisher, R. P., & Hughes-Scholes, C. H. (2008b). The effect of using trained versus untrained adult respondents in simulated practice interviews about child abuse. *Child Abuse and Neglect, 32*(11), 1007–1016. https://doi.org/10.1016/j.chiabu.2008.05.005

Powell, M. B., Guadagno, B., & Benson, M. (2014). Improving child investigative interviewer performance through computer-based learning activities. *Policing and Society, 26*(4), 365–374. https://doi.org/10.1080/10439463.2014.942850

R Core Team. (2020). *R: A Language and environment for statistical computing (3.6.3) [software]*. R Foundation for Statistical Computing. https://www.R-project.org/.

Revelle, W. (2022). *psych: Procedures for psychological, psychometric, and personality research*. Northwestern University. https://CRAN.R-project.org/package=psych.

Rheingold, A. A., Zajac, K., Chapman, J. E., Patton, M., de Arellano, M., Saunders, B., & Kilpatrick, D. (2014). Child sexual abuse prevention training for childcare professionals: An independent multi-site randomized controlled trial of stewards of children. *Prevention Science, 16*(3), 374–385. https://doi.org/10.1007/s11121-014-0499-6

Røed, R. K., Baugerud, G. A., Hassan, S. Z., Sabet, S. S., Salehi, P., Powell, M. B., … Johnson, M. S. (2023). Enhancing questioning skills through child avatar chatbot training with feedback. *Frontiers in Psychology, 14*, 1198235. https://doi.org/10.3389/fpsyg.2023.1198235

Røed, R. K., Powell, M. B., Riegler, M. A., & Baugerud, G. A. (2023). A field assessment of child abuse investigators' engagement with a child-avatar to develop interviewing skills. *Child Abuse & Neglect, 143*, Article 106324. https://doi.org/10.1016/j.chiabu.2023.106324

RStudio Team. (2022). *RStudio: Integrated development environment for R*. RStudio, PBC. http://www.rstudio.com/.

Salehi, P., Hassan, S. Z., Lammerse, M., Sabet, S. S., Riiser, I., Røed, R. K., … Riegler, M. A. (2022). Synthesizing a talking child avatar to train interviewers working with maltreated children. *Big Data and Cognitive Computing, 6*(2), 62. https://doi.org/10.3390/bdcc6020062

Saywitz, K. J., Larson, R. P., Hobbs, S. D., & Wells, C. R. (2015). Developing rapport with children in forensic interviews: Systematic review of experimental research: Developing rapport with children. *Behavioral Sciences & the Law, 33*(4), 372–389. https://doi.org/10.1002/bsl.2186

Schols, M. W., de Ruiter, C., & Öry, F. G. (2013). How do public child healthcare professionals and primary school teachers identify and handle child abuse cases? A qualitative study. *BMC Public Health, 13*, 807.

Schubert, T., Friedmann, F., & Regenbrecht, H. (2001). The experience of presence: Factor analytic insights. *Presence: Teleoperators and Virtual Environments, 10*(3), 266–281. https://doi.org/10.1162/105474601300343603

Schultheis, M. T., & Rizzo, A. A. (2001). The application of virtual reality technology in rehabilitation. *Rehabilitation Psychology, 46*(3), 296–311. https://doi.org/10.1037/0090-5550.46.3.296

Sedlak, A. J., Mettenburg, J., Basena, M., Petta, I., McPherson, K., Greene, A., & Li, S. (2010). Fourth national incidence study of child abuse and neglect (NIS – 4): Report to congress. In *National data archive on child abuse and neglect*. https://doi.org/10.1037/e659872010-001

Sternberg, K. J., Lamb, M. E., Hershkowitz, I., Esplin, P. W., Redlich, A., & Sunshine, N. (1996). The relation between investigative utterance types and the informativeness of child witnesses. *Journal of Applied Developmental Psychology, 17*(3), 439–451. https://doi.org/10.1016/S0193-3973(96)90036-2

Talbot, T., & Rizzo, A. S. (2019). Virtual Human Standardized Patients for Clinical Training. In A. S. Rizzo, & S. Bouchard (Eds.), *Virtual Reality for Psychological and Neurocognitive Interventions. Virtual Reality Technologies for Health and Clinical Applications*. New York, NY: Springer. https://doi.org/10.1007/978-1-4939-9482-3_17.

Tamm, A., Otzipka, J., & Volbert, R. (2021). Assessing the individual interviewer rapport-building and supportive techniques of the R-NICHD protocol. *Frontiers in Psychology, 12*, 2987. https://doi.org/10.3389/fpsyg.2021.659438

Tener, D., & Sigad, L. (2019). "I felt like I was thrown into a deep well": Educators coping with child sexual abuse disclosure. *Children and Youth Services Review, 106*, Article 104465. https://doi.org/10.1016/j.childyouth.2019.104465

Topping, K. J., & Barron, I. G. (2009). School-based child sexual abuse prevention programs: A review of effectiveness. *Review of Educational Research, 79*(1), 431–463. https://doi.org/10.3102/0034654308325582

Volbert, R. (2015). Gesprächsführung mit von sexuellem Missbrauch betroffenen Kindern und Jugendlichen. Hrsg.. In J. M. Fegert, U. Hoffmann, E. König, J. Niehues, & H. Liebhardt (Eds.), *Sexueller Missbrauch von Kindern und Jugendlichen* (pp. S. 185–194). Springer-Verlag

Walsh, K., Mathews, B., Rassafiani, M., Farrell, A., & Butler, D. (2012). Understanding teachers' reporting of child sexual abuse: Measurement methods matter. *Children and Youth Services Review, 34*(9), 1937–1946. https://doi.org/10.1016/j.childyouth.2012.06.004

Wilcox, B., & Wilcox, S. (2013). Making it real: Loebner-winning chatbot design. *ARBOR Ciencia, Pensamiento y Cultura, 189*(764), Article a086.

Secondary Analysis to Article 3:

Changing Cognitions and Emotions about Child Sexual Abuse through a Seminar Intervention

Participants of the ViContact evaluation study filled in the 23 items of the questionnaire on Cognitions and Emotions about Child Sexual Abuse (CECSA) before and after the interventions. However, for Article 3, we decided to reduce the number of outcome variables to limit alpha inflation, and therefore included only the Naive Confidence scale into the analyses. NC was chosen because the issues covered by this scale were explicitly discussed in the seminar training (e.g., non-diagnosticity of children's behavior for CSA, children's disclosure and reporting patterns, suggestion, false allegations), while the topics of the other two scales were not explicitly targeted. As expected, the seminar training led to a reduced NC score to a significantly larger extend than in the control group. The other expected contrasts (advantages of combined training over the control and VR groups; advantage of seminar training over VR training) were not found. This secondary analysis assesses how - next to NC - the other two CECSA scales Emotional Reactivity and Justice System Distrust were influenced by the training interventions.

In general, changing cognitive and emotional patterns through seminar training as provided in ViContact is promising because knowledge transfer (expertise and meta-knowledge about cognitive biases) is a foundation of debiasing strategies (Oberlader et al., 2024). This is because it can correct the false beliefs that underlie cognitive biases according to process-oriented conceptualizations of bias (Oeberst & Imhoff, 2023).

The issues of JSD and ER were less explicitly discussed in the seminar training, but they were touched on various occasions, such that influence on these scales seems plausible as well. For example, Justice System Distrust may have been influenced from a generally positive

perspective on the justice system that was transferred in the training (e.g., legal definitions were used, documentation useful for juridical procedures was taught). Emotions about CSA were not explicitly addressed, but the training generally aimed at bringing a certain level-headedness into the often very emotional topic of CSA through providing empirical information and concrete action strategies and reducing uncertainty. Open discussions provided the opportunity to ask questions and discuss emotional topics. In addition, cognitive appraisal theory suggests that cognitive evaluations of an event shape the emotional responses to it (Siemer et al., 2007), which allows for optimism about a seminar training effect on Emotional Reactivity.

On the other hand, knowledge transfer has also often proven insufficient for fundamental change in attitudes (Forscher et al., 2019), emotions (Smith & Neumann, 2005), or bias in forensic judgement (Neal et al., 2022). Although the ViContact seminar training also included exercises and discussions, these were mostly focused on conversational skills, not on biasing cognitions or emotions. Similarly little is known about how other interview training programs influence cognitive and emotional patterns (Akca et al., 2021).

The following analyses compare the effects of the ViContact training components on the CECSA scales. Participants who received the seminar training – either uniquely or combined with the VR training – are expected to show larger decreases on the CECSA scales compared to both the control participants and the participants receiving only the VR training. For ease of interpretation and to apply a common alpha correction, all three CECSA scales are integrated into the following analyses, although the results for NC are already provided in Article 3.

Method

A secondary analysis of the dataset described in Article 3 of this dissertation (Krause et al., 2024) was conducted within R (v1.4.1717; R Core Team, 2021). For information on study

design and participants, see Article 3. Data and code are openly available via the Open Science Framework (https://osf.io/jzaqv/). Using the rstatix package (Kassambara, 2021), three global 4 (group, between) by 2 (pre-post, within) mixed ANOVAs, and – for significant interaction terms - following pairwise comparisons with 2 (group) by 2 (pre-post) mixed ANOVAs were conducted to assess differences between pre-post changes of the four groups (Seminar Training, VR Training, Combined Training, Control Group) for each of the CECSA scales. As in Article 3, Bonferroni correction was used to conservatively control for multiple testing, setting significance levels to $\alpha = .05/3 = .017$ for the global ANOVAs and $\alpha = .05/12 = .0042$ for the pairwise 2 by 2 ANOVAs.

Results

Internal consistencies of the CECSA scales at baseline and post-test varied between being excellent ($\alpha = .91$) and just below acceptable for early stages of research ($\alpha = .68$; Lance et al., 2006), with ER showing the highest and NC the lowest values. See Table 1 for further psychometric information. Mean CECSA scale scores per intervention group and timepoint are shown in Table 2 and Fig. 1. To ease comparisons with all studies of this thesis, Table 1 reports both mean sum scores of each scale (as used for this analysis and as reported in Article 3) and mean item scores (as reported in Articles 1 and 2).

Significant group by pre-post global interaction terms were observed for the three CECSA scales NC ($F = 4.63$, $p = .004$, $\eta_p^2 = .12$), ER ($F = 5.32$, $p = .002$, $\eta_p^2 = .13$), and JSD ($F = 6.54$, $p < .001$, $\eta_p^2 = .16$). Table 3 shows the interaction terms of the pairwise comparisons, lending partial support to the hypotheses: As expected, the seminar training alone led to a significant reduction of Naive Confidence and Emotional Reactivity compared to the control-group. For ER, but not for NC, this effect was also observed for the combination of seminar and

the VR training. Justice System Distrust was reduced only by the combined training. None of the contrasts comparing the seminar or combined training to the VR training was significant.

Table 1

Descriptive Results for the CECSA Scales at Baseline and Post-test Timepoints

Subscale	Timepoint	*M*	*SD*	*Mdn*	skew	kurtosis	α
Naive Confidence (NC)	baseline	3.11	0.53	3.09	0.01	-0.1	.68
	post-test	2.84	0.64	2.82	-0.04	0.61	.74
Emotional Reactivity (ER)	baseline	4.6	0.8	4.67	-0.71	1.22	.82
	post-test	4.36	1.01	4.33	-0.52	0.53	.91
Justice System Distrust (JSD)	baseline	3.42	0.79	3.5	0.04	0.17	.78
	post-test	3.23	0.84	3.25	0.05	-0.38	.86

Note. $N = 110$

Table 2

Means and Standard Deviations of the CECSA Scales per Group and Timepoint

		Naive Confidence				Emotional Reactivity				Justice System Distrust			
		Sum Score		Item Score		Sum Score		Item Score		Sum Score		Item Score	
		M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Control Group (n = 28)	baseline	33.5	5.69	3.05	.52	28.43	4.03	4.74	.67	19.93	4.42	3.32	.74
	post-test	32.71	6.32	2.97	.57	28.75	5.02	4.79	.84	20.54	5.32	3.42	.89
Seminar Training (n = 26)	baseline	33.46	6.2	3.04	.56	25.19	5.4	4.2	.9	19.88	4.36	3.31	.73
	post-test	28	6.59	2.55	.6	22.69	6.45	3.78	1.08	18	4.97	3	.83
Combined Training (n = 29)	baseline	34.66	5	3.15	.45	27.28	4.53	4.55	.76	21.1	4.85	3.52	.81
	post-test	30.24	6.33	2.75	.58	24.93	5.64	4.16	.94	18.31	4.7	3.05	.78
VR Training (n = 27)	baseline	35.26	6.67	3.21	.61	29.33	4.41	4.89	.74	21.19	5.44	3.53	.91
	post-test	33.93	7.91	3.08	.72	28.15	5.43	4.69	.91	2.7	4.91	3.45	0.82

Figure 1

Means of Main Outcomes at Baseline and Post-Test by Training Group.

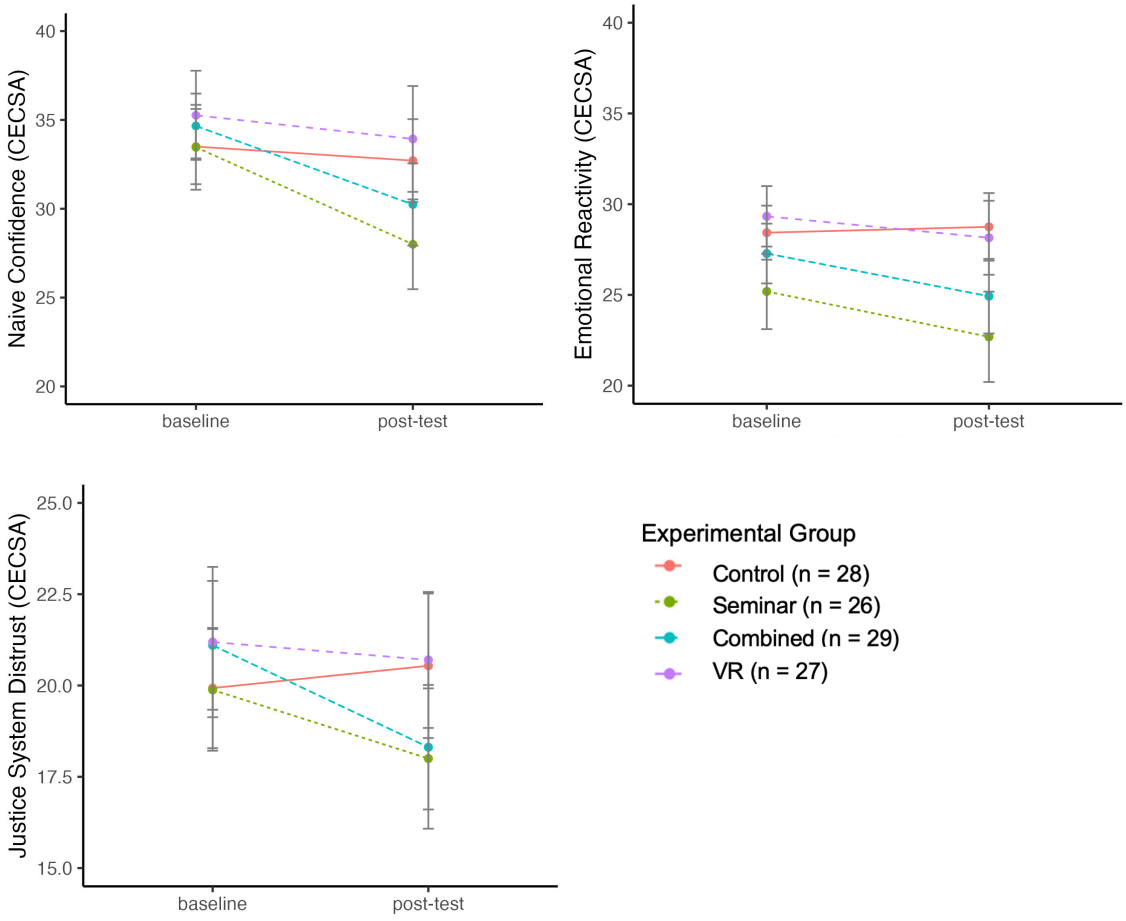


Table 3

Interaction Terms of 2 (Group) x 2 (Pre-Post) ANOVAs on the CECSA scales

Scale	Hypothesis	*F*	*p*	η_p^2
Naive Confidence	ST > CG	10.576	.002*	.169
	ST+VR > CG	5.969	.018	.098
	ST > VR	7.493	.009	.128
	ST+VR > VR	3.947	.052	.068
Emotional Reactivity	ST > CG	11.148	.002*	.177
	ST+VR > CG	11.306	.001*	.171
	ST > VR	2.615	.112	.049
	ST+VR > VR	2.315	.134	.041
Justice System Distrust	ST > CG	7.727	.008	.129
	ST+VR > CG	13.255	<.001*	.194
	ST > VR	3.828	.056	.070
	ST+VR > VR	8.630	.005	.138

Note. *indicates significant results for the Bonferroni-corrected threshold of $\alpha = .0042$

Discussion

The results overall show that a seminar training – either alone or in combination with a VR training – has the potential to decrease biasing cognitions and emotions about child sexual abuse compared to a no-training control group. This is particularly evident for Emotional Reactivity, where both groups receiving seminar training showed significant score reductions (of around 2.5 points, from mean scale scores of 25.2 to 22.7, and from 27.3 to 24.9). Although not directly addressed by the seminar, participants emotionality regarding CSA seems to have calmed down through receiving expert knowledge about CSA allegations, discussing, and practicing conversational skills. For Naive Confidence, only the seminar group showed a significant reduction (~ 5.5 points, from 33.5 to 28), while the combined group showed no significant effect. Potentially, improving one's conversational skills and conducting successful virtual conversations with children, as most largely observed in the combined group, not only

raises well-justified *self*-confidence (Krause et al., 2024) but also unjustified *naive* confidence about handling CSA suspicions. This is in line with research showing that practical experience and expertise does not lower, and sometimes even increases bias in different (forensic) domains (Neal et al., 2022; Oeberst & Imhoff, 2023). Future combined trainings could include discussions on how practical experience and perceptions of success in eliciting children's disclosure can mislead professionals into believing that they can innately solve CSA cases and thus increase their susceptibility to bias. For JSD, the effects were reversed for the training conditions, with the combined training leading to a significant but small reduction (~ 3 points, from 21.1 to 18.3), while the seminar training alone yielded no significant effect. Potentially, experiencing the difficulty of questioning children and trying to find out the truth about their experiences in the combined group somewhat raised participants' understanding about the challenges of juridical professionals who work with such cases in real life.

Overall, the impact of adding a VR or other practical interview training to a seminar should be reassessed with a larger sample size. The differences found here between the seminar and combined group are descriptively only small and the differences in significance may be due to the small sample size and the decision for a conservative Bonferroni correction. Less conservative procedures (e.g., Bonferroni-Holm correction), might have led to the non-significant effects, that is, the effect of the combined training for NC ($\alpha = .018$), and the effect of the seminar training for JSD ($\alpha = .008$), being interpreted as significant.

When comparing the effects of the seminar and combined training with the effects of the VR training, instead of the control group, none of the expected advantages in reducing CECSA scores were found to be significant. The reason seems to be that VR training alone already led to a slight descriptive decrease on most of the CECSA scales, while the control group usually

showed no change or even a small increase. As a result, the effect sizes of the contrasts may have been too small to be detected with the sample size of this study. Studies with larger sample sizes are needed to shed light on the different effects of seminar vs. VR training in reducing participants' biasing cognitions and emotions.

From a practical perspective, the significant effect sizes contrasting the interventions to the control group appear small at first sight. In the original metric ranging from 1 (= *fully disagree*) to 6 (= *fully agree*), the effect sizes correspond to mean item score reductions of 0.4 to 0.5. On the other hand, for NC and JSD, these small effects lowered the mean item score from the area of agreement (mean values above 3) to non-agreement (mean values below 3), which indicates practical relevance. Developers of seminar trainings could build on these findings and try to increase their trainings' effects on biasing cognitions and emotions by targeting the topics from the CECSA scales more directly, especially ER and JSD. For example, modules on emotional coping strategies might be valuable to reduce Emotional Reactivity, and fostering knowledge on juridical procedures and challenges, e.g., the presumption of innocence and the frequent lack of external evidence in CSA cases, might lower Justice System Distrust.

This study was conducted with a sample of teaching students. Future studies should assess whether the effect of a seminar trainings on reducing CECSA scores also holds true for working professionals and for other professions that (prospectively) conduct conversations with children (e.g., childcare professionals, police officers). Another important task is to assess the longevity of the training effects on CECSA scores.

To sum, the provision of a two-day seminar training successfully reduced participants' Naive Confidence and Emotional Reactivity, while Justice System Distrust was influenced only

in combination with a VR training. The role of JSD and adding a VR component to a seminar training overall yielded nuanced results that need to be reassessed with a larger sample size.

Epilogue

This thesis aimed at investigating individual differences in cognitive and emotional patters as sources of bias and suggestive questioning in child sexual abuse investigations. It is based on a series of empirical studies, including the development of self-report scales on Cognitions and Emotions about Child Sexual Abuse (CECSA), investigations of the relationship between the CECSA scales with bias and suggestive questioning, and a randomized controlled trial aiming to reduce CECSA scores through a two-day seminar training.

Article 1 (Gewehr et al., 2023) focused on the development of the CECSA scales and their validation as a tool to predict bias when judging CSA allegations. Data on a pool of 66 items was collected from 801 human sciences students, and the questionnaire's structure was derived using an Ant Colony Optimization (ACO) algorithm. The final CECSA structure demonstrated acceptable model fit and good internal consistencies for the three scales (1) Naive Confidence, which reflects an overestimation of both one's ability to recognize abuse and the reliability of children's reports, (2) Emotional Reactivity, which measures the intensity of emotional responses to the issue of CSA, and (3) Justice System Distrust, which assesses distrust in the justice system's competence to prosecute CSA cases. Each of the CECSA scales predicted biased evaluations toward the abuse hypothesis when participants were faced with scenarios of children displaying unspecific behavioral problems, such as mood issues or seemingly sexualized behavior. This article established the CECSA questionnaire as a reliable tool for identifying cognitive and emotional predispositions that can influence bias in CSA evaluations.

Article 2 (Gewehr, Merschhemke, et al., 2024) explored the relationship between the CECSA scales and the degree of suggestive questioning in conversations or interviews with children about CSA. In three studies using mock case paradigms, participants posed questions to children to clarify vague suspicions of CSA in three different formats: a single-choice format, a

free-writing format, and a natural speech format within a (VR) simulation. The questions varied or were coded regarding their suggestiveness. In addition, a biased mindset was measured by having participants rate how strongly the case material indicated CSA in two of the studies. Across the three studies and a meta-analytical integration, Naive Confidence and Emotional Reactivity predicted both suggestive questioning and bias robustly (i.e., in most of the formats and meta-analytically), while Justice System Distrust showed small significant effects in only few of the analyses. Overall, the studies of Article 2 found the CECSA scales NC and ER, but not JSD, to be useful measures of individual difference that predict bias and suggestive questioning of children.

Article 3 (Krause et al., 2024) consisted of a randomized controlled trial to evaluate 'ViContact', a training program to teach childcare professionals how to conduct conversations with children about CSA suspicions. Comparing intervention groups that received a classical seminar training, a practical virtual reality training, or both, with a control group, the study showed that the combined training of a classical seminar and practical VR training was most beneficial in improving appropriate questioning, the provision of rapport and socio-emotional support, and self-efficacy. The focus of the present thesis was the influence of the seminar training on the CECSA scales, which were also included in the data collection. Because for Article 3, only the NC scale was integrated into the analyses, a secondary analysis was conducted for this thesis to equally assess training effects for ER and JSD. Overall, the seminar training succeeded in reducing all three CECSA scale scores to a significantly larger extend than observed in the control group. This was achieved by the seminar training alone for NC and ER, while JSD was significantly reduced only in the combined training group, which included the seminar and the VR component. The expected advantage of the seminar over the VR training in

reducing CECSA scores was descriptively observable, but not found to be significant. Overall, a seminar training including knowledge transfer, exercises and discussions on how to handle CSA suspicions and how to talk to the children involved showed potential to significantly reduce Naive Confidence and Emotional Reactivity, while the role of Justice System Distrust and a VR training needs to be reassessed with larger samples.

Collectively, the findings underscore the importance and value of addressing differential suggestiveness and bias – that is, the variation in individuals' susceptibility to suggestive questioning and biased judgements. The results show that Naive Confidence and Emotional Reactivity, as reliably measured by the CECSA instrument, function as individual-level predictors of bias and suggestion, but that they can be somewhat reduced through a seminar training that transfers expert knowledge and lets participants engage in discussions and exercises. For the Justice System Distrust scale, results for both the suggestiveness and bias prediction and the training receptivity were inconclusive, warranting caution and further research before practical application.

Theoretical and Practical Contributions

Most of previous research on bias and suggestion in CSA or other forensic investigations has focused on general processes or on situational influences (Neal et al., 2022; O'Donohue & Cirlugea, 2021). This dissertation adds a differential perspective by showing that individual differences, such as the cognitive and emotional predispositions measured by the CECSA scales, may play a role in individuals' susceptibility to bias and suggestion. It thereby responses to recent calls from the field of legal psychology to investigate individual-level sources of bias (Neal et al., 2022; Talwar et al., 2024; Zapf & Dror, 2017) in order to explore debiasing strategies on diverse levels. By focusing not only on forensic interviews but also on the informal

conversations that childcare, health or education staff often have with children, the dissertation contributes to combating bias and suggestion at the various stages at which allegations and disclosures can occur (Korkman et al., 2024).

The development and validation of CECSA scales provides future research with the possibility to further inquire the role of NC, ER, and JSD in different areas of CSA investigations and explore how they can be mitigated in practice. By demonstrating that these characteristics are not fully static and are receptive to change through a seminar-style training intervention, this thesis also adds a differential perspective to the discussion on debiasing strategies and effective interviewer training: Next to the usually discussed behavioral strategies to mitigate bias and suggestiveness (e.g., considering alternative hypotheses; training question formulation), considering individual's unique predispositions may additionally mitigate bias at the levels of personal selection, allocation of interview tasks to suitable personal, and the identification of individual training needs.

Psychometric Properties of the CECSA Scales

The psychometric evidence from the five studies in this dissertation allows for a comprehensive evaluation of the CECSA scales regarding reliability, validity, and utility. Other studies that have already made use of the CECSA scales provide additional insights.

Reliability. In the CECSA development sample from Article 1, the final factor solution resulting from the ant colony optimization (ACO) procedure showed acceptable model fit, and all scales achieved good internal consistencies (α = .82–.88), which indicates good reliability in a large sample (N = 801). In the first study of Article 2, the CECSA scales showed similarly good internal consistencies (α = .79–.90), but this was a subsample from Article 1, thus does not provide novel psychometric information. Across the two further studies from Article 2, which

used independent samples (total $N = 356$), reliability results were more nuanced, with ER still proving good internal consistency ($\alpha = .83, .82$), and values for NC ($\alpha = .69, 69$) and JSD ($\alpha = .68, .76$) showing values at the border of what is contemporarily discussed as being acceptable for early stages of research ($\alpha \geq .7$; Lance et al., 2006; Nunnally & Bernstein, 1994). Article 3 reports data from a subsample of the Virtual Reality sample already reported in Article 2 but Article 3 includes a repeated measurement of the CECSA scales after the interventions. For this second assessment, reliability was slightly higher with acceptable to good values for all scales ($\alpha =.74–.91$).

Overall, the ER scale repeatedly showed good internal consistency, while NC and JSD values were good in the development sample and somewhat lower in the independent samples, although still at the border of acceptability for early or new areas of research. As these studies are the first independent applications of a newly developed measure, they can be considered as early research. Thus, all three scales may be applied for research purposes, but more through considerations are warranted for practical applications. In general, decisions about applying psychometric instruments in practice should not be made solely based on the (often arbitrary) cut-off scores for reliability but should take the importance of the practical decision that will follow from the assessment into account. The higher the stakes of the consequences of an assessment, the higher the reliability should be (Cho & Kim, 2015). For example, while the internal consistency values of NC and JSD may be sufficient to identify individual training needs of professionals and individualize training curricula according to their test scores, they may not be sufficient to base overall personal selection heavily on their scores. A related issue is the degree of discriminability that is required of an instrument. Instruments with lower reliability have a limited ability to discriminate at a fine-grained level but may well be able to tell

individuals with stronger score deviances apart. For example, the NC and JSD scales may be unable to differentiate interviewers with slightly diverging test results around the middle range of the distribution but may be able to identify individuals with more extreme test scores. For practical purposes, identifying extreme test scores is of relevance, for example, identifying individuals that are at high risk for bias and suggestive questioning to allocate training resources accordingly, or identifying individuals with a very low risk, to entrust them with interviewing a child. Ultimately, in practice, lower reliability translates to wider confidence intervals of individual test scores (i.e., higher uncertainty), which should always be taken into account when interpreting test results.

When aiming to increase the reliability of a scale, it needs to be considered that higher reliability often comes at the expense of criterion validity, especially when broader constructs are to be measured (attenuation paradox; Cho & Kim, 2015; Lance et al., 2006). In the case of the CECSA scales, the Naive Confidence scale covers a broad construct with including both naive confidence in one's own innate ability to recognize abused children, and the naive confidence in the accuracy of children's abuse reports. These sub-constructs appear to be strongly related, as the underlying items were better represented by a common factor than by two separate factors in the ACO analysis in Article 1. Similarly, Justice System Distrust covers a broad construct with including mistrust in both the competence and the willingness of different representatives of the judicial system (e.g., police, judges). Thus, the modest reliability of NC and JSD may be the price for the scales' ability to depict broad constructs. To increase the reliability of these scale, while retaining their validity, future studies could increase the number of items but should aim to maintain their diversity. For example, for the NC scale, developing further items that describe

the overall meaning of the NC construct and further items that summarize each of the two subconstructs might be fruitful to increase reliability.

Validity. In Article 1, convergent validity of the CECSA scales was demonstrated by expected variance overlaps with most of the theoretically selected self-report instruments. The most comprehensive convergent validation was conducted for the Emotional Reactivity scale, which overlapped with negative emotionality, empathy, sensitivity to injustice done to others, and anger about sexual assault, which overall underscores a general emotional-empathic component of the scale, as well as specific emotionality for sexual delinquency. The Naive Confidence scale was positively related to a preference for intuitive decisions, but not, as expected, negatively to deliberate decision-making, which might indicate that people who generally prefer deliberate decisions don't necessarily do so in the context of CSA investigations. The Justice System Distrust scale was associated to punitive attitudes towards sexual offenders, but not with a general belief in a just world or justice sensitivity, suggesting that a distrust in the justice system to handle sex crimes may be independent of more general justice attitudes. Overall, results of the convergent validation indicates that Emotional Reactivity regarding CSA allegations might be related to general emotionality, while the constructs of Naive Confidence and Justice System Distrust might be more specific to the issue of CSA and not generalizable to other contexts.

Turning to predictive validity, a major goal of this thesis was to develop scales that predict biased mindsets and suggestiveness when handling CSA allegations. In Article 1, all three CECSA scales were interpreted as successfully predicting bias, as they were associated with rating children's unspecific behavioral issues as indicative of sexual abuse (NC and ER) or with perceiving a suspect's acquittal in such cases as incorrect (NC and JSD). By leveraging the

flexibility of Ant Colony Optimization (ACO), the prediction of bias was already included as a criterion for the item selection. Thus, revisiting these predictions in independent samples was of importance. Across the three studies of Article 2, the prediction of a biased mindset was again robustly found for the NC and ER scales, but not for JSD (bias was again operationalized by CSA indicativity ratings, but the acquittal rating task was not used again). Similarly, suggestiveness was robustly predicted by NC and ER, but not JSD across the three studies. Across all four studies of Articles 1 and 2, the effect-sizes for NC and ER were in the medium range for predicting both bias (b = .13–.37) and suggestive questioning (b = .15–.26), which corresponds to a potentially large practical impact, when considering how the effects of bias and suggestion can accumulate across repeated situations.

Re-evaluating the findings on JSD in Article 1 with the insights from the nonsignificant results in Article 2, the conclusion that JSD predicted bias in Article 1 seems to be challenged. This conclusion was based on the association between JSD and the acquittal rating task, while no association was found with CSA likelihood ratings. An alternative explanation for the predicted acquittal rating is that individuals who are more distrustful of the justice system are also more sceptical of any court decision, whether it is an acquittal or a conviction. In the light of Article 2, this alternative explanation seems more plausible.

Utility. The CECSA scales are of utility for both scientific and practical purposes. All three scales provide reliable measurement, and the Naive Confidence and Emotional Reliability scales also serve as a tool for measuring predisposition to bias and suggestive questioning. Legal psychology researchers may utilize the scales to assess the role of cognitive and emotional individual differences in, for example, police work, jury decisions, expert witness evaluations, interview performance, or the handling of CSA allegations in more informal settings. The scales

may also be used to evaluate interview training programs or more broad educational programs on handling CSA allegations in formal or informal settings. As shown in Article 3 of this thesis, the ER and NC scales – but not necessarily JSD – are receptive to deliberate change and can be reduced through a seminar training that transfers expert knowledge.

A number of research teams have used the CECSA scales since their preprint publication, which provides insight into their actual utility for the field and helps to further evaluate their psychometric properties: Segal et al. (2023; 2022) used the Emotional Reactivity scale to investigate the role of students' ($N = 30$, $N = 60$) emotions when interviewing child avatars. They replicated the good internal consistency of the ER scale ($\alpha = .8$) and provided a Lithuanian translation (Segal et al., 2022). Providing further construct validation, they reported that the ER score related to participants' emotional reactions when interviewing avatars with a history of CSA (anger, disgust) or no CSA (relief), as measured by self-report and facial expressions.

Imhoff (2024) reported insights from a student sample ($N = 475$), in which the CECSA scales exploratorily correlated with conspiracy mentality ($r = .1–.3$; Bruder et al., 2013), with the beliefs in absolute evil ($r = .19–.38$; Campbell & Vollhardt, 2014), absolute good (only NC and ER; $r \approx .22$), and organized ritual sexual abuse ($r = .28–.41$), the latter of which has been labelled as highly unlikely from scientific and juridical perspectives (Mokros et al., 2024) and is classified as a conspiracy theory by Imhoff (2024). These results provide further construct validation for the CECSA scales as they show that the CECSA scores associate not only to biased judgments of CSA allegations, but also to other indicators of biased mindsets such as conspiracy theories and black-and-white thinking.

Some studies applied or plan to apply the CECSA scales to evaluate interview training programs: The effect of reducing Naive Confidence scores through the ViContact seminar

training reported in Article 3 was replicated with a sample of psychology students (N = 28; Gewehr, Tamm, et al., 2024). A variant of the combined ViContact training using e-learning instead of a seminar reduced NC scores among a small sample of child protection workers (N = 15; Buchwald, in prep.; Gewehr, Tamm, et al., 2024). A swizz research team plans to use the CECSA scales to assess differential training effects of a training system in which police officers practice interviewing with a virtual child that is role-played by a large language model (Virtual Kids; Tuggener et al., 2024; T. Schneider, personal communication, September 27, 2024).

Bayer et al. (2024) used the CECSA scales to evaluate differential training needs of school professionals (teachers, social workers, and headmasters; $N = 276$), aiming to inform the development of a modular training system to handle CSA allegations. They found increased ER scores for teachers and unexperienced professionals, and now plan to develop an optional stress regulation module for these professionals.

Overall, the psychometric evidence from the five studies of this dissertation, complemented by the additional research, supports the CECSA scales Naive Confidence and Emotional Reactivity as reliable and valid scales to assess individual differences that predispose bias and suggestion when dealing with CSA allegations. The applications of the CECSA scales since their publication underscore their utility for diverse research purposes.

Measuring Suggestive Questioning

Article 2 presents newly developed mock cases and three variants to measure a suggestive questioning style in mock conversations: paper-pencil single choice selection of questions, paper-pencil free-writing of questions, and verbal posing of question in a dynamic virtual reality simulation. The results of the studies provide insights into advantages and disadvantages of the different approaches for future studies. First, the non-diagnostic mock cases

developed for Studies 1 and 2 effectively elicited a moderate level of suspicion about sexual abuse, indicating appropriateness for further studies on bias and suggestion. Second, the suggestiveness of the questions selected or posed in Studies 1 and 2 showed good fit for a unidimensional model, indicating appropriateness to be summarized into a common measure for a suggestive questioning style. Third, participants formulated few suggestive questions across all approaches, which is in line with former research (Cyr et al., 2021; Kask et al., 2022; Pompedda et al., 2015; Sternberg et al., 2001) but poses a limit to reliability. Future studies must select or further develop a measure that provokes an increased number of suggestive questions in order to reliably capture a suggestive questioning style. The free writing approach from Study 2 elicited the highest percentage of suggestive questions (32% vs. 17% and 10%), but because Study 2 also used a different sample (police students vs. human sciences and teaching students), the differences cannot be attributed with certainty to the different measures. Fourth, the three approaches vary in their level of ecological validity, which must be weighed together with the reliability of the measures. Fifth, for each of the approaches, increasing the number of questions (i.e., suggestiveness indicators) and fine-grading the suggestiveness scales might further increase reliability.

Implications for Interview Training, Legal Processes, and the Children Involved

For practical purposes, the main message from this dissertation is that children who are questioned about abuse suspicions are exposed to an increased risk of suggestive pressure and biased evaluations if the adults who question them a) tend to react strongly emotional to the topic of CSA, or b) believe that they can innately or intuitively tell abused from non-abused children apart and that children's abuse reports are purely accurate. Adults' degree of mistrust in the

justice system regarding the prosecution of CSA allegations is however not associated to an increased risk for bias or suggestion.

Because children are susceptible to suggestive pressure, biased and suggestive conversations or interviews – at any level of the investigative process and especially if accumulated – can impair the accuracy of children's reports and memories, harm the children's long-term well-being, and threaten the just prosecution of CSA allegations (Baldwin et al., 2024b; Brown & Lamb, 2015; Ceci et al., 2016; Howe & Knott, 2015; Scoboria et al., 2017). Decision-makers who wish to minimize the risk for suggestive and biased child interviews or conversations may take the individual differences measured by the CECSA scales NC and ER into account when wanting to select suitable employees to conduct interviews or conversations with children, to take on roles of trust (e.g., liaison teacher), or to receive specialized interview training.

Article 3 and the accompanying secondary analysis in this dissertation showed that the CECSA scales were responsive to intervention efforts. Specifically, a two-day seminar-style training significantly reduced student's NC and ER scores. A similar effect was observed in small sample of child-protection workers who underwent e-learning and VR-based training (Gewehr, Tamm, et al., 2024). Existing training programs for improving the questioning of children typically focus on practicing open-ended, non-suggestive questions and the suppression of suggestive utterances. Expanding these curricula to also address underlying cognitive and emotional factors could help to further reduce bias and suggestiveness. For instance, integrating expert knowledge about biases in human judgement, children's disclosure patterns, and the absence of reliable behavioral indicators of abuse into a training program may help reduce Naive Confidence. Adding modules on emotional regulation strategies when facing distressing CSA

allegations could lower Emotional Reactivity – though ViContact's training achieved this reduction without explicitly addressing emotional coping. Lower scores on NC and ER may, in turn, reduce bias and suggestive questioning, as indicated by the correlational findings in Article 2. However, the causal relationship between these factors remains to be demonstrated. Note that the findings from this dissertation do not support the notion that increasing trust in the justice system through training efforts can reduce bias or suggestiveness.

To distribute training resources more efficiently, modular training curricula could be customized according to participants' individual training needs. For example, extended modules aiming to reduce Naive Confidence and Emotional Reactivity could be offered only for participants with high values on these scales. Individual case supervision, as sometimes conducted in forensic or child-protection settings, might equally benefit from taking participants NC or ER scores into account when discussing individual sources of bias and suggestive questioning or individual counterstrategies.

Limitations and Future Research

Several limitations need to be considered when interpreting and drawing conclusions from this work.

Conceptual Limitations

The CECSA instrument is not a comprehensive measure of differential suggestiveness and bias in all their potential facets. Rather, it is an approach of beginning to identify individual differences that predispose bias and suggestiveness and making them psychometrically accessible. The scales were developed based on prior research on attitudes and beliefs towards CSA (Finnilä-Tuohimaa et al., 2008), and expanded by the emotional component based on theoretical considerations. Future research may further shape and extended the scales.

The biased mindset discussed in this thesis, which reflects a tendency for excessive sensitivity and overcalling of abuse (risking false positives), is only one of two potential biases, and the flip side of over-focusing on specifity and underestimating CSA (risking false negatives), is equally concerning. If adults tailor their questioning of children according to an a priori assumption that sexual abuse has *not* occurred, children who have actually experienced abuse may not feel supported enough to disclose their experiences (Cromer & Goldsmith, 2010). Here, too, open mindedness and open-ended questioning are crucial to allow children to report from their autobiographical experiences instead of leading them to follow adult's presumptions.

The CECSA scales are not intended to be interpreted normatively or as a reflection of incorrect knowledge. Some items indeed contradict empirical evidence (e.g., "suggestive interview techniques only affect children's memories of trivial details"; Scoboria et al., 2017), but others lack sufficient research for a valid assessment (e.g., "courts do not take children seriously enough in cases of child sexual abuse"), can be both correct and incorrect, depending, for example, on regional differences (e.g., perceptions of judicial fairness; Cross et al., 2003; Ernberg et al., 2018), or are inherently non-normative, such as the Emotional Reactivity scale or items describing judgement or information-processing styles (e.g., "I would trust my first impression when assessing whether a child was sexually abused or not").

The development of the Emotional Reactivity scale was based on the notions that the emotional valence of a situation (positive vs. negative) directly informs judgment and decisions (Feeling-as-information theory; Schwarz, 2012) and that emotionality associates to confirmation bias (Jonas et al., 2006). However, other theories (e.g., cognitive-appraisal theories and the model of emotion-imbued choice) argue that discrete emotions within the same valence category (e.g., anger and sadness) can have different effects on judgment and choice (Lerner et

al., 2015, Lench et al., 2011), which has been backed by meta-analytical evidence (Angie et al., 2011). Research on the role of emotions in suspect interviewing has followed this approach, but so far yielded mixed results. For example, Sambrano (2020) found that sad compared to angry participants preferred benevolent interrogations tactics; but did not find differences for hostile tactics. Ask and Granhag (2007) reported more heuristic information-processing for angry compared to sad police officers, but a meta-analysis on general depth of information processing (McKasy, 2020) found no effect of anger vs. sadness. Salerno (2021) concludes in a review that anger and disgust lead to greater confidence in own opinions which may increase heuristic information processing. Albeit the somewhat inconclusive results, these findings raise the question of whether the CECSA Emotional Reactivity scale, which compiles diverse emotions based on their negative valence (i.e., anger, sadness, disgust), represents an oversimplification of distinct, potentially contradicting, emotional effects on bias and suggestion. This is in line with Segal et al. (2023; 2023), who found partially distinct effects of anger and sadness in avatar CSA interviews (e.g., anger was associated with closed vs. open questions, but to less belief-consistent reinterpretation of children's reports), and who also argue for differentiating emotions beyond valence. Although the ER scale showed robust effects on bias and suggestion in the studies of this dissertation, future studies could investigate whether distinguishing distinct emotions can further harness their potential to predict bias and suggestion. The results of the studies described above suggest that anger may have a particular role to play in predicting bias.

Methodological Limitations

The CECSA scales do not include reversed-scores items. This makes it more difficult to detect or prevent response biases (e.g., agreement bias, extreme responding, careless responding) and can inflate internal consistency. It can also inflate associations with other self-report

instruments that lack reversed-scored items, which was not the case for most of the self-report instruments used to validate the CECSA scales. Future research aiming to improve the psychometric properties or extend the CECSA scales should consider adding reverse-scored items to each of the scales. Another important next step is to assess the long-term stability of the CECSA scores, which is a prerequisite for assuming stable individual differences. Promisingly, the preliminary self-report instrument that was the base for developing the cognitive CECSA scales showed good retest reliability (r = .82 to .91) among a small sample of 26 students over a three-week period (Finnilä-Tuohimaa et al., 2008). Finally, the CECSA scales currently lack standardized norm values, which are necessary to interpret an individual's score in comparison to a reference sample.

The suggestiveness measures introduced in Article 2 were limited in reliability, and, as discussed in the section on psychometric properties, this can be improved by increasing the number of indicators. Just as the CECSA scales, the long-term stability of the measures on suggestiveness should be assessed by future studies. Assessing the ecological validity of the measures through associations with suggestiveness in real interviews would be another valuable contribution.

Regarding the prediction of suggestion and bias through the CECSA scales, it needs to be stressed that these are correlational, not causal findings. It can thus not be concluded with certainty that changes in CECSA scores, for example trough a training intervention, go along with changes in suggestive questioning or biased judgements. Randomized controlled studies could inquire this issue. As a first approximative step, parallel changes of both variables could be assessed using the data from the ViContact evaluation study: Positive correlations between pre-post changes in the CECSA scores with pre-post changes in suggestive questioning of

participants who received the seminar training would – albeit not implying causality - point towards a parallel effect on both variables into the same direction. A further step could be the assessment of differential training effects, i.e., investigating whether participants profit differently from the training depending of their CECSA scores.

The three studies in Article 2 report three different measures to assess suggestive questioning, but because the samples also varied between the three studies, differences between the results cannot with certainty be attributed to either of these factors. In particular, the free-writing paradigm used in Study 2 was advised as the most promising tool to reliably measure suggestive questioning, because it yielded the largest number of suggestive questions. However, Study 2 was also the only study using a police student sample, which can equally be the reason for increased suggestive questions. Due to their job description, police students might, compared to human sciences or teaching students, feel a higher need to solve a case and obtain information from a child and thus be at higher risk for suggestive questioning.

Regarding the effects of training interventions on the CECSA scores reported in the additional analysis to Article 3, the small sample sizes of the intervention groups (n = 26–28) might have caused the partially mixed pattern of results, especially regarding the Justice System Distrust Scale and the effect of adding the VR training to the seminar intervention. Larger sample sizes would additionally allow to detect smaller effects.

Practical Limitations

The studies of this thesis have been conducted with human sciences, teaching, and police students, which are relevant samples as those students will make up the professionals that question children in different contexts in the future. However, generalizability of the findings to working practitioners still needs to be established. The same accounts for the longevity of the

training effects. Promisingly, a first assessment of the ViContact training with a small sample of child protection workers (Gewehr, Tamm, et al., 2024) replicated the training effect on CECSA scores found in the student sample of Article 3. Similarly, Lahtinen et al. (2017), reported that training of investigative interviewers reduced scores on the attitudinal measure preceding the cognitive CECSA scales, which sustained at a one-year follow-up.

General Directions for Future Research on Differential Bias and Suggestiveness

Future research might aim to identify further individual differences that are not covered from the CECSA scales but also associate to bias and suggestion in CSA investigations. Promising candidates may be cognitive thinking styles or preferences, such as cognitive flexibility (Martin & Anderson, 1998), the need for cognitive closure (Webster & Kruglanski, 1994), reflexive vs. reflective thinking styles (Martire et al., 2020), ambiguity tolerance (Furnham & Marks, 2013), motivated reasoning (Kahan, 2013), or apophenia (i.e., the disposition to false positives; Blain et al., 2020). Based on the two-step process model of cognitive biases proposed by Oeberst & Imhoff (2023), one might also distinguish the search for individual differences into those that a) foster false beliefs or b) associate to belief-consistent information processing. Alternatively, a bottom-up approach of using a large sample to exploratorily correlate a wide range of individual difference measures with suggestiveness and bias might be fruitful to derive hypotheses for concrete associations.

Conclusion

This thesis explored how individual differences in cognitive and emotional patterns contribute to bias and suggestive questioning in child sexual abuse investigations. In a series of five empirical studies, summarized in three articles, the scales on Cognitions and Emotions about

Child Sexual Abuse (CECSA) scales were developed, and their relationship to biased judgments and suggestive questioning as well as their responsiveness to training were tested.

The first Article focuses on the development and initial validation of the three CECSA scales Naive Confidence, Emotional Reactivity, and Justice System Distrust, which showed acceptable to good reliability and predicted biased mindsets toward the abuse hypothesis in evaluations of CSA suspicions. The second article reports how Naive Confidence and Emotional Reactivity, but not Justice System Distrust, robustly predicted biased judgements and a suggestive questioning style across three studies and a meta-analytical integration. The third study evaluated the effectiveness of a training program consisting of a seminar and a Virtual Reality (VR) component designed to improve professionals' questioning skills. A secondary analysis showed that the seminar component alone significantly reduced Naive Confidence and Emotional Reactivity scores, while Justice System Distrust was reduced only in combination with the VR training.

In summary, this thesis presents the CECSA scales Naive Confidence and Emotional Reactivity as reliable measures for predicting differential bias and suggestiveness and shows that both scales are receptive to change through training efforts. The scales can be of value across a variety of scientific and practical contexts, such as the development and customization of interview training curricula, the evaluation of training programs, or the selection of adequate personal to interview children. While a differential perspective has been largely absent from research on interviewer bias and suggestion so far, the CECSA scales can be used and extended to further explore individual differences in child sexual abuse investigations.

Ultimately, the work of this dissertation ought to contribute to an improved practice of questioning children in a supportive but unbiased and non-suggestive manner, which not only

improves the accuracy and fairness of legal proceedings but also fosters the welfare of the children involved.

References

(for Prologue, Secondary Analysis to Article 3, and Epilogue)

Akca, D., & Eastwood, J. (2021). The impact of individual differences on investigative interviewing performance: A test of the police interviewing competencies inventory and the five factor model. *Police Practice and Research*, *22*(1), 1027–1045. https://doi.org/10.1080/15614263.2019.1644177

Akca, D., Larivière, C. D., & Eastwood, J. (2021). Assessing the efficacy of investigative interviewing training courses: A systematic review. *International Journal of Police Science & Management*, *23*(1), 73–84. https://doi.org/10.1177/14613557211008470

Andrews, S. J., & Lamb, M. E. (2021). Lawyers' Question Repetition and Children's Responses in Scottish Criminal Courts. *Journal of Interpersonal Violence*, *36*(1–2), 276–296. https://doi.org/10.1177/0886260517725739

Ask, K., & Granhag, P. A. (2007). Hot cognition in investigative judgments: The differential influence of anger and sadness. *Law and Human Behavior*, *31*(6), 537–551. https://doi.org/10.1007/s10979-006-9075-3

Baginsky, M. (2003). Newly qualified teachers and child protection: A survey of their views, training and experiences. *Child Abuse Review*, *12*(2), 119–127. https://doi.org/10.1002/car.783

Baldwin, J. R., Coleman, O., Francis, E. R., & Danese, A. (2024a). Prospective and Retrospective Measures of Child Maltreatment and Their Association With Psychopathology: A Systematic Review and Meta-Analysis. *JAMA Psychiatry*, *81*(8), 769–781. https://doi.org/10.1001/jamapsychiatry.2024.0818

Baldwin, J. R., Coleman, O., Francis, E. R., & Danese, A. (2024b). Prospective and Retrospective Measures of Child Maltreatment and Their Association With Psychopathology: A Systematic Review and Meta-Analysis. *JAMA Psychiatry*. https://doi.org/10.1001/jamapsychiatry.2024.0818

Bayer, L., Maike, C., Eifgang, J., Mensing, F., & Pülschen, S. (2024). *Cognitions and emotions regarding child sexual abuse Implications for designing training and interview training for school staff*. Conference of the European Association of Psychology and Law (EAPL), Caporica, Portugal.

Blain, S. D., Longenecker, J. M., Grazioplene, R. G., Klimes-Dougan, B., & DeYoung, C. G. (2020). Apophenia as the disposition to false positives: A unifying framework for openness and psychoticism. *Journal of Abnormal Psychology*, *129*(3), 279–292. https://doi.org/10.1037/abn0000504

Brewin, C. R., & Andrews, B. (2017). Creating Memories for False Autobiographical Events in Childhood: A Systematic Review: Creating false childhood memories. *Applied Cognitive Psychology*, *31*(1), 2–23. https://doi.org/10.1002/acp.3220

Brown, D. A., & Lamb, M. E. (2015). Can Children Be Useful Witnesses? It Depends How They Are Questioned. *Child Development Perspectives*, *9*(4), 250–255. https://doi.org/10.1111/cdep.12142

Brubacher, S. P., Powell, M. B., Snow, P. C., Skouteris, H., & Manger, B. (2016). Guidelines for teachers to elicit detailed and accurate narrative accounts from children. *Children and Youth Services Review*, *63*, 83–92. https://doi.org/10.1016/j.childyouth.2016.02.018

Brubacher, S. P., Powell, M., Skouteris, H., & Guadagno, B. (2014). An Investigation of the Question-Types Teachers Use to Elicit Information From Children. *The Australian*

Educational and Developmental Psychologist, *31*(2), 125–140. https://doi.org/10.1017/edp.2014.5

Brubacher, S. P., Shulman, E. P., Bearman, M. J., & Powell, M. B. (2022). Teaching child investigative interviewing skills: Long-term retention requires cumulative training. *Psychology, Public Policy, and Law*, *28*(1), 123–136. https://doi.org/10.1037/law0000332

Bruck, M., & Ceci, S. J. (1997). The suggestibility of young children. *Current Directions in Psychological Science*, *6*(3), 75–79.

Bruder, M., Haffke, P., Neave, N., Nouripanah, N., & Imhoff, R. (2013). Measuring Individual Differences in Generic Beliefs in Conspiracy Theories Across Cultures: Conspiracy Mentality Questionnaire. *Frontiers in Psychology*, *4*. https://doi.org/10.3389/fpsyg.2013.00225

Buchwald, M. (in prep.). *ViContact 2.0 in der Praxis. Erprobung eines interaktiven Trainingssystems für Gespräche mit Kindern bei Missbrauchsverdacht im Kinderschutz* [Unpublished master's thesis].

Campbell, M., & Vollhardt, J. R. (2014). Fighting the Good Fight: The Relationship Between Belief in Evil and Support for Violent Policies. *Personality and Social Psychology Bulletin*, *40*(1), 16–33. https://doi.org/10.1177/0146167213500997

Ceci, S. J., & Bruck, M. (1993). Suggestibility of the Child Witness: A Historical Review and Synthesis. *Psychological Bulletin*, *113*(3), 403–439.

Ceci, S. J., & Bruck, M. (1995). *Jeopardy in the courtroom: A scientific analysis of children's testimony*. American Psychological Association.

Ceci, S. J., & Friedman. (2000). The suggestibility of children: Scientific research and legal implications. *Cornell Law Review*, *86*(1), 34–108.

Ceci, S. J., Hritz, A., & Royer, C. (2016). Understanding Suggestibility. In W. O'Donohue & M. Fanetti (Eds.), *Forensic Interviews Regarding Child Sexual Abuse: A Guide to Evidence-Based Practice.* (pp. 141–153). Springer.

Cho, E., & Kim, S. (2015). Cronbach's Coefficient Alpha: Well Known but Poorly Understood. *Organizational Research Methods*, *18*(2), 207–230. https://doi.org/10.1177/1094428114555994

Cirlugea, O., & O'Donohue, W. T. (2016). Review of psychometrics of forensic interview protocols. In W. T. O'Donohue & M. Fanetti (Eds.), *Forensic interviews regarding child sexual abuse: A guide to evidence-based practice* (pp. 237–255). Springer.

Cromer, L. D., & Goldsmith, R. E. (2010). Child Sexual Abuse Myths: Attitudes, Beliefs, and Individual Differences. *Journal of Child Sexual Abuse*, *19*(6), 618–647. https://doi.org/10.1080/10538712.2010.522493

Cross, T. P., Walsh, W. A., Simone, M., & Jones, L. M. (2003). Prosecution of Child Abuse: A Meta-Analysis of Rates of Criminal Justice Decisions. *Trauma, Violence, & Abuse*, *4*(4), 323–340. https://doi.org/10.1177/1524838003256561

Cyr, M., Dion, J., Gendron, A., Powell, M., & Brubacher, S. (2021). A test of three refresher modalities on child forensic interviewers' posttraining performance. *Psychology, Public Policy, and Law*. https://doi.org/10.1037/law0000300

Duke, M., Elisabeth, R., & Price, H. (2016). Avoiding problems in child abuse interviews and investigations, in Forensic Interviews Regarding Child Sexual Abuse. In W. O'Donohue

& M. Fanetti (Eds.), *Forensic Interviews Regarding Child Sexual Abuse: A Guide to Evidence-Based Practice.* (pp. 179–195). Springer.

Ernberg, E., Magnusson, M., & Landström, S. (2018). Prosecution of Child Sexual Abuse Cases Involving Preschool-Aged Children: A Study of Swedish Cases from 2010 to 2014. *Journal of Child Sexual Abuse*, *27*(7), 832–851. https://doi.org/10.1080/10538712.2018.1501786

Everson, M. D., & Sandoval, J. M. (2011). Forensic child sexual abuse evaluations: Assessing subjectivity and bias in professional judgements. *Child Abuse & Neglect*, *35*(4), 287–298. https://doi.org/10.1016/j.chiabu.2011.01.001

Fessinger, M. B., & McAuliff, B. D. (2020). A national survey of child forensic interviewers: Implications for research, practice, and law. *Law and Human Behavior*, *44*(2), 113–127. https://doi.org/10.1037/lhb0000368

Finnilä-Tuohimaa, K., Santtila, P., Björnberg, L., Hakala, N., Niemi, P., & Sandnabba, K. (2008). Attitudes related to child sexual abuse: Scale construction and explorative study among psychologists. *Scandinavian Journal of Psychology*, *49*(4), 311–323. https://doi.org/10.1111/j.1467-9450.2008.00635.x

Finnilä-Tuohimaa, K., Santtila, P., Sainio, M., Niemi, P., & Sandnabba, K. (2009). Expert judgment in cases of alleged child sexual abuse: Clinicians' sensitivity to suggestive influences, pre-existing beliefs and base rate estimates. *Scandinavian Journal of Psychology*, *50*(2), 129–142. https://doi.org/10.1111/j.1467-9450.2008.00687.x

Forscher, P. S., Lai, C. K., Axt, J. R., Ebersole, C. R., Herman, M., Devine, P. G., & Nosek, B. A. (2019). A meta-analysis of procedures to change implicit measures. *Journal of*

Personality and Social Psychology, *117*(3), 522–559. https://doi.org/10.1037/pspa0000160

Furnham, A., & Marks, J. (2013). Tolerance of Ambiguity: A Review of the Recent Literature. *Psychology*, *04*(09), 717–728. https://doi.org/10.4236/psych.2013.49102

Gewehr, E., Merschhemke, M., Pülschen, S., Pülschen, D., & Volbert, R. (2024). *Predicting Suggestive Questioning from Cognitions and Emotions about Child Sexual Abuse across Three Study Paradigms*. https://doi.org/10.31234/osf.io/qyfd9

Gewehr, E., Tamm, A., Buchwald, M., Schneider, J., Barbe, H., Fromberger, P., Krause, N., Mensing, F., Mershhemke, M., Müller, J., Pülschen, S., Siegel, B., & Volbert, R. (2024). *ViContact 2.0 Kinderschutz – Ein interaktives Trainingssystem für Gespräche mit Kindern bei Missbrauchsverdacht.* 53. Tagung der Deutschen Gesellschaft für Psychologie (DGPs) / 15. Tagung der Österreichischen Gesellschaft für Psychologie (ÖGP), Wien, Österreich.

Gewehr, E., Volbert, R., Merschhemke, M., Santtila, P. O., & Pülschen, S. (2023). *Cognitions and Emotions about Child Sexual Abuse (CECSA): Development of a Self-Report Measure to Predict Interviewer Bias* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/qcfvb

Goldman, J. D. G. (2007). Primary school student-teachers' knowledge and understandings of child sexual abuse and its mandatory reporting. *International Journal of Educational Research*, *46*(6), 368–381. https://doi.org/10.1016/j.ijer.2007.09.002

Greytak, E. A. (2009). *Are Teachers Prepared? Predictors of Teachers' Readiness to Serve as Mandated Reporters of Child Abuse* [Doctoral dissertation].

Gumpert, C. H., & Lindblad, F. (2000). Expert Testimony on Child Sexual Abuse:a Qualitative Study of the Swedish Approach toStatement Analysis. *Expert Evidence*, *7*(4), 279–314.

Howe, M. L., & Knott, L. M. (2015). The fallibility of memory in judicial processes: Lessons from the past and their modern consequences. *Memory*, *23*(5), 633–656. https://doi.org/10.1080/09658211.2015.1010709

Huang, C.-Y., & Bull, R. (2021). Applying Hierarchy of Expert Performance (HEP) to investigative interview evaluation: Strengths, challenges and future directions. *Psychiatry, Psychology and Law*, *28*(2), 255–273. https://doi.org/10.1080/13218719.2020.1770634

Imhoff, R. (2024). *Ist der Glaube an ritualisierten Kindesmissbrauch eine Verschwörungstheorie und was hieße das?* 53. Tagung der Deutschen Gesellschaft für Psychologie (DGPs) / 15. Tagung der Österreichischen Gesellschaft für Psychologie (ÖGP), Wien, Austria.

Johnson, J., McWillians, K., Goodman, G. S., Shellley, A., & Piper, B. (2016). Basic Principles of Interviewing the Child Eyewitness. In W. O'Donohue & M. Fanetti (Eds.), *Forensic Interviews Regarding Child Sexual Abuse* (pp. 179–195). Springer.

Johnson, M., Magnussen, S., Thoresen, C., Lønnum, K., Burrell, L. V., & Melinder, A. (2015). Best Practice Recommendations Still Fail to Result in Action: A National 10-Year Follow-up Study of Investigative Interviews in CSA Cases: Follow-up study of investigative interviews. *Applied Cognitive Psychology*, *29*(5), 661–668. https://doi.org/10.1002/acp.3147

Jonas, E., Graupmann, V., & Frey, D. (2006). The Influence of Mood on the Search for Supporting Versus Conflicting Information: Dissonance Reduction as a Means of Mood

Regulation? *Personality and Social Psychology Bulletin*, *32*(1), 3–15. https://doi.org/10.1177/0146167205276118

Kahan, D. M. (2013). Ideology, motivated reasoning, and cognitive reflection. *Judgment and Decision Making*, *8*(4), 407–424. https://doi.org/10.1017/S1930297500005271

Kask, K., Pompedda, F., Palu, A., Schiff, K., Mägi, M.-L., & Santtila, P. (2022). Transfer of Avatar Training Effects to Investigative Field Interviews of Children Conducted by Police Officers. *Frontiers in Psychology*, *13*. https://www.frontiersin.org/articles/10.3389/fpsyg.2022.753111

Kassambara, A. (2021). *rstatix: Pipe-Friendly Framework for Basic Statistical Tests* (Version 0.7.0) [Computer software]. https://CRAN.R-project.org/package=rstatix

Kendall-Tucket, K., Williams, L. M., & Finkelhor, D. (1993). Impact of sexual abuse on children: A review and synthesis of recent empirical studies. *Psychological Bulletin*, *113*(1), 164–180. https://doi.org/10.1037/0033-2909.113.1.164

Korkman, J., Juusola, A., & Santtila, P. (2014). Who made the disclosure? Recorded discussions between children and caretakers suspecting child abuse. *Psychology, Crime & Law*, *20*(10), 994–1004. https://doi.org/10.1080/1068316X.2014.902455

Korkman, J., Otgaar, H., Geven, L. M., Bull, R., Cyr, M., Hershkowitz, I., Mäkelä, J.-M., Mattison, M., Milne, R., Santtila, P., van Koppen, P., Memon, A., Danby, M., Filipovic, L., Garcia, F. J., Gewehr, E., Gomes Bell, O., Järvilehto, L., Kask, K., … Volbert, R. (2024). White paper on forensic child interviewing: Research-based recommendations by the European Association of Psychology and Law. *Psychology, Crime & Law*, *0*(0), 1–44. https://doi.org/10.1080/1068316X.2024.2324098

Krause, N., Gewehr, E., Barbe, H., Merschhemke, M., Mensing, F., Siegel, B., Müller, J. L., Volbert, R., Fromberger, P., Tamm, A., & Pülschen, S. (2024). How to prepare for conversations with children about suspicions of sexual abuse? Evaluation of an interactive virtual reality training for student teachers. *Child Abuse & Neglect*, *149*, 106677. https://doi.org/10.1016/j.chiabu.2024.106677

La Rooy, D., Brubacher, S. P., Aromäki-Stratos, A., Cyr, M., Hershkowitz, I., Korkman, J., Myklebust, T., Naka, M., Peixoto, C. E., Roberts, K. P., Stewart, H., & Lamb, M. E. (2015). The NICHD protocol: A review of an internationally-used evidence-based tool for training child forensic interviewers. *Journal of Criminological Research, Policy and Practice*, *1*(2), 76–89. https://doi.org/10.1108/JCRPP-01-2015-0001

Lahtinen, H.-M., Korkman, J., Laitila, A., & Mehtätalo, L. (2017). The effect of training on investigative interviewers' attitudes and beliefs related to child sexual abuse. *Investigative Interviewing: Research and Practice*, *8*(1), 16–30.

Lamb, M. E., La Rooy, D., Malloy, L., & Katz, C. (Eds.). (2011). *Children's testimony: A handbook of psychological research and forensic practice* (2. ed.). Wiley-Blackwell.

Lamb, M. E., Orbach, Y., Hershkowitz, I., Esplin, P. W., & Horowitz, D. (2007). Structured forensic interview protocols improve the quality and informativeness of investigative interviews with children: A review of research using the NICHD Investigative Interview Protocol. *Child Abuse & Neglect*, *31*(11–12), 1201–1231. https://doi.org/10.1016/j.chiabu.2007.03.021

Lance, C. E., Butts, M. M., & Michels, L. C. (2006). The Sources of Four Commonly Reported Cutoff Criteria: What Did They Really Say? *Organizational Research Methods*, *9*(2), 202–220. https://doi.org/10.1177/1094428105284919

Lavoie, J., Wyman, J., Crossman, A. M., & Talwar, V. (2021). Meta-analysis of the effects of two interviewing practices on children's disclosures of sensitive information: Rapport practices and question type. *Child Abuse & Neglect*, *113*, 104930. https://doi.org/10.1016/j.chiabu.2021.104930

Lemaigre, C., Taylor, E. P., & Gittoes, C. (2017). Barriers and facilitators to disclosing sexual abuse in childhood and adolescence: A systematic review. *Child Abuse & Neglect*, *70*, 39–52. https://doi.org/10.1016/j.chiabu.2017.05.009

Lerner, J. S., Li, Y., Valdesolo, P., & Kassam, K. S. (2015). Emotion and Decision Making. *Annual Review of Psychology*, *66*(1), 799–823. https://doi.org/10.1146/annurev-psych-010213-115043

Lewis, T., McElroy, E., Harlaar, N., & Runyan, D. (2016). Does the impact of child sexual abuse differ from maltreated but non-sexually abused children? A prospective examination of the impact of child sexual abuse on internalizing and externalizing behavior problems. *Child Abuse & Neglect*, *51*, 31–40. https://doi.org/10.1016/j.chiabu.2015.11.016

Lilienfeld, S. O. (2016). Forensic Interviewing for Child Sexual Abuse: Why Psychometrics Matters. In W. T. O'Donohue & M. Fanetti (Eds.), *Forensic Interviews Regarding Child Sexual Abuse: A Guide to Evidence-Based Practice* (pp. 155–178). Springer International Publishing. https://doi.org/10.1007/978-3-319-21097-1_9

Magnusson, M., Joleby, M., Luke, T. J., Ask, K., & Lefsaker Sakrisvold, M. (2021). Swedish and Norwegian Police Interviewers' Goals, Tactics, and Emotions When Interviewing Suspects of Child Sexual Abuse. *Frontiers in Psychology*, *12*. https://doi.org/10.3389/fpsyg.2021.606774

Marchant, R., Carter, J., & Fairhurst, C. (2020). Opening doors: Suggested practice for medical professionals for when a child might be close to telling about abuse. *Archives of Disease in Childhood*, archdischild-2020-320093. https://doi.org/10.1136/archdischild-2020-320093

Martin, M. M., & Anderson, C. M. (1998). The cognitive flexibility scale: Three validity studies. *Communication Reports*, *11*(1), 1–9. https://doi.org/10.1080/08934219809367680

Martire, K. A., Growns, B., Bali, A. S., Montgomery-Farrer, B., Summersby, S., & Younan, M. (2020). Limited not lazy: A quasi-experimental secondary analysis of evidence quality evaluations by those who hold implausible beliefs. *Cognitive Research: Principles and Implications*, *5*(1), 65. https://doi.org/10.1186/s41235-020-00264-z

McCartan, K. F., Kemshall, H., & Tabachnick, J. (2015). The construction of community understandings of sexual violence: Rethinking public, practitioner and policy discourses. *Journal of Sexual Aggression*, *21*(1), 100–116. https://doi.org/10.1080/13552600.2014.945976

McKasy, M. (2020). A discrete emotion with discrete effects: Effects of anger on depth of information processing. *Cognitive Processing*, *21*(4), 555–573. https://doi.org/10.1007/s10339-020-00982-8

Mokros, A., Schemmel, J., Körner, A., Oeberst, A., Imhoff, R., Suchotzki, K., Oberlader, V., Banse, R., Kannegießer, A., Gubi-Kelm, S., Lehmann, R., & Volbert, R. (2024). Rituelle sexuelle Gewalt. Eine kritische Auseinandersetzung mit fragwürdigen empirischen Belegen für ein fragliches Phänomen. *Psychologische Rundschau*. https://doi.org/10.1026/0033-3042/a000663

Neal, T. M. S., Lienert, P., Denne, E., & Singh, J. P. (2022). A general model of cognitive bias in human judgment and systematic review specific to forensic mental health. *Law and Human Behavior*, *46*(2), 99. https://doi.org/10.1037/lhb0000482

Nickerson, R. S. (1998). Confirmation Bias: A Ubiquitous Phenomenon in Many Guises. *Review of General Psychology*, *2*(2), 175–220. https://doi.org/10.1037/1089-2680.2.2.175

Nunnally, J., & Bernstein, I. (1994). The assessment of reliability. In *Psychometric Theory* (3nd ed., pp. 248–292). McGraw-Hill.

Oberlader, V., Schmidt, A. F., Banse, R., & Quinten, L. (2024). *How can I reduce bias in my work? Discussing debiasing strategies for forensic psychological assessments*. https://doi.org/10.31234/osf.io/6e5md

O'Donohue, W., Benuto, L. T., & Cirlugea, O. (2013). Analyzing Child Sexual Abuse Allegations. *Journal of Forensic Psychology Practice*, *13*(4), 296–314. https://doi.org/10.1080/15228932.2013.822245

O'Donohue, W., & Cirlugea, O. (2021). Controlling for Confirmation Bias in Child Sexual Abuse Interviews. *The Journal of the American Academy of Psychiatry and the Law*, *49*(3), 371–380. https://doi.org/10.29158/JAAPL.200109-20

Oeberst, A., & Imhoff, R. (2023). Toward Parsimony in Bias Research: A Proposed Common Framework of Belief-Consistent Information Processing for a Set of Biases. *Perspectives on Psychological Science*, *18*(6), 1–24. https://doi.org/10.1177/17456916221148147

Oeberst, A., Wachendörfer, M. M., Imhoff, R., & Blank, H. (2021). Rich false memories of autobiographical events can be reversed. *Proceedings of the National Academy of Sciences*, *118*(13), e2026447118. https://doi.org/10.1073/pnas.2026447118

Olaguez, A. P., Peplak, J., Lundon, G., & Klemfuss, J. Z. (2023). The role of discrete emotional reactions to child sexual abuse (CSA) testimony in mock juror decision-making. *Psychology, Crime & Law*. https://www.tandfonline.com/doi/full/10.1080/1068316X.2023.2292516

Otgaar, H., de Ruiter, C., Howe, M. L., Hoetmer, L., & van Reekum, P. (2017). A Case Concerning Children's False Memories of Abuse: Recommendations Regarding Expert Witness Work. *Psychiatry, Psychology and Law*, *24*(3), 365–378. https://doi.org/10.1080/13218719.2016.1230924

Patihis, L., & Pendergrast, M. H. (2019). Reports of Recovered Memories of Abuse in Therapy in a Large Age-Representative U.S. National Sample: Therapy Type and Decade Comparisons. *Clinical Psychological Science*, *7*(1), 3–21. https://doi.org/10.1177/2167702618773315

Pompedda, F., Zappalà, A., & Santtila, P. (2015). Simulations of child sexual abuse interviews using avatars paired with feedback improves interview quality. *Psychology, Crime & Law*, *21*(1), 28–52. https://doi.org/10.1080/1068316X.2014.915323

Pompedda, F., Zhang, Y., Haginoya, S., & Santtila, P. (2022). A Mega-Analysis of the Effects of Feedback on the Quality of Simulated Child Sexual Abuse Interviews with Avatars. *Journal of Police and Criminal Psychology*. https://doi.org/10.1007/s11896-022-09509-7

Porter, S., Peace, K. A., & Emmett, K. A. (2007). You Protest Too Much, Methinks: Investigating the Features of Truthful and Fa... *Canadian Journal of Behavioural Science;*, *39*(2), 79–91. https://doi.org/10.1037.cjbs2007007

Powell, M. B. (2008). Designing Effective Training Programs for Investigative Interviewers of Children. *Current Issues in Criminal Justice*, *20*(2), 189–208. https://doi.org/10.1080/10345329.2008.12035804

Powell, M. B., Brubacher, S. P., & Baugerud, G. A. (2022). An overview of mock interviews as a training tool for interviewers of children. *Child Abuse & Neglect*, *129*, 105685. https://doi.org/10.1016/j.chiabu.2022.105685

Powell, M. B., Hughes-Scholes, C. H., & Sharman, S. J. (2012). Skill in Interviewing Reduces Confirmation Bias: Confirmation bias and interviews. *Journal of Investigative Psychology and Offender Profiling*, *9*(2), 126–134. https://doi.org/10.1002/jip.1357

R Core Team. (2021). *R: A language and environment for statistical computing.* [Computer software]. R Foundation for Statistical Computing. https://www.R-project.org/.

Rohrabaugh, M., London, K., & Hall, A. K. (2016). Planning the Forensic Interview. In W. O'Donohue & M. Fanetti (Eds.), *Forensic Interviews Regarding Child Sexual Abuse: A Guide to Evidence-Based Practice* (pp. 197–218). Springer International Publishing. https://doi.org/10.1007/978-3-319-21097-1_11

Rosenthal, R. (1976). *Experimenter Effects in Behavioral Research*. Irvington.

Salerno, J. M. (2021). The Impact of Experienced and Expressed Emotion on Legal Factfinding. *Annual Review of Law and Social Science*, *17*, 181-203. https://doi.org/10.1146/annurev-lawsocsci-021721-072326

Sambrano, D., Masip, J., & Blandón-Gitlin, I. (2020). How emotions affect judgement and decision making in an interrogation scenario. *Legal and Criminological Psychology*, lcrp.12181. https://doi.org/10.1111/lcrp.12181

Saywitz, K. J., Larson, R. P., Hobbs, S. D., & Wells, C. R. (2015). Developing Rapport with Children in Forensic Interviews: Systematic Review of Experimental Research: Developing rapport with children. *Behavioral Sciences & the Law*, *33*(4), 372–389. https://doi.org/10.1002/bsl.2186

Schwarz, N. (2012). Feelings-as-information theory. In P. A. M. Van Lange, A. W. Kruglanski, & E. T. Higgins (Eds.), *Handbook of theories of social psychology, Vol. 1* (pp. 289–308). Sage Publications Ltd. https://doi.org/10.4135/9781446249215.n15

Scoboria, A., Wade, K. A., Lindsay, D. S., Azad, T., Strange, D., Ost, J., & Hyman, I. E. (2017). A mega-analysis of memory reports from eight peer-reviewed false memory implantation studies. *Memory*, *25*(2), 146–163. https://doi.org/10.1080/09658211.2016.1260747

Segal, A., Bakaitytė, A., Kaniušonytė, G., Ustinavičiūtė-Klenauskė, L., Haginoya, S., Zhang, Y., Pompedda, F., Žukauskienė, R., & Santtila, P. (2023). Associations between emotions and psychophysiological states and confirmation bias in question formulation in ongoing simulated investigative interviews of child sexual abuse. *Frontiers in Psychology*, *14*. https://www.frontiersin.org/articles/10.3389/fpsyg.2023.1085567

Segal, A., Kaniušonytė, G., Bakaitytė, A., Žukauskienė, R., & Santtila, P. (2023). The Effects of Emotions on the Assessment of Child Sexual Abuse Interviews. *Journal of Police and Criminal Psychology*, *38*(4), 826–837. https://doi.org/10.1007/s11896-022-09571-1

Segal, A., Pompedda, F., Haginoya, S., Kaniušonytė, G., & Santtila, P. (2022). Avatars with child sexual abuse (vs. No abuse) scenarios elicit different emotional reactions. *Psychology, Crime & Law*, 1–21. https://doi.org/10.1080/1068316X.2022.2082422

Siemer, M., Mauss, I., & Gross, J. J. (2007). Same situation--Different emotions: How appraisals shape our emotions. *Emotion*, *7*(3), 592–600. https://doi.org/10.1037/1528-3542.7.3.592

Smith, E. R., & Neumann, R. (2005). Emotion processes considered from the perspective of dual-process models. In L. Feldman Barrett, P. Niedenthal, & P. Winkelnam (Eds.), *Emotion and Consciousness* (pp. 287–311). Guilford Press.

Sternberg, K. J., Lamb, M. E., Davies, G. M., & Westcott, H. L. (2001). The Memorandum of Good Practice: Theory versus application. *Child Abuse & Neglect*, *25*(5), 669–681. https://doi.org/10.1016/S0145-2134(01)00232-0

Stoltenborgh, M., Bakermans-Kranenburg, M. J., Alink, L. R. A., & van IJzendoorn, M. H. (2015). The Prevalence of Child Maltreatment across the Globe: Review of a Series of Meta-Analyses: Prevalence of Child Maltreatment across the Globe. *Child Abuse Review*, *24*(1), 37–50. https://doi.org/10.1002/car.2353

Talwar, V., Crossman, A. M., Block, S., Brubacher, S., Dianiska, R., Espinosa Becerra, A. K., Goodman, G., Huffman, M. L., Lamb, M. E., London, K., La Rooy, D., Lyon, T. D., Malloy, L. C., Maltby, L., Greco, V. P. N., Powell, M., Quas, J., Rood, C. J., Spyksma, S. D., … Wylie, B. (2024). Urgent issues and prospects on investigative interviews with children and adolescents. *Legal and Criminological Psychology*, lcrp.12269. https://doi.org/10.1111/lcrp.12269

Tuggener, D., Schneider, T., Huwiler, A., Kreienbühl, T., & Hischier, S. (2024, September). *Role-Playing LLMs in Professional Communication Training: The Case of Investigative Interviews with Children*. Proceedings of the 20th Conference on Natural Language Processing (KONVENS 2024), Vienna.

Volbert, R., & Kuhle, L. F. (2019). Sexueller Kindesmissbrauch. In R. Volbert, A. Huber, A. Jacob, & A. Kannegießer (Eds.), *Empirische Grundlagen der familienrechtlichen Begutachtung* (pp. 233–262). Hogrefe.

Wachendörfer, M. M., & Oeberst, A. (2023). Differences Between True and False Autobiographical Memories: A Scoping Review. *European Psychologist*, *28*(4), 247–264. https://doi.org/10.1027/1016-9040/a000513

Webster, D. M., & Kruglanski, A. W. (1994). Individual Differences in Need for Cognitive Closure. *Journal of Personality and Social Psychology*, *67*(6), 1049.

Zajac, R., & Brown, D. A. (2018). Conducting Successful Memory Interviews with Children. *Child and Adolescent Social Work Journal*, *35*(3), 297–308. https://doi.org/10.1007/s10560-017-0527-z

Zapf, P. A., & Dror, I. E. (2017). Understanding and Mitigating Bias in Forensic Evaluation: Lessons from Forensic Science. *International Journal of Forensic Mental Health*, *16*(3), 227–238. https://doi.org/10.1080/14999013.2017.1317302

Appendix

Appendix A: Danksagung (Acknowledgement)

Eine Promotion ist ein lehrreiches Unterfangen und gen Ende ein ziemlich zähes Kaugummi. Ich möchte mich bei den Menschen bedanken, die mich in den interessanten und zähen Phasen dieses Projekts begleitet haben. Ohne euch wäre die Promotionszeit öde und das Ergebnis schlecht geworden.

Mein Dank gilt zuvorderst Prof. Dr. Renate Volbert, die diese Promotion, so wie viele andere meiner Projekte, betreut hat. Liebe Renate, Ich danke dir für die Unermüdlichkeit, mit der mich und andere unterstützt. Dafür, dass du uns an deinem Wissensschatz teilhaben lässt, dich immer wieder für unsere Ideen interessierst, schlechte Ideen in ihre Schranken verweist und ohnehin noch selbst so voller Ideen bist. Danke für deine Inspiration.

Ich danke meinen teils langjährigen Kolleginnen und Kollegen Anett Tamm, Niels Krause, Dr. Jonas Schemmel, Mona Leve, Jana Otzipka, Marie Merschhemke, und Dr. Kristin Jankowsky für ihre Unterstützung, die gute Laune im Büro, und die gemeinsame Arbeit an Projekten, die entweder in diese Dissertation einfließen oder mich davon ablenken konnten. Auch meinen zeitweiligen Vorgesetzten Prof. Dr. Simone Pülschen und Prof. Dr. Ulrich Schroeders sowie meinem Zweitbetreuer Prof. Dr. Stefan Krumm danke ich für die inhaltliche, zeitliche und organisatorische Unterstützung bei meiner Promotion. Mein aufrichtiger Dank gilt außerdem den Studierenden, die im Rahmen von Praktika, Masterarbeiten oder als Wissenschaftliche Hilfskräfte an den Datenerhebungen für diese Promotion mitgewirkt haben.

Work und Life sind in der Academia bekanntermaßen nicht sonderlich getrennt. Danke Johannes, dass du bei all dem an meiner Seite bist. Danke für die vielen Gespräche über die Details meiner Arbeit, fürs Aushalten und Rückenfreihalten und für die Leichtigkeit, die du mich

lehrst. Mein lieber Mats, danke, dass du mir die Irrelevanz akademischer Würden aufzeigst. Was ist ein Doktortitel gegen die Tatsache, dass wir morgens gemeinsam am Fenster sitzen und dem Müllmann winken und er manchmal sogar zurückwinkt. Danke, Mama, für die Grundsteinlegung für all das und alles darüber hinaus.

Appendix B: Zusammenfassung in deutscher Sprache (German Abstract)

In dieser Arbeit wird der Einfluss von kognitiven und emotionalen Mustern auf Voreingenommenheit (Bias) und Suggestivität in der Befragung von Kindern zu Verdacht auf sexuellen Missbrauch untersucht. Im Rahmen von fünf empirischen Studien, die in drei Artikeln zusammengefasst sind, wurden Skalen zu „Kognitionen und emotionalen Reaktionen im Umgang mit sexuellem Missbrauch von Kindern“ (Cognitions and Emotions about Child Sexual Abuse [CECSA]) entwickelt, validiert, und Zusammenhänge zu Bias und suggestivem Befragungsstil sowie die Veränderbarkeit der Skalen untersucht.

Artikel 1 beschreibt die Entwicklung und erste Validierung der drei CECSA-Skalen Unreflektierte Gewissheit (Naive Confidence [NC]), Emotionale Reaktivität (Emotional Reactivity [ER]) und Misstrauen in das Justizsystem (Justice System Distrust [JSD]) an einer Stichprobe von 801 Studierenden der Humanwissenschaften. Die Skalen wiesen gute Modellanpassung und akzeptable bis gute Reliabilitätswerte auf und sagten bei der Beurteilung vager Missbrauchsverdachtsfälle einen Bias in Richtung der Missbrauchshypothese hervor. Artikel 2 umfasst drei Studien zur Vorhersage von Bias und suggestivem Befragungsstil in fingierten Gesprächen mit Kindern, für die unterschiedliche Formate zum Stellen der Fragen entwickelt wurden: ein Single-Choice-Format, ein Freitextformat und ein Format für natürliche Sprache in einer Virtual Reality (VR) Simulation. Die Ergebnisse der drei Studien und einer meta-analytischen Integration zeigen für insgesamt 674 Studierende aus verschiedenen Disziplinen (Humanwissenschaften, Lehramt, Polizeistudium) eine robuste Vorhersage von Bias und Suggestivität durch die Skalen NC und ER, nicht jedoch durch JSD. Artikel 3 evaluiert ein Trainingsprogramm zur Verbesserung der Gesprächsführung mit Kindern in Missbrauchsverdachtsfällen. Eine Sekundäranalyse der Daten zeigte, dass eine zweitägige

Schulung mit Lehramtsstudierenden zu einer signifikanten Reduktion der NC- und ER-Werte führte, während die Ergebnisse für JSD uneindeutiger ausfielen.

Die Ergebnisse dieser Dissertation zeigen, dass hohe Ausprägungen auf den CECSA Skalen NC und ER Bias und Suggestivität in der Befragung von Kindern hervorsagen, die Werte jedoch durch Schulungen reduziert werden können. Die Skalen können diagnostisch oder zu Evaluationszwecken eingesetzt werden, etwa für die Entwicklung von Befragungstrainings, die Auswahl geeigneter Befragungspersonen oder zur weiteren Erforschung der Rolle individueller Unterschiede in der Abklärung und Ermittlung von Missbrauchsverdachtsfällen.

Appendix C: Eigenständigkeitserklärung (Declaration of Authenticity)

Hiermit versichere ich, Elsa Gewehr, dass ich

- die vorliegende Dissertation selbstständig verfasst und ohne unerlaubte Hilfe angefertigt habe
- die Stellen der Arbeit, die dem Wortlaut oder dem Sinn nach anderen Werken (dazu zählen auch Internetquellen und KI-basierte Tools) entnommen sind, unter Angabe der Quelle kenntlich gemacht habe
- alle Hilfsmittel, die verwendet wurden, angegeben habe.

Die Dissertation ist in keinem früheren Promotionsverfahren angenommen oder abgelehnt worden.

Ort, Datum

Elsa Gewehr