DISSERTATION

Valerie Hauch

2015

Meta-analyses

on the detection of deception

with linguistic and verbal content cues.

Inaugural-Dissertation

zur

Erlangung des Doktorgrades

der Philosophie des Fachbereiches 06 Psychologie

der Justus-Liebig-Universität Gießen

vorgelegt von

Valerie Hauch

aus Dortmund

2015

Dekan: Prof. Dr. Dr. Jürgen Hennig

1. Berichterstatter: Prof. Siegfried Ludwig Sporer, Ph.D.

2. Berichterstatter: Prof. Dr. Renate Volbert

Tag der Disputation: 10.11.2016

DANK

An dieser Stelle möchte ich mich bei einigen Personen bedanken, die mich während der Erstellung der Dissertation unterstützt haben.

Mein besonderer Dank gilt meinem Doktorvater Prof. Siegfried Ludwig Sporer, Ph.D., ohne dessen Ideen und unermüdlichen Forschungsdrang diese Arbeit und die daraus resultierende Forschungskooperation nicht entstanden wären. Ebenfalls zolle ich meinen herzlichen Dank für die hilfreiche fachlich-kompetente und motivierende Unterstützung Dr. Jaume Masip und Dr. Iris Blandón-Gitlin.

Ich bedanke mich bei dem Gleichstellungskonzept der Frauenbeauftragten der JLU Gießen für das zweieinhalb Jahre andauernde Promotionsstipendium. Des Weiteren danke ich der Zweitgutachterin Frau Prof. Dr. Renate Volbert, die sich spontan bereit erklärt hat, diese Arbeit zu begutachten.

Für die akribische Kontrolle der kodierten Daten möchte ich mich herzlich bei Andreas Reis und für die Unterstützung der Kodierung einzelner Studien bei Emma Halfmann bedanken. Bei Dr. Stephen Michael, Dr. Kristina Kaminski, meiner Kollegin Franziska Rudzik, meiner Kindergartenfreundin Jennifer Busch und meinem besten Freund Alexander Müller möchte ich mich ganz herzlich für die hilfreichen Rückmeldungen und Anregungen im Hinblick auf diese Arbeit bedanken. Nicht zuletzt gilt mein aufrichtiger Dank meinen lieben Freunden und Freundinnen, insbesondere meinem Partner Michael Heintz, Kristina Kaminski, Alexander Müller, Jennifer Busch und Nadja Schulte, sowie meiner Schwester Sandra Neumann und meinen Eltern Astrid und Jens-Jürgen Hauch, die mich alle über die Jahre stets motiviert, emotional unterstützt und immer an mich geglaubt haben.

Weil am Rhein, im Dezember 2015 *Valerie Hauch*

ABSTRACT

This dissertation reports two meta-analyses on verbal cues to deception. Whereas the first synthesis focuses on the validity of linguistic cues to deception, the second article focuses on the inter-rater reliability of verbal content cues. In general, the validity deals with the question if and to what extent a certain indicator of deception distinguishes truthful from deceptive statements. On the other side, the inter-rater reliability describes the amount of agreement that can be reached from several evaluators when rating specific verbal content cues.

More specifically, the *first meta-analysis* investigates the validity of linguistic cues to deception that are assessed with computer programs. From 44 studies meeting the inclusion criteria, operational definitions for 79 linguistic cues were identified and allocated to six broader research questions. As predicted, meta-analyses showed that relative to truth-tellers, liars experienced greater cognitive load, expressed more negative emotions, and distanced themselves more from events. On the other side, liars expressed fewer sensory-perceptual words, and referred less often to cognitive processes. However, compared to liars, truth-tellers slightly used more terms related to uncertainty. Most main effects were moderated by several important independent variables such as event type, personal involvement, emotional valence, intensity of interaction, motivation, production mode, type of computer program and publication status. Although the average effect size was small, theoretical predictions were partially supported indicating that (a) liars and truth-tellers seem to use different words in a specific context and (b) computer programs can be designed to count some of these linguistic differences. However, at this point, computer programs are far from being applied in real life deception detection contexts. These findings not only further our knowledge about the

usefulness of linguistic cues to detect deception with computers in applied settings but also elucidate the relationship between language and deception.

The *second meta-analysis* examines the inter-rater reliability of a different kind of verbal content criteria, the so-called Criteria-based Content Analysis (CBCA). CBCA consists of 19 credibility criteria and constitutes an important component of Statement Validity Assessment (SVA). SVA is a forensic assessment procedure used in many countries to evaluate whether statements (e.g., of sexual abuse) are based on experienced or fabricated events. Furthermore, these criteria have frequently been adapted for research on the detection of deception as a “credibility assessment tool”. A total of 82 hypothesis tests from 52 English and 22 German studies were included and revealed high inter-rater reliabilities for most CBCA criteria as measured with several reliability indices. Due to large heterogeneity, moderator analyses and meta-regression were conducted on Pearson’s *r*. Significant findings occurred for research paradigm, intensity of rater training, type of rating scale used, and the frequency of occurrence of CBCA criteria (base rates) for some criteria. Implications for future research and forensic practice are discussed.

In summary, these meta-analyses suggest that human language is probably the most promising source to differentiate liars from truth-tellers. Moreover, these results show that several linguistic and verbal content cues fulfilled psychometric quality standards like validity and inter-rater reliability to some extent and under specific conditions. Taken several limitations into account, implications for research and practice are discussed.

TABLE OF CONTENTS

INTRODUCTION

In a recent German trial, a well-known weather forecaster named Jörg Kachelmann was accused by the German prosecution of having severely raped his former beloved (Claudia D., also being the accessory prosecution) in coincidence with a grievous bodily harm in February 2010 (Spiegel Online, 2010, May). Kachelmann denied this serious accusation (Doerris, 2010, March), and a long and complex trial known as the “Kachelmann-Prozess” (Kachelmann-trial) started and attracted great national and international public interest (e.g., Connolly, 2010, September). In general, in the inquisitorial legal system in Germany, expert (and lay) judges’ main task is to establish the truth by evaluating and weighting evidence presented by prosecutors and defense lawyers and finally to pass a sentence. Alike many cases of sexual abuse or rape brought to court, no external or independent evidence, such as a videotape, existed (Steller & Köhnken, 1989; Undeutsch, 1982; Vrij, 2008). As a consequence, often, the only evidence that can be consulted is testimonies of the accused, the victim, or other witnesses. Precisely this happened in the Kachelmann-trial: In principal, her statement stood against his statement and they unsurprisingly did not correspond to each other.

In view of this adverse starting position, how did the judges come to a decision? Within the scope of two main evidence lines of this trial, expert witnesses were called upon. First, medical forensic experts were heard to evaluate forensic evidence (DNA traces on panties and knife; physical injuries of Claudia D.). From their analyses, no conclusive findings were presented. Hence, in a second stage, two experienced and reputable German forensic psychologists (Prof. Dr. Luise Greuel and Prof. Dr. Günther Köhnken) evaluated the credibility of Claudia D.’s testimony. The opportunity of the judge to mandate qualified expert witnesses is

widely adopted in many European countries in complex cases where explicit evidence does not exist (e.g., Austria, Germany, Netherlands, Sweden, Switzerland; Sporer, 1983; Steller & Köhnken, 1989; Vrij, 2008). Here, psychological experts attempt to assess the credibility of a statement - not the general credibility of the person - with a clinical assessment procedure called Statement Validity Assessment (SVA, Köhnken, 2004; Steller & Köhnken, 1989). In the Kachelmann-trail, both experts concluded (in two different ways) that Claudia D.'s statement was probably not based on real experience. One reason for this assumption as cited in the media was a less differentiated and undetailed statement of the critical action in comparison to the victim's generally more detailed narrative style (Albrecht-Heider, 2011, May; Friedrichsen, 2011, May). However, her accusation could neither be verified nor falsified, and eventually the genuine truth (also called "ground truth") remained hidden. Finally, the district court ("Landgericht") Mannheim acquitted Kachelmann (Bock, 2011, May) in terms of the juridical principle "in dubio pro reo" (European Court of Human Rights, 2010, June). Recently, in a civil proceeding denounced by Jörg Kachelmann, the district court Cologne suited the most widely read German newspaper "Bild" for a payment of compensation due to several infringements of Kachelmann's personal rights (Landgericht Köln, 28 O 2/14, 28 O 7/14; Zeit Online 2015, September).

This famous German trial is an example of hundreds of related cases of (alleged) sexual abuse or rape lacking unambiguous evidence that were and need to be negotiated (Arntzen, 1992; Sporer, 1983; Steller, 2013). Out of the courtroom, in everyday life, there are also many social situations in which a mere statement stands against another statement and no proof of the truth exists. To be more specific, DePaulo and her colleagues found that people tell one to two lies on average per

day (DePaulo, Kashy, Kirkendol, Wyer, & Epstein, 1996). For example, a manager comes too late to a business meeting and tells he was stuck in traffic, or a little girl asserts her mother that she has not eaten the missing piece of cake. These are situations in which the listener of the story faces himself or herself with the following question: How can you tell when people are lying? Exactly this question was asked to lay persons in a large-scale investigation by several deception research experts in 75 different countries (Global Deception Research Team, 2010). Interestingly, regardless of the type of question asked (open or closed-ended), a worldwide, pan-cultural stereotype of a liar was uncovered: Most people subjectively assume that liars avoid eye contact. Also, liars are considered to be nervous, shift postures, touch and scratch themselves, or have a flawed language (e.g., more pauses, stuttering, inconsistent) compared to truth-tellers (Global Deception Research Team, 2010).

These findings unclose the next question: Can these stereotypes indeed help in distinguishing deceptive from true stories? Put differently, do these *subjective* assumptions actually correspond to *objective* indicators (or cues) to deception? A first insight comes from two meta-analyses on the general ability to detect deception (Aamodt & Custer, 2006; Bond & DePaulo, 2006). Before presenting the main results, a brief definition of meta-analysis is given. A meta-analysis represents a quantitative integration or synthesis of empirical studies investigating the same research question. Opposite to a mere literature review, meta-analyses quantify study outcomes with predetermined methodological and statistical methods, in particular by means of effect sizes (APA, 2008; Cooper, Hedges, & Valentine, 2009; Lipsey & Wilson, 2001, Sporer & Cohn, 2011). However, both meta-analyses found a general average detection accuracy of 54% in more than 24,000 judgments (Bond & DePaulo, 2006). Alarmingly, this finding is only slightly higher (due to the large

sample size) than the probability of flipping a coin (50%). In other words, in general, people are poor lie detectors. Anyway, further analyses suggested that people's ability to detect true stories (61.34%) is significantly higher than the ability to detect lies (47.55%). From these findings it can be hypothesized that a discrepancy between subjective lay assumptions and objective indicators of deception may exist (see Sporer & Schwandt, 2007). Contrary to this hypothesis are the findings of Hartwig and Bond's (2011) large-scale meta-analyses: They found that people do actually not rely on the wrong cues to deception. Rather, the authors attributed the generally low detection accuracy to a limited validity of objective behavioral cues to deception. Actually, research on objective cues to deception (DePaulo, Lindsay, Malone, Muhlenbruck, Charlton, & Cooper, 2003; Sporer & Schwandt, 2006, 2007) suggested that on average, nonverbal (e.g., gestures, adaptors, eye contact) and paraverbal (e.g., speech rate, pauses, voice pitch) cues do not show impressive effect sizes. Therefore, these cues to deception may indeed not be useful for deciding whether a person is lying or telling the truth.

If only weak objective cues to deception exist and if people generally make judgments on chance level - how *can* you tell when people are lying or telling the truth? This dissertation is an attempt to come closer to an answer to this important question. More specifically, it focuses on empirical studies that examine content and language differences of deceptive and true statements. Before introducing the specific investigations that are presented in this dissertation, important research findings will be identified that justify these attempts.

First, the meta-analysis by DePaulo et al. (2003) suggested that objective indicators to deception related to the content of a statement reveal somewhat higher effect sizes than nonverbal or paraverbal cues to deception (see also Sporer &

Schwandt, 2006, 2007). The only shortcoming of these findings is the small number of studies examining verbal content cues that were included in this meta-analysis. However, a recent meta-analysis (Amado, Arce, & Fariña, 2015) on 19 specific verbal content cues, called *Criteria-based Content Analysis* (CBCA, Steller & Köhnken, 1989) - as an important part of Statement Validity Assessment - synthesized more studies than DePaulo and her colleagues (DePaulo et al., 2003). The authors found medium to large effect sizes for almost all CBCA criteria when assessed in children's statements in 18 published empirical studies. In other words, these verbal content (CBCA) criteria are on average more frequently present in true than in deceptive statements.

Second, Bond and DePaulo (2006) showed that discrimination accuracy was lowest when judges are exposed to videotapes rather than presented with audiovisual or audio only stimuli. The authors interpreted this finding with their double-standard framework (i.e., people tend to overestimate other people's lies and underestimate their own lies), that a liar's stereotype is predominantly visual and thus mostly activated when seeing a video. From a different perspective, this finding could be another hint to the assumed superiority of verbal versus nonverbal and paraverbal cues to deception.

Third, another line of research confirms this assumption. As detection accuracy is consistently found to be around chance level, for decades, researchers from different areas have trained people in order to increase their detection accuracy. Recently, 30 studies implementing an experimental (i.e. training) group versus control group design were meta-analyzed (Hauch, Sporer, Michaels, & Meissner, 2014). Obviously, training programs with verbal content cues resulted in significantly higher training effects than training programs with nonverbal and/or

paraverbal cues to deception, or with giving feedback to judges (versus no feedback control group).

Taken together, these empirical findings from different approaches suggest that the content of a statement is probably more diagnostic than behavioral cues in differentiating lies from true accounts. Therefore, more detailed and specialized analyses of the existing research on these promising verbal cues is warranted and focus of this thesis.

Essentially, this dissertation presents two quantitative syntheses on verbal cues (or criteria) to deception. The *first meta-analysis* deals with an apparently curious method to detect deception: Besides the aforementioned approaches of investigating people's behavior or content of the statement (e.g., CBCA) in previous meta-analyses, a different method was inspected. More precisely, researchers and practitioners have developed computer programs to analyze the verbal content, more specifically, linguistic markers (or cues; i.e. words, or word categories) in transcribed statements. The intention is clear: A computer program is supposed to distinguish lies from true statements based on linguistic differences. Surprisingly, already forty years ago, the first study was designed to analyze linguistic markers with a computer program (Knapp, Hart, & Dennis, 1974). Since then, a large amount of studies from various research disciplines, such as psychology and law, social psychology, communication, linguistics, or computer science, were conducted on this topic. Additionally, a number of different linguistic cues were investigated (e.g., Newman, Pennebaker, Berry, & Richards, 2003; Zhou, Burgoon, Nunamaker, & Twitchell, 2004). Therefore, a systematic review is clearly warranted.

Meta-analysis is the first attempt to quantitatively summarize and categorize these studies and linguistic cues to deception - assessed by computer programs.

This enterprise is conducted with support of basic theoretical frameworks from the deception literature within main research questions. The aim of this meta-analysis was to test directional hypothesis and to present operational definitions for linguistic cues to deception. Furthermore, several important independent variables, such as the type of the event, degree of personal investment, emotional valence, or extrinsic motivation of the storyteller are analyzed to shed light on the relevance of the context of a statement for linguistic differences of liars and truth-tellers. The findings will be discussed in regard to their theoretical background, limitations and their usefulness in applied settings.

As the first meta-analysis deals with the validity of linguistic cues to deception, the *second meta-analysis* focuses on the inter-rater reliability of a different set of verbal content cues to deception. More specifically, the aforementioned set of 19 CBCA criteria, credibility criteria, or verbal content criteria, is object of the second meta-analysis (Steller & Köhnken, 1989). As noted in the example case (Kachelmann-trial), these criteria feature an important component of the credibility assessment procedure called Statement Validity Assessment (e.g., Köhnken, 2004; Steller, 1989), which is widely applied in psychological expert testimonies in many countries. Opposite to linguistic cues assessed by computer programs, human judges rate these criteria in statements with regard to their presence (or strength of presence). With respect to its validity, true accounts are assumed to contain more criteria compared to false accounts due to qualitative differences, and the presence of a criterion is an indicator (not evidence) that the statement is based on real experience (e.g., Undeutsch, 1967; Steller & Köhnken, 1989).

A large amount of research on credibility assessment and the detection of deception frequently investigate the validity of CBCA criteria as an attempt to

distinguish true from false statements (e.g., Vrij, 2008). As mentioned earlier, a recent meta-analysis on a subset of published CBCA studies with children (Amado et al., 2015) estimated the validity of almost all criteria as present to some extent. Preliminary results of a large-scale meta-analysis (Sporer, Hauch, Blandón-Gitlin, & Masip, 2015, August) also showed that most criteria significantly differ between deceptive and true statements, although some important variables (e.g., type of experimental paradigm, type of rating scale used, or the age of senders) moderate the effect sizes.

Not less important than its validity is the question of its inter-rater reliability, especially for its legal application (e.g., Köhnken, 2004; Steller & Köhnken, 1989). Inter-rater reliability represents the amount of agreement that can be reached in subjective ratings from several judges. Therefore, the inter-rater reliability constitutes a prerequisite of its validity and thus is important to be quantified (Küpper & Sporer, 1995). Put differently, with two rubber tape measures with varying elastic properties - how can you correctly determine your exact height? This is probably an impossible endeavor. Back to CBCA, the following research question guides the second meta-analysis: To what extent do different evaluators agree on their presence (or strength of presence) ratings of CBCA criteria? Therefore, in an attempt to synthesize all published and unpublished studies written in English and German, this meta-analysis is the first quantitative review on several inter-rater reliability indices of individual CBCA criteria. The second aim is to quantify the association of inter-rater reliability with important independent variables, such as the frequency of occurrence (i.e., base rate), the rating scale used, different research paradigms, and the training of the raters. The findings will be discussed in light of their implications for future research and practice.

The purpose of the following meta-analyses in this dissertation is (a) to quantify linguistic differences between deceptive and true accounts by integrating empirical studies from different research areas, and (b) to assess the amount of inter-rater reliability of verbal content cues as a prerequisite of its validity. Taken together, this dissertation aims to add scientific knowledge on deception detection research from a verbal perspective.

References

Aamodt, M. G. & Custer, H. (2006). Who can best catch a liar? A meta-analysis of individual differences in detecting deception. *Forensic Examiner, 15*, 6-11.

Albrecht-Heider, C. (2011, May). *Plädoyer für den Angeklagten Kachelmann [Pleading for the defendant Kachelmann]*. Retrieved from http://www.fr-online.de/panorama/prozess-plaedoyer-fuer-den-angeklagten-kachelmann,1472782,8426576.html on November 14th 2015.

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

APA Publications and Communications Board Working Group on Journal Article Reporting Standards (2008). Reporting standards for research in Psychology. Why do we need them? What might they be? *American Psychologist, 63*, 839-851. doi:10.1037/0003-066X.63.9.839

Arntzen, F. (1992). Die Situation der forensischen Aussagepsychologie in der BRD. [Current state of psychology of testimony in the Federal Republic of Germany]. In A. Trankell (Ed.), *Reconstructing the past: The role of psychologists in criminal trials* (pp. 107-120). Deventer, The Netherlands: Kluwer.

Bock, J. (2011, May). *Pressemitteilung vom 31.05.2011 - Freispruch für Jörg Kachelmann [Press release from 05/31/2011 - Acquittal of Jörg Kachelmann].* Retrieved from http://www.landgericht-mannheim.de/pb/,Lde/1167947?QUERYSTRING =Kachelmann on November 14th 2015.

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

Connolly, K. (2010, September). *German weatherman faces rape trial.* The Guardian. Retrieved from http://www.theguardian.com/world/2010/sep/05/germany-weatherman-rape-trial on November 14th 2015.

Cooper, H., Hedges, L. V., & Valentine, J. C. (2009). (Eds.) *The handbook of research synthesis and meta-analysis*. New York: Russell Sage Foundation.

DePaulo, B. M., Kashy, D. A., Kirkendol, S. E., Wyer, M. M., & Epstein, J. A. (1996). Lying in everyday life. *Journal of Personality and Social Psychology, 70*, 979-995. doi:10.1037/0022-3514.70.5.979

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

Doerries, B. (2010, March). *Ich bin unschuldig [I am innocent].* Süddeutsche Zeitung. Retrieved from http://www.sueddeutsche.de/panorama/kachelmann-bleibt-in-haft-ich-bin-unschuldig-1.12069 on November 14th 2015.

European Court of Human Rights (2010, June). *European Convention on Human Rights, Section 1, Article 6 (2)*. Retrieved from http://www.echr.coe.int/Documents/Convention_ENG.pdf on November 14th 2015.

Friedrichsen, G. (2011, May). *Gutachter im Kachelmann-Prozess: "Vielleicht hat sie das Messer nur gefühlt?" [Expert witnesses in Kachelmann-trial: "Perhaps she has just felt the knife?"]* Retrieved from http://www.spiegel.de/panorama/justiz/gutachter-in-kachelmann-prozess-vielleicht-hat-sie-das-messer-nur-gefuehlt-a-761541.html on November 14th 2015.

Global Deception Research Team (2006). A world of lies. *Journal of Cross-Cultural Psychology, 37*, 60–74. doi:10.1080/14789940412331337353

Hartwig, M. & Bond, C. F. (2011). Why do lie-catchers fail? A lens model meta-analysis of human lie judgments. *Psychological Bulletin, 137*, 643-59. doi:10.1037/a0023589.

Hauch, V., Sporer, S. L., Michael, S. W., & Meissner, C. A. (2014). *Does training improve the detection of deception? A meta-analysis.* Communication Research. Advance online publication. doi:10.1177/0093650214534974

Knapp, M. L., Hart, R. P., & Dennis H. S. (1974). An exploration of deception as a communication construct. *Communication Research, 1*, 15-29. doi:10.1111/j.1468-2958.1974.tb00250.x

Köhnken, G. (2004). Statement Validity Analysis and the "detection of the truth". In P. A. Granhag & L. A. Strömwall (Eds.*), The detection of deception in forensic contexts* (pp. 41-63). Cambridge, UK: Cambridge University Press.

Küpper, B., & Sporer, S. L. (1995). Beurteilerübereinstimmung bei Glaubwürdig-keitsmerkmalen: Eine empirische Studie [Inter-rater reliability of content credibility criteria: An empirical study]. In G. Bierbrauer, W. Gottwald, & B. Birnbreier-Stahlberger (Eds.), *Verfahrensgerechtigkeit* (pp. 187-213). Köln, Germany: Otto Schmidt.

Landgericht Köln (2015, September). *Urteil 28 O 2/14 [Sentence No. 28 O 2/14].* Retrieved from http://www.justiz.nrw.de/nrwe/lgs/koeln/lg_koeln/j2015/28_O_2_14_Urteil_20150930.html on November 14th 2015.

Landgericht Köln (2015, September). *Urteil 28 O 7/14 [Sentence No. 28 O 7/14].* Retrieved from http://www.justiz.nrw.de/nrwe/lgs/koeln/lg_koeln/j2015/28_O_7_14_Urteil_20150930.html on November 14th 2015.

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis.* Thousand Oaks: Sage Publications.

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

Spiegel Online (2010, May). *Vorwurf der Vergewaltigung. Staatsanwälte klagen Moderator Kachelmann an [Accusation of rape. Prosecutors accused moderator Kachelmann].* Retrieved from http://www.spiegel.de/panorama/justiz/vorwurf-der-vergewaltigung-staatsanwaelte-klagen-moderator-kachelmann-an-a-695568.html on November 14th 2015.

Sporer, S. L. (1983, August). *Content criteria of credibility: The German approach to eyewitness testimony*. Paper presented at the 91st Annual Meeting of the American Psychological Association, Anaheim, CA.

Sporer, S. L., & Cohn, L. D. (2011). Meta-analysis. In B. D. Rosenfeld, & S. D. Penrod (Eds.), *Research methods in forensic psychology* (pp. 43-62). New York: Wiley.

Sporer, S. L., Hauch, V., Blandón-Gitlin, I., & Masip, J. (2015, August). *Content cues to veracity: A meta-analysis of the validity of Criteria-based Content Analysis.* Paper presented at the European Association of Psychology and Law Conference in Nuremberg, Germany.

Sporer, S. L., & Schwandt, B. (2006). Paraverbal correlates of deception: A meta-analysis. *Applied Cognitive Psychology, 20,* 421-446. doi:10.1002/acp.1190

Sporer, S. L., & Schwandt, B. (2007). Moderators of nonverbal indicators of deception: A meta-analytic synthesis. *Psychology, Public Policy, and Law, 13,* 1-34. doi:10.1037/1076-8971.13.1.1

Steller, M. (1989). Recent developments in statement analysis. In J. C. Yuille (Ed.), *Credibility assessment* (pp. 135-154). Dordrecht, Netherlands: Kluwer Academic Publishers.

Steller, M. (2013). Vier Jahrzehnte forensische Aussagepsychologie: Eine nicht nur persönliche Geschichte [Four decades of forensic psychology of testimony: Not only a personal story]. *Praxis der Rechtspsychologie, 23*, 11-32.

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

Undeutsch, U. (1967). Beurteilung der Glaubhaftigkeit von Zeugenaussagen. In U. Undeutsch (Ed.), *Handbuch der Psychologie, Band 11: Forensische Psychologie* (pp. 26-181). Göttingen, Germany: Hogrefe.

Undeutsch, U. (1982). Statement reality analysis. In A. Trankell (Ed.), *Reconstructing the past: The role of psychologists in criminal trials* (pp. 27-56). Deventer, Netherlands: Kluwer.

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

Zeit Online (2015, September). *Springer-Verlag muss Rekordentschädigung zahlen [Springer publishers has to pay record compensation].* Retrieved from http://www.zeit.de/gesellschaft/zeitgeschehen/2015-09/joerg-kachelmann-bildzeitung-urteil-schmerzensgeld on November 14th 2015.

Zhou, L., Burgoon, J. K., Nunamaker, J. F., & Twitchell, D. (2004). Automating linguistics-based cues for detecting deception in text-based asynchronous computer-mediated communication. *Group Decision and Negotiation, 13*, 81-106. doi:10.1023/B:GRUP.0000011944.62889.6f

META-ANALYSIS I:

Are Computers Effective Lie Detectors?

A Meta-Analysis of Linguistic Cues to Deception

Deception is an ubiquitous phenomenon, and people at all times have sought to find ways to detect it. Humans have searched for indicators of deception in physiological, nonverbal and paraverbal behavior, and the very content of what people are saying. Since the beginning of experimental psychology, researchers have systematically investigated different types of cues assumed to reveal deception (Benussi, 1914; Freud, 1905; Wertheimer & Klein, 1904; see Bunn, 2012; Grubin & Madsen, 2005; Sporer, 2008, for historical reviews).Despite these efforts, meta-analyses indicate that humans are not very good at discriminating between truths and lies (Bond & DePaulo, 2006). Reasons may lie in the complexity and difficulty of the task, incorrect beliefs about cues and the use of invalid cues, as well as the pervasive biases in decision making (Global Detection Research Team, 2006; Reinhard, Sporer, Scharmach, & Marksteiner, 2011; Vrij, 2008b).

In this meta-analysis, we focus on the use of computers to overcome these limitations. However, we unpretentiously believe the present contribution goes far beyond this goal. Based on a series of theoretical frameworks rooted in cognitive and social psychology, we posed (and tested) specific directional hypotheses concerning the potential utility to detect deception with a number of linguistic cues. Our findings are relevant not only in terms of the potential practical utility of computers to detect deception, but also in terms of basic knowledge about the language of deception and the underlying theories predicting specific linguistic differences between truths and lies.

Human Judgmental Biases

Humans are biased lie detectors. Biases include a reliance on cognitive heuristics (Levine & McCornack, 2001), overestimation of dispositional factors (O'Sullivan, 2003), and an exaggerated focus on nonverbal relative to verbal content cues (Reinhard et al., 2011; Vrij, 2008b). Other researchers have shown that humans are prone to truth or lie biases (Levine, Park, & McCornack, 1999; Meissner & Kassin, 2002; Zuckerman, Koestner, Collela, & Alton, 1984), which are the tendency to judge statements as truthful--or as deceptive--regardless of their actual veracity. It has also been shown that observers' veracity judgments are affected by factors unrelated to the veracity of particular statements, such as the sender's facial appearance (Masip, Garrido, & Herrero, 2003). Likewise, Bond and DePaulo (2006) argue that people hold the stereotype that liars are "tormented, anxious, and conscience stricken" (p. 216), and that they may draw on this stereotype when judging the veracity of other people.

As a possible remedy to overcome these deficiencies in human judgments, physiological psychologists and brain researchers have utilized "machines" like the polygraph, voice stress analyzer, pupillometry, electromyogram, and brain imagery (e.g., EEG, fMRI) to detect deception. In the last 40 years, but particularly most recently, scientists from various fields have also sought to detect deception by analyzing speech content with computers, looking for specific word cues or sentence structures to reveal deception.

A computer system would arguably be less prone to the influence of biases and stereotypes than human judges. There would be virtually no top-down processing. Additionally, online assessment of various deception cues from ongoing interactions or videos can tax the cognitive capacity of human judges and lead to

errors. Computers can quickly analyze large amounts of information and provide more reliable data. These are the principal reasons for the appeal of the automatization of lie detection. However, we must not forget that computers do not make choices about definitions of word categories nor about the specific words to be contained in broader categories. Most importantly, computers do not make choices about the *direction* of any particular cue as a lie or truth indicator. It is important to stress that, for a computer to be able to detect deception, the linguistic characteristics to be analyzed must be revealing of deception. Here, in examining what linguistic cues identified with computers differ between truths and lies, we also contribute to our basic understanding about linguistic markers of deception.

Can Computers be Useful to Detect Deception?

In an attempt to identify and quantify linguistic cues to deception, researchers had an (unrealistic) dream: Enter peoples' words into a computer to find out if they are telling the truth or not. In an early study, Knapp, Hart, and Dennis (1974) assessed several linguistic cues using a program called TEXAN on a CDC 6500 mainframe computer. The program analyzed word frequencies without taking contextual meaning into account. Most of the investigated cues significantly differed in the expected direction between truths and lies.

Many years passed until similar but more modern word frequency count approaches were used regularly to deception detection (at least in research contexts). The most common program, called Linguistic Inquiry and Word Count (LIWC; Pennebaker, Francis, & Booth, 2001), was developed to count words in psychology-relevant dimensions across multiple text files. LIWC has been used in numerous domains like personality, health, or psychological adjustment (see Tausczik & Pennebaker, 2010, for a review). LIWC analyzes typed or transcribed

accounts on a word-by-word basis, where each word is compared against a dictionary of 2000 pre-selected words allocated to 72 linguistic categories. Although LIWC was not specifically designed to assess deception, Newman, Pennebaker, Berry, and Richards (2003) used it to calculate the percentages of specific linguistic cues in true versus deceptive statements, yielding above-chance accuracy of classifications for different types of lies. Subsequently, researchers from a variety of fields have also applied LIWC with the same purpose (see Appendix C).

Other researchers realized that the methods used ought to be more complex. As a result, specialized programs and algorithms have been developed which are oriented more directly to detecting deception. For example, Agent99Analyzer was created to specifically detect (linguistic cues to) deception in texts and videos (Fuller, Biros, Burgoon, Adkins, & Twitchell, 2006). One of its sub-tools is a natural language processing unit called “GATE” (General Architecture for Text Engineering; Cunningham, 2002; Qin, Burgoon, Blair, & Nunamaker, 2005). Other related automated text-based tools used were “iSkim” or “CueCal” (Zhou, Booker, & Zhang, 2002; Zhou, Burgoon, Nunamaker, & Twitchell, 2004). More specifically, smaller text units are analyzed and integrated in the context of the whole text through examining different levels of human language (e.g., sub-sentential, sentential and discourse processing; see also Zhou et al., 2004). Recently, a growing body of research using machine-learning approaches of natural language processing emerged to detect linguistic cues to deception (Nunamaker, Burgoon, Twyman, Proudfoot, Schuetzler, & Giboney, 2012).

A highly sophisticated program of this kind called “Coh-Metrix” (Graesser, McNamara, Louwerse, & Cai, 2004; McNamara & Graesser, 2012), goes beyond word frequency analysis. Specifically, in analyzing “cohesion relations”, Coh-Metrix

takes into account meaning and context in which words or phrases occur in texts (http://cohmetrix.memphis.edu). Although not specifically developed to detect deception, Coh-Metrix was recently applied for this purpose (e.g., Bedwell, Gallagher, Whitten, & Fiore, 2011). A somewhat different detection deception software called Automated Deception Analysis Machine ("ADAM"; Derrick, Meservy, Burgoon, & Nunamaker, 2012) focuses on editing processes while typing messages (e.g., backspace, delete, or spacebar) and measures response latencies. The program includes an automated interviewer asking questions from an internal script.

Taken together, various computer programs from different research areas and labs originated in the last 15 years that were either applied to detecting deception or specifically developed for this purpose. The effectiveness of such programs can be better determined with a comprehensive and integrative quantitative analysis of the results on various linguistic cues to deception. This is the focus of the current meta-analysis.

The Importance of Theory

Is this dream of automated lie detection realistic? A quick preview of our results hints to the fragmented nature of the findings from computer studies. Effect sizes in our meta-analysis were coded in a way that *positive* g_us are indicative of *truth*, while *negative* g_us are indicative of *deception*. For 1,093 effect sizes we calculated for 79 linguistic cues, we obtained an approximately normal distribution centering on a mean effect size of g_u = -0.01 (*SD* = 0.37), and a *Mdn* of 0.02. The effect sizes ranged from -1.95 to 1.43 and the first and third quartiles were -0.17, and 0.20, respectively. To get a more accurate picture of the diagnostic usefulness of linguistic markers of deception, we calculated the *absolute* magnitude of all effect sizes, assuming that all were in the expected direction as predicted by a-priori

specified hypotheses (Figure 1). The average absolute effect size was 0.26 (*SD* = 0.26) with a *Mdn* of 0.19 (first quartile = 0.09, third quartile = 0.34). This average effect size denotes the maximum possible mean of all cues *if* the results had actually been in the direction predicted. This mean effect size implies that across all studies and cues only small effect sizes were obtained. This suggests that without a-priori theoretical predictions, computer analyses of linguistic cues to deception are a futile exercise. Can larger effect sizes be observed if we classify cues into theoretically meaningful categories and consider possible moderators?

Theoretical Approaches Used to Predict Linguistic Cues to Deception

We cannot provide an exhaustive review of all approaches taken by different research groups. Some authors may prefer to emphasize the role of emotion, arousal and motivation, while communication researchers may look at deception as strategic behavior. We will address some of these alternative interpretations where appropriate. Instead, we focus more on a cognitive and memory-oriented approach, supplemented by social psychological considerations and self-presentation, which help us to pin down the differences in processes involved in telling true stories vs. lies. Hence, we focus on four viewpoints resulting in six research questions: (1) Recalling an experience from episodic memory vs. constructing a lie from semantic memory. Constructing a lie may be more cognitively taxing (Research Question 1) and reduces the certainty with which lies are delivered (Research Question 2). (2) Again drawing on the literature on memory, we discuss the role of emotion and affect in recall of true experiences vs. reporting lies (Research Question 3). (3) We discuss the role of the self as an organizational principle as well as self-presentational strategies and the role of immediacy in communication (Research Question 4). (4) We draw on the reality monitoring framework to derive predictions about sensory and

perceptual cues (Research Question 5) and cognitive operations (Research Question 6).

For each question we noted those linguistic cues that would elucidate differences between accounts of truth-tellers and liars, clearly specifying the direction of effect for each cue. Some of the theoretical approaches we discuss elaborate retrieval and construction processes truth-tellers engage in when reporting an event while others focus on lie construction. Furthermore, we developed clear operational definitions for each cue in order to provide consistency in the names and definitions used in different research areas (see Appendices A and B). Most cues investigated could be allocated to one of the six research questions. However, because some cues did not clearly fit in any theory or research question, they were relegated to the miscellaneous question category. Following are the principal research questions.

Research Question 1: Do Liars Experience Greater Cognitive Load?

Telling a lie can be more cognitively demanding than truth-telling, because it involves the execution of a number of concurrent tasks requiring a great deal of mental resources. In general, both liars and truth-tellers must tell a plausible and coherent story that does not contradict their own former statements or facts the observer/interviewer may know about. Also, in some cases lying requires suppressing thoughts about the truth (Gombos, 2006); this may inadvertently preoccupy the speaker’s thinking (Pennebaker & Chew, 1985; see also Lane & Wegner's, 1995, model of secrecy). Further, as communication researchers have emphasized, storytellers must monitor their own behaviors and observers' reactions (Buller & Burgoon, 1996). Truth-tellers may also engage in some of these cognitive processes but for liars this task is more difficult because they cannot easily draw on episodic memories. Instead, they must rely on the semantic memory system or on

rather nonspecific scripts or schemata (Schank & Abelson, 1977; Sporer & Küpper, 1995).

When constructing a lie, a convincing scenario has to be communicated. However, due to the demands for cognitive resources, a lie may not include the complexities and richness of information that characterize reports of real experiences. In contrast, telling a story about a true event relies on retrieval of experienced events. Although this typically involves reconstruction, and may at times even take increased effort, recall of episodic memories and supporting details is generally rather automatic.

Much research on the cognitive load approach has not been grounded on well-articulated cognitive models of deception (Blandón-Gitlin, Fenn, Masip, & Yoo, 2014). Yet, a few such models have been proposed to specify cognitive processes involved in lie production (for reviews, see Gombos, 2006, and Walczyk, Igou, Dixon, & Tcholakian, 2013). Some of these models (Sporer & Schwandt, 2006, 2007; Walczyk, Schwartz, Clifton, Adams, Wei, & Zha, 2005; Walczyk, Harris, Duck, & Mulay, 2014; Walczyk et al., 2013) have invoked Baddeley's (2000, 2006) working memory model, which involves transferring information from long-term memory to an episodic buffer in working memory. While this should facilitate truth-telling, it should also make lying more difficult (see, e.g., Walczyk et al., 2005, 2013, 2014).

Does research support the cognitive load assumptions? Numerous recent studies (for review, see Vrij & Granhag, 2012) have provided indirect evidence by experimentally increasing a storyteller's task demands. This has elicited more discernable cues to deception than in control, lower cognitive load conditions. Note, however, that manipulating "cognitive load" is not equivalent to assessing the cognitive mechanisms postulated as a function of such manipulations (Blandón-Gitlin

et al., 2014). More direct (and revealing) evidence comes from behavioral studies using response latencies and other indices of cognitive load (e.g., Debey, Verschuere, & Crombez, 2012; Johnson, Barnhardt, & Xhu, 2004; Walczyk et al., 2005; for a summary, see Walczyk et al., 2013). There is even evidence from brain imaging studies (e.g., Abe, 2009; Christ, Van Essen, Watson, Brubaker, & McDermott, 2009) showing that telling lies, particularly those involving short responses, requires greater involvement of and access to key mental resources than truth-telling (Gamer, Bauermann, Stoeter, & Vosse, 2008).

Cues to deception theoretically connected to the cognitive load perspective have been found in previous meta-analyses, particularly for nonverbal and paraverbal behaviors (DePaulo, Lindsay, Malone, Muhlenbruck, Charlton, & Cooper, 2003; Sporer & Schwandt, 2006, 2007). In comparison with truth-tellers, liars had longer response latencies, tended to communicate shorter stories, made more speech errors, nodded less, and displayed fewer hand, foot, and leg movements. Particularly relevant for the analysis of linguistic markers are findings on verbal content cues that demonstrate that compared to true accounts, deceptive accounts appear less plausible, coherent and detailed while including more phrase and word repetitions. These indices can be signs of the experience of cognitive load either from a taxed system (e.g., longer response latencies) or because of liars' strategies to reduce cognitive load (Walczyk, Mahoney, Doverspike, & Griffith-Ross, 2009).

Predictions. From a cognitive load/working memory perspective, we predict that compared to true accounts, false accounts will be (a) shorter as indicted by *word* and *sentence quantity* cues, (b) less precisely elaborated as indicated by fewer *content words* (expressing lexical meaning), a lower *type-token ratio* (number of distinct content words, e.g., *house, walk, mother*) divided by total number of words),

and *shorter words* (i.e., less than six letters; average word length), (c) involve less complex stories as indicated by fewer *verbs,* fewer *causation* words (*because*, *effect*, *hence*) and fewer *exclusive words* (*but*, *except*, *without*), and (d) include more *writing errors* (possibly moderated by mode of production [orally telling a lie, hand writing, or typing]). (For a list of the operational definition of all cues included see Appendices A and B.)

From a different perspective, based on DePaulo's self-presentational perspective (DePaulo et al., 2003), one would expect that liars are less likely than truth-tellers to take their credibility for granted and therefore may take a greater effort and deliberately edit their communication (cf. Derrick et al., 2012). Note, however, that this editing process will also usurp cognitive resources detracting from successful lie constructions.

Research Question 2: Are Liars Less Certain Than Truth-Tellers?

DePaulo et al. (2003) contend that deceptive self-presentations are not as convincingly embraced as truthful ones. This may be a result either of the speakers' moral scruples, which may lead them to feel guilty or ashamed when lying, or of liars not having as much personal investment in their claims as truth-tellers. The psychological closeness or distance between a speaker and his or her message might be reflected in language (Wiener & Mehrabian, 1968). Liars should display more linguistic markers indicative of psychological detachment than truth-tellers (Buller, Burgoon, Busling, & Roiger, 1996; Kuiken, 1981; Wagner and Pease, 1976; Zhou, Burgoon, Nunamaker, & Twitchell, 2004; Zhou, Burgoon, Twitchell, Qin, & Nunamaker, 2004). Indeed, in their meta-analysis DePaulo et al. (2003) found that liars were verbally and vocally less involved and more verbally and vocally uncertain than truth-tellers but observed no reliable differences for tentative constructs and

shrugs. Uncertainty words have been proposed as markers of psychological distance between a speaker and his or her account (e.g., Kuiken, 1981). Thus, liars' accounts should contain more uncertainty words than truth-tellers' accounts.

It may also be the case that deceivers withhold information not to give their lies away. Indeed, research shows that when lying to conceal their transgressions, people indicate that they try not to provide incriminating details (Hartwig, Granhag, & Strömwall, 2007; Masip & Herrero, 2013), and try to keep the story simple (Strömwall, Hartwig, & Granhag, 2006) or vague (Vrij, Mann, Leal, & Granhag, 2010). DePaulo et al. (2003) found liars to be significantly more discrepant/ambivalent than truth-tellers. Therefore, liars might provide vague, ambiguous, or uncertain replies in order not to expose their lies (Buller et al., 1996; Cody, Marston, & Foster, 1984).

Predictions. From these perspectives, it is expected that liars will be less certain and definite than truth-tellers. Consequently, deceptive accounts should contain fewer *certainty words* (*always*, *clear*, *never*) and more *tentative words* (*guess*, *maybe*, *perhaps*, *seem*) and *modal verbs* (*can*, *shall*, *should*) than truthful accounts. (It should be noted that *modal verbs* also include the verb "must" that expresses more certainty and purposiveness whereas all other modal verbs indicate more uncertainty).

It may be argued that liars are aware that uncertainty indicates deception and thus may strategically incorporate certainty indicators to evade detection (e.g., Bender, 1987). However, research does not support this contention. To our knowledge, around ten reports have been published so far on liars' and truth-tellers' strategies to be convincing (for a brief review, see Masip & Herrero, 2013). Only rarely has certainty (or any related construct) emerged as a strategy, and in these

instances it has been mentioned (a) only infrequently, and (b) equally often by liars and truth-tellers (e.g., Hines, Colwell, Hiscock-Anisman, Garrett, Ansarra, & Montalvo, 2010: “admit uncertainty”; for an exception see Strömwall et al., 2006).

Research Question 3a: Do Liars Use More Negations and Negative Emotion Words?

Emotional approach.[1] When people lie, they may experience feelings of guilt and fear of getting caught (Ekman, 1988, 2001).[2] Even when telling everyday lies of little consequence, people report feeling uncomfortable (DePaulo et al., 2003). Vrij (2008a) also noted that liars might make negative comments or use negative words that reflect negative affect induced by guilt and fear.

Numerous studies have shown that arousal is associated with specific emotions (see the meta-analysis by Lench, Flores, & Bench, 2011), some of which are likely to be experienced by liars, such as guilt and fear of punishment (Ekman, 2001; Zuckerman, DePaulo, & Rosenthal, 1981). These emotional states may elicit specific nonverbal and verbal cues to deception (see DePaulo et al., 2003; Sporer & Schwandt, 2006; Vrij, 2008a). Recent studies have used brain-imaging technology to specifically investigate the role of emotion in deception (for a review see Abe, 2011). For example, Abe, Suzuki, Mori, Itoh, and Fujii (2007) found that neural structures associated with heightened emotions were also uniquely associated with deceiving an interrogator, and that self-reported feelings of immorality (sense of sin) and anxiety were higher in deceptive conditions than in truth-telling conditions. These results support the notion that deception is associated with negative emotions.

Predictions. From an emotional approach perspective, we predict that compared to true accounts, lies will include (a) more *negation* words (*no*, *never*, *not*)

because these reveal a more defensive tone or denial of wrongdoing, which is likely to be accompanied by negative emotions of the liar, and (b) more words denoting overall *negative emotions (enemy, worthless, skeptic), anger (hate, kill, weapon), anxiety (unsure, vulnerable) and sadness (tears, useless, unhappy)*.

Research Question 3b: Do Liars Use Fewer Positive Emotion Words?

Research on autobiographical memory suggests that people's emotional appraisal of past events tends to be positively biased (Walker, Vogl, & Thompson, 1997). One mechanism by which this bias occurs is a tendency for emotions associated with negative-event memories to fade faster than emotions associated with positive-event memories (Walker, Skowronski, & Thompson, 2003). In a review of this research, Walker and Skowronski (2009) suggest that this *fading affect bias* leads people to generally remember events less negatively regardless of the original affect associated with the event. This effect is not due to forgetting of event details, as the accuracy of the memories is comparable for negative and positive events. It is the memory of the emotional intensity associated with the event that fades, with negative events fading at a faster rate than positive events.

Predictions. Because truth-tellers have a specific memory of the event, whereas liars cannot draw on such an episodic memory, we predict that compared to true accounts, lies will contain fewer words denoting *positive emotions* (happy, pretty, good) or *feelings (luck, joy)*.

Research Question 3c: Do Liars Express More or Less Unspecified Emotion Words?

Many researchers from different fields, such as social psychology, psychology and law, or computer linguistics (e.g., Ali & Levine 2008; Fuller et al., 2006, Newman

et al., 2003), have investigated the frequency of occurrence of emotional and affective terms in true and deceptive accounts without taking the valence of these emotions into account. Therefore, we decided to also investigate the cues of *unspecified emotions* (positive and negative) and *pleasantness* or *unpleasantness* of the story despite the lack of theoretical specification of the direction in the original studies. Predictions could be derived from a social psychological perspective. Depending on the seriousness of a lie, from a trivial lie in everyday life to high stake lies, the situation may become increasingly emotional. Hence, one would predict higher frequencies of *unspecified emotion* words in lies than in truths.

Research Question 4: Do Liars Distance Themselves More From Events?

In the preceding section, we have assumed that people are more likely to experience different types of negative emotions when telling a lie. Given such negative experiences and emotions, from DePaulo et al.'s (2003) self-presentational perspective we further assume that liars will distance themselves more from the story being told, and, relatedly, will be less forthcoming than truth-tellers (see also Research Question 2 on certainty cues above). Possible linguistic indicators for this assumption are personal pronouns, cues to responsibility and verb tense shifts. To clarify the predictions of specific cues we present them within the theoretical accounts of immediacy, self-organization, egocentric bias, and narrative conventions.

Immediacy. A possible way to express ownership and take responsibility for an action or event is to tell a story from a first-person perspective, where the sender is reporting an event where he/she is the actor, not an observer-bystander. Evidence for this assumption comes from the long tradition of research on verbal and nonverbal communication which has investigated *immediacy* as a cue to truthful

messages (Cody et al., 1984; Knapp et al., 1974; Kuiken, 1981; Mehrabian, 1972; Wagner & Pease, 1976; Wiener & Mehrabian, 1968; Zhou, Burgoon, Nunamaker et al., 2004; Zhou, Burgoon, Twitchell et al., 2004). In these studies, one aspect of immediacy has been operationalized as the psychological distance between the speaker and his/her communication. More specifically, immediacy can indicate the degree to which there is directness and intensity between the communicator and the event being communicated (Wiener & Mehrabian, 1968, p. 4). Taking this aspect of the definition of immediacy, deception researchers consider nonimmediacy as an indicator of deceptive communication by way of the speaker distancing from his/her own statement (e.g., Buller et al., 1996; Kuiken, 1981; Wagner & Pease, 1976; Zhou et al., 2004).

However, evidence for nonverbal and verbal indicators of the relationship between immediacy and deception is mixed. In the meta-analysis by DePaulo et al. (2003) there were no significant effects for self- or other-references, but more general indices of *verbal* immediacy (all categories) as well as verbal and vocal immediacy (impressions) were observed significantly more frequently or to a higher extent in truthful than fabricated messages. This latter effect appeared to be stronger when immediacy was measured subjectively than when assessed via more objective measures.

The self as an organizational structure. Another line of research we consider is social psychological theorizing on social memory, which has emphasized the role of the self as an organizational structure. In fact, one of the primary distinctions between episodic and autobiographical memory is that the self provides an organizing principle, which relates experiences to one's self-schema. Experimental evidence comes from research on the self-reference effect (Rogers,

Kuiper, & Kirker, 1977), which demonstrated that information is particularly well remembered when it has been encoded in relation to oneself, or when the person plays an active, rather than passive role (e.g., Slamecka & Graf, 1978). Variations on this theme are discussed under ego-defensive, self-serving, egocentric or egotistic biases (see Greenwald, 1980). Greenwald (1980) has gone as far as referring to the self as a "totalitarian ego" that puts itself in the foreground, assuming a central role and ownership when talking about self-experienced past events and actions. This prevailing tendency should lead to more frequent uses of first-person pronouns (*I*, *me*, *we*, *us*, *our*, etc.) when telling the truth relative to lying.

However, while the *egocentric bias* may play a role when reporting (complex) autobiographical events, it may be restricted to positive outcomes, and reversed for negative outcomes (Greenwald, 1980). Also, the so-called “better than average effect” refers to the tendency to evaluate oneself more favorably than an average peer (e.g., Brown, 2012). For instance, 70% of high school seniors estimated that they had above average leadership skills, whereas only 2% said their leadership skills were below average (College Board, 1976–1977). Another example of the positive outcome bias is a classic study by Bahrick, Hall, and Berger (1996; see also Bahrick, 1996) who found that students accurately recalled better high school grades than worse ones. Relatedly, in a classical study on the self-enhancing bias by Cialdini, Borden, Thorne, Walker, Freeman, and Sloan (1976, Experiment 2) college students not only donned their school colors on Monday after their team had won, but also identified, or distanced, themselves by use of different personal pronouns ("we won"; "they lost"). This suggests that first-person pronouns and statements of personal responsibility will be more prevalent among truth-tellers than liars, particularly for positive outcomes.

Predictions. In summary, from different theoretical perspectives we assume more frequent use of *first-person pronouns*, and less frequent use of *third-person pronouns* for reports of self-experienced events. Self-experienced events should also be characterized by more statements of own responsibility, at least for positive outcomes. This prediction is more likely to hold for *first-person singular* than *first-person plural* because the plural may designate both the group the storyteller belongs to, and identifies with, as well as a communication partner who acts as an antagonist in an interaction (e.g., "we quarreled"). Thus, with plural pronouns, ownership and responsibility are less clear-cut than with singular pronouns. On the other hand, *passive voice* or *generalizing terms* in phrases like "one has to..." or “everybody does this...” signal less personal involvement and hence should be found more frequently in lies than truthful accounts.

Narrative conventions and verb tense shifts. Communication about past events follow narrative conventions (acquired during childhood) that require the storyteller to talk about who, what, when, where, and why (Brown & Kulik, 1977; Neisser, 1982) and to adhere to a temporal structure (Bruner, 1990). Anecdotal evidence from research on autobiographical memory for significant life events shows that people sometimes switch from telling a story in the past tense to the present tense at crucial moments of the event (Pillemer, Desrochers, & Ebanks, 1998). In many of these examples, it appears that the protagonist is reliving the past event, describing his or her sensory and perceptual experiences, making the accounts to appear more vivid (cf. the reality monitoring approach described in Research Question 5). Although present tense may be less concrete than past tense when it refers to repeated or routine actions (e.g., “I [usually] go to church on Sunday” versus “I went to church on Sunday”), when talking about a specific past event

present tense is more vivid than past tense. Whether verb tense shifts occur involuntarily or unconsciously, or are strategically used by skillful storytellers (like fiction writers) to communicate intensity and feeling to a recipient, cannot be answered by these archival type studies, nor by our meta-analyses.

Predictions. We expect reports of true events to be more likely to contain *present tense verbs* than lies, at least in accounts of personally significant events. For other types of lies, this prediction may not hold. The live character of these narratives may also diminish with repeated retellings of a story. Conversely, lies should contain more *past tense verbs* than true accounts.

Research Question 5: Do Liars Use Fewer (Sensory and Contextual) Details?

Reality monitoring framework applied to deception. The reality monitoring model by Johnson and Raye (1981) describes how individuals differentiate between externally generated memories of actual experiences versus memories of internally generated events that involve thoughts, fantasies, or dreams. In contrast to imagined events, experienced events are encoded and embedded in memory within an elaborate network of information that typically includes more perceptual details, contextual and semantic information. Conversely, internally generated memories are characterized by cognitive inferences or reasoning processes.

People differentiate between their own external and internal memories on the basis of these phenomenal characteristics (Johnson, Hashtroudi, & Lindsay, 1993), and similar features are also useful to differentiate between accounts of external and internal memories of *other people* (an attribution process that has been tagged “interpersonal reality monitoring”; Johnson, Bush, & Mitchell, 1998; Johnson & Suengas, 1989; Sporer, 2004; Sporer & Sharman, 2006).

Deceptive accounts can be characterized as representing internally generated memories, because in a deceptive situation people imagine the event at the time of its construction (Sporer, 2004). Even if people lie by borrowing from actual experience, the time and place or the context in which the event occurred may be changed during construction (Sporer, 2004; Vrij, 2008a). Therefore, even partially true deceptive accounts may lack the typical characteristics of true accounts. With these considerations in mind, researchers have extrapolated from the reality monitoring model to make predictions about specific sets of criteria that may discriminate between true and deceptive accounts (e.g., Granhag, Strömwall, & Olsson, 2001; Sporer, 1997; for reviews see Masip, Sporer, Garrido, & Herrero, 2005; Sporer, 2004; Vrij, 2008a). DePaulo et al.'s (2003) meta-analysis, which only included a few studies available then, showed small and nonsignificant effects sizes for reality monitoring criteria. However, in a more comprehensive review of studies, Masip et al. (2005) found that some of the reality monitoring criteria involving perceptual processes, contextual (including time) information, and realism/plausibility of the story were useful to discriminate between truth and deception.

Predictions. From a reality monitoring perspective, we predict that compared to true accounts, false accounts will (a) contain fewer perceptual details as indicated by *sensory and perceptual word* cues (*taste, touch, smell*), (b) be less contextually embedded as indicated by *space* (*around, under*) *and time word* cues (*hour, year*), and (c) include fewer descriptive words as indicated by *prepositions* (*on, to*), *numbers* (*first, three*), *quantifiers* (*all, bit, few*), *modifiers* (*adverbs and adjectives*), *and motion verbs* (*walk, run, go*). This latter set of cues involves words that describe events and actions in the story in more specific terms (e.g., "I took every short cut to get to work"). The lack of these words (e.g., "I went to work") would make the

account seem less real or vivid as would be predicted from the reality monitoring perspective (Sporer, 1997, 2004).

Research Question 6: Do Liars Refer Less (yes, Less!) Often to Cognitive Processes?

The reality monitoring approach, unlike other verbal-content cues based credibility assessment procedures, such as Criteria-Based Content Analysis (CBCA, Steller & Köhnken, 1989), does not only contain "truth criteria" (e.g., spatial and time details), but also one lie criterion. Specifically, reality monitoring predicts that references to internal processes at the time of the event (cognitive operations like reasoning processes) should be more likely contained in imagined than in self-experienced events. Applied to detecting deception, researchers have consequently postulated that references to cognitive operations can be used as a lie criterion (Sporer, 1997; Vrij, 2008a).

However, empirical evidence regarding this proposition is mixed. Perhaps, depending on the operationalization of this construct, some studies have found more references to cognitive operations in *lies* (e.g., Vrij, Akehurst, Soukara, & Bull, 2004), many studies have found *no differences* (e.g., Sporer & Sharman, 2006; 14 out of 19 studies reviewed in Vrij, 2008a), and some studies have found reliably more references to internal processes (like memory processes and rehearsal as well as thoughts) in *true* accounts (Granhag, Strömwall, & Olsson, 2001; Sporer, 1998; Sporer & Walther, 2006; Vrij, Edward, Roberts, & Bull, 2000).

From a different perspective, some thirty years of research on autobiographical memory has emphasized the associative nature of memories. Recollecting (personally significant) life events involves not only the conscious utilization of retrieval cues but also cross-referencing to supporting memories related

to the event in question. It also involves rehearsal processes, which are important determinants of remembering (Conway, 1990). These processes can also be subsumed under cognitive operations. To the extent that studies on deception involve complex (autobiographical) events, like being questioned about a crime or reporting an alibi, such retrieval processes and supporting memories (cf. the Criteria-Based Content Analysis criterion "External Associations") are likely to be used and mentioned when recalling true events (e.g. “I know it was the day before Easter because Good Friday was my birthday.”).

Finally, there is empirical evidence from several studies that cognitive operations are *positively* correlated not only with other reality monitoring criteria (Sporer, 1997, 2013) but also with many Criteria-Based Content Analysis criteria like "External Associations", "Own Psychological Processes", "Spontaneous Corrections" or "Doubts about one's own Testimony", loading on a common underlying factor (Sporer, 2004, Table 4.4). All of these criteria are assumed to indicate truthfulness.

Predictions. Consequently, we predict that linguistic cues referring to cognitive operations including memory processes are more likely to be found in truths than in lies. The two cues under this research question are *cognitive processes* (*cause, ought*), and *insight words* (*think, know, consider*).

Miscellaneous Category

Because many linguistic cues were investigated without a specific theoretical background or directed predictions, we created a miscellaneous category including linguistic cues analyzed in more than five studies (e.g., *inhibition*, *social processes*, *health, sports*; see Appendix B).

Hypotheses for Moderator Variables

It would be unwise to assume that the above predictions will hold across all types of lies, motivation, level of interaction, production mode, and other contextual factors. Hence, we conducted a series of moderator analyses within the theoretical frameworks provided above.

Event type and personal involvement. Across studies, senders described events or attitudes that differed in terms of personal involvement. We organized the studies into three categories. In the "Attitude/liking" paradigm, senders described their attitude towards a specific topic or person they like or dislike. In the "First-person experience" paradigm senders experienced a staged event or mock crime, described a personal life event, or were involved in a real criminal case. Lastly, the "Miscellaneous" category included studies where participants solved a problem, performed a specific task, or described a video scene.[3] We do acknowledge, however, that some attitudes/liking studies may also reflect high involvement but this would work against our hypothesis.

We argue that the higher the personal involvement in the event the higher the cognitive load (for example, due to a preoccupation with an interaction partner's reactions) and arousal (negative or unspecified emotions) will be when telling a lie. Also, liars might express more uncertainty terms or try to distance themselves more from events when their personal involvement is high. In other words, we expect the effects under Research Questions 1, 2, 3a, 3c, and 4 to be larger for the "First-person experience" compared to the "Attitude/liking" or the "Miscellaneous" paradigms in the aforementioned direction.

Emotional valence. The topics or events senders were asked to talk about were classified as positive (e.g., holidays), neutral (e.g., task performance), or

negative (e.g., confession of wrongdoing) in nature. If we assume that more negative emotions accompany telling a negative rather than a neutral event, liars should express even more negative emotion words when the event is negative (Research Question 3a). Also, we assume that the amount of unspecified emotion words (Research Question 3a) will be higher when the event is *not* neutral. Moreover, cognitive load might also be higher because senders have to deal with additional negative emotions that may induce concern, leading to a decrease in *word count* and *diverse* and *exclusive words* (Research Question 1--Cognitive Load: cues 01, 02, 03).

Also, if liars are more negatively involved in their story, they could appear more uncertain (Research Question 2--Certainty) and try to distance themselves more using less *self-* and more *other-references* (Research Question 4--Distancing). In summary, we hypothesized that effect sizes under Research Questions 1, 2, 3a, 3c, and 4 would be highest (in the expected direction) if the emotional valence was negative rather than neutral (or positive).

Intensity of interaction. The degree of interactions between the storyteller and another person varies widely in deception detection research (Vrij & Granhag, 2012). We differentiated four interaction levels: (a) no interaction: participants are only given a written or spoken instruction; (b) computer-mediated communication: participants are communicating via connected computers (e.g., only by *typing* words in studies included); (c) interview: interviewees are simply responding to questions from an interviewer (one-way direction); and (d) person to person interactions: sender and receiver are present in person and interacting bidirectionally.[4] We hypothesized that with increasing intensity of interactions from (a) to (d) (cf. Buller &

Burgoon, 1996), effects would become stronger under Research Questions 1, 2, 3a, 3c, and 4.

Motivation. Researchers varied the level of motivation for their senders to appear credible. Some researchers did not motivate their senders at all, some others tried to motivate them with incentives or written instructions, and still others used accounts from real criminal cases, where the motivation to appear credible must have been high due to real consequences for getting caught (*high-stake lies*; cf. DePaulo et al., 2003).

DePaulo and Kirkendol (1989, p. 54) postulated the *motivational impairment effect*, according to which highly motivated liars try to control their expressive behaviors to appear credible, but they are only successful in doing so with their *verbal* behavior, while their *nonverbal* behavior appears disrupted. In other words, liars' nonverbal behavior should be impaired whereas their verbal behavior (i.e., the content of messages) should be improved. DePaulo, Lanier, and Davies (1983) provided support for these hypotheses, as highly motivated liars were easier to detect in the visual or audiovisual conditions, but less successfully detected in the verbal (transcript) condition (there was no difference in the audio-only condition).

Assuming that the motivational impairment effect also applies to linguistic cues as a form of verbal behavior, we hypothesized that highly motivated liars might try harder to control their words, so differences between liars and truth-tellers should become smaller under Research Questions 1, 2, 3a, 3c, and 5.

Production mode. Participants' accounts were either handwritten, typed on a keyboard, or spoken (and audio- or videotaped). Horowitz and Newman (1964) proposed that, in general, speaking is easier than writing, because speakers have more liberty and feel less inhibited than writers. Also, writing involves more

deliberateness (see also Hancock, Woodworth, & Goorha, 2010) and more serious commitment. Horowitz and Newman found support for their hypothesis in that speaking is more productive and elaborative than writing. This resulted in more words, more phrases and more sentences when speaking than when writing. More recently, Kellogg (2007) hypothesized that writing is slower and less practiced than speaking and thus results in higher demands on working memory. He found that accounts of a recalled story were more complete and more accurate when spoken than written (cf. also Sauerland & Sporer, 2011).

Hence, we hypothesized that liars produce even fewer *words*, *diverse words*, and *sentences* (Research Question 1--Cognitive Load) when writing than speaking due to an increased cognitive load and decreased working memory capacity. Furthermore, liars should also use fewer (*sensory* and *contextual*) *details* when writing than speaking compared to truth-tellers (Research Question 5; see Elntib, Wagstaff, & Wheatcroft, 2014, for a recent empirical investigation of this issue). Regarding emotion-related cues (under Research Questions 3a, and 3c), we hypothesized that liars use more *negative* and *unspecified emotion words* than truth-tellers when speaking than when writing, because emotions might be expressed more directly and frequently in direct speech.

An empirical issue for studies involving writing is whether handwriting or typing comes easier. Therefore, we separated written accounts into hand-written vs. typed for our moderator analysis. Unfortunately, we do not know the level of typing skill of participants.

To sum up, differences between liars and truth-tellers should be more pronounced in written (typed or handwritten) compared to orally given accounts for linguistic cues under Research Questions 1 (cognitive load) and 5 (details), whereas

for emotion-related cues (Research Questions 3a, and 3c), the effect sizes should be larger if stories were spoken than written.

Program type. Researchers from various fields used different computer programs to analyze deceptive and truthful accounts. The most common one is *LIWC*. Although it is a general program (i.e., not specifically designed to detect deception), we separated it from other general programs such as *Coh-Metrix* or *WordScan*. This is because *LIWC* was used in a disproportionally large number of studies. Other software, such as *Agent99Analyzer* or *Automated Deception Analysis Machine*, were specifically developed to detect deception. We hypothesized that studies applying deception-specific programs should yield stronger effects for any linguistic cue than studies using *LIWC* or any other general program based on simple word counts.

Publication status. The tendency that studies with nonsignificant findings are less likely to be written, submitted, and accepted for publication in peer-reviewed journals, is referred to as *publication bias* (Cooper, 2010; Sutton, 2009). In short, the publication of a study may partially depend on its results rather than on its theoretical or methodological quality (Rothstein, Sutton, & Borenstein, 2005). One method to statistically quantify a publication bias is to compare the effect sizes of published and unpublished studies (see Appendix E in supplemental online materials); another is to test for the association between effect sizes and sample sizes (Levine, Asada, & Carpenter, 2009).

Experimental design*.* We also assessed *experimental design* as a moderator (between- vs. within-participants), assuming larger effects for the latter (see results in Appendix F in supplemental online materials).

Goals of the Meta-Analysis

The main goals of our meta-analysis were (a) to provide a comprehensive set of operational definitions for each linguistic marker, (b) to offer an elaborate theoretical background in order to specify directed predictions for each cue, (c) to provide a quantitative and comprehensive synthesis of linguistic cues to deception assessed with computer programs obtained from interdisciplinary research areas, and (d) to analyze the influence of important theoretical and methodological moderator variables on the outcome of linguistic cues to deception.

Method

Inclusion and Exclusion Criteria

Studies had to meet the following eligibility criteria to be included in our meta-analysis: (1) Use of software to locate linguistic cues; (2) Reports of specific linguistic cues (not just paraverbal/paralinguistic or nonverbal or physiological cues); studies that reported word counts only, but no other linguistic cues were excluded[5]; (3) Independence of data sets: when analyses of the same data set of transcripts and cues were reported in multiple publications, we only included the source published in the journal with the highest publication standard [e.g., peer review] and excluded the other source(s) to ensure independence of all data sets; and (4) Sufficiency of data to calculate effect sizes (see *Effect Size Measure* section below). Furthermore, (5) whenever a field study with statements from real criminal cases met the aforementioned criteria (e.g., ten Brinke & Porter, 2012), special care was taken to assure ground truth had not been established solely on the basis of a court verdict, but in addition from more than one type of external and independent source of evidence (e.g., physical evidence, witness statements, confessions, etc.). However,

these studies should be treated with caution because linguistic aspects of the account may have affected the final case disposition (e.g., lie or truth).

Literature Search and Study Retrieval

As a first step, we searched through reference lists of most relevant studies or reviews (e.g., DePaulo et al., 2003; Newman et al., 2003; Tausczik & Pennebaker, 2010; Zhou et al., 2004). Next, several exhaustive literature searches were conducted from September 2011 to February 2012 in the most important psychological research literature databases, such as the Social Sciences Citation Index (with cited reference search), PsycInfo, Dissertation Abstracts, and Google Scholar, examining articles published between 1945 and February 2012.

The combination and permutations of four keyword clusters were used: (a) decept*, deceit, lie; (b) verbal, linguistic, language; (c) automatic, computer, software, artificial. These searches resulted in 948 published and unpublished articles, which were reduced to 394 after removing duplicates. Then, the inclusion and exclusion criteria were carefully applied. This reduced the number of articles to 99, from which we still had to exclude 54 for different reasons (Appendix G in supplemental online materials), mostly incomplete reporting of data necessary for our analysis. This resulted in 44 relevant data sets that met all inclusion criteria.

Linguistic Cues to Deception

A total of 202 linguistic cues were extracted from the articles and sorted based on their name and operational definition (if available). In some cases, we merged cues with different names that had very similar operational definitions. For example, *type-token ratio*, *unique words*, *lexical diversity*, or *different words*, were all

similarly operationally defined and refer to the same construct. We chose the name most commonly used (e.g., *type-token ratio* in the prior example).

All linguistic cues had to be calculated as a ratio of all other words (except raw frequencies of words, verbs and sentences), and had to be investigated in at least $k = 4$ hypothesis tests. This resulted in 79 linguistic cues of which 50 were allocated to one of the six research questions based on their content and theoretical meaning. The remaining 29 cues could not be allocated to a theory or one of the research questions, and were assigned to the Miscellaneous category. All linguistic cues, with all of their names and final operational definitions, are listed in Appendices A and B.

Effect Size Measure

As an effect size measure we used Hedges's g_u (Hedges, 1981; Borenstein, 2009; Lipsey & Wilson, 2001), an unbiased estimator of the standardized mean difference (Cohen's *d*). Here, it is the standardized mean difference of the average frequency or ratio for each linguistic cue between deceptive and true accounts. If a specific linguistic cue occurred more often during deception than truth, g_u has a *negative* sign. If it occurred more often during truth than deception, g_u was assigned a *positive* value. To calculate g_u, we coded means, standard deviations, and *n*s separately for deceptive and true stories. If this information was not given, other appropriate measures (*t*- or *F*-values with 1 degree of freedom in the numerator, or *p*-values) were coded (for formula collections see Borenstein, 2009; Lipsey & Wilson, 2001).

If no relevant statistical data were available, we e-mailed the researchers to request them. In some instances, there may be discrepancies between the effect sizes reported here and those in the original articles. Reasons for such differences are that some authors provided us with more (differentiated) data, that we

sometimes chose specific subgroups for the analyses, or calculated the average effect size across subgroups, as explained in more detail under *Meta-Analytic Techniques* below.

Independent Variables and Moderator Variables

After coding typical study characteristics (e.g., study ID, author names, year of publication, number of senders and gender, etc.), we coded for information that defined the moderator variables or further independent variables of potential interest. These were: Publication status (e.g., published, thesis, etc.), type of computer program (LIWC; other general programs like Wordscan, Microsoft Word, or Coh-Metrix; or specific programs like ADAM (Automated Deception Analysis Machine), Agent99Analyzer, GATE (General Architecture for Text Engineering), iSkim, CueCal, or Connexor), language of accounts, theory presented (if any), cue selection (a priori, reported all or significant cues only), age of the senders, experimental design (between- or within-participants), preparation time, event type, event valence, interaction between sender and receiver, mode of production, and type/level of motivation to lie successful.

Coding Procedures and Intercoder Reliability

Two trained raters coded all dependent and independent variables from the articles with a standardized coding manual. After discussing two articles as examples and agreeing on order of article review, each coder worked independently. For eleven continuous variables, inter-coder reliabilities were highly satisfactory, with all coefficients ranging from Pearson's r = .86 to r = 1.0 (except for preparation time: r = .77). For eight categorical variables, inter-coder reliabilities were excellent, with Cohen's *kappa* (Cohen, 1960) ranging from .75 to 1.0. For six additional categorical

variables, Cohen's *kappa* ranged from .51 to .67, which was still a fair to good agreement (according to Fleiss, 1981). The few disagreements were resolved by discussion between the two coders. Final coding decisions of the moderator variables for each study are displayed in Appendix C.

Meta-Analytic Techniques

Dependencies of effect sizes. In some studies, in addition to accounts' truth status, other independent variables were manipulated as *between*- or *within*-participants factors and the data were reported separately for these subgroups (e.g., Schelleman-Offermans & Merckelbach, 2010: high- vs. low-fantasy-proneness). In studies with additional *within*-participants factors, dependency was avoided by calculating effect sizes separately for each subgroup and averaging them to ensure that only one effect size per study per linguistic cue was included (Lipsey & Wilson, 2001). In two other studies, a second *between*-participants factor (Ali & Levine, 2008: denials or confessions; Qin et al., 2005: text-chat, audio, face-to-face) was examined; here we included each of these subgroups (with *different* stimulus persons) as independent data sets.

Superordinate categories and sub-cues. Sometimes a linguistic category of cues had differentiated effect sizes that seemed to represent a single construct. As an example, we defined cue 19 with the superordinate category ("umbrella term") *positive emotions and feelings* including results from *positive emotions only* and *positive feelings only*. In studies using LIWC 2001, positive feelings and positive emotions/affects are treated as two different linguistic cues--and the data are reported separately for each (in LIWC 2007, they are combined). To ensure that only one effect size per construct is included, we combined sub-cues to a superordinate category (by averaging their effect sizes). However, we also calculated separate

meta-analyses for each of these sub-cues (here: cue 19.1 *positive emotions only* and 19.2 *positive feelings only*) to investigate whether the results are more differentiated, or if merging these cues was justifiable. The same procedure was applied to cue 18 *negative emotions* and to cue 28 *sensory-perceptual processes* (see Table 1). These superordinate categories make results from LIWC more comparable with studies using other computer programs that did not differentiate between different sub-cues (e.g., *anger*, *anxiety*, *sadness*).

Weighted average effect size. For each of the 79 linguistic cues, individual meta-analyses under the fixed-effects model (Lipsey & Wilson, 2001; Shadish & Haddock, 2009; Sporer & Cohn, 2011) were calculated. Average effect sizes were weighted by the inverse of the variance to give more weight to studies with larger samples. For six studies the total number of accounts was extremely large. To avoid unjustified extra-ordinary large weights we adjusted the number of total accounts for these studies (see *Results* section).

Homogeneity of effect sizes. We report both the homogeneity test statistic *Q* (Lipsey & Wilson, 2001) and the descriptive homogeneity statistic I^2 (Higgins & Thompson, 2002; Shadish & Haddock, 2009). In rare cases where I^2 resulted in a negative value, it was set to 0. In case of heterogeneity, outlier and moderator analyses were conducted.

Outlier analysis. To test for the presence of outliers, we applied the two methods recommended by Hedges and Olkin (1985, Chapter 12, and programmed by the fourth author). The number of outliers did not exceed 15% of the total number of effect sizes to avoid an artificial restriction of the variance between effect sizes. If outliers were detected, we calculated each meta-analysis with and without the outliers as sensitivity analyses (Greenhouse & Iyengar, 2009). Due to space

limitations, we only report results without outliers in Table 1 (results with and without outliers are displayed in Appendix H in supplemental online materials).

Moderator analyses. We used categorical variables as potential moderators with Hedges's analogue to ANOVA (Hedges, 1982; Lipsey & Wilson, 2001). Moderator analyses were only conducted if the homogeneity statistic was significant and if an individual meta-analysis of a specific linguistic cue contained enough hypothesis tests to avoid empty cell sizes and to increase power. Moderator analyses were only conducted without outliers to prevent biased results. To clarify potential confounds between moderator variables, we calculated their intercorrelations as well as all two-way and three-way cross-tabulations for each variable combination, to avoid empty or low frequency cells. As a consequence, only moderator analyses for $k \geq 13$ hypothesis tests are reported.

Computer-software for calculations. For computing individual effect sizes, weights and confidence intervals, formulae were programmed in Microsoft Office Excel (2011) spreadsheets by the first and fourth author. Calculations of meta-analyses and outlier analyses were conducted using Excel spreadsheets programmed by the fourth author and cross-validated using Lipsey and Wilson's (2001) SPSS macros (Wilson, 2002). Moderator analyses were also conducted using these macros.

Results and Discussion

Study Characteristics

We included $k = 44$ independent studies or data sets (see Appendix C for all individual coding decisions), dated between 2002 and February 2012. Most studies

were published (k = 27), 11 were conference presentations (poster or paper), and the rest were 4 Dissertations, 1 Master's Thesis, and 1 submitted manuscript.

Computer program. More than half of the studies (58.1%) used LIWC (2001 or 2007), 23.3% used other general programs, and 18.6% applied a program specifically developed to detect deception. Three studies, where the type of program was not specifically described or labeled (e.g., "automated analysis method", "natural language processing tool", "message analyzing software"), were categorized under other general programs.

Senders. There were a total of 3,780 senders (k = 43) with an average of 87.91 (SD = 19.60, Mdn = 53) senders per study, ranging from eight to 800. Information about senders' gender was provided in 30 studies, with more male than female participants in total (N_{male} = 1,254; N_{female} = 895), and on average per study (M_{male} = 41.80; SD_{male} = 9.22; M_{female} = 29.83; SD_{female} = 5.76). Exact information about senders' age was reported in only 29.5% of the studies. Across all age groups, senders' mean age was 19.33 years (SD = 8.45), ranging from 4 to 58 years. The mean age of N = 1,015 adults only was 24.17 (SD = 4.11) with a range of 17 to 58 years, whereas the mean age of N = 218 children (k = 4) was 8.45 years (SD = 1.57), ranging from 4 to 14 years.

Accounts. There were a total of 11,680 (N_{truth} = 5,650, N_{lie} = 6,030) accounts originally. However, six studies contained an extremely large number of accounts, ranging from N = 608 (Schafer, 2007, Experiment 1) to N = 3,162 (Derrick et al., 2012), with a mean of 1295.17 accounts (SD = 948.98). In the other 38 studies, the mean was M = 102.87 (SD = 73.17), ranging from N = 13 (Ali & Levine, 2008, confessions) to N = 322 (Cooper, 2008). Therefore, we decided to adjust the number of total accounts for these six studies to N = 500 (n_{truth} = 250, n_{lie} = 250) to avoid

extra-ordinary high weights. Consequently, the final average number of accounts per study was *M* = 157.02 (*SD* = 153.66, *Mdn* = 103), with *M* = 82.02 (*SD* = 80.71) for truths and *M* = 75.00 (*SD* = 76.68) for lies. All accounts were provided in English except for four studies (two Spanish, one Dutch, one Arabic).

Preparation. Only eight studies provided information about how long senders had time to prepare their accounts. In four of these, senders had no opportunity, for the other four studies, senders had on average 1.31 minutes (*SD* = 0.71; range: 1 to 5 minutes) to prepare.

Theoretical background. Twelve studies referred to Newman et al.'s (2003) explanations ("LIWC approach") to predict the outcome of specific linguistic cues, three used Interpersonal Deception Theory (IDT, Buller & Burgoon, 1996), two reality monitoring (RM; Sporer, 2004), and 12 a combination of IDT and reality monitoring. Twelve additional studies referred to other theoretical backgrounds, for example, Media Richness Theory (Daft & Lengel, 1986), or Verbal Immediacy (Mehrabian & Wiener, 1966), and three studies did not mention any theory at all. A-priori selections of linguistic cues were made for 37 studies while seven reported only significant findings.

Interpretation of Effect Sizes

As a rule of thumb, Cohen (1988) classified the effect size *d* into three categories, with *d* = 0.20 as small, *d* = 0.50 as medium and *d* = 0.80 as large effect sizes. However, in meta-analyses about cues to deception, effect sizes are often much smaller (DePaulo et al., 2003: *Mdn* g_u = 0.10; similarly low for Sporer & Schwandt, 2006, 2007). Richard, Bond, and Stokes-Zota (2003) examined 322 meta-analyses in social psychology and provided an empirically based effect size distribution that might serve as a good comparison for our results (cf. Sporer & Cohn,

2011). It should be noted that in DePaulo et al.'s (2003) meta-analysis positive effect sizes refer to stronger or more frequent cues in lies.

Research Questions

In this section, we present results for 50 linguistic cues to deception grouped according to six research questions (see Table 1). The weighted average g_u, with the 95% confidence interval (CI), is reported for all analyses. Recall that positive effect sizes denote stronger presence in true accounts (similarly to Sporer & Schwandt, 2006, 2007, but contrary to DePaulo et al., 2003). A data file with all dependent and predictor variables coded is available in supplemental online materials.

Research Question 1: Do Liars Experience Greater Cognitive Load?

(a) Are liars' accounts shorter in terms of *number of words* (cue 01), *number of sentences* (cue 07), and *average sentence length* (cue 08)? As expected, liars used fewer words than truth-tellers (*word quantity*, 0.24 [0.19, 0.29]), with g_us ranging from -1.25 to 1.43, but no shorter sentences than truth-tellers (*average sentence length*, 0.05 [-0.03, 0.13]). Contrary to our prediction, liars used more *sentences* than truth-tellers (-0.33 [-0.44, -0.21]), although the distribution of effect sizes was also quite heterogeneous. The effect size for *sentence quantity* was derived from a small subset of nine studies compared to 42 studies serving data for *word quantity*. Therefore, *word quantity* is a more precise estimate for statement length.

Note that DePaulo et al. (2003) did not examine number of words per se but only *response length* defined as "length or duration" (cue 01, d = -0.03, k = 49, *ns*), or as *talking time* (cue 02, d = -0.35, k = 4, $p < .05$). Sporer and Schwandt (2006) found no reliable associations for number of words (d = -0.018, k = 8), nor for

message duration (*d* = -0.078, *k* = 23). These differences in findings may be due to the stimulus accounts used. More recent studies analyzing verbal content cues to deception sometimes do (e.g., Ansarra, Colwell, Hiscock-Anisman, Hines, Fleck, Cole, & Belarde, 2011) and sometimes do not find (e.g., Leal, Vrij, Warmelink, Vernham, & Fisher, 2013) differences between liars' and truth-tellers' length of accounts operationalized by the number of words.

(b) Are deceptive accounts less elaborated in terms of *content word diversity* (cue 02), *type-token ratio* (cue 03), or word length cues (cues 04, 05)? Indeed, liars used *fewer diverse content words* (0.48 [0.34, 0.61]) and distinct words (*type-token ratio*: 0.14 [0.07, 0.21]) than truth-tellers. These findings could be attributed to liars' increased cognitive load and reduced working memory capacity (relative to truth-tellers), which in turn is associated with a limitation of creative word production in speaking or writing. These findings also favor a cognitive over a social psychological explanation, as it is unlikely that liars strategically use fewer diverse content and distinct words. However, the prediction that liars would provide shorter words was not supported (see Table 1). Presumably, the number of distinct words and word diversity indices are more sensitive to cognitive load and working memory capacity than word length.

(c) Are deceptive accounts less complex than true accounts, as indicated by fewer *verbs* (cue 06), *causation* (cue 09) and *exclusive words* (cue 10)? Liars indeed used fewer *exclusive words* like *but*, *except*, or *without*, than truth-tellers (0.24 [0.17, 0.31]). Using few exclusive words results in simpler stories (Newman et al., 2003). Liars may resort to telling simple stories because their cognitive system is more taxed than that of truth tellers. Our predictions that liars

would use fewer words assigning a cause to his or her behavior (*causation*), or use fewer *verbs* than truth-tellers, were not confirmed (Table 1).

(d) Do liars commit more *writing errors* (cue 11) than truth-tellers? No support was found for this hypothesis with or without two outliers (Lee, Walker & Odom, 2009; Zhou & Zhang, 2004). This can be reconciled with DePaulo's self-presentational perspective (DePaulo, 1992; DePaulo et al., 2003). Liars might be more self-aware and deliberate than truth-tellers; hence, they may edit their typing errors. Derrick et al. (2012) showed that liars were significantly more likely to edit their words on the keyboard (e.g., in using the backspace and delete button) than truth-tellers (-0.12 [-0.19; -0.05]). Whether or not their edits were aimed at correcting explicit typing errors or not, was not investigated and should be examined more closely. In six of the ten studies exploring *writing errors*, participants typed their stories on a computer keyboard; unfortunately, they did not measure editing behavior (with the exception of Derrick et al., 2012).

Research Question 2: Are Liars Less Certain Than Truth-Tellers?

Effects for *certainty* and *modal verbs* were not significant. The difference between DePaulo et al.'s (2003) findings (who found liars to appear more verbally and vocally uncertain: cue 31, $k = 10$, $d = 0.30$, $p < .05$) and ours could be due to different operationalizations. Whereas we included studies that automatically counted words expressing *certainty*, DePaulo et al. considered the *subjective* impression of uncertainty ("the speaker seems uncertain, insecure, (...)", p. 114). The opposing findings suggest that (a) there is a difference between objective and subjective assessments of (un)certainty, and/or (b) liars may nonverbally give the impression of being uncertain without using fewer certainty words than truth-tellers.

Contrary to our prediction, deceptive accounts contained slightly *fewer tentative words* (such as *may*, *seem*, *perhaps*) than truthful accounts (0.13 [0.06, 0.20] for an exception, see ten Brinke & Porter, 2012). A reason for this unexpected finding could be that liars think that tentative expressions diminish their credibility and therefore try to avoid them, although we are not aware of any empirical evidence that liars pursue this strategy to appear more credible. Note that DePaulo et al. (2003) also reported less "*tentative constructions*" (cue 30, k = 3, d = -0.16, *ns*) in lies. A different explanation for this finding could be derived from the literature on credibility assessment (e.g., Steller & Köhnken, 1989). The underlying assumption is that due to their motivation to appear credible, liars (here: alleged victims of sexual abuse) would probably not correct themselves spontaneously, admit a lack of memory or raise doubts about their own statement. These criteria relate to uncertainty or tentative words to the extent that liars try to hide any kind of deficiencies or ambiguities in their statement in order to appear or stay credible (Sporer, 2004). Especially the criterion "admitting lack of memory" is less often expressed by liars than truth-tellers (DePaulo et al., 2003, cue 73: k = 5, d = -0.42, p < .05; Vrij, 2005). Research also shows that guilty suspects attempt to be firm in their denial of guilt (Hartwig et al., 2007); this is contrary to showing uncertainty.

Research Question 3a: Do Liars Use More Negations and Negative Emotion Words?

(a) To the extent that liars defend themselves or deny something they have done, do they use more *negation terms* such as *no*, *never*, or *not* (cue 17)? This prediction was supported, with a significant negative effect of -0.15 [-0.22, -0.09] based on 20 studies (but large heterogeneity). Our results contradict Hancock, Curry, Goorha and Woodworth's (2008) view, who considered *negations* as a form of

distinction marker (in addition to *exclusive terms*) expected to occur *less* frequently in deceptive accounts, presumably to avoid contradictions by being less specific than truth-tellers.

Our findings concur with those of DePaulo et al. (2003), who found a significant effect for *negative statements and complaints* (cue 52: d = 0.21, k = 9, p < .05) showing that liars use slightly more negative utterances than truth-tellers.

(b) Do liars use more *negative emotion words* in general (cues 18, 18.1), as well as more specific negative-emotion words, such as *anger* (cue 18.2), *anxiety* (cue 18.3), or *sadness* (cue 18.4), than truth-tellers? Contrary to the prediction that people might feel negative emotions while lying (Ekman, 2001; Zuckerman et al., 1981), liars did not use more negative emotion words (cue 18; -0.07 [-0.15, 0.01]). However, the sub-cue *negative emotions only* revealed a small but reliable negative effect (-0.18 [-0.24, -0.12]). The difference between these results can be explained with their different operationalization. Whereas the superordinate category *negative emotions* (cue 18) contained all types of negative emotions (including *anger*, *anxiety*, and *sadness*), cue 18.1 encompassed only a reduced set of negative emotion words (e.g., *hate*, *worthless*, *enemy*).

A more differentiated picture of various negative emotions under investigation emerged when we look at the more specific type of emotion words used. Liars used more *anger* terms than truth-tellers (cue 18.2, -0.27 [-0.35, -0.19]), although no significant differences were found for *anxiety* (cue 18.3) or *sadness* (cue 18.4, see Table 1). Newman et al.'s (2003, p. 672) assertion that "anxiety words are more predictive than overall negative emotion" was not supported. Rather, the present findings indicate that there are differences in words expressing feelings and/or

different negative emotions while lying. Liars might not feel anxious or sad but rather feel angry, and this might be manifested in words like *worthless*, or *annoyed*.

Research Question 3b: Do Liars Use Fewer Positive Emotion Words?

Did truth-tellers express more *positive emotion* (cue 19.1) or *positive feeling* (cue 19.2) words than liars? While the effect for *positive emotions only* just missed significance (-0.07 [-0.15, 0.00], overall, there was no support for this prediction (Table 1). DePaulo and colleagues (2003) also did not find a significant effect for being friendly and pleasant (cue 49: d = -0.16, k = 6, *ns*).

Research Question 3c: Do Liars Express More or Less Unspecified Emotion Words?

For 21 studies investigating unspecified *emotion words* (cue 15), liars used more unspecified emotion words than truth-tellers (-0.11 [-0.19, -0.04]). However, liars and truth-tellers did not differ in words expressing *unpleasantness* or *pleasantness* (cue 16, -0.10 [-0.25, 0.06]). DePaulo et al. (2003) also found no significant difference for being "*friendly and pleasant*" (cue 49: d = -0.16, k = 6, *ns*). Conversely, DePaulo et al.'s findings for two other subjectively rated cues associated with pleasantness, namely "*cooperation*" (cue 50: d = -0.66, k = 3, p < .05), and "*facial pleasantness*" (cue 54: d = -0.12, k = 13, p < .05), showed that truth-tellers appeared more pleasant than liars. These differences might indicate that the pleasantness construct tracked by DePaulo et al.'s human-rated cues (subjective impressions) is different from the one operationalized in computer-based studies (objective word count). Alternatively, truth-tellers might only appear more pleasant than liars in their nonverbal or paraverbal behavior, but not in their choice of words.

Research Question 4: Do Liars Distance Themselves More From Events?

(a) Do liars use fewer *first-person pronouns* (cues 21, 22, 23) and more *second-person* (cue 24) *and third-person pronouns* (cue 25) than truth-tellers? Although no significant differences were found for *first-person singular*, or *first-person plural references* (see Table 1), the weighted average effect size for *total first-person pronouns* was significant in the expected direction, that is, liars used fewer *total first-person pronouns* than truth-tellers (0.14 [0.06, 0.22], when the extreme negative effect size found by Brunet, 2009, both conditions: -1.63 [-1.98, -1.29] was excluded).

On the other side of the coin, we predicted *second-*and *third-person pronouns* to occur more often in liars' than truth-tellers' accounts. Our meta-analyses supported this prediction, with a negative g_u = -0.10 (Table 1). The results indicated that liars in general tried to redirect the focus of attention to other people by using more references to their interaction partner(s) (*you*), or to (a) third person(s) (*he*, *she, they*) than truth-tellers.

Overall pronoun use. As researchers seem to be interested in the use of any type of pronouns (*total pronouns*, cue 20), we aggregated all of the pronoun effect sizes. The resulting effect size was not significant (0.06 [-0.02, 0.14]).

(b) Do deceptive accounts contain more *passive voice verbs* (cue 26) and *generalizing terms* (cue 27) than truthful accounts? Although effect sizes for *passive voice verbs* varied considerably (see Table 1), all were nonsignificant. This is probably due to small sample sizes or a generally low frequency of occurrence (*floor effect*). *Generalizing terms* had a medium negative effect size (-0.37 [-0.79, 0.05]) that was nevertheless not significant because of the small number of studies

and large heterogeneity. Similarly, DePaulo et al. (2003) did not find a significant effect for *generalizing terms* (cue 21: d = 0.10, k = 5, *ns*).

(c) Do lies include more *past tense* verbs (cue 47) and fewer *present tense* verbs (cue 48) than true accounts? Significant differences were neither found for *past tense verbs* nor for *present tense verbs* (Table 1). A potential reason why the data did not support our predictions could be the way the dependent variable was operationalized. It is important to note that Pillemer et al.'s (1998) hypothesis stated that verb tense *shifts* occur more often in critical parts of experienced (i.e., true) autobiographical events. Here we did not consider verb tense *shifts*, but *absolute number* of present and past tense verbs. Future research could construct a more suitable linguistic cue than counting the number of verbs only.

Research Question 5: Do Liars Use Fewer (Sensory and Contextual) Details?

(a) Do liars use fewer *sensory* and *perceptual* details than truth-tellers? They did, according to our findings for *sensory-perceptual processes only* (cue 28.1), although the average effect size was very small (0.06 [0.00, 0.13]). For the variable *sensory-perceptual processes overall* (cue 28), the effect size was not significant (0.05 [-0.01, 0.12], after two outliers were excluded).

Some support came from the more specialized cue *hearing* (cue 28.4, 0.17 [0.09, 0.25]), showing that liars used fewer words expressing their acoustic impressions (like *listen*, *sound*, or *speak*) than truth-tellers. Indeed, in case of entirely fabricated lies (compared to partially fabricated lies or lies of omission), persons may not experience any audio(visual) impressions at all and do not seem to deliberately include these words in their lies. However, the cues *seeing* (cue 28.2) and *feeling* (cue 28.3) yielded nonsignificant results (see Table 1). Although DePaulo et al. (2003) also found no significant effects for *sensory information* (cue 05: d = -0.17, k

= 4, *ns*) there have been many new reality monitoring studies we are currently synthesizing in an updated large scale meta-analysis.

(b) Are liars' accounts less contextually embedded than those of truth-tellers, as indicated by fewer *time* and *space* words? No significant effects emerged for *time* (cue 29), *space* (cue 30), or the combination of *spatial and temporal details* (cue 31). Our results for *temporal* and *spatial details* are in line with DePaulo et al.'s (2003) nonsignificant finding for *contextual embedding* (cue 76: *d* = -0.21, *k* = 6, *ns*), though it should be noted that *contextual embedding* goes beyond *temporal* and *spatial details* in that the event has to be connected to everyday occurrences, habits, relationships, and so forth (e.g., Steller & Köhnken, 1989). Again, many newer Criteria-Based Content Analysis and reality monitoring studies found support for this assumption (see Masip et al., 2005; Vrij, 2005, 2008a) but linguistic analyses by computers do not seem to capture them.

(c) Relative to truth-tellers, do liars use fewer descriptive words, such as *prepositions* (cue 32), *numbers* (cue 33), *quantifiers* (cue 34), *modifiers* (adverbs and adjectives, cue 35), and *motion verbs* (cue 36)? The only significant effect size was obtained for *quantifiers* (0.14 [0.02, 0.25]) indicating a slightly lower use of words such as *all, bit, few, less*, among liars. However, this finding was synthesized from four studies only, so we should not make strong conclusions for this cue in general.

Interestingly liars produced more *motion verbs* (such as *walk, go,* or *move*) than truth-tellers (-0.09 [-0.17, -0.01]) after removing the only significant positive effect size (Liu, Hancock, Zhang, Xu, Markowitz, & Bazarova, 2012; 0.38 [0.21, 0.56]), which was found to be an outlier. This finding is contrary to our prediction but is in line with the cognitive load approach (Research Question 1) and Newman et

al.'s (2003) assumption that, when constructing a lie, "simple, concrete actions are easier to string together than false evaluations" (p. 667). Therefore, liars, who are cognitively taxed by the act of lying, "should use more *motion verbs* and fewer *exclusive words*" (Newman et al., 2003, p. 667).

Research Question 6: Do Liars Refer Less Often to Cognitive Processes?

As predicted, weighted average effect sizes for both cues (37 and 38) were significantly positive (see Table 1), indicating that liars expressed words relating to their inner thoughts (*insight*) and *cognitive processes* less often than truth-tellers.

Miscellaneous Category

Twenty-nine cues that could not be allocated to any research question were subsumed under the miscellaneous category. As displayed in Appendix D (in supplemental online materials), significant positive effect sizes (without outliers) were obtained for *inhibition*, *humans*, and for three cues expressing biological processes, namely: *biology*, *physical states*, and *eating*. Liars used fewer words from all of these word classes than truth-tellers. In contrast, negative effect sizes for *future tense* and *leisure terms* indicated that these terms occurred more frequently in deceptive than truthful accounts.

Moderator Analyses

Due to the large number of potential moderator analyses for all linguistic cues, we only report significant findings (all Q_Bs were significant at $p < .05$) for both theoretically and methodologically important moderator variables. Specifically, we examined six moderator variables for 25 linguistic cues, with each analysis containing at least 13 studies.[7] Note that the overall number of studies is smaller for

the moderator analyses as many studies did not report enough information to code them. Analyses of two additional moderators, experimental design (between- vs. within-participants) and publication status are available in supplemental online materials (Appendices E and F). Also, it must be acknowledged that blocking groups of studies in meta-analyses analogous to ANOVA often introduces confounds (see Pigott, 2012) although we have taken great care to minimize them (see *Method*).

Event type and personal involvement. We hypothesized that larger effect sizes would be found for Research Questions 1, 2, 3a, 3c and 4 if the event was personally relevant to the participant ("First-person experience", k = 21) than in the "Attitude/liking" paradigm (k = 7) or the "Miscellaneous" paradigms (k = 14; see Table 2). First, concerning cognitive load (Research Question 1), event type affected *average sentence length* only. Liars used shorter sentences than truth-tellers when articulating attitudes (0.17), but not under the other two paradigms. Second, regarding negative emotions and negations (Research Question 3a), liars used more *negative emotion words* than truth-tellers only if they had to tell a personally relevant story (-0.37, -0.57), and expressed more *negations* only in miscellaneous paradigms (-0.59). Thus, although liars might experience and express more negative emotions when the topic is personally relevant, they do not necessarily use more negations. Third, as expected, liars also expressed more *unspecified emotions* (Research Question 3c; -0.45) when talking about a personal experience than when having performed other tasks. Fourth, concerning distancing (Research Question 4), liars used fewer *first-person plural pronouns* primarily when describing a video (0.38), but fewer *total first-person pronouns* when talking about attitudes (0.31). This unexpected finding suggests that it may be especially hard for liars to refer to themselves while articulating a false attitude; however, liars may still use self-

references while telling a personal event because it is common (in the English language) to refer to oneself as the actor. Also, it would be hard to avoid self-references when telling a story with oneself as the acting person, even when lying.

Liars used more *total second-person pronouns* only when talking about attitudes (-0.18), and more *total third-person pronouns* in all kinds of events except the miscellaneous paradigms. In general, thus, the predicted differences for Distancing (Research Question 4) between liars and truth-tellers appear enhanced in the attitude/liking paradigm--compared to the other two paradigms.

Emotional valence. We predicted effects for cues under Research Questions 1 to 4 to be larger for negative (k = 18) than for neutral events or themes (k = 17; see Table 3).[8] Indeed, liars used fewer *words* (Research Question 1--Cognitive Load; 0.54) only when the event was negative. In terms of negative emotions (Research Question 3a), liars also used considerably more *negations* (-0.42) and *negative emotions* (-0.39, -0.65) than truth-tellers, most notably when the event was negative. This supported the notion that telling a negatively toned lie might be accompanied by negative emotions.

However, contrary to our predictions regarding the cognitive load cues (Research Question 1), differences between lies and truths for *type-token ratio* (0.32) and *exclusive words* (0.47) were larger when telling a neutral event rather than a negative event. Perhaps truth-tellers reporting a negative event are as emotionally involved as liars. This may imply using less elaborate language (compared to neutral events), which would explain the lack of difference in *type-token ratio* between liars and truth-tellers. Finally, no difference in the use of *unspecified emotions* (Research Question 3c) was found between neutral and negative topics: Liars used more *emotion words* overall for both (-0.54; -0.45).

Regarding distancing (Research Question 4), somewhat contradictory findings occurred for self-references. When telling neutral events, liars used more *first-person singular* pronouns (-0.25) but fewer *total first-person* pronouns (0.22) than truth-tellers. Also, when telling negative events liars used fewer *first-person singular* pronouns than truth-tellers (0.27) but about the same amount of *total first-person* pronouns (-0.13). These findings clearly show that (a) differences exist between liars and truth-tellers in terms of referring solely to oneself or to oneself in addition to one's group, and (b) these differences depend on the valence of the event. If we think about examples of wrongdoing as typical negative events, it perfectly makes sense to distribute responsibility to "we" (or "you and me", "they and me") than to take it on one's own shoulders ("I"). Finally, liars expressed more *total second-person pronouns* only when the event was neutral (-0.40).

Intensity of interaction. We predicted that the higher the interaction level, the larger the effect sizes would be (Table 4). Indeed, effect sizes for *word count* (Research Question 1--Cognitive Load; 0.69), *negative emotions* (Research Question 3a; -0.48, -0.79), *unspecified emotions* (Research Question 3c; -0.63), and *first-person singular pronouns* (Research Question 4--Distancing; 0.34) were largest in person to person interactions. Note also that for computer-mediated communication the direction of effect (-0.41) reversed compared to other conditions. Furthermore, in the interview condition (which was considered as the second intense interaction category), effect sizes for *word count*, *exclusive words*, and *negative emotions only*, were in the expected direction. Interestingly, when no interaction took place, *liars* used significantly more *first-person singular pronouns* (-0.22) and *total third-person pronouns* than truth-tellers (-0.31).

Together, this evidence suggests that some verbal differences between liars and truth-tellers manifest themselves most when a bidirectional interaction between two persons--not only a one-way interview--took place.

Motivation. In support for our hypotheses, larger effects occurred for not-motivated liars, who used fewer *words* (Research Question 1--Cognitive Load) than truth-tellers (0.47), compared to moderately (0.19), or highly motivated liars (0.18; see Table 5). Also, liars used fewer *temporal details* (Research Question 5 regarding details) only when no motivation was induced (0.20). These findings support the notion that highly motivated liars are more successful than unmotivated liars in controlling their verbal behavior (at least in terms of number of *words* and *temporal details*). Note that liars seem to be less able to control their paraverbal behavior under high motivation (e.g., for pitch, response latency; Sporer & Schwandt, 2006) nor their visual nonverbal behavior (e.g., for eye contact; DePaulo et al., 2003).

Other linguistic cues under various research questions showed findings contrary to our hypothesis (see Table 5): (a) only highly motivated liars used fewer *different words* (*type-token ratio*: 0.67); (b) only moderately motivated liars built slightly shorter sentences (*average sentence length*: 0.15) than truth-tellers; (c) liars expressed more *negative emotional* words than truth-tellers only when they were highly (-0.56, -1.03) or not motivated (-0.20, -0.19); (d) liars expressed more *unspecified emotions* (-0.53) than truth-tellers when highly motivated; and (e) both highly (0.21, 0.25) and moderately motivated (both 0.12) liars reported fewer *sensory-perceptual processes* than truth-tellers, whereas not motivated liars tended to refer *more* often to these processes than truth-tellers (-0.22, -0.29).

Taken together, these results show a mixed picture. Our prediction that highly motivated liars would control their verbal behavior better than less motivated liars

was confirmed for only two cues. However, our findings should not be over-interpreted because the number of studies with highly motivated participants was very small--calling for more research with highly motivated liars.

Production mode. Moderator analyses showed mixed findings (see Table 6). Liars used fewer *words* (Research Question 1--Cognitive Load) than truth-tellers under all production modes, though effects were larger for handwritten texts (0.33) and for transcripts from spoken accounts (0.26) than for typed texts (0.10). It seems that storytellers may use the opportunity to edit their accounts when typing, thus reducing the number of errors. More direct evidence for this point comes from a study by Derrick et al. (2012) who developed a specialized computer applet that clandestinely recorded edits and revisions during real time synchronous communication between a computer interviewer and senders. They found that when deceiving, people were significantly more likely to take longer and perform a greater number of edits to their responses (more frequently using the delete and backspace keys) than when telling the truth. To the extent that deceptive individuals are more likely to engage in such editing, differences in writing errors between true and false statements may be obscured. This might explain why the effect for number of *typed words* is smaller than for number of handwritten or spoken words.

In line with our hypothesis concerning details (Research Question 5), liars expressed fewer *sensory-perceptual words* than truth-tellers only when writing their accounts by hand (0.33, 0.34), whereas liars used fewer *spatial details* than truth-tellers only in typed accounts (0.13). Contrary to our hypothesis (but in line with Newman et al.'s (2003) assumption), liars used more *motion verbs* than truth-tellers when handwriting (-0.28) or speaking (-0.16), but not when typing them (0.00).

Our hypothesis concerning negative emotions (Research Question 3a) was not supported: Liars expressed more *negations* and *negative emotions* than truth-tellers when handwriting (-0.60 -0.28, respectively) rather than when speaking or typing. A potential reason for the larger effect in the handwriting condition could be that a writer might take more time to re-experience a negative emotion linked to the process of lying (see Ekman, 1988). Also, the special advantage to edit *typed* words could be a reason why the difference between liars and truth-tellers disappeared under this condition. Interestingly, regarding unspecified emotions (Research Question 3c), liars' spoken messages--compared to truth-tellers'--showed no differences (-0.04) in *unspecified emotion words* but more when typing (-0.44) or handwriting (-0.25).

In conclusion, the question of how the mode of production affects the language of lying is not sufficiently answered. Again, other moderators such as interaction type or motivation may be confounded in these analyses. The pattern of findings that typed accounts showed smallest effects also converges with the finding that computer-mediated communication showed smallest effects (Table 4 above). Future studies should investigate interaction intensity and production mode in more detail, perhaps controlling for language proficiency and typing skill.

Computer program. The hypothesis that effects would be larger if statements were analyzed with programs specifically designed to detect deception (k = 8) rather than with LIWC (k = 26) or general programs (k = 11) was only confirmed for *first-person plural pronouns* (-0.31; see Table 7). Specific programs were more sensitive than LIWC or other general programs to differences in *first-person plural* pronouns, finding more of these words among liars than among truth-tellers. Contrary to our hypothesis, four linguistic cues were found to have larger effects if

LIWC or other general programs were used than with more specific programs. The direction of the effect for word quantity was even reversed if specific lie detection software was used. A parsimonious explanation may be that these specific programs were developed and used for different types of accounts. It also demonstrates that the validity of linguistic cues to deception depends on the kind of program used. However, this conclusion is limited by the fact that we had to exclude quite a few studies using specialized software as these did not contain sufficient information to calculate effect sizes. Journal editors and grant agencies should emphasize completeness of data reporting including effect sizes (APA, 2008).

Publication bias. The correlation between sample sizes (number of accounts) and the absolute value of all effect sizes (excluding extremely large samples to avoid skewed distributions) was $r(904) = -.11$, $p < .001$. This negative correlation could be due to a publication bias, that is, a tendency for significant findings to be more likely to be published than unpublished (Levine et al., 2009). On the other hand, our moderator analyses showed that for 7 of 12 cues, for which there was a significant difference between published and unpublished studies, effects were actually greater in *unpublished* studies (see Appendix E in supplemental online materials). Thus, publication bias is unlikely to be a threat to the validity of our conclusions.

General Discussion

Setting some of the exceptions discussed under the moderator analyses aside we venture some take home message to our research questions, taking also rival theoretical approaches into consideration.

Research Question 1--Cognitive Load. Taken together, the notion that liars experience greater cognitive load was mainly supported. As predicted from the

working memory model and the cognitive load approach, lies were shorter (fewer words and fewer sentences), less elaborated (fewer different words), and less complex (fewer exclusive terms) than true stories. Even if liars were to strategically withhold information that could give them away, doing so would heighten their working memory burden, thus indirectly also supporting the cognitive load approach.

Research Question 2--Certainty. Because only three cues were investigated here and they yielded contradictory results, this question could hardly be answered. In general, the prediction for this research question that liars look linguistically less certain than truth-tellers due to a lack of personal investment or feelings of ambiguity or guilt was not supported. Contrary to our prediction, truth-tellers used more tentative words than liars.

Research Question 3a--Negative Emotions. Altogether, the prediction that liars express more negative emotion words and defend themselves to a greater extent than truth-tellers due to the experience of negative emotions when lying was corroborated. More specifically, liars expressed more terms of anger (rather than other negative emotions like anxiety and sadness) and denied accusations more often than truth-tellers.

Research Question 3b--Positive Emotions. Our assumptions based on the fading-affect bias that truth-tellers express more positive emotions than liars was not supported. While this result may be dependent on the type of lie being told it does run counter to our assumptions from the autobiographical memory literature (as well as Criteria-Based Content Analysis and reality monitoring research): Taken together, it appears wise to differentiate specific emotions and feelings and separate them according to their valence.

Research Question 3c--Unspecified Emotions. In general, liars expressed more unspecified emotions (i.e., negative and positive emotions undifferentiated) than truth-tellers. Given the results for different types of negative emotions, linguistic researchers should revisit their analyses to separate different types of emotions.

Research Question 4--Distancing. As expected, liars distanced themselves from events more than truth-tellers by using fewer self-references (total first-person) and more other-references (total second- and total third-person). On the other hand, liars and truth-tellers did not differ in terms of generalizing terms, use of passive voice or verb tenses.

Research Question 5--Details. Overall, the reality monitoring approach was only partially supported. We only found small effects for some cues (sensory-perceptual processes only, particularly when motivation is high or the account is handwritten, hearing words and quantifiers) but null-findings for most other cues. In their review on international reality monitoring research, Masip et al. (2005) concluded that visual and auditory details, contextual and temporal information were the most discriminative criteria. The discrepancies may either be due to the fact that the reality monitoring criteria cannot easily be captured by word-counting programs like LIWC, or the fact that the LIWC categories were not created on the basis of reality monitoring theory (see Vrij, Mann, Kristen, & Fisher, 2007). Coding reality monitoring criteria and indicators involves much more than mere word counting, and only well-trained human raters, who also take the context of specific words or sentences, as well as the background or motivation of a statement into account, can do it.

Research Question 6--Cognitions. We found that truth-tellers used more words indicating cognitive processes than liars. The findings support our predictions

from autobiographical memory theory that persons refer more often to retrieval processes, supporting memories, and cognitive operations when talking about true events but contradicts the assumption of many reality monitoring deception researchers who postulate the opposite (e.g. Vrij, 2008a).

Limitations

Several limitations restrict the generalizability of our findings. First, we had to exclude more than 50 studies for different reasons (see Appendix G in supplemental online materials). Most of these studies did not provide sufficient statistical data, or calculated linguistic patterns in a way not suitable for our analysis (e.g., Keila & Skillicorn, 2005). While we are grateful to all authors who provided us with additional data, journal editors should emphasize the reporting of all results, not just significant ones, along with effect sizes like Cohen's *d*.

Second, we were able to find significant effects for many linguistic cues (see Table 1). These effects were generally very small according to Cohen (1988), but not much smaller than those for nonverbal and paraverbal cues meta-analyzed by DePaulo et al. (2003) and Sporer and Schwandt (2006, 2007). However, even if all cues had been in the predicted direction, the mean g_u = 0.26 is rather disappointing compared to mean effect sizes in the social psychological literature (Richard et al., 2003: mean g_u = 0.43; r = 0.21, SD = 0.15).

Third, for those linguistic cues where effect size distributions were quite heterogeneous, and sensitive to moderator variables, general conclusions can only be very tentative. Specific circumstances of individual studies documented in Appendix C should be considered for specific types of lies, topics, paradigms or production modes.

Fourth, most findings were only available for the English language. As Newman et al. (2003) discussed, deception may be manifested through different linguistic cues in different languages. For example, Romanic languages do not require the use of specific personal pronouns, because pronouns are already expressed by the verb form (e.g., Masip, Bethencourt, Lucas, Sánchez-San Segundo, & Herrero, 2012). Unfortunately, no moderator analysis could be conducted for language, because only four studies analyzed accounts in languages other than English. Besides language, culture might also make a difference. For example, Taylor, Tomblin, Conchie, and Menacere (2011) found that North African participants used first-person pronouns most frequently when lying, whereas White British participants used them most frequently when telling the truth.

Fifth, differences between children and adults became evident as three out of four studies conducted with children were detected as significant outliers, though not for the same cues. This underscores the need to investigate differences between adults' and children's linguistic cues to deception separately. Further, not all children are equal. Linguistic skills develop during childhood, and this may presumably influence the frequency of some potential linguistic deception markers. Children of different ages may show different linguistic cues to deception.

Despite these limitations, our meta-analyses were the first large effort to quantitatively synthesize research in this area. Therefore, they can be seen as the most accurate estimate to date of linguistic differences between liars and truth-tellers assessed by computer programs.

Conclusions and Implications for Future Research

The main goal of the present meta-analysis was to assess the extent to which computer programs are valid and useful tools to detect deception in verbal accounts.

We provided clear operational definitions for each cue, derived from an analysis and integration of definitions from different research domains. We then posed theoretically based hypotheses as to the direction of effects for all cues, as well as concerning potential moderator variables. While not all results could be reported due to space limitations, additional appendices and analyses as well as all our raw data are available as supplemental online materials. Researchers are invited to peruse our rich database for additional analyses. Future research should also look at the intercorrelations between linguistic cues to arrive at a better theoretical understanding.

In addition, future research should consider the context of deceptive vs. truthful utterances. A potential reason why only small to medium effect sizes were found in general could be that most computer programs simply count single words without considering the semantic context. If this suggestion goes beyond what computer programs can do at this time, perhaps the linguistic cues with greater effect sizes (for the respective paradigms) should be weighted more heavily than those with smaller or nonsignificant effects. A recent attempt in this direction was made by Chandramouli, Chen, and Subbalakshmi (2011), who employed several weighting mechanisms. They applied for an international patent for their invention of this weighing mechanism.

Ultimately, researchers should directly compare the performance of computers versus human raters. Since context is relevant in analyzing and judging a statement, human raters' assessments of certain linguistic cues might lead to more pronounced differences than objective computer-based codings. On the other hand, the advantages of computer-based coding should not be overlooked. Humans and computers are best at different skills. Humans are less accurate in manual counting

of specific cues or in rendering accurate judgments of complex syntactic relationships, whereas computers cannot provide subjective, gestalt like judgments or capture the meaning or intention of what people are saying (for an example of computer-assisted subjective codings see Sporer, 2012).

Finally, we encourage researchers to further investigate the impact of (a) different interview and interaction conditions, (b) mode of production, (c) types of events, (d) age of sender, and (e) language on linguistic markers of deception. In line with Hancock and Woodworth (2013), we found that linguistic cues to deception are sensitive to contextual factors (see moderator variables). These variables are relevant in applied contexts (forensic, work and organizational settings). Researchers should strive to design experiments containing psychological features analogous to real world deceptive situations to enhance ecological validity (e.g., opportunity for preparation, or high motivation).

In sum, to answer the question whether computer programs are effective lie detectors, our answer must be rather skeptical at this time. The effects were not significant for many of the variables studied or small in magnitude, or moderated by situational variables. Alternative theoretical approaches may find other cues or moderators to be important. At this time, researchers', and particularly practitioners', (unrealistic) dream has yet to come true.

References

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

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

Abe, N. (2011). How the brain shapes deception: An integrated review of the literature. *Neuroscientist, 17,* 560-574. doi:10.1177/1073858410393359

Abe, N., Suzuki M., Mori E., Itoh M., & Fujii, T. (2007). Deceiving others: Distinct neural responses of the prefrontal cortex and amygdala in simple fabrication and deception with social interactions. *Journal of Cognitive Neuroscience, 19,* 287-295. doi:10.1162/jocn.2007.19.2.287

Adams, S. H. (2002). *Communication under stress: Indicators of veracity and deception in written narratives* (Doctoral dissertation). Retrieved from http://scholar.lib.vt.edu/ (URN: etd-04262002-164813).

Adams, S. H., & Jarvis, J. P. (2006). Indicators of veracity and deception: An analysis of written statements made to police. *Speech, Language, and the Law, 13,* 1-21. doi:10.1558/ijsll.v13i1.2

*Ali, M. & Levine, T. (2008). The language of truthful and deceptive denials and confessions. *Communication Reports, 21,* 82–91. doi:10.1080/08934210802381862

*Almela Sanchez-Lafuente, A., Valencia-Garcia, R., & Cantos Gomez, P. (2012). Detectando la mentira en lenguaje escrito [Detecting deception in written language]. *Procesamiento de Lenguaje Natural, 48,* 65-72.

Ansarra, R., Colwell, K., Hiscock-Anisman, C., Hines, A., Fleck, R., Cole, L., & Belarde, D. (2011). Augmenting ACID with affective details to assess credibility. *The European Journal of Psychology Applied to Legal Context, 3,* 141-158.

APA Publications and Communications Board Working Group on Journal Article Reporting Standards (2008). Reporting standards for research in Psychology. Why do we need them? What might they be? *American Psychologist, 63,* 839-851. doi:10.1037/0003-066X.63.9.839

Bachenko, J., Fitzpatrick, E., & Schonwetter, M. (2008, August). Verification and implementation of language-based deception indicators in civil and criminal narratives. *Proceedings of the 22nd International Conference on Computational Linguistics* (pp. 41-48), Manchester.

Baddeley, A. (2000). The episodic buffer: A new component of working memory? *Trends in Cognitive Sciences, 4,* 417–423. doi:10.1016/S1364-6613(00)01538-2

Baddeley, A. (2006) Working memory: An overview. In Pickering S. (Ed.), *Working memory and education* (pp. 1–31). New York: Academic Press.

Bahrick, H. P. (1996). The relation between reproductive and reconstructive processing of memory content. *Behavioral and Brain Sciences, 19,* 191. doi:10.1017/S0140525X00042151

Bahrick, H. P., Hall, L. K., & Berger, S. A. (1996). Accuracy and distortion in memory for high school grades. *Psychological Science, 7,* 265-271. doi:10.1111/j.1467-9280.1996.tb00372.x

*Bedwell, J. S., Gallagher, S., Whitten, S. N., & Fiore, S. M. (2011). Linguistic correlates of self in deceptive oral autobiographical narratives. *Consciousness and Cognition, 20,* 547-555. doi:10.1016/j.concog.2010.10.001

Bender, H.-U. (1987). *Merkmalskombinationen in Aussagen* [Criteria combinations in eyewitness statements]. Tübingen: J. C. B. Mohr.

Benussi, V. (1914). Die Atmungssymptome der Lüge [Breathing symptoms of lying]. In E. Meumann (Ed.), *Sammlung von Abhandlungen zur Psychologischen Pädagogik* (Vol. 3, pp. 513-542). Leipzig: Engelmann.

Blandón-Gitlin, I., Fenn, I., Masip, J., & Yoo, A. (2014). Cognitive-load approaches to detect deception: Searching for cognitive mechanisms. *Trends in Cognitive Sciences, 18*, 441-444. doi:10.1016/j.tics.2014.05.004

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

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

Borenstein, M. (2009). Effect sizes for continuous data. In H. Cooper, L. V. Harris, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (pp. 221-235). New York: Russell Sage Foundation.

Brown, J. D. (2012). Understanding the better than average effect: Motives (still) matter. *Personality and Social Psychology Bulletin, 38,* 209-219. doi:10.1177/0146167211432763

Brown, R., & Kulik, J. (1977). Flashbulb memories. *Cognition, 5,* 73-99. doi:10.1016/0010-0277(77)90018-X

Bruner, J. (1990). *Acts of meaning.* Cambridge: Harvard University Press.

*Brunet, M. K. (2009). *Why bullying victims are not believed: Differentiating between children's true and fabricated reports of stressful and non-stressful events* (Master's thesis). Retrieved from https://tspace.library.utoronto.ca/.

Brunet, M. K., Evans, A. D., Talwar, V., Bala, N., Lindsay, R. C., & Lee, K. (2013). How children report true and fabricated stressful and non-stressful events. *Psychiatry, Psychology and Law, 20*, 867-881. doi:10.1080/13218719.2012.750896

Buller, D. B., & Burgoon, J. K. (1996). Interpersonal deception theory. *Communication Theory, 3,* 203-242.

Buller, D. B., Burgoon, J. K., Busling, A., & Roiger, J. (1996). Testing Interpersonal Deception Theory: The language of interpersonal deception. *Communication Theory, 6,* 268-289. doi:10.1111/j.1468-2885.1996.tb00129.x

Bunn, G. C. (2012). *The truth machine: A social history of lie detection.* Baltimore, MD: The Johns Hopkins University Press.

Burgoon, J., Blair, J. P., Qin, T., & Nunamaker, J. F. (2003). Detecting deception through linguistic analysis. In H. Chen (Ed.), *Lecture Notes in Computer Sciences 2665. Intelligence and Security Informatics* (pp. 91–101). Berlin, Germany: Springer-Verlag. doi:10.1007/3-540-44853-5_7

*Burgoon, J. K., & Qin, T. T. (2006). The dynamic nature of deceptive verbal communication. *Journal of Language and Social Psychology, 25,* 76-96. doi:10.1177/0261927X05284482

*Chen, X. (2010). *Psycho-linguistic forensic analysis of internet text data* (Unpublished doctoral dissertation). Faculty of the Stevens Institute of Technology, Hoboken, New Jersey.

Chandramouli, R., Chen, X., & Subbalakshmi, K. P. (2011). *Psycho-linguistic statistical deception detection from text content* (Report No. WO/2011/085108). Retrieved from: http://patentscope.wipo.int/search/en/WO2011085108

Christ, S. E., Van Essen, D. C., Watson, J. M., Brubaker, L. E., & McDermott, K. B. (2009). The contributions of prefrontal cortex and executive control to deception: Evidence from activation likelihood estimate meta-analysis. *Cerebral Cortex, 19,* 1557-66. doi:10.1093/cercor/bhn189

Churyk, N. T., Lee, C.-C., Clinton, D. (2008, October). Can we detect fraud earlier? A technique called content analysis raises the possibility. *Strategic Finance.* Retrieved from www.thefreelibrary.com

Cialdini, R. B., Borden, R. J., Thorne, A., Walker, M. R., Freeman, S., & Sloan, R. L. (1976). Basking in reflected glory: Three (football) field studies. *Journal of Personality and Social Psychology, 34,* 366-375. doi:10.1037/0022-3514.34.3.366

Cody, M. J., Marston, P. J., & Foster, M. (1984). *Deception: Paralinguistic and verbal leakage.* In R. N. Bostrom (Ed.), *Communication yearbook 8* (pp. 464-490). Beverly Hills, CA: Sage.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20,* 37-46. doi:10.1177/001316446002000104

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

College Board (1976–1977). *Student descriptive questionnaire*. Princeton, NJ: Educational Testing Service.

*Colwell, K., Hiscock, C. K., & Memon, A. (2002). Interviewing techniques and the assessment of statement credibility. *Applied Cognitive Psychology, 16,* 287–300. doi:10.1002/acp.78

Conway, M. A. (1990). *Autobiographical memory. An introduction.* Buckingham: Open University Press.

*Cooper, J. E. (2008). *Using natural language processing to identify the rhetoric of deception in business and competitive intelligence email communications* (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses database (UMI No. 3374690).

Cooper, H. (Ed.) (2010). *Research synthesis and meta-analysis: A step-by-step approach* (4th ed.). Los Angeles: Sage Publications.

Cunningham, H. (2002). GATE, a General Architecture for Text Engineering. *Computers and the Humanities, 36,* 223–254. doi:10.1023/A:1014348124664

Daft, R. L., & Lengel, R. H. (1986). Organizational information requirements, media richness and structural design. *Management Science, 32,* 554-571. doi:10.1287/mnsc.32.5.554

Debey, E., Verschuere, B., & Crombez, G. (2012). Lying and executive control: An experimental investigation using ego depletion and goal neglect. *Acta Psychologica, 140,* 133-141. doi:10.1016/j.actpsy.2012.03.004

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

DePaulo, B. M., & Kirkendol, S. E. (1989). The motivational impairment effect in the communication of deception. In J. C. Yuille (Ed.), *Credibility assessment* (pp. 51-70). Dordrecht, The Netherlands: Kluwer. doi:10.1007/BF00987487

DePaulo, B. M., Lanier, K., & Davis, T. (1983). Detecting the deceit of the motivated liar. *Journal of Personality and Social Psychology, 45,* 1096-1103. doi:10.1037//0022-3514.45.5.1096

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

DePaulo, B. M., & Morris, W. L. (2004). Discerning lies from truths: Behavioural cues to deception and the indirect pathway of intuition. In P. A. Granhag, & L. A. Strömwall (Eds.), *Deception detection in forensic contexts* (pp. 15-40). Cambridge, England: Cambridge University Press. doi:10.1017/CBO9780511490071

*Derrick, D., Meservy, T., Burgoon, J. & Nunamaker, J. (2012, January). An experimental agent for detecting deceit in chat-based communication. *Proceedings of the Rapid Screening Technologies,* Deception Detection and Credibility Assessment Symposium, Grand Wilea, Hawaii. doi:10.1109/HICSS.2003.1173793

Dilmon, R. (2009). Between thinking and speaking: Linguistic tools for detecting a fabrication. *Journal of Pragmatics, 41,* 1152-1170. doi:10.1016/j.pragma.2008.09.032

Dulaney, E. F. (1982). Changes in language behavior as a function of veracity. *Human Communication Research, 9,* 75-82. doi:10.1111/j.1468-2958.1982.tb00684.x

Duran, N. D., Crossley, S. A., Hall, C., McCarthy, P. M., & McNamara, D. S. (2009). Expanding a catalogue of deceptive linguistic features with NLP technologies.

Proceedings of the Twenty-Second International Flairs Conference, Sanibel Island, FL, 243-248.

*Duran, N. D., Hall, C., McCarthy, P. M., & McNamara, D. S. (2010). The linguistic correlates of conversational deception: Comparing natural language processing technologies. *Applied Psycholinguistics, 31,* 439-462.doi:10.1017/S0142716410000068

Dzindolet, M. T., & Pierce, L. G. (2004). A computerized text analysis can detect deception. *Proceedings of the Human Factors and Ergonomics Society 48th Annual Meeting, New Orleans, Louisiana.* doi:10.1177/154193120404800377

*Dzindolet, M. T., & Pierce, L. G. (2005). Using a linguistic analysis tool to detect deception. *Proceedings of the Human Factors and Ergonomics Society 49th Annual Meeting, Santa Monica, California.* doi:10.1177/154193120504900374

Ekman, P. (1988). Lying and nonverbal behavior: Theoretical issues and new findings. *Journal of Nonverbal Behavior, 12,* 163–176. doi:10.1007/BF00987486

Ekman, P. (2001). *Telling Lies: Clues to Deceit in the Marketplace, Politics, and Marriage.* New York: W. W. Norton.

Elkins, A. L. (2011). *Vocalic markers of deception and cognitive dissonance for automated emotion detection systems* (Doctoral dissertation). Retrieved from http://arizona.openrepository.com/arizona.

Elntib, S., Wagstaff, G. H., and Wheatcroft, J. M. (2014). The role of account length in detecting deception in written and orally produces autobiographical accounts using reality monitoring. *Journal of Investigative Psychology and Offender Profiling.* Advance online publication. doi:10.1002/jip.1420

Enos, F. (2009). *Detecting deception in speech* (Doctoral dissertation). Retrieved from http://www.cs.columbia.edu/. (UMI No. 3348430).

Enos, F., Shribergy, E., Graciarena, M., Hirschberg, J., & Stolcke, A. (2007, August). Detecting deception using critical segments. *Proceedings of Interspeech, Antwerpen, Belgium.*

*Evans, A. D., Brunet, M. K., Talwar, V., Bala, N., Lindsay, R. C. L., Lee, K. (2012). The effects of repetition on children's true and false reports. *Psychiatry, Psychology and Law, 19,* 517–529. doi:10.1080/13218719.2011.615808

Fleiss, J. L. (1981). *Statistical methods for rates and proportions* (2nd ed.). New York: John Wiley.

Fornaciari, T., & Poesio, M. (2011, June). Lexical vs. surface features in deceptive language analysis. *Proceedings of the ICAIL 2011 Workshop: Applying Human Language Technology to the Law,* Pittsburgh, USA.

Freud, S. (1905/1959). *Bruchstück einer Hysterieanalyse* [Fragments of an analysis of a case of hysteria]. Collected papers. New York: Basic Books. Reprinted in 1959.

Fuller, C. M. (2008). *High-stakes, real-world deception: An examination of the process of deception and deception detection using linguistic-based cues* (Doctoral dissertation). Retrieved from http://dc.library.okstate.edu/cdm/.

*Fuller, C. M., Biros, D. P., Burgoon, J. K., Adkins, M. Twitchell, D. P. (2006). An analysis of text-based deception detection tools. *Proceedings of the 12th Americas Conference on Information Systems, Acapulco, Mexico.*

Fuller, C. M., Biros, D. P., & Delen, D. (2008, January). Exploration of feature selection and advanced classification models for high-stakes deception

detection. *Proceedings of the 41st Hawaii International Conference on System Sciences.* doi:10.1109/HICSS.2008.158

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

Fuller, C. M., Biros, D. P., & Wilson (2009). Decision support for determining veracity via linguistic-based cues. *Decision support systems, 46,* 695-703. doi:10.1016/j.dss.2008.11.001

Gamer, M., Bauermann, T., Stoeter, P., & Vosse, G. (2008). Covariations among fMRI, skin conductance, and behavioral data during processing of concealed information. *Human Brain Mapping, 28,* 1287-1301. doi:10.1002/hbm.20343

Global Deception Research Team (2006). A world of lies. *Journal of Cross-Cultural Psychology, 37,* 60–74. doi:10.1080/14789940412331337353

Gombos, V. A. (2006). The cognition of deception: The role of executive processes in producing lies. *Genetic, Social, and General Psychology Monographs, 132,* 197-214. doi:10.3200/MONO.132.3.197-214

Graciarena, M., Shriberg, E., Stolcke, A., Enos, F., Hirschberg, J., & Kajarekar, S. (2006, May). Combining prosodic, lexical and cepstral systems for deceptive speech detection. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing,* Toulouse, France. doi:10.1109/ICASSP.2006.1660200

Graesser, A. C., McNamara, D. S., Louwerse, M., & Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavioral Research Methods, Instruments, and Computers, 36*, 193–202. doi:10.3758/BF03195564

Granhag, P. A., Strömwall, L. A., & Olsson, C. (2001, June). *Fact or fiction? Adults's abibility to assess children's veracity.* Paper presented at the I1th European Conference on Psychology and Law, Lisbon, Portugal.

Greenhouse, J. B., & Iyengar, S. (2009). Handling missing data. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (pp. 417-433). New York: Russell Sage.

Greenwald, A. G. (1980). The totalitarian ego: Fabrication and revision of personal history. *American Psychologist, 35,* 603-618. doi:10.1037/0003-066X.35.7.603

Grubin, D., & Madsen, L. (2005). Lie detection and the polygraph: A historical review. *Journal of Forensic Psychiatry & Psychology, 16*, 357-369.

Gupta, S. (2007) *Modelling deception detection in text* (Master's thesis). Retrieved from http://qspace.library.queensu.ca/handle/1974/922.

Hancock, J. T., Curry, L. E., Goorha, S., & Woodworth, M. (2004, July). Lies in conversation: An examination of deception using automated linguistic analysis. *Proceedings of the 26th Annual Conference of the Cognitive Science Society, Vancouver, Canada.* doi:10.1.1.87.9371

Hancock, J. T., Curry, L. E., Goorha, S., & Woodworth, M. (2005, January). Automated linguistic analysis of deceptive and truthful synchronous computer-mediated communication. *Proceedings of the 38th Hawaii International Conference on System Sciences,* Hawaii. doi:10.1109/HICSS.2005.111

*Hancock, J. T., Curry, L. E., Goorha, S., & Woodworth, M. (2008). On lying and being lied to: A linguistic analysis of deception in computer-mediated communication. *Discourse Processes, 45,* 1-23. doi:10.1080/01638530701739181

Hancock, J., & Woodworth, M. (2013). An “eye” for an “I”: The challenges and opportunities for spotting credibility in a digital world. In B. S. Cooper. D. Griesel, & M. Ternes (Eds.), *Applied issues in investigative interviewing, eyewitness memory, and credibility assessment* (pp. 325-340). New York: Springer.

Hancock, J. T., Woodworth, M., & Goorha, S. (2010). See no evil: The effect of communication medium and motivation on deception detection. *Group Decision and Negotiation, 19,* 327-336. doi:10.1007/s10726-009-9169-7

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

Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics, 6,* 107-128. doi:10.3102/10769986006002107

Hedges, L. V. (1982). Estimation of effect size from series of independent experiments. *Psychological Bulletin, 92,* 490-499. doi:10.1037/0033-2909.92.2.490

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis.* New York: Academic Press.

Higgins, J. P. T., & Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine, 21,* 1539-1558. doi:10.1002/sim.1186

Hines, A., Colwell, K., Hiscock-Anisman, C., Garrett, E., Ansarra, R., & Montalvo, L. (2010). Impression management strategies of deceivers and honest reporters in an investigative interview. *European Journal of Psychology Applied to Legal Context, 2,* 73-90.

Hirschberg, J., Beus, S., Brenier, J. M., Enos, F., Friedman, S., Gilman, S., Girand, C., Graciarena, M., Kathol, A., Michaelis, L., Pellom, B., Shriberg, E., & Stolcke, A. (2005, September). Distinguishing deceptive from non-deceptive speech. *Proceedings of the 9th European Conference on Speech Communication and Technology, Interspeech,* 1833-1836, Lisboa, Portugal. doi:10.1.1.59.8634

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

*Humpherys, S. L., Moffitt, K. C., Burns, M. B., Burgoon, J. K., & Felix, W. F. (2011). Identification of fraudulent financial statements using linguistic credibility analysis. *Decision Support Systems, 50,* 585–594. doi:10.1016/j.dss.2010.08.009

*Jensen, M. L., Bessarabova, E., Adame, B., Burgoon, J., & Slowik, S. M. (2011). Deceptive language by innocent and guilty criminal suspects: The influence of dominance, question, and guilt on interview responses. *Journal of Language and Social Psychology, 30,* 357-375. doi:10.1177/0261927X11416201

Jensen, M. L., Burgoon, J. K., & Nunamaker, J. F. (2010). Judging the credibility of information gathered from face-to-face interactions. *ACM Journal of Data and Information Quality, 2,* Article 3. doi:10.1145/1805286.1805289.

Jensen, M. L., Lowry, P. B., Burgoon, J. K., & Nunamaker, J. F. (2010). Technology dominance in complex decision making: The case of aided credibility assessment. *Journal of Management Information Systems, 27,* 175–201. doi:10.2753/MIS0742-1222270108

Jensen, M. L., Lowry, P. B., & Jenkins, J. L. (2011). Effects of automated and participative decision support in computer-aided credibility assessment. *Journal*

of Management Information Systems, 28, 201–233. doi:10.2753/MIS0742-1222280107

Jensen, M. L., Meservy, T. O., Burgoon, J. K., & Nunamaker, J. F. (2010). Automatic, multimodal evaluation of human interaction. *Group Decision and Negotiation, 19,* 367-389. doi:10.1007/s10726-009-9171-0

Johnson, R., Barnhardt, J., & Xhu, J (2004). The contribution of executive processes to deceptive responding. *Neuropsychologia, 42,* 878-901. doi:10.1016/j.neuropsychologia.2003.12.005

Johnson, M. K., Bush, J. C., & Mitchell, K. J. (1998). Interpersonal reality monitoring: Judging the sources of other people's memories. *Social Cognition, 16,* 199-224. doi:10.1521/soco.1998.16.2.199

Johnson, M. K., Hashtroudi, S., & Lindsay, D. S. (1993). Source Monitoring. *Psychological Bulletin, 114,* 3-28. doi:10.1037//0033-2909.114.1.3

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

Johnson, M. K., & Suengas, A. G. (1989). Reality monitoring judgments of other people's memories. *Bulletin of the Psychonomic Society, 27,* 107–110.

Keila, P. S., & Skillicorn, D. B. (2005). *Detecting unusual and deceptive communication in email* (ISSN-0836-0227-.2005-498). External Technical Report, School of Computing, Queen's University. Retrieved from http://research.cs.queensu.ca/. doi:10.1.1.68.3131

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

Knapp, M. L., Hart, R. P., & Dennis H. S. (1974). An exploration of deception as a communication construct. *Communication Research, 1,* 15-29. doi:10.1111/j.1468-2958.1974.tb00250.x

*Koyanagi, J. & Blandón-Gitlin, I. (2011, March). *Analysis of children's deception with the Linguistic Inquiry and Word Count approach.* Poster session presented at the 4th International Congress on Psychology and Law / 2011 Annual Meeting of the American Psychology-Law Society, Miami, Florida.

*Krackow, E. (2010). Narratives distinguish experienced from imagined childhood events. *The American Journal of Psychology, 123,* 71-80. Retrieved from: http://www.jstor.org/stable/10.5406/amerjpsyc.123.1.0071

Kuiken, D. (1981). Nonimmediate language style and inconsistency between private and expressed evaluations. *Journal of Experimental Social Psychology, 17,* 183-196. doi:10.1016/0022-1031(81)90013-5

Lane, J. D., & Wegner, D. M. (1995). The cognitive consequences of secrecy. *Journal of Personality and Social Psychology, 69,* 237-253. doi:10.1037/0022-3514.69.2.237

Leal, S., Vrij, A., Warmelink, L., Vernham, Z., & Fisher, R. P. (2013). You cannot hide your telephone lies: Providing a model statement as an aid to detect deception in insurance telephone calls. *Legal and Criminological Psychology.* Advance online publication. doi:10.1111/lcrp.12017

*Lee, C.-C., Welker, R. B., Odom, M. B. (2009). Features of computer-mediated, text-based messages that support automatable, linguistics-based indicators for deception detection. *Journal of Information Systems, 23,* 5-24. doi:10.2308/jis.2009.23.1.24

Lench, H. C., Flores, S. A., & Bench, S. W. (2011). Discrete emotions predict changes in cognition, judgment, experience, behavior, and physiology: A meta-analysis of experimental emotion elicitations. *Psychological Bulletin, 137,* 834-855. doi:10.1037/a0024244

Levine, T., Asada, K. J., & Carpenter, C. (2009). Sample sizes and effect sizes are negatively correlated in meta-analyses: Evidence and implications of a publication bias against nonsignificant findings. *Communication Monographs, 76*, 286-302. doi:10.1080/03637750903074685

Levine, T. R., & McCornack, S. A. (2001). Behavioral adaptation, confidence, and heuristic-based explanations of the probing effect. *Human Communication Research, 27,* 471-502. doi:10.1111/j.1468-2958.2001.tb00790.x

Levine, T. R., Park, H. S., & McCornack, S. A. (1999). Accuracy in detecting truths and lies: Documenting the "veracity effect". *Communication Monographs, 66,* 125-144. doi:10.1080/03637759909376468

Leuprecht, C. (2011). *Deception in speeches of candidates for public office.* Retrieved from Social Sciences Research Network http://papers.ssrn.com/sol3/papers.cfm?abstract_id=173928. doi:10.2139/ssrn.1739282

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis.* Thousand Oaks: Sage Publications.

*Liu, X., Hancock, J., Zhang, G., Xu, R., Markowitz, D., & Bazarova, N. (2012, January). Exploring linguistic features for deception detection in unstructured text. *Proceedings of the Rapid Screening Technologies, Deception Detection and Credibility Assessment Symposium, Grand Wilea, Hawaii.*

*Masip, J., Bethencourt, M., Lucas, G., Sánchez-San Segundo, M., & Herrero, C. (2012). Deception detection from written accounts. *Scandinavian Journal of Psychology, 53,* 103-111. doi:10.1111/j.1467-9450.2011.00931.x

Masip, J., Garrido, E., & Herrero, C. (2003). Facial appearance and judgments of credibility: The effects of facial babyishness and age on statement credibility. *Genetic Social and General Psychology Monographs, 129,* 269-311.

Masip, J., & Herrero, C. (2013). 'What would you say if you were guilty?' Suspects' strategies during a hypothetical behavior analysis interview concerning a serious crime. *Applied Cognitive Psychology, 27,* 60-70. doi:10.1002/acp.2872

Masip, J., Sporer, S. L., Garrido, E., & Herrero, C. (2005). The detection of deception with the reality monitoring approach: A review of the empirical evidence. *Psychology, Crime, & Law, 11,* 99-122. doi:10.1080/10683160410001726356

McNamara, D. S. & Graesser, A. C. (2012). Coh-Metrix: An automated tool for theoretical and applied natural language processing. In P. M. McCarthy & C. Boonthum-Denecke (Eds.), *Applied natural language processing: Identification, investigation, and resolution* (pp. 188-205). Hershey, PA: Information Science Reference. doi: 10.4018/978-1-60960-741-8

Mehrabian, A. (1972). *Nonverbal communication.* Chicago: Aldine Publishing Company.

Mehrabian, A., & Wiener, M. (1966). Non-immediacy between communicator and object of communication in a verbal message: Application to inference of attitudes. *Journal of Consulting Psychology, 30,* 420-425. doi:10.1037/h0023813

Meissner, C. A., & Kassin, S. M. (2002). "He's guilty!": Investigator bias in judgments of truth and deception. *Law and Human Behavior, 5,* 469-480. doi:10.1023/A:1020278620751

Mihalcea, R., & Strappavara, C. (2009, August). The lie detector: Explorations in the automatic recognition of deceptive language. *Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing Conference Short Papers, 309–312,* Suntec, Singapore.

*Morgan, C. A., Colwell, K., & Hazlett, G. A. (2011). Efficacy of forensic statement analysis in distinguishing truthful from deceptive eyewitness accounts of highly stressful events. *Journal of Forensic Sciences, 56,* 1227-1234. doi:10.1111/j.1556-4029.2011.01896.x

*Morgan, C. A., Mishara, A., Christian, J., & Hazlett, G. A. (2008, August). Detecting deception through automated analysis of translated speech: Credibility assessment of Arabic-speaking interviewees. *Journal of Intelligence Community Research and Development, 1-22.*

Morgan, C. A., Rabinowitz, Y., Christian, J., & Hazlett, G. A. (2009, January). Detecting deception in Vietnamese: Efficacy of forensic statement analysis when interviewing via an interpreter. *Journal of Intelligence Community Research and Development, 1-16.*

Morgan, C. A., Steffian, G., Clark, W., Coric, V., & Hazlett, G. A. (2008). *Efficacy of Verbal and global judgment cues in the detection of deception in persons interviewed via an interpreter.* Unpublished manuscript.

Neisser, U. (1982). *Memory observed: Remembering in natural contexts.* San Fransico: W. H. Freeman and Company.

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

Nunamaker, J. F., Burgoon, J. K., Twyman, N. W., Proudfoot, J. G., Schuetzler, R., & Giboney, J. S. (2012, January). Establishing a foundation for automated human credibility screening. *IEEE International Conference on Intelligence and Security Informatics.* Retrieved from: http://arizona.openrepository.com/arizona/bitstream/10150/222874/1/azu_etd_12045_sip1_m.pdf

*Ott, M., Choi, Y., Cardie, C., & Hancock, J. T. (2011, June). Finding deceptive opinion spam by any stretch of the imagination. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics in Portland, Oregon.*

O'Sullivan, M. (2003). The fundamental attribution error in detecting deception: The boy-who-cried-wolf effect. *Personality and Social Psychology Bulletin, 29,* 1316-1327. doi:10.1177/0146167203254610

Pennebaker, J. W., & Chew, C. H. (1985). Behavioral inhibition and electrodermal activity during deception. *Journal of Personality and Social Psychology, 49,* 1427-1433. doi:10.1037/0022-3514.49.5.1427

Pennebaker, J. W., Francis, M.E., Booth, R.J. (2001). *Linguistic Inquiry and Word Count (LIWC): LIWC 2001.* Mahwah, NJ: Erlbaum.

Pigott, T. D. (2012). *Advances in meta-analysis*. New York: Springer.

Pillemer, D. B., Desrochers, A. B., & Ebanks, C. M. (1998). Remembering the past in the present: Verb tense shifts in autobiographical memory narratives. In C. P. Thompson, D. J. Herrmann, D. Bruce., J. D. Read, D. G. Payne, & M. P. Toglia

(Eds.), *Autobiographical memory: Theoretical and applied perspectives* (pp. 145-162). Malwah, NJ: Erlbaum.

Qin, T., & Burgoon, J. (2007, May). An investigation of heuristics of human judgments in detecting deception and potential implications in countering social engineering. *Proceedings of the IEEE Intelligence and Security Informatics,* 152-159. doi:10.1109/ISI.2007.379548

*Qin, T., Burgoon, J., Blair, J. P., & Nunamaker, J. F. (2005). Modality effects in deception detection and applications in automatic deception detection. *Proceedings of the 38th Hawaii International Conference on System Sciences, Hawaii.* doi:10.1109/HICSS.2005.436

Qin, T., Burgoon, J., & Nunamaker, J. F. (2004). An exploratory study on promising cues in deception detection and application of decision tree. *Proceedings of the 37th Hawaii International Conference on System Sciences.* doi:10.1109/HICSS.2004.1265083

Reinhard, M.-A., Sporer, S. L., Scharmach, M., & Marksteiner, T. (2011). Listening, not watching: Situational familiarity and the ability to detect deception. *Journal of Personality and Social Psychology, 101,* 467-484. doi:10.1037/a0023726

Richard, F. D., Bond, C. F., & Stokes-Zoota, J. J. (2003). One hundred years of social psychology quantitatively described. *Review of General Psychology, 7,* 331-363. doi:10.1037/1089-2680.7.4.331

Rogers, T. B., Kuiper, N. A., & Kirker, W. S., (1977). Self-reference and the encoding of personal information. *Journal of Personality and Social Psychology, 35,* 677-688. doi:10.1037/0022-3514.35.9.677

Rothstein, H. R., Sutton, A. J., & Borenstein, M. (Eds.). (2005). *Publication bias in meta-analysis: Prevention, assessement and adjustments.* Chichester: John Wiley. doi:10.1002/0470870168.ch1

*Rowe, K. & Blandón-Gitlin, I. (2008, March). *Discriminating true, suggested, and fabricated statements with the Linguistic Inquiry and Word Count approach.* Poster presented at the Annual Meeting of the American Psychology-Law Society, Jacksonville, Florida.

Rubin, V. L., & Conroy, N. (2012, March). Discerning truth from deception: Human judgments of automation efforts. *First Monday, 17,* 3. Retrieved from http://firstmonday.org/htbin/

Sauerland, M., & Sporer, S. L. (2011). Written vs. spoken eyewitness accounts: Does modality of testing matter? *Behavioral Sciences and the Law, 29,* 846-857. doi:10.1002/bsl.1013

*Schafer, J. R. (2007). *Grammatical differences between truthful and deceptive narratives* (Unpublished doctoral dissertation). Fielding Graduate University, Santa Barbara, CA.

Schank, R. & Abelson, R. (1977). *Scripts, plans, goals, and understanding: An inquiry into human knowledge structure.* Hillsdale, NJ: Lawrence Erlbaum Associates.

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

Shadish, W. R., & Haddock, C. K. (2009). Combining estimates of effect size. In H. Cooper, L. V. Harris, & J. C. Valentine (Eds.), *The handbook of research*

synthesis and meta-analysis (pp. 257-277). New York: Russell Sage Foundation.

Slamecka, N. J., & Graf, P. (1978). The generation effect: Delineation of a phenomenon. *Journal of Experimental Psychology: Human Learning and Memory, 4,* 592-604. doi:10.1037//0278-7393.4.6.592

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

Sporer, S. L. (1998, March). *Detecting deception with the Aberdeen Report Judgment Scales (ARJS): Theoretical development, reliability and validity.* Paper presented at the Biennial Meeting of the American Psychology-Law Society in Redondo Beach, CA.

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

Sporer, S. L. (2008). Lessons from the origins of eyewitness testimony research in Europe. *Applied Cognitive Psychology, 22,* 737-757. doi:10.1002/acp.1479

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

Sporer, S. L. (2013, March). *Content-criteria to detect deception: Methodological pitfalls and solutions.* Paper presented at the Annual Meeting of the American Psychology-Law Society, Portland, USA.

Sporer, S. L., & Cohn, L. D. (2011). Meta-analysis. In B. D. Rosenfeld, & S. D. Penrod (Eds.), *Research methods in forensic psychology* (pp. 43-62). New York: Wiley.

Sporer, S. L., & Küpper, B. (1995). Realitaetsueberwachung und die Beurteilung des Wahrheitsgehaltes von Erzaehlungen: Eine experimentelle Studie [Reality monitoring and the judgment of credibility of stories: An experimental investigation]. *Zeitschrift für Sozialpsychologie, 26,* 173-193.

Sporer, S. L., & Schwandt, B. (2006). Paraverbal correlates of deception: A meta-analysis. *Applied Cognitive Psychology, 20,* 421-446. doi:10.1002/acp.1190

Sporer, S. L., & Schwandt, B. (2007). Moderators of nonverbal indicators of deception: A meta-analytic synthesis. *Psychology, Public Policy, and Law, 13,* 1-34. doi:10.1037/1076-8971.13.1.1

Sporer, S. L., & Sharman, S. J. (2006). Should I believe this? Reality monitoring of accounts of self-experienced and invented recent and distant autobiographical events. *Applied Cognitive Psychology, 20,* 985-1001. doi:10.1002/acp1234

Sporer, S. L., & Walther, A. (2006, March). *Truth Detection by Content Cues: General vs. Specific Questions.* Paper presented at the Meeting of the American Psychology-Law Society in Petersburg, FL.

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

Strömwall, L. A., Hartwig, M., & Granhag, P. A. (2006). To act truthfully: Nonverbal behavior and strategies during a police interrogation. *Psychology, Crime & Law, 12,* 207-219. doi:10.1080/10683160512331331328

*Suckle-Nelson, J. A., Colwell, K., Hiscock-Anisman, C., Florence, S., Youschak, K. E., & Duarte, A. (2010). Assessment Criteria Indicative of Deception (ACID): Replication and gender differences. *The Open Criminology Journal, 3,* 23-30.

Sutton, A. J. (2009). Publication bias. In H. Cooper, L. V. Harris, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (pp. 435-452). New York: Russell Sage Foundation.

Tausczik, Y., & Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology, 29,* 24-54. doi:10.1177/0261927X09351676

Taylor, P. J., Tomblin, S., Conchie, S. M., Menacere, T. (2011, March). *Verbal indicators of deception in some cultures are indicators of truth in others.* Paper presented at the 4th International Congress on Psychology and Law and the Annual Meeting of the American Psychology-Law Society, Miami, USA.

*ten Brinke, L., & Porter, S. (2012). Cry me a river: Identifying the behavioural consequences of extremely high-stakes interpersonal deception. *Law and Human Behavior, 36,* 469-477. doi:10.1037/h0093929

Toma, C. L., & Hancock, J. T. (2010, February). Reading between the lines: Linguistic cues to deception in online dating profiles. *Proceedings of the ACM Conference on Computer Supported Cooperative Work, 5-8, Savannah, USA.* doi:10.1145/1718918.1718921

Toma, C. L., & Hancock, J. T. (2012). What lies beneath: The linguistic traces of deception in online dating profiles. *Journal of Communication, 62,* 78-97. doi:10.1111/j.1460-2466.2011.01619.x

Twitchell, D. P. (2005). *Automated analysis techniques for online conversations with application in deception detection* (Doctoral dissertation). Retrieved from http://arizona.openrepository.com/arizona/handle/10150/194997

Twitchell, D. P., Biros, D. P., Adkins, M., Forsgren, N., Burgoon, J. K., & Nunamaker, J. F. (2006, January). Automated determination of the veracity of interview statements from people of interest to an operational security force. *Proceedings of the 39th Hawaii International Conference on System Sciences.* doi:10.1109/HICSS.2006.70

Twitchell, D. P., Nunamaker, J. F., & Burgoon, J. K. (2004). Using speech act profiling for deception detection. In H. Chen (Ed.), *Lecture Notes in Computer Sciences 3073. Intelligence and Security Informatics* (pp. 403–410). Berlin, Germany: Springer-Verlag.

*Van Swol, L. M., Braun, M. T., & Malhotra, D. (2012). Evidence for the Pinocchio Effect: Linguistic differences between lies, deception by omissions, and truths. *Discourse Processes, 49,* 79-106. doi:10.1080/0163853X.2011.633331

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

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

Vrij, A. (2008b). Nonverbal dominance versus verbal accuracy in lie detection. A plea to change police practice. *Criminal Justice and Behavior, 35,* 1323-1336. doi:10.1177/0093854808321530

Vrij, A., Akehurst, L., Soukara, R., & Bull, R. (2004). Detecting deceit via analyses of verbal and nonverbal behavior in adults and children. *Human Communication Research, 30,* 8–41. doi:10.1111/j.1468-2958.2004.tb00723.x

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

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

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

Vrij, A., Mann, S., Leal, S., & Granhag, P. A. (2010). Getting into the minds of pairs of liars and truth-tellers: An examination of their strategies. *The Open Criminology Journal, 3,* 17-22. doi:10.2174/1874917801003010017

Wagner, H., & Pease. K. (1976). The verbal communication of inconsistency between attitudes held and attitudes expressed. *Journal of Personality, 44,* 1-15. doi:10.1111/j.1467-6494.1976.tb00580.x

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

Walczyk, J. J., Igou, F. P., Dixon, A. P., & Tcholakian, T. (2013). Advancing lie detection by inducing cognitive load on liars: A review of relevant theories and techniques guided by lessons from polygraph-based approaches. *Frontiers in Psychology, 4*(14), 1-13. doi:10.3389/fpsyg.2013.00014

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

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

Walker, W. R. & Skowronski, J. J. (2009). The fading affect bias: But what the hell is it for? *Applied Cognitive Psychology, 23,* 1122–1136. doi:10.1002/acp.1614

Walker, W. R., Skowronski, J. J., & Thompson, C. P. (2003). Life is pleasant and memory helps to keep it that way. *Review of General Psychology, 7,* 203–210. doi:10.1037/1089-2680.7.2.203

Walker, W. R., Vogl, R. J., & Thompson, C. P. (1997). Autobiographical memory: Unpleasantness fades faster than pleasantness over time. *Applied Cognitive Psychology, 11,* 399–413. doi:10.1002/(SICI)1099-0720(199710)11:5<399::AID-ACP462>3.0.CO;2-E

Watson, K. W. (1981). *Oral and written linguistic indices of deception during employment interviews* (Unpublished doctoral dissertation). Graduate Faculty of the Louisiana State University, Baton Rouge, Louisiana, USA.

Wertheimer, M., & Klein, J. (1904). Psychologische Tatbestandsdiagnostik [Psychological assessment of facts about an event]. *Archiv für Kriminologie, Anthropologie und Kriminalistik, 15,* 72–113.

Wiener, M., & Mehrabian, A. (1968). *Language within language: Immediacy, a channel in verbal communication.* New York: Appleton-Century-Crofts.

*Williams, S. M., Talwar, V., Lindsay, R. C., Bala, N. C., & Lee, K. (2012). Is the truth in your words? Distinguishing children's deceptive and truthful statements. *Journal of Criminology, 2014.* doi:10.1155/2014/547519

Wilson, D. B. (2002). *Meta-analysis macros for SAS, SPSS, and Stata.* Retrieved from http://mason.gmu.edu/~dwilsonb/ma.html

*Zhou, L. (2005). An empirical investigation of deception behavior in instant messaging. *IEEE Transactions on Professional Communication, 48,* 147-160. doi:10.1109/TPC.2005.849652

Zhou, L., Booker, Q. E., & Zhang, D. (2002). ROD: Towards rapid ontology development for underdeveloped domains. *Proceedings of the 35th Hawaii International Conference on System Sciences.* doi:10.1109/HICSS.2002.994046

*Zhou, L., Burgoon, J. K., Nunamaker, J. F., & Twitchell, D. (2004). Automating linguistics-based cues for detecting deception in text-based asynchronous computer-mediated communication. *Group Decision and Negotiation, 13,* 81-106. doi:10.1023/B:GRUP.0000011944.62889.6f

Zhou, L., Burgoon, J. K., Twitchell, D. P., Qin, T. T., & Nunamaker, J. F. (2004). A comparison of classification methods for predicting deception in computer-mediated communication. *Journal of Management Information Systems, 20,* 139-165.

Zhou, L., Burgoon, J. K., Zhang, D. S., & Nunamaker, J. F. (2004). Language dominance in interpersonal deception in computer-mediated communication. *Computers in Human Behavior, 20,* 381-402. doi:10.1016/S0747-5632(03)00051-7

Zhou, L., Shi, Y. M., & Zhang, D. S. (2008). A statistical language modeling approach to online deception detection. *IEEE Transactions on Knowledge and Data Engineering, 20,* 1077-1081. doi:10.1109/TKDE.2007.190624

Zhou, L., & Sung, Y.-W. (2008). Cues to deception in online Chinese groups. *Proceedings of the 41st Hawaii International Conference on System Sciences.* doi:10.1109/HICSS.2008.109

Zhou, L., Twitchell, D. P., Qin, T., Burgoon, J. K., & Nunamaker, J. F. (2003). An exploratory study into deception detection in text-based computer-mediated communication. *Proceedings of the 36th Hawaii International Conference on System Sciences.* doi:0.1109/HICSS.2003.1173793

*Zhou, L., & Zhang, D. (2004). Can online behavior unveil deceivers? An exploratory investigation of deception in instant messaging. *Proceedings of the 37th Hawaii International Conference on System Sciences.* doi:10.1109/HICSS.2004.1265079

Zhou, L., & Zhang, D. S. (2006). A comparison of deception behavior in dyad and triadic group decision making in synchronous computer-mediated communication. *Small Group Research, 37,* 140-164. doi:10.1177/1046496405285125

Zhou, L., & Zenebe, A. (2005). Modeling and handling uncertainty in deception detection. *Proceedings of the 38th Hawaii International Conference on System Sciences,* Hawaii. doi:10.1109/HICSS.2005.438

Zhou, L., & Zenebe, A. (2008, April). Representation and reasoning under uncertainty in deception detection: A neuro-fuzzy approach. *IEEE Transactions on Fuzzy Systems, 16,* 2, 442-454. doi:10.1109/TFUZZ.2006.889914

Zuckerman, M., DePaulo, B. M., & Rosenthal, R. (1981). Verbal and nonverbal communication of deception. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 14, pp. 1–60). New York: Academic.

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

Zuckerman, M., Koestner, R. E., Colella, M. J., & Alton, A. O. (1984). Anchoring in the detection of deception and leakage. *Journal of Personality and Social Psychology, 47,* 301-311. doi:10.1037/0022-3514.47.2.301

Footnotes

[1] Similar to Vrij (2008a), we use this term to denote theory regarding both emotions or feelings *and* arousal. While differences between these states have been noted, their overlap has also been acknowledged (Zuckerman et al., 1981).

[2] Ekman (2001) noted that a liar may experience joy ("duping delight"). However, the link between this emotion and verbal cues to deception is not clear (Vrij, 2008a). Therefore, we do not consider it further.

[3] Due to empty cells or small cell sizes in each category, we had to merge previously more differentiated categories to broader categories (see Appendix C).

[4] Although we are aware of some potential confounding variables, such as production mode, communication medium, perspective of sender (e.g., actor or observer), or length of interaction, we developed this moderator variable to find subgroups of studies that were similar in terms of the intensity of interaction between sender and another person. Originally, the categories were more sophisticated, but due to small cell sizes, we had to collapse some related categories.

[5] Although we consider word count an important variable, this variable has been investigated in several other meta-analyses (e.g., Sporer & Schwandt, 2006; DePaulo et al. 2003, and Zuckerman & Driver, 1985: combination of *duration* and *number of words*), plus in all the studies that investigated linguistic cues summarized here. Also, many studies on content cues to deception assessed by humans have reported on word count, usually by using a word processor. To review all these studies (likely to be several hundred) where the main focus was not on computer-aided detection of deception would constitute a meta-analysis of its own and is beyond the scope of this paper.

[6] Even when four outliers (with two positive and two negative values) were excluded for *negative emotions only*, the effect remained significant (*k* = 20, -0.12 [-0.19, -0.04]).

[7] Due to the fairly large number of potential pairwise comparisons for each moderator variable and linguistic cue, we did not calculate these specific comparisons. More differentiated results for homogeneity test statistics between and within groups as well as all other (e.g., nonsignificant) moderator-analytic results can be requested from the first author.

[8] Because purely positive events (*k* = 3) as well as a combination of positive and negative events (*k* = 6) were quite rarely used, they were excluded from moderator analyses.

Table 1

Meta-Analyses of Linguistic Cues under Research Questions 1 to 6

Linguistic Cue	Pred. DOE	*k*	*N*	Min g_u	Max g_u	g_u	CI-low	CI-high	*Q*	I^2
Research Question 1: Do Liars Experience Greater Cognitive Load?										
(a) Length of Accounts										
01 Word Quantity[M]	T	42	6,713	-1.25	1.43	**0.24**	0.19,	0.29	**315.85**	87.02
07 Sentence Quantity	T	9	1,334	-1.31	0.28	**-0.33**	-0.44,	-0.21	**104.01**	93.31
08 Average Sentence Length[wO,M]	T	15/16	2,704	-0.37	0.43	0.05	-0.03,	0.13	20.46	31.58
(b) Elaboration of Accounts										
02 Content Word Diversity[wO]	T	7/9	1,076	0.27	1.00	**0.48**	0.34,	0.61	8.13	26.22
03 Type-Token Ratio[M]	T	22	3,589	-1.40	1.09	**0.14**	0.07,	0.21	**171.95**	87.79
04 Six Letter Words	T	10	1,617	-0.28	0.25	-0.05	0.14,	-0.05	5.19	0.00
05 Average Word Length[wO]	T	7/8	954	-0.42	0.69	0.11	-0.03,	0.25	6.85	12.41
(c) Complexity of Accounts										
06 Verb Quantity	T	12	2,356	-1.21	0.44	-0.03	0.11,	-0.60	**78.91**	86.06
09 Causation	T	17	2,773	-0.68	0.25	-0.07	-0.14,	0.01	18.61	14.03

10 Exclusive Words[wO,M]	T	18/20	2,783	-0.24	0.81	**0.24**	0.17,	0.31	25.87	25.66
(d) Errors in Production										
11 Writing Errors[wO]	D	8/10	990	-0.65	0.43	-0.01	-0.15,	0.12	13.36	47.61
Research Question 2: Are Liars Less Certain Than Truth-Tellers?										
12 Tentative Words[wO]	D	19/20	3,145	-1.27	0.36	**0.13**	0.06,	0.20	20.13	10.56
13 Modal Verbs	D	25	3,889	-0.42	0.80	0.00	-0.07,	0.06	32.73	26.62
14 Certainty	T	18	2,823	-0.25	0.94	-0.06	-0.14,	0.01	25.15	32.40
Research Question 3a: Do Liars Use More Negations and Negative Emotion Words?										
17 Negations[M]	D	20	3,659	-0.98	0.53	**-0.15**	-0.22,	-0.09	**155.53**	87.78
18 Negative Emotions[+,wO,M]	D	21/24	2,593	-0.39	0.87	-0.07	-0.15,	0.01	21.03	4.88
18.1 Negative Emotions Only[M]	D	24	3,641	-1.90	0.87	**-0.18**	-0.24,	-0.12	**214.57**	89.28
18.2 Anger	D	12	2,452	-1.32	0.38	**-0.27**	-0.35,	-0.19	**165.03**	93.34
18.3 Anxiety[wO]	D	11/12	1,952	-0.30	0.44	0.07	-0.02,	0.02	13.55	26.19
18.4 Sadness	D	12	2,452	-0.34	0.25	0.04	-0.04,	0.12	13.01	15.46
Research Question 3b: Do Liars Use Fewer Positive Emotion Words?										

19 Positive Emotions and Feelings[+,wO]	T	20/21	2,703	-0.84	0.37	-0.05	-0.12,	0.03	29.59	35.79
19.1 Positive Emotions Only[wO,M]	T	20/21	2,703	-0.84	0.35	-0.07	-0.15,	0.00	27.98	32.08
19.2 Positive Feelings Only	T	9	1,422	-0.47	0.40	0.07	-0.03,	0.18	14.88	46.25
Research Question 3c: Do Liars Express More or Less Unspecified Emotion Words?										
15 Emotions[wO,M]	?	21/25	2,941	-0.63	0.48	**-0.11**	-0.19,	-0.04	28.92	30.85
16 Pleasantness and Unpleasantness	?	6	806	-0.35	0.30	-0.10	-0.25,	0.06	8.43	40.68
Research Question 4: Do Liars Distance Themselves More From Events?										
(a) Personal Pronouns										
21 First-Person Singular[M]	T	22	3,761	-1.00	0.61	-0.06	-0.13,	0.00	**274.38**	92.35
22 First-Person Plural[wO,M]	T	22/25	3,224	-0.72	0.39	0.06	-0.01,	0.13	27.98	24.93
23 Total First-Person[wO,M]	T	22/23	2,541	-0.39	0.57	**0.14**	0.06,	0.22	32.26	34.91
24 Total Second-Person[wO,M]	D	21/23	3,072	-0.61	0.33	**-0.10**	-0.17,	-0.02	28.64	30.16
25 Total Third-Person[wO,M]	D	26/29	3,848	-0.41	0.55	**-0.10**	-0.16,	-0.04	37.20	32.79
20 Total Pronouns[wO,M]	-	18/19	2,460	-0.36	0.65	0.06	-0.02,	0.14	19.32	12.02
(b) Passive Voice and Generalizing Terms										

26 Passive Voice Verbs	D	11	1,221	-0.47	0.49	0.06	-0.06,	0.18	9.52	0.00
27 Generalizing Terms	D	4	93	-1.63	0.44	-0.37	-0.79,	0.05	**15.78**	80.99
(c) Past and Present Tense										
47 Past Tense	D	16	3,047	-0.53	0.41	0.06	-0.01,	0.14	22.67	33.83
48 Present Tense^{wO,M}	T	16/17	2,607	-0.51	0.60	0.01	-0.07,	0.09	19.38	22.61
Research Question 5: Do Liars Use Fewer (Sensory and Contextual) Details?										
(a) Sensory and Perceptual Details										
28 Sens.-Perceptual Processes$^{+,wO,M}$	T	25/27	3,957	-0.70	0.70	0.05	-0.01,	0.12	36.00	33.33
28.1 Sens.-Percept. Processes Only^{M}	T	27	4,177	-0.90	0.70	**0.06**	0.00,	0.13	**89.26**	70.87
28.2 Seeing^{wO}	T	9/11	1,740	-0.17	0.34	0.03	-0.06,	0.13	13.34	40.03
28.3 Feeling^{wO}	T	11/12	2,304	-0.49	0.27	-0.03	-0.11,	0.05	15.00	33.31
28.4 Hearing	T	11	2,344	-0.41	0.48	**0.17**	0.09,	0.25	14.15	29.35
(b) Contextual Embedding										
29 Time^{wO}	T	23/24	3,296	-1.25	0.53	0.03	-0.04,	0.10	28.95	24.00
30 Space^{wO,M}	T	22/24	3,199	-0.36	0.58	-0.04	-0.13,	0.03	31.74	33.74

31 Space & Time	T	5	634	-0.25	0.61	-0.04	-0.19,	0.12	**10.48**	61.84
(c) Descriptive Words										
32 Prepositions	T	14	2,479	-0.55	0.48	0.02	-0.06,	0.10	16.54	21.38
33 Numbers	T	12	2,452	-0.28	0.23	0.05	-0.03,	0.13	9.37	0.00
34 Quantifier	T	4	1,198	0.06	0.22	**0.14**	0.02,	0.25	1.22	0.00
35 Modifier	T	11	1,361	-1.04	0.43	-0.08	-0.20,	0.03	**77.46**	87.09
36 Motion Verbs^{wO,M}	T	16/17	2,359	-0.72	0.13	**-0.09**	-0.17,	-0.01	16.84	10.92
Research Question 6: Do Liars Refer Less Often to Cognitive Processes?										
37 Cognitive Processes^{wO}	T	18/19	2,915	-0.25	0.36	**0.09**	0.01,	0.16	19.66	13.54
38 Insight^{M}	T	15	2,539	-0.41	0.59	**0.13**	0.05,	0.21	**35.65**	60.73

Notes. g_u = effect size Hedges' g_u, positive g_us indicate higher frequencies in true accounts, negative g_us indicate higher frequencies in deceptive accounts; ^{wO} = without outlier: Results after removal of outliers detected using Hedges and Olkin's (1985) procedure; ^{M} = Moderator analyses conducted; $^{+}$indicates that the specific linguistic cue is an umbrella term; Pred. DOE = predicted direction of effect; T = occurs more often in true accounts: D = occurs more often in deceptive accounts; *k* = number of hypothesis tests, where the second value behind the slash indicates the number of hypothesis tests with outliers; *N* = total number of accounts; *Min* = minimum; *Max* = maximum; *Q* = homogeneity test statistic; CI = 95% confidence interval; I^2 = descriptive measure of heterogeneity; **values in bold** indicate significance ($p < .05$).

Table 2

Effect Sizes of Linguistic Cues to Deception when Studies used Different Type of Events and Personal Involvement

Linguistic Cue (Research Question, RQ)	*k*	Overall g_u [CI]	k_1	Attitude/Liking Paradigm	k_3	First-Person Experience	k_2	Miscellaneous Paradigms
08 Average Sentence Length^{wO} (RQ1)	15	0.04 [-0.04, 0.12]	2	***0.17 [0.05, 0.29]***	8	-0.07 [-0.20, 0.06]	5	-0.07 [-0.17, 0.13]
17 Negations (RQ3a)	20	**-0.15 [-0.22, -0.08]**	7	0.08 [-0.02, 0.18]	7	-0.08 [-0.20, 0.05]	6	***-0.59 [-0.71, -0.47]***
18 Neg. Emotions^{+} (RQ3a)	24	**-0.13 [-0.19, -0.06]**	7	0.03 [-0.06, 0.13]	10	***-0.37 [-0.48, -0.25]***	7	-0.10 [-0.24, 0.03]
18.1 Neg. Emotions Only (RQ3a)	24	**-0.17 [-0.24, -0.11]**	7	0.06 [-0.04, 0.15]	10	***-0.57 [-0.69, -0.45]***	7	-0.11 [-0.25, 0.03]
15 Emotions (RQ3c)	24	**-0.21 [-0.27, -0.14]**	7	-0.08 [-0.17, 0.02]	12	***-0.45 [-0.56, -0.34]***	5	-0.02 [-0.19, 0.15]
22 First-Person Plural (RQ4)	24	**0.08 [0.01, 0.14]**	7	0.09 [-0.01, 0.19]	12	-0.08 [-0.18, 0.03]	5	***0.38 [0.23, 0.53]***
23 Total First-Person^{wO} (RQ4)	22	**0.14 [0.06, 0.22]**	6	**0.31 [0.19, 0.42]**	11	-0.11 [-0.27, 0.05]	5	0.10 [-0.05, 0.26]
24 Total Second-Person (RQ4)	22	-0.04 [-0.01, 0.03]	7	***-0.18 [-0.27, -0.08]***	10	0.09 [-0.02, 0.20]	5	0.09 [-0.09, 0.26]
25 Total Third-Person (RQ4)	28	**-0.11 [-0.17, -0.05]**	7	**-0.12 [-0.21, -0.02]**	**13**	***-0.22 [-0.32, -0.12]***	8	0.06 [-0.07, 0.18]

Notes. k = number of hypothesis tests; g_u = effect size (ES) Hedges' g_u, positive g_us indicate higher frequencies in true accounts, negative g_us indicate higher frequencies in deceptive accounts; CI = 95% confidence interval; ^{wO} = without Outlier; $^{+}$ indicates that the specific linguistic cue is an umbrella term; Neg.= Negative; **bold** ES indicate significant difference from zero; ES in *italics* correspond to the largest difference between liars and truth tellers (in the predicted direction according to the RQ) for a specific cue across the moderator variable levels; this indicates under what level of the moderator variable our hypotheses were most strongly supported.

Table 3

Effect Sizes of Linguistic Cues to Deception when the Emotional Valence of the Event was Neutral versus Negative

Linguistic Cue (Research Question, RQ)	*k*	Overall g_u [CI]	k_1	Neutral	k_2	Negative
01 Word Quantity (RQ1)	33	**0.25 [0.19, 0.31]**	17	0.04 [-0.03, 0.12]	16	***0.54 [0.45, 0.62]***
03 Type-Token Ratio (RQ1)	20	**0.26 [0.18, 0.33]**	11	***0.32 [0.24, 0.41]***	9	0.04 [-0.12, 0.19]
10 Exclusive Words (RQ1)	14	**0.38 [0.29, 0.46]**	8	***0.47 [0.36, 0.59]***	6	**0.26 [0.11, 0.38]**
17 Negations (RQ3a)	14	**-0.30 [-0.38, -0.21]**	8	**-0.13 [-0.26, 0.01]**	6	***-0.42 [-0.53, -0.31]***
18 Neg. Emotions (RQ3a)	17	**-0.26 [-0.35, -0.18]**	10	**-0.16 [-0.28, -0.05]**	7	***-0.39 [-0.52, -0.26]***
18.1 Neg. Emotions Only^{wO} (RQ3a)	17	**-0.41 [-0.50, -0.32]**	10	**-0.22 [-0.34, -0.10]**	7	***-0.65 [-0.79, -0.52]***
15 Emotions (RQ3c)	18	**-0.50 [-0.58, -0.42]**	8	**-0.54 [-0.65, -0.43]**	10	**-0.45 [-0.57, -0.33]**
21 First-Person Singular^{wO} (RQ4)	15	-0.04 [-0.12, 0.05]	9	**-0.25 [-0.36, -0.13]**	6	***0.27 [0.14, 0.40]***
23 Total First-Person^{wO} (RQ4)	17	**0.10 [0.01, 0.20]**	8	***0.22 [0.10, 0.34]***	9	-0.13 [-0.30, 0.05]
24 Total Second-Person^{wO} (RQ4)	15	**-0.20 [-0.28, -0.12]**	9	***-0.40 [-0.50, -0.29]***	6	0.09 [-0.05, 0.22]

Notes. Please consult notes from Table 2.

Table 4

Effect Sizes of Linguistic Cues to Deception when Studies Applied Different Type of Interaction Levels (between Sender and Receiver)

Linguistic Cue (Research Question, RQ)	*k*	Overall g_u [CI]	k_1	No Interaction	k_2	Computer-Mediated Communication	k_3	Interview	k_4	Person to Person Interaction
01 Word Quantity (RQ1)	37	**0.22 [0.17, 0.28]**	13	**0.14 [0.07, 0.21]**	5	**-0.41 [-0.65, -0.18]**	16	**0.35 [0.23, 0.46]**	3	***0.69 [0.53, 0.85]***
10 Exclusive Words (RQ1)	19	**0.30 [0.23, 0.36]**	9	***0.37 [0.29, 0.45]***	2	-0.02 [-0.31, 0.26]	6	**0.25 [0.04, 0.46]**	2	**0.17 [0.02, 0.33]**
18 Neg. Emotions^{+} (RQ3a)	22	**-0.14 [-0.21, -0.07]**	8	0.03 [-0.07, 0.12]	3	-0.18 [-0.46, 0.10]	7	-0.16 [-0.34, 0.01]	4	***-0.48 [-0.63, -0.34]***
18.1 Neg. Emotions Only (RQ3a)	22	**-0.20 [-0.27, -0.13]**	8	0.05 [-0.05, 0.14]	3	-0.18 [-0.46, 0.10]	7	**-0.18 [-0.36, -0.01]**	4	***-0.79 [-0.94, -0.64]***
15 Emotions (RQ3c)	24	**-0.34 [-0.41, -0.28]**	11	**-0.36 [-0.44, -0.28]**	0		11	-0.04 [-0.19, 0.11]	2	***-0.63 [-0.79, -0.47]***
21 First-Person Singular^{wO} (RQ4)	21	-0.06 [-0.12, 0.01]	9	**-0.22 [-0.30, -0.13]**	2	-0.04 [-0.33, 0.24]	7	0.01 [-0.16, 0.19]	3	***0.34 [0.20, 0.49]***
25 Total Third-Person (RQ4)	27	**-0.21 [-0.27, -0.15]**	10	***-0.31 [-0.40, -0.23]***	2	-0.21 [-0.52, 0.09]	12	-0.04 [-0.18, 0.09]	3	-0.09 [-0.23, 0.06]

Notes. Please consult notes from Table 2.

Table 5

Effect Sizes of Linguistic Cues to Deception when Studies Induced Different Levels of Motivation

Linguistic Cue (Research Question, RQ)	*k*	Overall g_u [CI]	k_1	No Motivation	k_2	Low to Medium Motivation	k_3	High Motivation
01 Word Quantity (RQ1)	37	**0.27 [0.21, 0.32]**	11	***0.47 [0.36, 0.57]***	22	**0.19 [0.12, 0.26]**	4	**0.18 [0.04, 0.31]**
03 Type-Token Ratio (RQ1)	19	0.00 [-0.08, 0.08]	4	-0.17 [-0.39, 0.05]	12	**-0.12 [-0.21, -0.02]**	3	***0.67 [0.47, 0.87]***
08 Average Sentence Length^{wO} (RQ1)	13	0.08 [-0.01, 0.16]	5	0.09 [-0.09, 0.28]	6	**0.15 [0.04, 0.26]**	2	-0.21 [-0.42, 0.01]
18 Neg. Emotions^{+} (RQ3a)	21	**-0.14 [-0.21, -0.07]**	6	**-0.19 [-0.36, -0.03]**	13	-0.01 [-0.10, 0.07]	2	***-0.56 [-0.74, -0.39]***
18.1 Neg. Emotions Only (RQ3a)	21	**-0.20 [-0.27, -0.13]**	6	**-0.20 [-0.37, -0.03]**	13	0.00 [-0.09, 0.09]	2	***-1.03 [-1.21, -0.86]***
15 Emotions (RQ3c)	23	**-0.22 [-0.29, -0.15]**	4	-0.20 [-0.42, 0.01]	15	**-0.10 [-0.19, -0.02]**	4	***-0.53 [-0.66, -0.39]***
28 Sens.-Perc. Processes^{+} (RQ5)	25	**0.08 [0.02, 0.15]**	7	**-0.22 [-0.39, -0.06]**	15	**0.12 [0.03, 0.20]**	3	***0.21 [0.07, 0.35]***
28.1 Sens.-Perc. Processes Only (RQ5)	25	**0.08 [0.02, 0.15]**	7	**-0.29 [-0.45, -0.13]**	15	**0.12 [0.03, 0.20]**	3	***0.25 [0.12, 0.38]***
29 Time^{wO} (RQ5)	21	0.03 [-0.05, 0.10]	4	***0.20 [0.01, 0.40]***	15	-0.02 [-0.10, 0.07]	2	0.09 [-0.12, 0.30]

Notes. Please consult notes from Table 2.

Table 6

Effect Sizes of Linguistic Cues to Deception when Studies used Different Modes of Producing an Account

Linguistic Cue (Research Question, RQ)	*k*	Overall g_u [CI]	k_1	Handwritten	k_2	Typed	k_3	Spoken
01 Word Quantity (RQ1)	38	**0.19 [0.13, 0.24]**	6	***0.33 [0.21, 0.44]***	14	**0.10 [0.03, 0.17]**	18	**0.26 [0.15, 0.36]**
17 Negations (RQ3a)	19	**-0.19 [-0.26, -0.12]**	5	***-0.60 [-0.72, -0.46]***	5	0.06 [-0.05, 0.17]	9	-0.14 [-0.28, -0.01]
18.1 Neg. Emotions Only (RQ3a)	23	-0.04 [-0.11, 0.03]	3	***-0.28 [-0.51, -0.05]***	8	0.07 [-0.03, 0.16]	12	**-0.14 [-0.25, -0.02]**
15 Emotions (RQ3c)	21	**-0.29 [-0.36, -0.22]**	4	**-0.25 [-0.44, -0.07]**	6	***-0.44 [-0.54, -0.35]***	11	-0.04 [-0.16, 0.08]
28 Sens.-Perc. Processes^{+} (RQ5)	24	0.06 [-0.01, 0.13]	3	***0.33 [0.11, 0.54]***	8	0.01 [-0.08, 0.11]	13	0.06 [-0.06, 0.17]
28.1 Sens.-Perc. Processes Only (RQ5)	24	0.05 [-0.01, 0.12]	3	***0.34 [0.12, 0.56]***	8	0.00 [-0.09, 0.10]	13	0.05 [-0.07, 0.16]
30 Space (RQ5)	22	0.04 [-0.03, 0.11]	3	0.03 [-0.19, 0.24]	6	***0.13 [0.02, 0.23]***	13	-0.06 [-0.17, 0.05]
36 Motion Verbs^{wO} (RQ5)	16	**-0.09 [-0.17, -0.01]**	2	**-0.28 [-0.57, 0.00]**	4	0.00 [-0.11, 0.11]	10	**-0.16 [-0.29, -0.04]**

Notes. Please consult notes from Table 2.

Table 7

Effect Sizes of Linguistic Cues to Deception when Studies used LIWC, a General Program or a Specific Program

Linguistic Cue	*k*	Overall g_u [CI]	k_1	LIWC	k_2	General Program	k_3	Specific Program
01 Word Quantity	41	**0.25 [0.20, 0.30]**	23	**0.28 [0.21, 0.34]**	10	***0.53 [0.43, 0.63]***	8	**-0.19 [-0.30, -0.07]**
15 Emotions	25	**-0.25 [-0.41, -0.28]**	19	**-0.39 [-0.45, -0.32]**	-	-	6	-0.14 [-0.29, 0.02]
17 Negations	20	**-0.15 [-0.22, -0.08]**	17	0.05 [-0.03, 0.12]	3	***-0.82 [-0.96, -0.69]***	-	-
20 Total Pronouns[wO]	18	**0.29 [0.20, 0.38]**	13	**0.38 [0.28, 0.49]**	2	0.06 [-0.17, 0.28]	3	-0.13 [-0.44, 0.17]
			k_1	LIWC + General Program			k_2	Specific Program
22 First-Person Plural	25	-0.04 [-0.10, 0.02]	18	0.01 [-0.05, 0.08]			7	**-0.31 [-0.46, -0.15]**

Notes. Please consult notes from Table 2.

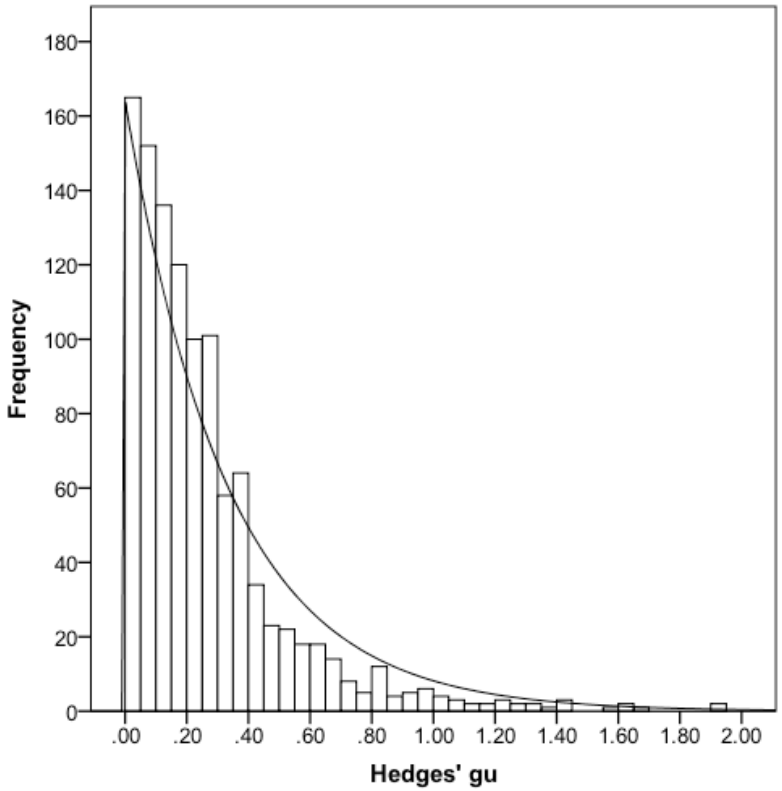

Figure 1. Distribution of all absolute values effect sizes (*k* = 1,093).

Appendix A

Definition of Linguistic Cues to Deception assigned to Research Questions

	Linguistic Cue	Final Operational Definition
	Research Question 1: Do Liars Experience Greater Cognitive Load?	
01*	**Word Quantity** // Word Count // Number of Words // Productivity	Total number of words.
02	**Content Word Diversity** // Diversity // Content Diversity	Total number of *different* content words divided by total number of content words, where content words express lexical meaning.
03	**Type-Token Ratio** // Unique Words // Lexical Diversity // Different Words	% of distinct words divided by total number of words.
04	**Six letter words** // Percentage Words longer than six letters	% of words that are longer than six letters.
05	**Average Word Length** (AWL; Complexity) // Lexical complexity	Total number of letters divided by the total number of words.
06*	**Verb Quantity** // Verb Count	Total number of verbs.
07*	**Sentence Quantity** // Number of Sentences	Total number of sentences.
08	**Average Sentence Length** (Complexity Measure) // Words per Sentence	Total number of words divided by total numbers of sentences.
09	**Causation**	% of words that try to assign a cause to whatever the person is describing (e.g., because, effect, hence).

10	**Exclusive**	% of words that make a distinction what is in a category and what is not (e.g., without, except, but).
11	**Writing Errors** // Typographical error ratio (Informality) // Typo ratio // Misspelled Words	% of writing errors or misspelled words divided by number of words.
	Research Question 2: Are Liars Less Certain Than Truth-Tellers?	
12	**Tentative**	% of tentative words (e.g., maybe, perhaps, see).
13	**Modal Verbs** // Uncertainty // Discrepancy	% of modal verbs or auxiliary verbs or words expressing uncertainty (e.g., should, would, could).
14	**Certainty**	% of words that express certainty (e.g., always, never).
	Research Question 3a: Do Liars Use More Negations and Negative Emotion Words?	
17	**Negations** // Less Positive Tone // Spontaneous Negations // Negation Connectives	% of words that express negations (e.g., no, never, not).
18**+**	**Negative Emotions** // Negative Affect // Anger // Anxiety, Fear // Sadness	% of words that express negative emotion / affect (e.g., hate, worthless, enemy) AND anger (e.g., hate, kill, annoyed) AND anxiety (e.g., worried, fearful, nervous) AND sadness (e.g., crying, grief, sad).
18.1	**Negative Emotions (only)** // Negative Affect	% of words that express negative emotion / affect (e.g., hate, worthless, enemy).
18.2	**Anger**	% of words that express anger (e.g., hate, kill, annoyed).
18.3	**Anxiety**	% of words that express anxiety (e.g., worried, fearful, nervous).

18.4	**Sadness**	% of words that express sadness (e.g., crying, grief, sad).
	Research Question 3b: Do Liars Use Less Positive Emotion Words?	
19**+**	**Positive Emotions and Feelings** // Positive Emotions // Positive Affects // Positive Feelings	% of words that express positive emotion / affect (e.g., happy, pretty, good) AND positive feelings (e.g., joy, love).
19.1	**Positive Emotions (only)**// Positive Affect	% of words that express positive emotion / affect (e.g., happy, pretty, good).
19.2	**Positive Feelings (only)**	% of words that express positive feelings (e.g., joy, love).
	Research Question 3c: Do Liars Express More or Less Unspecified Emotion Words?	
15	**Emotions //** Emotional / Affective Processes // Affect (Ratio) // Positive and Negative Affect	% of words that express any type of emotions / affects (e.g., happy, ugly, bitter).
16	**Pleasantness and Unpleasantness**	% of words that express pleasantness / unpleasantness.
	Research Question 4: Do Liars Distance Themselves More From Events?	
20	**Total Pronouns** // Personal Pronouns	% of all personal (e.g., I, our, they) or total pronouns (e.g., that, somebody, the).
21	**First-Person Singular**	% of first-person singular pronouns (e.g., I, my, me).
22	**First-Person Plural**	% of first-person plural pronouns (e.g., we, us, our).
23	**Total First-Person**	% of first-person singular and first-person plural pronouns (e.g., I, we, me).

24	**Total Second-Person**	% of second-person pronouns (e.g., you, you'll).
25	**Total Third-Person** // Other References // Third-Person Singular // Third-Person Plural	% of third-person pronouns (e.g., she, their, them).
26	**Passive Voice Verbs** // Verbal Nonimmediacy	% of passive voice verbs (e.g., “it was searched for”).
27	**Generalizing Terms** // Leveling terms	% of generalizing terms (e.g., everybody, all, anybody).
47	Past Tense Verb	% of past tense verbs (e.g., went, drove, ate).
48	Present Tense Verb	% of present tense verbs of all words (e.g., walk, run, cry).

Research Question 5: Do Liars Use Fewer (Sensory and Contextual) Details?

28**+**	**Sensory-Perceptual Processes** // Perceptual Processes/Information // Perceptions and Sense // Sensory ratio // See // Hear // Feel	% of words that express sensory-perceptual processes (e.g. taste, touch, feel) AND visual (e.g., view, saw, seen) AND haptical (e.g., feels, touch) AND aural (e.g., listen, hearing) sensory-perceptual processes.
28.1	**Sensory-Perceptual Processes (only)** // Perceptual Processes // Perceptual Information // Perceptions and Sense // Sensory ratio	% of words that express sensory-perceptual processes (e.g. taste, touch, feel).
28.2	**Seeing**	% of words that express visual sensory-perceptual processes (e.g., view, saw, seen).
28.3	**Feeling**	% of words that express tactile sensory-perceptual processes (e.g., feels, touch).

28.4	**Hearing**	% of words that express aural sensory-perceptual processes (e.g., listen, hearing).
29	**Time** // Temporal ratio // Temporal specificity // Temporal cohesion	% of temporal words (e.g., hour, day, o'clock).
30	**Space** // Spatial Terms // Spatial Ratio // Spatial Specificity // Spatial Cohesion	% of spatial words (e.g., around, over, up).
31	**Temporal-spatial Terms** // Temporal and Spatial Details Total // Spatio-Temporal Information // Space and Time	% of temporal (e.g., hour, day, o'clock) AND spatial words (e.g., around, over, up).
32	**Prepositions**	% of prepositions (e.g., on, to, from).
33	**Numbers**	% of numbers (e.g., first, one, thousand).
34	**Quantifier**	% of quantifier (e.g., all, bit, few, less).
35	**Modifiers** (Adverbs & Adjectives) // Rate of Adjectives and Adverbs (Specificity and Expressiveness)	% of modifier: adverbs & adjectives (e.g., here, much, few, very).
36	**Motion Verbs** // Motion Terms	% of words that describe movements (e.g., walk, move, go).
	Research Question 6: Do liars refer less often to cognitive processes?	
37	**Cognitive Processes** // All Connectives	% of words related to cognitive processes (e.g., cause, know, ought).
38	**Insight**	% of words related to a person's insight (e.g., think, know, consider).

Notes. **Bold** font indicate the name of the linguistic cue chosen for this meta-analysis; ***** No ratio; + indicates that the specific linguistic cue is an umbrella term; % = number of specific words divided by total number of words.

Appendix B

Definition of Linguistic Cues to Deception--Miscellaneous Category

	Linguistic Cue	Final Operational Definition
39	Redundancy	Ratio of function words to number of sentences. Function words, such as articles and pronouns, are used to form grammatical relationships between other words. // The ratio of the number of function words to the number of messages. // Repetitive words. // Argument overlap: Explicit overlap between two sentences by tracking the common nouns in either single or plural form.
40	Assent	% of words that express an assent (e.g., agree, ok, yes).
41	Articles	% of articles (e.g., a, lot, an, the).
42	Inhibition	% of words that express inhibition (e.g., block, constrain, stop).
43	Social Processes	% of words that express social processes e.g., (talk, us, friend).
44	Friends	% of words that are related to friends (e.g., buddy, friend, neighbor).
45	Family	% of words that are related to family (e.g., daughter, husband, aunt).
46	Humans	% of words that are related to humans (e.g., adult, baby, boy).
49	Future Tense Verb	% of future tense verbs (e.g., will, going to).
50	Inclusive	% of inclusive words (e.g., with, and, include).
51	Achievement	% of words that express achievement (e.g., earn, hero, win).
52	Leisure	% of words that express leisure activities (e.g., cook, chat, movie).

53	Emotiveness	Total number of adjectives and total numbers of adverbs divided by total number of nouns and total numbers of verbs.
54	Pausality	Total number of punctuation marks divided by total number of sentences.
55	Swear Words	% of swear words (e.g., ass, heck, shit).
56	Biology	% of words that express biological processes/states (e.g., eat, pain, wash).
57	Health	% of words that express health issues (e.g., hospital, pill, flu).
58	Sexual	% of words that express sexual activities/states (e.g., passion, rape, sex).
59	Optimism	% of words that express optimism (e.g., certainty, pride, win).
60	Communication	% of words that express communication (e.g., talk, share, converse).
61	Occupation	% of words that express occupation (e.g., work, class, boss).
62	School	% of words that express school issues (e.g., class, student, college).
63	Job / Work	% of words that express job issues (e.g., employ, boss, career).
64	Home	% of words that express home issues (e.g., bed, home, room).
65	Sports	% of words that express sport (e.g., football, game, play).
66	Money	% of words that express money and financial issues (e.g., cash, taxes, income).
67	Physical	% of words that express physical states and functions (e.g., ache, breast, sleep).
68	Body	% of words that express body states and symptoms (e.g., asleep, heart, cough).
69	Eating	% of words that express eating, drinking, dieting issues (e.g., eat, swallow, taste).

Notes. % = number of specific words divided by total number of words.

Appendix C

Coding Decisions for Moderator Variables for Each Study

Authors (Year)	Publ. Type	Program	Lang.	Theory	Select.	Age	Prepa-ration	Event Type	Vale-nce	Inter-action	Moti-vation	Mode
Ali & Levine (2008, denials)	publ.	LIWC01	E	IDT/RM	a-priori	adults	n/a	mock crime	neg.	interview	low	spoken
Ali & Levine (2008, confess.)	publ.	LIWC01	E	IDT/RM	a-priori	adults	n/a	mock crime	neg.	interview	low	spoken
Almela et al. (2012)	publ.	LIWC01	S	none	a-priori	adults	n/a	att./liking	neg./pos.	none	low	typed
Bedwell et al. (2011)	publ.	Coh-Metrix	E	other	sign.	adults	prep.	trivial LE	neutral	instruct.	low	spoken
Bond & Lee (2005)	publ.	LIWC01	E	IDT/RM	a-priori	adults	prep.	video	neg.	interact.	low	spoken
Brunet (2009)*	Thesis	LIWC01	E	LIWC	a-priori	child.	n/a	sign. LE	neg./pos.	interview	none	spoken
Burgoon & Qin (2006)	publ.	GATE	E	IDT/RM	a-priori	adults	n/a	other	neutral	interview	low	spoken
Chen (2010; Dataset 3)	Diss.	LIWC01	E	IDT/RM	a-priori	adults	n/a	n/a	neutral	none	n/a	typed
Colwell et al. (2002)	publ.	Wordscan	E	other	a-priori	adults	n/a	live	neg.	interview	n/a	n/a
Cooper (2008)	Diss.	Connexor^{+}	E	IDT/RM	a-priori	adults	n/a	other	neutral	n/a	n/a	typed
Derrick et al. (2012)	pres.	ADAM	E	IDT/RM	a-priori	adults	n/a	other	neutral	none	low	typed

Dzindolet & Pierce (2005)	pres.	LIWC01	E	LIWC	a-priori	adults	n/a	att./liking	neg./pos.	instruct.	n/a	written
Evans et al. (2012, Interv. 1)	publ.	LIWC01	E	LIWC	a-priori	child.	n/a	other	pos.	interview	none	n/a
Fuller et al. (2006, Agent99A.)	pres.	Agent99A.+	E	IDT/RM	a-priori	adults	n/a	real case	neg.	n/a	high	written
Hancock et al. (2008) / Duran et al. (2010)	publ.	LIWC01/ Coh-Metrix	E	IDT/RM	a-priori	adults	prep.	trivial LE	neg./pos.	cmc	n/a	typed
Humpherys et al. (2011)	publ.	Agent99A.	E	IDT/RM	a-priori	adults	n/a	real case	neutral	none	high	typed
Jensen et al. (2011)	publ.	LIWC01	E	other	a-priori	adults	n/a	real case	neg.	interview	high	n/a
Koyanagi & Blandón-Gitlin (2011)	pres.	LIWC07	E	IDT/RM	a-priori	child.	no	other	neutral	interview	none	spoken
Krachow (2010)	publ.	LIWC07	E	IDT/RM	a-priori	adults	prep.	trivial LE	neg./pos.	instruct.	none	spoken
Lee et al. (2009)	publ.	LIWC01	E	other	a-priori	adults	n/a	other	neutral	cmc	low	typed
Liu et al. (2012)	pres.	LIWC07+	E	LIWC	a-priori	adults	n/a	real case	neg.	interact.	high	n/a
Masip et al. (2012)	publ.	LIWC07	S	IDT/RM	a-priori	adults	no	trivial LE	pos.	instruct.	low	written
Morgan et al. (2011, free recall)	publ.	“automated analysis method”	E	none	a-priori	adults	n/a	n/a	neg.	interview	low	spoken
Morgan et al. (2008, free recall)	publ.	n/a (general)	A	none	a-priori	adults	no	mock crime	neg.	interview	med.	spoken

Newman et al. (2003, Exp. 1)	publ.	LIWC01	E	LIWC	sign.	adults	n/a	att./liking	neutral	instruct.	low	spoken
Newman et al. (2003, Exp. 2)	publ.	LIWC01	E	LIWC	sign.	adults	n/a	att./liking	neutral	instruct.	low	typed
Newman et al. (2003, Exp. 3)	publ.	LIWC01	E	LIWC	sign.	adults	n/a	att./liking	neutral	none	low	written
Newman et al. (2003, Exp. 4)	publ.	LIWC01	E	LIWC	sign.	adults	n/a	att./liking	neg./pos.	instruct.	low	spoken
Newman et al. (2003, Exp. 5)	publ.	LIWC01	E	LIWC	sign.	adults	n/a	mock crime	neg.	interview	low	spoken
Ott et al. (2011)	pres.	LIWC07	E	IDT/RM	sign.	adults	n/a	att./liking	pos.	none	low	typed
Qin et al. (2005, audio)	pres.	GATE	E	other	a-priori	adults	n/a	mock crime	neg.	interview	low	spoken
Qin et al. (2005, face-to-face)	pres.	GATE	E	other	a-priori	adults	n/a	mock crime	neg.	interview	low	n/a
Qin et al. (2005, text chat)	pres.	GATE	E	other	a-priori	adults	n/a	mock crime	neg.	interview	low	typed
Rowe & Blandón-Gitlin (2008)	pres.	LIWC07	E	IDT/RM	a-priori	adults	no	other	neutral	interview	none	spoken
Schafer (2007, Exp. 1)	Diss.	MS Word	E	other	a-priori	adults	n/a	video	neg.	n/a	none	written
Schafer (2007, Exp. 2)	Diss.	MS Word	E	other	a-priori	adults	n/a	video	neg.	n/a	none	written
Schelleman-Offermans & Merckelbach (2010)	publ.	LIWC01	D	LIWC	a-priori	adults	n/a	sign. LE	neg.	n/a	none	typed
Suckle-Nelson et al. (2011, free recall)	publ.	Wordscan	E	IDT/RM	a-priori	adults	n/a	staged	neg.	interview	none	spoken

ten Brinke & Porter (2012)	publ.	LIWC01	E	LIWC	a-priori	adults	n/a	real case	neg.	interact.	high	spoken
Van Swol et al. (2012)	publ.	LIWC07[+]	E	IDT/RM	a-priori	adults	n/a	other	neutral	interact.	med.	spoken
Williams et al. (2012)	subm.°	LIWC07	E	LIWC	a-priori	child.	n/a	sign. LE	neutral	interview	none	spoken
Zhou (2005)	publ.	"NLP tool"	E	other	a-priori	adults	n/a	other	neutral	cmc	low	typed
Zhou et al. (2004)	publ.	iSkim/CueCal	E	other	a-priori	adults	n/a	other	neutral	cmc	none	typed
Zhou & Zhang (2004)	pres.	"Message analyzing software"	E	other	a-priori	adults	n/a	other	neutral	cmc	low	typed

Notes. confess. = confessions; Publ. = Publication; publ. = published; pres. = presented (Poster or Paper); subm. = submitted; Diss. = Dissertation; LIWC01 = LIWC 2001; LIWC07 = LIWC 2007; GATE = General Architecture for Text Engineering; ADAM = Automated Deception Analysis Machine; Agent99A. = Agent99Analyzer; MS Word = Microsoft Word; NLP = natural language processing; [+] = Study additionally applied a second program; Lang. = Language; A = Arabic; D = Dutch; E = English; S= Spanish; n/a = not available; IDT = Interpersonal Deception Theory; RM = reality monitoring; Select. = Selection; child. = children; Prepar./prep. = preparation; att. = attitude; staged = staged event; sign. = significant; LE = life events; cmc = computer-mediated communication; instruct. = instruction; med. = medium; * = In the meantime, Brunet, Evans, Talwar, Bala, Lindsay, and Lee (2013) formally published the data of Brunet's Thesis; ° = In the meantime, Williams, Talwar, Lindsay, Bala & Lee (2012) published their (at the time of conducting the meta-analyses unpublished) manuscript.

Appendix D

Meta-Analyses on Miscellaneous Linguistic Cues with and without Outliers

Linguistic Cue	*k*	*N*	Min g_u	Max g_u	g_u	CI-low	CI-high	*Q*	I^2
39 Redundancy	9	1,262	-0.33	0.42	0.00	-0.12,	0.12	9.12	12.30
40 Assent	12	2,452	-0.23	0.38	-0.01	-0.09,	0.07	12.50	12.02
41 Articles	14	2,479	-1.95	0.26	-0.01	-0.08,	0.08	**39.49**	67.08
41 Articles^{wO}	12	1,777	-1.95	0.15	-0.02	-0.11,	0.07	18.68	41.12
42 Inhibition	12	2,452	-0.37	0.31	**0.12**	0.04,	0.20	19.27	42.91
43 Social Processes	15	2,979	-1.69	0.26	**-0.26**	-0.33,	-0.18	**236.03**	94.07
43 Social	14	2,479	-0.48	0.26	-0.04	-0.12,	0.04	16.97	23.39
44 Friends	11	2,344	-0.47	0.51	0.08	-0.01,	0.16	**23.89**	58.15
44 Friends^{wO}	9	1,734	-0.47	0.39	0.00	-0.10,	0.09	11.95	33.07
45 Family	10	2,284	-0.39	0.36	-0.03	-0.11,	0.06	**30.77**	70.75
45 Family^{wO}	9	1,784	-0.16	0.36	0.07	-0.02,	0.17	10.09	20.70
46 Humans	11	2,344	-0.45	0.42	0.03	-0.06,	0.11	**27.85**	64.09
46 Humans^{wO}	9	1,734	-0.27	0.42	**0.12**	0.02,	0.21	12.82	37.58
49 Future Tense	15	2,979	-1.00	0.48	**-0.24**	-0.31,	-0.16	**98.26**	85.75
49 Future Tense^{wO}	14	2,497	-0.35	0.48	**-0.10**	-0.18,	-0.02	22.02	40.97
50 Inclusive	15	2,979	-1.00	**0.16**	-0.16	-0.23,	-0.09	**106.73**	86.88
50 Inclusive^{wO}	14	2,497	-0.43	0.16	-0.01	-0.01,	0.07	13.77	5.60
51 Achievement	11	2,344	-0.55	0.46	0.04	-0.04,	0.12	17.96	44.31
52 Leisure	11	2,344	-0.41	0.21	-0.05	-0.13,	0.03	**24.37**	58.97
52 Leisure^{wO}	10	1,844	-0.41	0.11	**-0.12**	-0.21,	-0.03	13.96	35.54

(*Appendix D continues*)

Appendix D (*continued*)

Linguistic Cue	*k*	*N*	Min g_u	Max g_u	g_u	CI-low	CI-high	*Q*	I^2
53 Emotiveness	9	1,158	-0.39	0.27	0.04	-0.09,	0.16	9.06	22.74
54 Pausality	8	1,158	-0.31	0.75	-0.09	-0.21,	0.04	**16.57**	57.76
54 Pausality^{wO}	7	1,128	-0.31	0.46	-0.11	-0.24,	0.01	11.23	46.59
55 Swear Words	10	2,284	-0.17	0.31	-0.04	-0.12,	0.04	5.11	0.00
56 Biology	4	1,198	-0.41	0.40	**0.16**	0.05,	0.28	**9.87**	69.61
57 Health	4	1,198	-0.11	0.22	0.05	-0.06,	0.16	6.51	53.89
58 Sexual	9	2,190	-0.31	0.57	0.08	-0.01,	0.16	12.02	33.42
59 Optimism	8	1,254	-0.28	0.30	0.01	-0.10,	0.12	9.02	22.40
60 Communication	8	1,254	-0.25	0.20	0.07	-0.04,	0.18	2.87	0.00
61 Occupation	7	1,146	-0.39	0.09	-0.07	-0.18,	0.05	4.87	0.00
62 School	7	1,146	-0.16	0.28	0.03	-0.08,	0.15	4.12	0.00
63 Job	11	2,344	-0.31	0.20	0.02	-0.06,	0.11	9.27	0.00
64 Home	11	2,344	-0.36	0.36	0.03	-0.05,	0.11	**24.06**	58.43
64 Home^{wO}	10	1,844	-0.36	0.36	-0.03	-0.12,	0.06	16.64	45.91
65 Sports	7	1,146	-0.55	0.08	-0.08	-0.19,	0.04	8.16	26.44
66 Money	12	2,446	-0.85	0.43	-0.01	-0.08,	0.08	**33.62**	67.28
66 Money^{wO}	11	2,344	-0.41	0.43	0.03	-0.05,	0.11	17.96	45.33
67 Physical	7	1,146	0.03	0.31	**0.15**	0.04,	0.27	1.52	0.00
68 Body	11	2,344	-0.39	0.44	0.02	-0.07,	0.10	13.40	25.38
69 Eating	7	1,146	-0.08	0.59	**0.12**	0.01,	0.24	7.85	23.54

Notes. Please consult notes from Table 1.

Appendix E

Effect Sizes of Linguistic Cues to Deception when Studies were Unpublished or Published

Linguistic Cue	*k*	Overall g_u [CI]	k_1	Unpublished	k_2	Published
01 Word Quantity	42	0.24 [0.19, 0.29]	17	***0.27 [0.21, 0.34]***	25	**0.19 [0.11, 0.27]**
03 Type-Token Ratio	22	0.14 [0.74, 0.21]	9	***0.24 [0.15, 0.33]***	13	0.03 [-0.07, 0.13]
15 Emotions	25	-0.34 [-0.41, -0.28]	10	***-0.56 [-0.65, -0.48]***	15	**-0.11 [-0.20, -0.02]**
18.1 Negative Emotions	24	-0.18 [-0.24, -0.11]	8	***-0.24 [-0.33, -0.14]***	16	**-0.12 [-0.21, -0.03]**
20 Total Pronouns^{wO}	18	0.29 [0.20, 0.37]	7	**0.45 [0.33, 0.58]**	11	**0.13 [0.01, 0.25]**
21 First-Person Singular	21	-0.06 [-0.13, 0.00]	8	**-0.29 [-0.38, -0.20]**	14	***0.19 [0.10, 0.29]***
23 Total First-Person^{wO}	22	0.14 [0.06, 0.22]	8	0.05 [-0.08, 0.18]	14	***0.20 [0.10, 0.30]***
24 Total Second-Person	23	-0.16 [-0.23, -0.10]	12	***-0.21 [-.29, -0.13]***	11	-0.09 [-0.19, -0.01]
25 Total Third-Person	29	-0.21 [-0.27, -0.15]	13	***-0.31 [-0.38, -0.23]***	16	-0.08 [-0.17, -0.01]
28.1 Sens.-Percept. Processes^{+}	27	0.06 [0.00, 0.13]	11	0.00 [-0.09, 0.09]	16	***0.13 [0.04, 0.22]***
30 Space	23	0.00 [-0.07, 0.06]	10	0.07 [-0.03, 0.16]	13	-0.07 [-0.16, 0.02]
36 Motion Verbs^{wO}	16	-0.10 [-0.17, 0.01]	5	0.01 [-0.12, 0.14]	11	**-0.15 [-0.26, -0.05]**

Notes. Please consult notes from Table 2.

Appendix F

Effect Sizes of Linguistic Cues to Deception when Studies Applied either a Between- or Within-Participants Design

Linguistic Cue	*k*	Overall g_u [CI]	k_1	Between-Participants	k_2	Within-Participants
01 Word Quantity	42	0.24 [0.19, 0.29]	24	0.07 [-0.01, 0.15]	18	***0.36 [0.30, 0.43]***
03 Type-Token Ratio	22	0.14 [0.07, 0.21]	16	***0.33 [0.24, 0.42]***	6	-0.08 [-0.18, 0.02]
12 Tentative Words^{wO+}	19	0.13 [0.06, 0.20]	9	0.03 [-0.09, 0.14]	10	***0.19 [0.11, 0.28]***
15 Emotions	25	-0.35 [-0.41, -0.28]	14	**-0.50 [-0.58, -0.40]**	11	**-0.22 [-0.30, -0.13]**
17 Negations	20	-0.15 [-0.22, -0.08]	7	**0.17 [0.03, 0.30]**	13	***-0.25 [-0.32, -0.17]***
18 Negative Emotions^{+}	24	-0.13 [-0.19, -0.06]	11	0.04 [-0.08, 0.15]	13	***-0.22 [-0.29, 0.13]***
18.1 Negative Emotions Only	24	-0.18 [-0.24, -0.11]	11	0.09 [-0.02, 0.20]	13	***-0.32 [-0.40, -0.24]***
19 Positive Emotions and Feelings^{+}	20	-0.04 [-0.13, 0.04]	11	**-0.16 [-0.27, -0.05]**	9	0.13 [-0.01, 0.26]
19.1 Positive Emotions Only^{wO}	20	-0.07 [-0.15, 0.00]	11	**-0.15 [-0.26, -0.04]**	9	0.00 [-0.11, 0.10]
20 Total Pronouns^{wO}	18	0.29 [0.10, 0.38]	8	-0.05 [-0.23, 0.14]	10	**0.38 [0.28, 0.48]**
21 First-Person Singular	22	-0.06 [-0.13, 0.00]	9	**-0.70 [-0.80, -0.59]**	13	***0.28 [0.20, 0.36]***
22 First-Person Plural	25	-0.04 [-0.10, 0.03]	15	**-0.16 [-0.25, -0.08]**	10	***0.09 [0.01, 0.18]***
23 Total First-Person^{wO}	22	0.14 [0.06, 0.22]	14	0.00 [-0.12, 0.12]	8	***0.26 [0.15, 0.37]***
24 Total Second-Person	23	-0.16 [-0.23, -0.10]	**10**	***-0.42 [-0.52, -0.32]***	13	0.00 [-0.08, 0.08]
25 Total Third-Person	29	-0.21 [-0.27, -0.15]	**15**	***-0.34 [-0.42, -0.25]***	14	**-0.11 [-0.19, -0.03]**
48 Present Tense^{wO}	16	0.01 [-0.06, 0.09]	7	-0.10 [-0.24, 0.04]	9	0.06 [-0.03, 0.15]
28.1 Sens.-Percept. Processes Only	27	0.06 [0.00, 0.13]	15	-0.04 [-0.13, 0.05]	12	***0.14 [0.06, 0.23]***
30 Space	24	0.00 [-0.06, 0.07]	**12**	***0.21 [0.11, 0.32]***	12	**-0.12 [-0.20, -0.04]**
37 Cognitive Processes^{wO}	18	0.09 [0.01, 0.16]	7	-0.03 [-0.17, 0.10]	11	***0.13 [0.05, 0.22]***
38 Insight	15	0.13 [0.05, 0.21]	7	-0.07 [-0.21, 0.06]	8	***0.23 [0.14, 0.32]***

Notes. Please consult notes from Table 2.

Appendix G

Excluded Studies and Reason for Exclusion

Authors	Reason for Exclusion
Adams (2002)	No additional data from authors
Adams & Jarvis (2006)	No additional data from authors
Bachenko, Fitzpatrick & Schonwetter (2008)	Data not applicable
Burgoon, Blair, Qin, & Nunamaker (2003)	Same data as Qin, Burgoon, Blair, & Nunamaker (2005)
Churyk, Lee, & Clinton (2008)	Truth status of management discussion not clear (could be fraud), no data
Dilmon (2009)	Response from author: No computer program used. A statistician conducted the coding.
Dulaney (1982)	Data (means, *F*-value) are not sufficient to calculate appropriate ES for within-participants design
Duran, Crossley, Hall, McCarthy, & Namara (2009)	Same data as Duran, Hall, McCarthy, & McNamara (2010)
Dzindolet & Pierce (2004)	First author could not provide data
Elkins (2011)	Data not applicable
Enos (2009)	Not enough data to calculate effect sizes
Enos, Shribergy, Graciarena, Hirschberg, & Stolcke (2007)	Not enough data to calculate effect sizes
Fornaciari & Poesio (2011)	Parts of speech instead of whole account. Not enough statistical data (use of "vectors")
Fuller (2008)	Data not applicable
Fuller, Biros, & Delen (2008)	Data not applicable

(*Appendix G continues*)

Appendix G (*continued*)

Authors	Reason for Exclusion
Fuller, Biros, & Delen (2011)	Not enough data to calculate effect sizes, no independent dataset (same as Fuller, Biros, & Wilson, 2009)
Fuller, Biros, & Wilson (2009)	Not enough data and no independent dataset
Graciarena, Shriberg, Stolcke, Enos, Hirschberg, & Kajarekar (2006)	No linguistic categories outlined specifically - used a superordinate category (Prosodic/Lexical)
Gupta (2007)	Not enough data to calculate effect sizes
Hancock, Curry, Goorha, & Woodworth (2004)	Exactly the same data as in Hancock et al. (2008)
Hancock, Curry, Goorha, & Woodworth (2005)	Exactly the same data as in Hancock et al. (2008)
Hirschberg, Beus, Brenier, Enos, Friedman, Gilman, Girand, Graciarena, Kathol, Michaelis, Pellom, Shriberg, & Stolcke (2005)	Not enough data to calculate effect sizes
Jensen, Burgoon, & Nunamaker (2010)	Hybrid detection system whereby humans interact with the program to aid them in making truth and deception decisions
Jensen, Lowry, Burgoon, & Nunamaker (2010)	Hybrid detection system whereby humans interact with the program to aid them in making truth and deception decisions
Jensen, Lowry, & Jenkins (2011)	Hybrid detection system whereby humans interact with the program to aid them in making truth and deception decisions
Jensen, Meservy, Burgoon, & Nunamaker (2010)	No independent data. Same transcripts as in Burgoon, et al. (2003)
Keila & Skillicorn (2005)	Data not applicable
Knapp, Hart, & Dennis (1974)	No standard deviations reported and first author could not provide them

(*Appendix G continues*)

Appendix G (*continued*)

Authors	Reason for Exclusion
Leuprecht (2011)	No data available
Little (2007)	Not enough data to calculate effect sizes
Mihalcea & Strappavara (2009)	Not enough data and no independent dataset
Morgan, Rabinowitz, Christian, & Hazlett (2009)	Analyses of interviewers speech only.
Morgan, Steffian, Clark, Coric, & Harzlett (2008)	No data available
Qin & Burgoon (2007)	Same dataset as Burgoon & Qin (2006)
Qin, Burgoon, & Nunamaker (2004)	Same data as Qin, Burgoon, Blair, & Nunamaker (2005)
Rubin & Conroy (2012)	Analysis of statements with various (continuous and self-determined) deception levels and no restricted topics. No data available
Taylor, Tomblin, Conchie, & Menacere (2011)	Not enough data to calculate effect sizes
Toma & Hancock (2010)	No truth condition, only low and high deceptive condition, and no data. Same dataset as Toma & Hancock (2012)
Toma & Hancock (2012)	No truth condition, only low and high deceptive condition, and no data
Twitchell (2005)	Same data as Twitchell, Adkins, Nunamaker, & Burgoon (2004)
Twitchell, Nunamaker, & Burgoon (2004)	Only one linguistic cue investigated
Twitchell, Biros, Adkins, Forsgren, Burgoon, Nunamaker (2006)	Data not applicable
Vrij, Mann, Kristen, & Fisher (2007)	First author could not provide data

(*Appendix G continues*)

Appendix G (*continued*)

Authors	Reason for Exclusion
Watson (1981)	Statistical data is not useful for computing effect sizes
Zhou, Burgoon, Twitchell, Qin, & Nunamaker (2004)	No additional data from authors
Zhou, Burgoon, Zhang, & Nunamaker (2004)	Same data as Zhou et al. (2004), but two additional dependent variables: "intensity" and "subjunctive language". Data for those cues is neither provided in the article, nor in data sent from the authors.
Zhou, Shi, & Zhang (2008)	Data not applicable
Zhou & Sung (2008)	First author recommend us to exclude this study because it does not have an independent data set
Zhou, Twitchell, Qin, Burgoon, & Nunamaker (2003)	First author recommend us to exclude this study because it does not have an independent data set
Zhou & Zenebe (2005)	Data not applicable
Zhou & Zenebe (2008)	Data not applicable
Zhou & Zhang (2006)	First author recommend us to exclude this study because it does not have an independent data set

Appendix H

Meta-Analyses of Linguistic Cues under Research Questions 1 to 6 with and without Outliers

Linguistic Cue	Pred. DOE	*k*	*N*	Min g_u	Max g_u	g_u	CI-low	CI-high	*Q*	I^2
Research Question 1: Do Liars Experience Greater Cognitive Load?										
(a) Length of Accounts										
01 Word Quantity[M]	T	42	6,713	-1.25	1.43	**0.24**	0.19,	0.29	**315.85**	87.02
07 Sentence Quantity	T	9	1,334	-1.31	0.28	**-0.33**	-0.44,	-0.21	**104.01**	93.31
08 Average Sentence Length	T	16	2,880	-0.37	0.81	**0.10**	0.02,	0.17	**42.83**	64.98
08 Average Sentence Length[wO,M]	T	15	2,704	-0.37	0.43	0.05	-0.03,	0.13	20.46	31.58
(b) Elaboration of Accounts										
02 Content Word Diversity	T	9	1,194	-0.30	1.00	**0.39**	0.26,	0.51	**22.77**	64.87
02 Content Word Diversity[wO]	T	7	1,076	0.27	1.00	**0.48**	0.34,	0.61	8.13	26.22
03 Type-Token Ratio[M]	T	22	3,589	-1.40	1.09	**0.14**	0.07,	0.21	**171.95**	87.79
04 Six Letter Words	T	10	1,617	-0.28	0.25	-0.05	0.14,	-0.05	5.19	0.00
05 Average Word Length	T	8	1,158	-0.59	0.69	-0.03	-0.16,	0.09	**25.95**	73.02
05 Average Word Length[wO]	T	7	954	-0.42	0.69	0.11	-0.03,	0.25	6.85	12.41
(c) Complexity of Accounts										
06 Verb Quantity	T	12	2,356	-1.21	0.44	-0.03	0.11,	-0.60	**78.91**	86.06
09 Causation	T	17	2,773	-0.68	0.25	-0.07	-0.14,	0.01	18.61	14.03
10 Exclusive Words	T	20	3,403	-0.24	0.81	**0.31**	0.24,	0.38	**43.91**	56.73
10 Exclusive Words[wO,M]	T	18	2,783	-0.24	0.81	**0.24**	0.17,	0.31	25.87	25.66

(*Appendix H continues*)

Appendix H (*continued*)

Linguistic Cue	Pred. DOE	*k*	*N*	Min g_u	Max g_u	g_u	CI-low	CI-high	*Q*	I^2
(d) Errors in Production										
11 Writing Errors	D	10	1,077	-0.65	0.86	-0.03	-0.16,	0.11	**22.10**	59.29
11 Writing Errors^{wO}	D	8	990	-0.65	0.43	-0.01	-0.15,	0.12	13.36	47.61
Research Question 2: Are Liars Less Certain Than Truth-Tellers?										
12 Tentative Words	D	20	3,197	-1.27	0.36	**0.11**	0.04,	0.18	**36.76**	48.31
12 Tentative Words^{wO}	D	19	3,145	-1.27	0.36	**0.13**	0.06,	0.20	20.13	10.56
13 Modal Verbs	D	25	3,889	-0.42	0.80	0.00	-0.07,	0.06	32.73	26.62
14 Certainty	T	18	2,823	-0.25	0.94	-0.06	-0.14,	0.01	25.15	32.40
Research Question 3a: Do Liars Use More Negations and Negative Emotion Words?										
17 Negations^{M}	D	20	3,659	-0.98	0.53	**-0.15**	-0.22,	-0.09	**155.53**	87.78
18 Negative Emotions^{+}	D	24	3,641	-1.90	0.87	**-0.13**	-0.19,	-0.06	**99.80**	76.95
18 Negative Emotions$^{+,wO,M}$	D	21	2,593	-0.39	0.87	-0.07	-0.15,	0.01	21.03	4.88
18.1 Negative Emotions Only^{M}	D	24	3,641	-1.90	0.87	**-0.18**	-0.24,	-0.12	**214.57**	89.28
18.2 Anger	D	12	2,452	-1.32	0.38	**-0.27**	-0.35,	-0.19	**165.03**	93.34
18.3 Anxiety^{wO}	D	11/12	1,952	-0.30	0.44	0.07	-0.02,	0.02	13.55	26.19
18.4 Sadness	D	12	2,452	-0.34	0.25	0.04	-0.04,	0.12	13.01	15.46
Research Question 3b: Do Liars Use Fewer Positive Emotion Words?										
19 Positive Emotions and Feelings^{+}	T	21	3,203	-0.84	0.45	0.03	-0.04,	0.10	**55.42**	63.91

(*Appendix H continues*)

Appendix H (*continued*)

Linguistic Cue	Pred. DOE	*k*	*N*	Min g_u	Max g_u	g_u	CI-low	CI-high	Q	I^2
19 Positive Emotions and Feelings[+,wO]	T	20	2,703	-0.84	0.37	-0.05	-0.12,	0.03	29.59	35.79
19.1 Positive Emotions Only	T	21	3,203	-0.84	0.45	0.01	-0.06,	0.08	**56.68**	64.71
19.1 Positive Emotions Only[wO,M]	T	20	2,703	-0.84	0.35	-0.07	-0.15,	0.00	27.98	32.08
19.2 Positive Feelings Only	T	9	1,422	-0.47	0.40	0.07	-0.03,	0.18	14.88	46.25
Research Question 3c: Do Liars Express More or Less Unspecified Emotion Words?										
15 Emotions	?	25	4,129	-1.57	0.48	**-0.34**	-0.41,	-0.28	**246.69**	90.27
15 Emotions[wO,M]	?	21	2,941	-0.63	0.48	**-0.11**	-0.19,	-0.04	28.92	30.85
16 Pleasantness and Unpleasantness	?	6	806	-0.35	0.30	-0.10	-0.25,	0.06	8.43	40.68
Research Question 4: Do Liars Distance Themselves More From Events?										
(a) Personal Pronouns										
21 First-Person Singular[M]	T	22	3,761	-1.00	0.61	-0.06	-0.13,	0.00	**274.38**	92.35
22 First-Person Plural	T	25	4,353	-1.12	0.79	-0.04	-0.10,	0.02	**228.85**	89.51
22 First-Person Plural[wO,M]	T	22	3,224	-0.72	0.39	0.06	-0.01,	0.13	27.98	24.93
23 Total First-Person	T	23	2,709	-1.63	0.57	0.05	-0.03,	0.13	**127.30**	82.72
23 Total First-Person[wO,M]	T	22	2,541	-0.39	0.57	**0.14**	0.06,	0.22	32.26	34.91
24 Total Second-Person	D	23	4,072	-1.18	0.33	**-0.16**	-0.23,	-0.10	**168.24**	86.92
24 Total Second-Person[wO,M]	D	21	3,072	-0.61	0.33	**-0.10**	-0.17,	-0.02	28.64	30.16
25 Total Third-Person	D	29	4,807	-1.18	0.55	**-0.21**	-0.27,	-0.15	**181.71**	84.59
25 Total Third-Person[wO,M]	D	26	3,848	-0.41	0.55	**-0.10**	-0.16,	-0.04	37.20	32.79
20 Total Pronouns	-	19	2,960	-0.64	0.65	-0.06	-0.13,	0.02	**68.18**	73.60
20 Total Pronouns[wO,M]	-	18	2,460	-0.36	0.65	0.06	-0.02,	0.14	19.32	12.02

Appendix H (*continued*)

Linguistic Cue	Pred. DOE	*k*	*N*	Min g_u	Max g_u	g_u	CI-low	CI-high	Q	I^2
(b) Passive Voice and Generalizing Terms										
26 Passive Voice Verbs	D	11	1,221	-0.47	0.49	0.06	-0.06,	0.18	9.52	0.00
27 Generalizing Terms	D	4	93	-1.63	0.44	-0.37	-0.79,	0.05	**15.78**	80.99
(c) Past and Present Tense										
47 Past Tense	D	16	3,047	-0.53	0.41	0.06	-0.01,	0.14	22.67	33.83
48 Present Tense	T	17	3,107	-1.41	0.60	**-0.18**	-0.25,	-0.11	**195.37**	91.81
48 Present Tense^{wO,M}	T	16	2,607	-0.51	0.60	0.01	-0.07,	0.09	19.38	22.61
Research Question 5: Do Liars Use Fewer (Sensory and Contextual) Details?										
28 Sens.-Perceptual Processes^{+}	T	27	4,177	-0.70	0.70	**0.06**	0.00,	0.13	**58.84**	55.81
28 Sens.-Perceptual Processes$^{+,wO,M}$	T	25	3,957	-0.70	0.70	0.05	-0.01,	0.12	36.00	33.33
28.1 Sens.-Percept. Processes Only^{M}	T	27	4,177	-0.90	0.70	**0.06**	0.00,	0.13	**89.26**	70.87
28.2 Seeing	T	11	2,344	-0.56	0.35	0.07	-0.01,	0.16	**33.25**	69.93
28.2 Seeing^{wO}	T	9	1,740	-0.17	0.34	0.03	-0.06,	0.13	13.34	40.03
28.3 Feeling	T	12	2,412	-0.49	0.40	-0.01	-0.09,	0.07	**19.64**	44.00
28.3 Feeling^{wO}	T	11	2,304	-0.49	0.27	-0.03	-0.11,	0.05	15.00	33.31
28.4 Hearing	T	11	2,344	-0.41	0.48	**0.17**	0.09,	0.25	14.15	29.35

(*Appendix H continues*)

Appendix H (*continued*)

Linguistic Cue	Pred. DOE	*k*	*N*	Min g_u	Max g_u	g_u	CI-low	CI-high	*Q*	I^2
(b) Contextual Embedding										
29 Time	T	24	3,796	-1.25	0.53	**0.10**	0.03,	0.16	**53.32**	56.86
29 Time^{wO}	T	23	3,296	-1.25	0.53	0.03	-0.04,	0.10	28.95	24.00
30 Space	T	24	3,851	-0.47	0.58	0.00	-0.06,	0.07	**60.12**	61.74
30 Space^{wO,M}	T	22	3,199	-0.36	0.58	-0.04	-0.13,	0.03	31.74	33.74
31 Space & Time	T	5	634	-0.25	0.61	-0.04	-0.19,	0.12	**10.48**	61.84
(c) Descriptive Words										
32 Prepositions	T	14	2,479	-0.55	0.48	0.02	-0.06,	0.10	16.54	21.38
33 Numbers	T	12	2,452	-0.28	0.23	0.05	-0.03,	0.13	9.37	0.00
34 Quantifier	T	4	1,198	0.06	0.22	**0.14**	0.02,	0.25	1.22	0.00
35 Modifier	T	11	1,361	-1.04	0.43	-0.08	-0.20,	0.03	**77.46**	87.09
36 Motion Verbs	T	17	2,859	-0.72	0.38	-0.01	-0.08,	0.07	**39.59**	59.59
36 Motion Verbs^{wO,M}	T	16	2,359	-0.72	0.13	**-0.09**	-0.17,	-0.01	16.84	10.92
Research Question 6: Do Liars Refer Less Often to Cognitive Processes?										
37 Cognitive Processes	T	19	2,995	-0.25	0.82	**0.10**	0.03,	0.18	**28.97**	37.88
37 Cognitive Processes^{wO}	T	18	2,915	-0.25	0.36	**0.09**	0.01,	0.16	19.66	13.54
38 Insight^{M}	T	15	2,539	-0.41	0.59	**0.13**	0.05,	0.21	**35.65**	60.73

Notes. Please consult notes from Table 1

META-ANALYSIS II:

Can Credibility Criteria be Assessed Reliably?

A Meta-Analysis of Criteria-based Content Analysis

Discerning between truths and lies is crucial in many settings. Therefore, it is no wonder that scholars from diverse disciplines have studied deception and its detection. These disciplines include (but are not limited to) psychophysiology and neurosciences (e.g., Farah, Hutchinson, Phelps, & Wagner, 2014; National Research Council, 2003); social, cognitive, developmental, clinical, forensic, evolutionary, and organizational psychology (e.g., Ekman, 2009; Granhag and Strömwall, 2004; Reinhard, Sporer, Scharmach, & Marksteiner, 2011; Vrij, 2008); communication (e.g., Levine, 2014; Miller & Stiff, 1993); behavioral economics (e.g., Ariely, 2012); and computational linguistics (see Hauch, Blandón-Gitlin, Masip, & Sporer, 2014).

Most research has focused on identifying valid indicators of deception--that is, observable behaviors displayed by communicators (senders) that are assumed to be correlated with the act of lying. However, meta-analyses reveal only few and small behavioral differences between truths and lies, and many of these differences are moderated by numerous moderator variables (DePaulo, Lindsay, Malone, Muhlenbruck, Charlton, & Cooper, 2003; Hartwig & Bond, 2011; Sporer & Schwandt, 2006, 2007).

However, not all behavioral cues are the same. Specifically, meta-analyses suggest that verbal content cues may be more revealing of veracity than nonverbal cues (DePaulo et al., 2003), that deception-detection training focused on verbal cues is more effective than training focused on nonverbal cues (Hauch, Sporer, Michael, &

Meissner, 2014), and that receivers judging veracity reach greater accuracy rates when they focus on verbal (rather than nonverbal) information (Bond & DePaulo, 2006; Reinhard et al., 2011).

In fact, verbal content cues have brought about a number of systematic and structured approaches to detect deception, corroborate truthfulness, or--more generally--to “assess credibility”. Best known among these are Criteria-based Content Analysis (CBCA; Steller & Köhnken, 1989) and the Reality Monitoring (RM) approach (e.g., Masip, Sporer, Garrido, & Herrero, 2005; Sporer, 2004; Vrij, 2008). Here we focus on CBCA.

Credibility Assessment and CBCA

In Germany, there is a long tradition of calling expert evaluators (usually psychologists) to help the court to assess the credibility of children or adults’ allegations of sexual abuse (Sporer, 1983; Steller & Köhnken, 1989; Wegener, 1989). Almost 60 years ago, the German Supreme Court (*Bundesgerichtshof:* BGH, 1954) ruled that expert witnesses, such as psychiatrists or psychologists, could be called to assess the credibility of statements in sexual abuse cases, particularly when no other evidence existed to confirm the truthfulness of the statement. This allowed expert witnesses to evaluate children’s credibility in probably more than 100,000 sexual abuse cases in Germany (Arntzen, 1992; Sporer, 1983; Steller, 2013). Authors like Undeutsch (1967, 1982) and Arntzen (1970, 1983), along with other experts from formerly East Germany (Dettenborn, Fröhlich, & Szewczyk, 1984; Szewczyk 1973) and Sweden (Trankell, 1972), developed individual verbal content criteria to systematically assess credibility. Köhnken (1982) and Steller and Köhnken (1989) integrated the criteria described in different sources under 19 “content criteria for statement analysis” subdivided into five major categories (see Table 1), termed

Criteria-based Content Analysis (CBCA; Steller, 1989; Köhnken, 2004). These criteria are not to be used in isolation but are embedded in a more general hypothesis testing approach referred to as Statement Validity Analysis (SVA; see below).

Although the necessity to call upon an expert has been debated over the years (for a comprehensive review, see Jansen, 2012), in 1999 the German Supreme Court (*BGH, 1999*) reached a milestone decision by declaring SVA as the standard procedure experts should follow in sexual abuse cases. Apparently, the decision became necessary because experts had provided inadequate expert testimony in several cases. In particular, in the case to be decided in that decision, the defense had obtained an additional expertise that questioned the credibility assessment procedure used by a court-appointed expert. In other words, two different "experts" did not reach the same conclusion. The *BGH* called upon (inter)nationally renowned experts to evaluate the scientific quality of this type of expert testimony. In a nutshell, the BGH appointed experts found flaws in the procedures and conclusions of the court-appointed "expert". So far, we have only described CBCA within the context of the (central European) inquisitorial system. We return to CBCA's potential role in an adversary legal system (e.g., in the U.S. or U.K.) in the discussion.

In this paper we focus on the *reliability* issue of assessing the presence of content criteria as part of a credibility as a prerequisite of its validity. If experts are unable to apply these credibility criteria in an objective and reliable manner, then their testimony is of no help to the courts (cf. Küpper & Sporer, 1995; van Koppen & Saks, 2003).

Steller and Köhnken's Content Criteria for Statement Analysis

The presence of each CBCA criterion in Table 1 is considered an indicator of the truthfulness of a statement, whereas its absence does not necessarily indicate that the statement is a lie (Raskin & Esplin, 1991; Steller, 1989). The first CBCA category, “general characteristics”, contains three criteria that apply to the whole statement without going into specific details. The second and third categories, *specific contents* and *peculiarities of the event*, contain criteria that would be cognitively difficult to purposefully include unless the narrator has actually experienced the event (Köhnken, 1990). In addition, the criteria in the third cluster “enhance concreteness and vividness” of a statement (Steller & Köhnken, 1989, p. 226). The fourth category of criteria, “motivation-related contents”, is based on the premise that a person making a false allegation would not include such contents because this would give the impression of a lack of veracity (Köhnken, 1990, 2004). The final category, “offense-specific elements”, consists of one single criterion examining whether the account contains descriptions of characteristics typical of this sort of event (e.g., sexual abuse cases) rather than misguided commonsense notions.

For a statement to be analyzed with CBCA, it has to be collected properly (e.g., Lamb, Sternberg, & Esplin, 1994; Lamb, Sternberg, Esplin, Hershkowitz, & Orbach, 1997a; Köhnken, 2004; Masip & Garrido, 2006; Raskin & Explin, 1991). The main part of the statement has to be given orally in a free report. Preferably, it should be the first time the narrator ever tells the story. It is important that diagnostically relevant parts of the statement--those which may distinguish between a fabricated and a truthful event--are of considerable length. After the free report, the expert may ask some more specific but open-ended, non-leading questions if necessary (“You

said something about the kitchen; please tell me more about it.”). Closed-ended (“Was his name Jim or John?”), leading or suggestive questions (“Were you afraid when he touched your leg?”), and yes/no questions should be avoided. The entire dialogue should be audio- or videotaped and transcribed later (Jansen, 2012). The CBCA analysis is to be performed on these transcripts.

When used in real criminal cases, CBCA is embedded in a more general and complex assessment procedure (SVA), which not only examines fabrication as an alternative hypothesis to accounts based on experience but also other problems like suggestive or repeated questioning, the origin of the first statement, or coaching. Likewise, personal characteristics of the witness--such as her or his cognitive, developmental, social and personality background--have to be taken into account (e.g., Köhnken, 2004; Steller, 1989; Volbert & Steller, 2014). In SVA, on the basis of an initial case-file analysis, rival hypotheses about the source of the statement are to be framed, tested, and falsified (Jansen, 2012; Steller, Volbert, & Wellershaus, 1993). Originally, SVA and CBCA were designed to assess the credibility of the statements of alleged victims of (child) sexual abuse. However, in the last three decades, many authors used CBCA criteria as a "tool" to detect deception, applying it not only to potential victims but also to statements of perpetrators, witnesses and other protagonists of complex (autobiographical) events in field and laboratory analogue simulations (Köhnken, 2004; Vrij, 2005, 2008).

Although SVA and CBCA are diagnostic and clinical assessment procedures rather than standardized psychometric tests, their inter-rater reliability and validity are of utmost importance for their practical application (Köhnken, 2004; Steller, 1989; Steller & Köhnken, 1989; Wells & Loftus, 1991). The validity of CBCA deals with the question of whether these criteria do indeed discriminate between truthful and

fabricated statements, and whether trained evaluators can assess the truth status above chance performance (Bond & DePaulo, 2006). But reliability is important, too: A known truism from psychometric theory is that the validity of an assessment procedure is limited by its reliability (Anastasi, 1990; Cronbach, 1990). Thus, here we examined CBCA reliability. Of the different forms of reliability, we specifically investigated the reliability of coding of CBCA by different assessors, that is, their intersubjective agreement.

Main Hypotheses

Individual CBCA criteria vary widely with respect to the precision with which they are operationalized. Also, some criteria are more global and/or more subjective than others (Anson, Golding, & Gully, 1993). Hence, criteria that have more straightforward or intuitive operationalizations, like *quantity of details* or *reproduction of conversation*, should result in higher reliabilities than criteria with more complicated or less clearly-defined operationalizations, such as *unstructured production*, *descriptions of interactions*, *unusual details* and *superfluous details* (see Table 1).

Also, we assume that different reliability indices (percentage agreement, Cohen's *kappa*, Pearson's *r*, etc.) may reflect different degrees of reliability. For example, due to the mathematical fact that *percentage agreement* does not correct for chance agreement (Cohen, 1960; Frick & Semmel, 1978), it will result in relatively high inter-rater agreement rates compared to other indices.

Hypotheses for Moderator Variables

We expected inter-rater reliability as measured with correlation coefficients to be associated with at least four predictor variables: The *research paradigm* of a

study, the *amount of training* raters were exposed to, the type of *rating scale* used, and the *frequency of occurrence* of each criterion, that is their *base rate*.

Research Paradigm. *Field studies* examine reports of alleged sexual abuse (or other criminal) cases which are recorded by police officers, psychologists, or other specialized occupational groups. On the other hand, in an *experiment*, situations, tasks, and procedures are set up by the researchers. Participants are randomly assigned to conditions (i.e., telling lies or the truth). For example, after having watched a video or having taken part in a mock crime, participants are asked to tell the truth or lie about what they have just experienced. In a *quasi-experiment*, participants are allocated to conditions on the basis of specific personal experiences (e.g., having experienced a traumatic life event) rather than being randomly assigned. For example, some participants are asked to tell a negative, stressful self-experienced life event (e.g., chosen from a prepared list of traumatic events), while other participants are asked to invent such an event they never experienced.

It can be argued that field studies and quasi-experiments share a number of characteristics that are absent in experiments. Having experienced a traumatic or stressful life event (the most frequent topic of field studies and quasi-experiments) has higher ecological validity than the manipulations used in laboratory experiments. These conceptual similarities between field studies and quasi-experiments, as well as the fact that CBCA was developed for autobiographical life events, led us to the hypothesis that inter-rater reliability would be higher in field studies and quasi-experiments than in laboratory experiments.

Rater training. Rating CBCA criteria is a complex and difficult task (Köhnken, 2004; Steller, 1989). Therefore, a prerequisite for the correct application of CBCA is that raters are well trained. Otherwise, the understanding that different raters will

have of the same criteria might differ, which will decrease reliability. In the studies reviewed, training took several forms: Raters had to read background literature on CBCA and SVA, heard a lecture held by an expert, were given definitions and examples for each CBCA criterion, and/or had the opportunity to practice and discuss example ratings prior to target ratings. In line with Steller's (1989) and Köhnken's (2004) reasoning that intensive training is required to correctly and reliably apply CBCA, we predicted that the higher the intensity of the training, the higher the inter-rater reliability for CBCA criteria.

Rating scale. CBCA criteria can be coded in different ways: Researchers or practioners either use dichotomous ratings (*0 = not present* vs. *1 = present*), an extended presence rating (*0 = not present*, *1 = present*, *2 = strongly present*), Likert scales (e.g., 1 to 5, 1 to 7, 1 to 10), or frequency counts. Because more fine-grained scales allow more possibilities to disagree (e.g., rater X assigning a "6", rater Y a "7" on a 7-point scale), we expected that more fine-grained scales should be associated with lower reliabilities. However, this may depend on the type of reliability coefficient used.

Base rates. Inter-rater reliability may be related to the base rates (frequency of occurrence) of specific CBCA criteria. Different CBCA criteria vary widely in frequency (e.g., Anson et al., 1993). For example, in his vote-counting review, Vrij (2005) found that criteria 01, 03, 04, and 19 occurred relatively often, whereas criteria 10, 13, 16, and 17 occurred only rarely. Sporer (1997a) observed that content criteria with either very low (e.g., *unstructured production*) or rather high (e.g., *logical consistency*) base rates were associated with particularly low inter-rater reliabilities (measured with Pearson's *r* or Cohen's *kappa*; presumably due to a restriction of range) but high percentage agreement rates (for similar findings, see Gödert,

Gamer, Rill, & Vossel, 2005). In the statistical literature, there is evidence that Cohen's *kappa* is generally lower when base rates are much lower or much higher than .50 (Shrout, Spitzer, & Fleiss, 1987; Spitznagel & Helzer, 1985).

We hypothesized that Pearson's *r* and Cohen's *kappa* should be lower with either very low (floor effects) or very high base rates (ceiling effects). Conversely, percentage agreement was expected to be rather high when base rates are either very low or very high.

Goals of the Meta-Analysis

The major aim was to provide the first comprehensive meta-analysis on the inter-rater reliability of CBCA criteria. The second goal was to explore potential differences between estimates provided by different reliability indices used in CBCA research (namely Pearson's *r*, percentage agreement (*PCA*), Cohen's *kappa*, *weighted kappa*, intra-class correlation coefficient (*ICC*), and Maxwell's *RE*). Our third goal was to examine the type of research paradigm, the amount of rater training, the type of rating scale used, and base rates as moderators of inter-rater reliability.

Method

Eligibility Criteria

Primary studies had to meet the following inclusion criteria to be integrated into the meta-analysis. First, studies had to analyze truthful or fabricated accounts of *past* experiences with at least one CBCA criterion. These accounts had to originate from spoken or written samples (transcripts) either in an experiment, quasi-experiment or field setting. Second, at least two raters had to evaluate some or all of the statements in the study. Third, raters had to be blind regarding truth status and

had to rate the statements independently. This was an important requirement to avoid any unintentional source of error such as confirmation bias. Fourth, raters had to be trained to identify CBCA criteria. Fifth, datasets had to be independent from each other to avoid dependencies within the data. If results from the same dataset judged by the same raters were published in several sources, we included the source with the highest publication standards (e.g., peer review). Sixth, the main study had to be written in English or German.

Studies meeting one or several of the following criteria were excluded: Senders had been coached in CBCA criteria before giving their statements; statements had been manipulated by the investigator (e.g., by changing specific sentences to manipulate the occurrence of CBCA criteria); only a single case was described in the study; or computer programs had been used to analyze the statements (see Hauch, Blandón-Gitlin, et al., 2014; Sporer, 2012).

Literature Search and Study Selection

In a first step, we located all 37 CBCA studies reviewed by Vrij (2005, 2008). Then, we searched through the reference lists of these studies to find more related papers and to create an author list. Then, we conducted several exhaustive literature searches up to January 2014 in the Web of Science, PsycInfo, WorldCat, Dissertation Abstracts, and Google Scholar. The authors' names, "criteria-based content analysis", or "CBCA" served as keywords. For the literature search in German databases (e.g., OPAC, PSYNDEX, ZPID-Datenbank Diplomarbeiten), the keyword "Glaubwürdigkeit" (in combination with "-merkmale", "-kriterien", "-diagnostik"), "Realkennzeichen", "Merkmals-", or "Kriterien-orientierte Inhaltsanalyse" were used. More than 800 sources published between 1982 and 2014 were located and examined with regard to the eligibility criteria.[1] To reduce publication bias,

special effort was made to locate and integrate unpublished studies. Authors or supervisors were asked personally, via telephone or email.

In total, 74 reports (52 in English and 22 in German) were included. Because three reports described more than one experiment (see Appendix A), the number of independent studies increased to 78.

Four reports provided reliability indices separately for different subgroups of senders or raters (Herrmann & Jena, 1995; Joffe, 1992; Petersen, 1997; Ruby & Brigham, 1998), increasing the final number of hypothesis tests to a maximum of k = 82 but was usually lower because many studies did not investigate all 19 CBCA criteria.

Independent Variables

Several study and sampling characteristics were coded, such as source of publication, research paradigm (experiment, quasi-experiment, field study), experimental design (*between*-participants design or *within*-participants design), language of statements, mode of production (spoken, handwritten, or typed), number of accounts per sender, number of senders, senders' gender, and senders' age. Also, some further descriptive variables were coded: status of the liar (witness, actor, victim, or perpetrator), type of event (watch video, observe an event, participate in event, attitude, mock crime, trivial life event, significant life event, sexual abuse, or other real crime), emotional valence (neutral, negative, positive), senders' motivation (none, low, medium, high), and interview style (free report only, semi-structured interview, structured interview, Cognitive Interview). Individual coding decisions for the aforementioned variables are displayed in Appendix A.

Furthermore, variables concerning the rater and the rating process were coded as follows (see Appendix B): number and occupation of raters (students or

experts), mode of presentation of statements (transcript, audio only, audiovisual), rating scale used (frequency count, dichotomous [0/1] , 0/1/2, 0 to 4, 1 to 5, 1 to 7, 1 to 10, or other Likert scale), training duration in hours on average per rater, and amount of training. Amount of training was coded as a summary score (0 to 6) of five separately coded (dichotomous [0 = No; 1 = Yes], or 0/1/2 rating) training components: (a) background information (0/1), (b) operational definitions (0/1), (c) example statements (0/1), (d) lecture (0/1), (e) example ratings, feedback, and/or homework (0/1/2). For component (e), when no sub-component was fulfilled we coded 0, when one component was fulfilled we coded 1, and when two or three components were fulfilled we coded 2. Finally, the base rate of each CBCA criterion was noted or calculated as the mean of the base rates of fabricated and truthful accounts.

Coding Procedure and Inter-Coder Reliability

Three expert coders who had read dozens of background articles and books on CBCA and had published articles on detection of deception in peer-reviewed journals served as coders. They were trained on example studies with a comprehensive coding manual. After a fair amount of agreement was established, each coder independently rated approximately 66% of the 52 English reports (randomly assigned), so that all independent and dependent variables of each English study were coded twice. Across the three coders, inter-rater reliability was highly satisfactory: For 20 categorical variables, *kappa* values ranged from .69 to 1.00 (*M* = .84, *Mdn* = .85), and ICC_{single} consistency values from .62 to 1.00 (*M* = .87, *Mdn* = .90). Furthermore, seven continuously coded variables yielded Pearson's *r* values ranging from .99 to 1.00, and ICC_{single} consistency values ranging between 0.98 and 1.00. After the independent coding process, each pair of raters compared

their coding decisions for all independent variables and resolved disagreements. For every disagreement, the original source was consulted again and the value corrected. The 22 German reports were coded by the first author and crosschecked by a research assistant.

Reliability Indices and Meta-Analytic Techniques

Unit of analysis. The number of statements that were rated by at least two independent raters in the original study served as the unit of analysis. We did not consider the total number of statements in a study, but only the portion of statements that had been rated by two or more independent raters. For example, in the study by Vrij, Mann, Kristen, and Fisher (2007), only 50% of their 120 transcripts were coded by a second rater. Thus, the number of statements was adjusted to $N = 60$ for that study (see Appendix B)..

Effect size measures. Inter-rater reliability indices analyzed were Pearson's *r* (or Spearman *rho*, or *phi*), percentage agreement (*PCA*), Cohen's *kappa*, *weighted kappa*, intraclass correlation (*ICC)*, and Maxwell's *RE*.

Pearson's *r*. Pearson's product-moment correlation coefficient is widely used to measure inter-rater reliability. According to Fleiss (1981), values between .40 and .74 designate good agreement between two raters, and values equal to or above .75 designate excellent agreement. As recommended in several standard meta-analysis handbooks (e.g., Cooper, Hedges, & Valentine, 2009; Lipsey & Wilson, 2001) we transformed Pearson's *r* (as well as Spearman's *rho*, or *phi*) into Fisher's Z_r as effect size measure. Because Fisher's Z_r is not defined for Pearson's $r = 1$, we reduced this *r* value to .999. For weighted analyses, N_{adj}-3 was used as the inverse variance weight. For ease of interpretation, Fisher's Z_r was back-transformed to Pearson's *r* (Lipsey & Wilson, 2001).

Separate meta-analyses were calculated for each of the 19 CBCA criteria under the fixed- and random-effects model (Lipsey & Wilson, 2001; Shadish & Haddock, 2009). Here, we report results of the random-effects model (REM) only because this model is seen as the state of the art (e.g., Hunter & Schmidt, 2004). Results of the fixed-effects model (FEM) are available from the first author. Furthermore, we tested whether the observed effect sizes estimated the same population parameter with *Q* (Lipsey & Wilson, 2001) and I^2 (Higgins & Thompson, 2002; Shadish & Haddock, 2009) as homogeneity test statistics. Whenever these tests indicated heterogeneity, we either calculated moderator analyses as an analogue to ANOVA (Hedges, 1982) or meta-regression analyses (Lipsey & Wilson, 2001; Pigott, 2012) under a FEM (due to small sample sizes). To guard against potential confounds, that is, systematic associations between predictor variables, we first calculated their inter-correlations as well as cross-classification tables of nominal predictors. Formulae to calculate effect sizes, weights, confidence intervals, and meta-analyses under the FEM and REM were programmed in Microsoft Office Excel (2011) spreadsheets by the first and second authors. For cross-validation, meta-analytic calculations, moderator analyses and meta-regression were additionally performed with Wilson's (2002) SPSS macros (see Lipsey & Wilson, 2001).

Percentage agreement. Percentage agreement (*PCA*) is another common measure of inter-rater agreement. The terms *percentage agreement* and *proportion agreement* are used interchangeably in article. Here, the number of agreements between two raters is divided by the total number of judgments. *PCA* ranges from 0% (no agreement) to 100% (perfect agreement). A highly problematic issue is that agreement by chance (e.g., with a dichotomous rating, chance agreement is 50%), is not taken into account (Cohen, 1960; Maxwell, 1977). Thus, *PCA* leads to an

overestimation of the true agreement between two raters. Many authors consider percentages higher than 70% as good agreement. However, many statisticians have criticized *PCA* because it does not correct for chance agreement (Cohen, 1960; Fleiss, Levin, & Paik, 2003).

As we expected the mean *PCA* to reach a value between 20% and 80%, we followed Lipsey and Wilson's (2001) recommendation to use *proportions* (i. e., *proportion agreement*) as effect sizes (*PCA* divided by 100, with values ranging from 0 to 1). Proportions of 1 were set to a value of .999 in order to calculate the standard error ($\sqrt{(p^*(1-p))/N}$) correctly. Also following Lipsey and Wilson (2001), we used the inverse of the squared standard error as the appropriate weight for each study. Due to methodological limitations inherent to *PCA*, we refrained from performing moderator analyses with this measure.

Other reliability indices. Many researchers have used (test-score) reliability indices as the main dependent variable in meta-analyses (see Vacha-Haase, 1998, for the so-called "reliability generalization studies"; for an overview see Vacha-Haase & Thompson, 2011). However, we were not able to locate any appropriate meta-analytic method to integrate inter-rater reliability indices other than Z_r and *PCA*. The main problem is that the authors of primary studies usually provided no information about the underlying distributions, nor did they report the necessary values to calculate the standard error of a specific measure amenable to meta-analysis (Bayerl & Paul, 2011; Donner & Klar, 1996). Because none of the studies that we included provided enough information to calculate standard errors, we could not calculate confidence intervals or weighted average effect sizes for Cohen's *kappa*, *weighted kappa*, *ICC*, or Maxwell's *RE*. Instead, we used each reliability index itself as the effect size and calculated only the unweighted average, median, and quartiles.

Cohen's kappa. *Kappa* (Cohen, 1960) is a chance-corrected measure of percentage agreement and is the measure of choice compared to *PCA* for categorical or ordinal data. Hereby, the difference between actual *PCA* and chance agreement is divided by the maximum agreement minus chance agreement. However, as noted above, *kappa* may depend on base rates. *Kappa* values range between -1 and +1, with +1 displaying perfect agreement, 0 chance agreement, and negative values denoting agreement worse than chance. According to Landis and Koch (1977), agreement as measured with Cohen's *kappa* can be described as "perfect" (*kappa* values between .81 and 1.00), "substantial" (.61 to .80), "moderate" (.41 to .60), "fair" (.21 to .40), "slight" (.01 to .20) or "poor" (negative values).

***Weighted kappa*.** Whereas *kappa* treats all disagreements between raters equally, *weighted kappa* gives different weights depending on the degree of deviation between the evaluators' ratings (Cohen, 1968, Formula 8). Like simple *kappa*, *weighted kappa* values can vary between -1 and +1. According to Fleiss (1981), the interpretation of *weighted kappa* values should be in line with correlation coefficients (below .40 = poor agreement; .40 to .74 = fair to good agreement, ≥ .75 = excellent agreement), because under restricted conditions (i.e., contingency table with equal marginal distributions and sufficient sample size), *weighted kappa* is equivalent to the product-moment correlation coefficient (Cohen, 1968).

Intraclass correlation. *Intraclass* correlation coefficients (*ICC*s) measure an association between variables of the *same* measurement construct (McGraw & Wong, 1996a, 1996b). *ICC*s measure not only the relationship between ratings of different raters but also take mean differences between raters into account (see Rosenthal, 1995). Six different intraclass correlation coefficients can be distinguished (Bartko, 1966; Gwet, 2012; Shrout & Fleiss, 1979). *ICC*s were used to measure

reliability in a number of primary studies included in this meta-analysis, but the reports often did not specify the type of *ICC* used. Therefore, we refrained from calculating an (unweighted) average in this meta-analysis. Instead, we only report the median and other descriptive statistics.

Weighted kappa and *ICC* are considered equivalent under certain conditions (Fleiss & Cohen, 1973; Krippendorff, 1970). Therefore, and because only a small number of studies used these indices, we combined these two measures into a single synthesis. Due to the similarity between *ICC* and *weighted kappa*, we used Fleiss's (1981) guidelines to interpret these coefficients.

***Maxwell's random error*.** Another chance-corrected measure of inter-rater reliability for dichotomous ratings is Maxwell's *random error* (*RE*; Maxwell, 1977). Maxwell assumed evaluators' decisions on doubtful cases to be randomly distributed--rather than depending on any known probabilities of an outcome. Maxwell's RE values range from -1 to 1. To our knowledge, Anson et al. (1993) introduced Maxwell's *RE* as a measure of inter-rater reliability for CBCA ratings. The authors praised its characteristics, especially the potential that Maxwell's *RE* does not overestimate the amount of agreement by chance with base rates differing from .50 (unlike Cohen's *kappa*). Anson et al. suggested the following classification: values larger than .50 indicate adequate reliability, between .30 and .50 marginal reliability, and below .30 inadequate reliability.

Missing Data. Whenever inter-rater reliability was not provided for individual criteria, or in case only a range or average was given, authors were contacted to provide the reliability values for each CBCA criterion. We are very much obliged to all authors who responded to our inquiries and took the effort to send the requested data.[2]

Results

Study Characteristics

The number of hypothesis tests for particular reliability indices varied between *k* = 5 (Maxwell's *RE*, *details characteristic of the offense*) and *k* = 35 (Pearson's *r*, *quantity of details*). Whenever data were reported, the following study characteristics are based on the total sample size of *k* = 82 hypothesis tests. An overview of studies reporting on various reliability indices can be obtained in Appendix C.

As shown in Table 2, about half of the studies were formally published in a scientific journal and more than 70% of the reports were composed in English. Although laboratory experiments were the most frequently used paradigm (53.66%), remarkably almost one half of the studies involved the more ecologically valid quasi-experiments or field studies. About two-thirds of studies used a between-participants design. In one half of the studies the statements were produced in English, and in one third they were in German. Researchers have tried to mirror the characteristics of real forensic cases for which CBCA was originally designed: (a) in nine out of every ten studies the statements were produced orally, (b) the interview styles used matched the kind of interviews recommended for forensic settings; (c) in almost one third of the studies, the narrator was a victim, and in a substantial proportion of studies he or she was either a witness or a perpetrator; (d) the event typically had a negative valence and involved direct participation of the sender, and (e) it was a significant life event (37% of studies) or a crime (20%: sexual abuse or other crimes). However, only a few senders were highly motivated, and most raters were students; these latter features are probably a result of most of these studies being conducted in universities. Usually, raters were fairly well trained, and in most studies there were two raters per study who made dichotomous decisions (criterion absent/present) or

used a three-point (*0 = criterion absent*; *1 = present*; *2 = strongly present*) classification scale (see Table 2).

In about 70% of the studies, all statements (100%) were evaluated by at least two independent raters. In the remaining studies, the percentage of multiply rated statements ranged from 12.5% to 83.0% (*M* = 31.2%, *SD* = 19.3%, *Mdn* = 22.5%, *Mode* = 20.0%). After adjustment, the mean number of accounts was 54 per study (see Table 2) and ranged from 4 to 200. More than one reliability index was reported in 37.8% of all studies. For those studies reporting only one reliability index, Pearson's *r* was most frequently used (see Table 2).

Pearson's *r*

For 18 out of 19 criteria, individual effect sizes had large heterogeneities, as shown by consistently large significant *Q* statistics, and by I^2 values ranging between 78.7% and 98.5% (see Table 3). Interquartile ranges ≥ .30 also indicated high heterogeneity for at least 14 criteria. Nevertheless, we calculated unweighted and weighted Pearson's *r*s for all criteria under the random effects model.

Inter-rater reliability was good to excellent for most criteria (Table 3). Six criteria (06, 10, 12, 13, 15, 17) had weighted inter-rater reliability values of .75 or higher. Eleven criteria reached quite good inter-rater reliability with values ranging from .60 to .74, and one criterion (09) reached also good reliability with values between .50 and .59. *Unstructured production* (02) showed the lowest reliability (.46). The CBCA sum score yielded a high inter-rater reliability of .90. Given the large heterogeneity and the wide confidence intervals of the results, further analyses are clearly warranted (see below).

Percentage Agreement

Our decision to set *PCA* values of 1.00 to .999 prior to calculation proved problematic: These values biased the meta-analytic results towards high average effect sizes. The more .999 values were included in a meta-analysis, the higher the probability that the weighted average effect size approximated .999. Therefore, we report only the more conservative calculations without “perfect” (1.00/.999) *PCA* values (Table 4). Thus, one to six values of perfect inter-rater agreement (M = 2.13, SD = 1.46) were excluded for sixteen criteria (with no exclusions for criteria 02, 11, 14, and the sum score).

Probably the most outstanding finding is the large heterogeneity between individual effect sizes for each CBCA criterion. As shown in Table 4, for almost all criteria, weighted individual *PCA* values ranged between poor (e.g., minimum value for 15: *admitting lack of memory*: .16) and high values (e.g., maximum value for several criteria: .990 for Criterion 10). Six criteria had interquartile ranges ≥ .30. Five criteria (10, 13, 16, 17, 18) reached high inter-rater agreement (≥ .80), and all other values ranged between .70 and .79, indicating good agreement except for *contextual embedding* (04: .68).

Cohen’s *Kappa*

Again, all estimates were highly heterogeneous, with individual reliability values ranging from -.16 to 1.00 (see Table 6). All criteria showed fairly large interquartile ranges ≥ .30. For most criteria, unweighted average inter-rater reliability was "moderate" using Landis and Koch’s (1977) guidelines, with six criteria (06, 12, 13, 15, 17, 18) reaching values higher than .50, and nine criteria (03, 04, 05, 07, 08, 09, 10, 11, 16) showing values between .40 and .49. The four remaining criteria (01:

logical structure, 02: *unstructured production*, 14: *spontaneous corrections*, 19: *characteristics of the offense*) exhibited values between .30 and .39, thus still indicating "fair" agreement (Landis & Koch, 1977).

Weighted Kappa and *ICC*

Values ranged from -.10 (14: *spontaneous corrections*) to 1.00 (for several criteria). Almost all criteria except 03: *quantity of details*, 07: *unexpected complications*, and 13: *attribution of perpetrator's mental state*, showed interquartile ranges ≥ .30 (see Table 6). Seven criteria showed median inter-rater reliability values of .60 or larger (03, 06, 08, 10, 12, 15, 17), five (02, 05, 11, 13, 14) yielded still good reliability, with medians between .50 and .59, and another five (04, 07, 09, 16, 18) reached adequate reliability, with medians between .40 and .49. Only *logical structure* (01: *Mdn* = .18) showed inadequate reliability.

The reliability of the CBCA sum score as measured with *ICCs* was reported only in two experiments (Leal et al., 2013, Experiments 1 and 2), showing high values for both (.91).

Maxwell's *RE*

Individual Maxwell's *RE* values (Table 7) varied widely from -.22 to 1.00 (e.g., 19: *characteristics of the offense*). For seven criteria, heterogeneity was quite high with interquartile ranges ≥ .30, whereas for thirteen other criteria, interquartile ranges varied from .05 (18: *pardoning the perpetrator*) to .27 (04: *contextual embedding*) indicating low to medium heterogeneity. Median reliability was adequate (> .50) for 16 criteria and the CBCA sum score. For four criteria, median reliability was even above .80 (Criteria 10, 16, 17, 18), and for three other criteria it was above .70

(Criteria 01, 06, 19). Maxwell's *RE* showed marginal values (.30 - .50) for two criteria: *unstructured production* (02: .40) and *quantity of details* (03: .40).

Problematic Issues

The misuse of CBCA summary scores. Although we have reported reliability estimates for CBCA summary scores in our tables, these reliability coefficients are highly problematic for several reasons, and we will therefore refrain from interpreting them. First, two raters may arrive at an identical summary score, without agreeing on a single individual criterion (see Sporer, 2012, for details). Suppose, rater X has scored criteria 01, 03, 05, 07, and 09 as present, while rater Y coded 02, 04, 06, 08, and 10 as present; both would receive a summary score of 5 (a "100% agreement"), even though both raters did not agree on a single criterion. This way, high *PCA*s as well as high *r*s can be obtained, without any inter-coder agreement at all. Second, in our meta-analysis, different authors used different numbers of criteria to build a summary score (e.g., summing criteria 01 to 14, or 01 to 19; see Appendix C). As we know from classic testing theory, longer tests containing more items yield higher reliability (Anastasi, 1990), provided the items measure a common underlying construct. Third, although the frequency distributions were most likely often skewed researchers used Pearson *r*s (instead of nonparametric coefficients).

Skewed distributions with frequency counts. Frequency counts in general are likely to result in skewed distributions, in particular when accounts vary widely in length. Because the presence in CBCA criteria is not corrected for account length (as is done with frequency counts in studies on nonverbal or linguistic cues to deception; DePaulo et al., 2003; Hauch, Blandón-Gitlin, et al., 2014) Pearson *r*s may be artificially inflated (e.g., by outliers).

Low and high base rates. To check for the possibility of artificially high reliabilities as a function of base rates we conducted preliminary regression analyses with base rates as predictors and *PCA*, *kappa*, and Pearson *r*s (or *phi*s) as to be predicted variables. Because studies reported base rates in different ways (e.g., as a proportion with dichotomous coding, or as an overall mean value for rating scales), all base rates were transformed to a standard format of "proportion present", with 0 for the lower limit and 1 as its maximum.

Figures 1 to 3 plot the weighted means of *PCA*s, Pearson *r*s, and *kappa*s of all CBCA criteria against the respective base rates for all studies that reported both values (which is less than the number of studies in our overall meta-analyses). Besides showing ceiling effects of *PCA*s for several variables (e.g., 10: *details misunderstood*), the *PCA* graph shows an U-curve relationship as a function of base rates (Figure 1). In other words, *PCA* was especially high for low and high base rates. Plots for Pearson *r*s (Figure 2) and *kappa*s (Figure 3), on the other hand show some evidence of the opposite pattern, that is, lower reliabilities for low and high base rates, respectively (i.e., an inverted U-shaped pattern). Besides the postulated curvilinear relationships, Figure 2 (Pearson *r*s) shows some evidence for a positive linear trend for lower base rates (< .40) while Figure 3 (*kappa*s) also appears to contain a negative linear trend as a function of base rates (> .30). For *PCA*s (Figure 1), there may also be a negative linear association because there are few criteria with high base rates (like 01: logical consistency).

Reviewing Shrout et al. (1987) it can be noted that their Figure 1 is quite similar in shape to our Figure 1. These authors used this relationship to demonstrate that reliabilities measured with *kappa* are likely to yield low values at the extremes of low and high base rates. Altogether, these preliminary analyses clearly show that

neither *PCA*s, Pearson *r*s, nor *kappa*s can meaningfully be interpreted without taking their base rates into account.

To more formally test for these relationships, we conducted weighted regression analyses with both the linear and squared components of the base rates as predictors, and the reliabilities as to be predicted variables (see Table 11 below).

Moderator Analyses for Pearson's *r*

The large heterogeneity of the different reliability coefficients across criteria and studies makes it necessary to look for potential moderator variables that may help understand the large variations. Many meta-analyses (in psychology and law) have used blocking analyses in analogy to ANOVA to find systematic differences. The problem with this approach is that by categorizing a given set of studies repeatedly for different moderator variables confounded variables are used as predictors, which may render the results uninterpretable. For example, if field studies were primarily conducted with little training, using dichotomous coding, while laboratory studies were conducted with elaborate training of coders using rating scales, any comparison between field and laboratory studies would be meaningless. To reduce (but not necessarily eliminate) this problem, we first cross-tabulated and/or correlated moderator variables with each other to assure that there were no empty or low frequency cells, or high correlations between pairs of predictors. Furthermore, to control for associations between moderator variables, we used meta-regression in addition to blocking analyses (see Pigott, 2012). We restricted significance tests of these moderator variables to CBCA criteria with subgroup *k*s ≥ 3.

Research Paradigm. As expected, inter-rater reliabilities in field studies and quasi-experiments were higher than in laboratory experiments for 15 (out of 19)

CBCA criteria (Table 8). Only for *unstructured production* (02), and *descriptions of interactions* (05), moderator analyses did not reveal significant differences. Reliability values for *pardoning the perpetrator* (18) were higher in experiments than in field studies or quasi-experiments.

Rater training. More intense training (operationalized as the number of training components: 5 to 6 vs. 2 to 4 components) yielded higher inter-rater reliabilities for 10 CBCA criteria (Table 9). No significant differences were found for five criteria (02, 06, 07, 09, 17). For four criteria (03, 05, 15, 18) results were contrary to expectation.

Rating Scale. Rating scales were classified into three categories: (a) presence rating (*not present* (0), *present* (1), and optionally *strongly present* (2)); (b) Likert scale (e.g., 1-5, 1-10) or weighting techniques; and (c) frequency counts. Due to the small number of studies ($k \leq 3$) using frequency counts, we excluded this category from blocked comparisons for 5 criteria (01, 02, 05, 11, 16), and refrained from calculating significance tests for 4 criteria (10, 17, 18, 19). Findings were rather mixed (Table 10): First, studies using frequency counts compared to other scoring options revealed the highest reliability values for four criteria (03, 04, 06, 12). Second, presence ratings showed higher reliability values for five criteria (01, 05, 11, 15, 16) compared to Likert- or weighting techniques. Third, for three criteria (08, 13, 14) studies using Likert scales tended to have lower reliability values compared to the other two rating categories. However, these blocking analyses ought to be interpreted with caution due to potential confounds--which are controlled for in the meta-regressions below.

Meta-Regression Analyses for Pearson's r

Base Rates. Fifteen criteria (01-09, 11-16) were included in these meta-regression analyses (Table 11). For eleven criteria, higher base rates were linearly associated with higher inter-rater reliabilities (indicated by positive *B*/*beta* weights). Curvilinear associations (B^2/$beta^2$ component) were observed for eight criteria. No associations were found for criteria 01, 05 and 06. The amount of explained variance varied from R^2 = .07 (04: *attribution of perpetrator's mental state*) to R^2 = .36 (11: *related external associations*). The residual model remained significant for all criteria, indicating that a large amount of heterogeneity was left unexplained.

Multiple Meta-Regression Analyses. Simultaneous multiple meta-regression analyses were conducted for 15 CBCA criteria with research paradigm, training, and rating scale (*without* frequency counts) as predictor variables (Table 12). Further analyses controlling for base rate (entered first as an additional predictor) were also calculated but omitted here because of small overall *k*s (available from the first author). Significant models were obtained for 12 criteria. First, research paradigm showed significant *positive B* weights for nine criteria: Quasi-experimental or field studies yielded higher reliabilities than laboratory experiments. Second, training intensity was associated with higher reliabilities for studies using 5 or 6 compared to 2, 3 or 4 training components for five CBCA criteria. Opposite results occurred for criteria 03, 05, and 06. Third, rating scale resulted in significant *negative B* weights for seven criteria. In other words, using presence ratings was associated with higher reliability values than using Likert scales (exception: 11: *related external associations*). Note that many of the *beta* values were rather high for this type of analysis, and that the significant regression models explained variance from R^2 = .04 (criterion 13) to R^2 = .49 (criterion 01).

Discussion

We provided a comprehensive and systematic meta-analysis of the extent to which each of the 19 CBCA criteria can be reliably assessed by different evaluators. The results are important both for theory, future research, and practice: Inter-rater reliability is seen as an essential prerequisite of validity (Anastasi, 1990; Cronbach, 1990; Steller, 1989), and reliable credibility assessments of alleged victims, witnesses, and suspects' statements are much needed in legal settings and in other applications of CBCA research.

Altogether, the present meta-analysis shows that most CBCA criteria can be rated with sufficient to good inter-rater reliability. Besides this overall picture, more differentiated findings emerged. First, regardless of the specific reliability index used, inter-rater agreement was excellent for some criteria, moderate to good for most others, and marginal for a few criteria. Second, the findings varied depending on the reliability index being used. Third, heterogeneous results could be partially explained with moderator variables.

Most and Least Reliable CBCA Criteria

Reliability was almost consistently high for five criteria, namely, *reproduction of conversation* (06), *accurately reported details misunderstood* (10), *raising doubts about one's own testimony* (16), *self-deprecation* (17), and *pardoning the perpetrator* (18).[3] One of these criteria, *reproduction of conversation*, was predicted to have high inter-rater agreement because of its straightforward definition--a literal replication of utterances of at least one person. The excellent reliability for this and the other four criteria may similarly be explained in terms of their relatively straightforward definitions.

On the other end of the reliability spectrum, *unstructured production* (02) and *superfluous details* (09) had low reliabilities (as predicted) regardless of the particular coefficient used. The result for *unstructured production* (overall, the criterion with the lowest reliability) replicated Vrij's (2005, 2008) conclusions. This is in line with our expectation that assessing whether a story is told in an unstructured or chaotic way, while at the same time still being a logical and coherent account with a clear storyline, is very subjective and hence a difficult enterprise (Anson et al., 1993). Concerning *superfluous details*, Anson et al. also concluded that this criterion "has a complex and possibly confusing definition" (p. 337), and Roma, Martini, Sabatello, Tatarelli, and Feracutti (2011) explained the low reliability of this criterion in terms of the difficulty for raters to differentiate between *superfluous* and *unusual details*. Hence, Roma et al. suggested combining these two criteria (see also Sporer, 2004).

Comparison of Different Reliability Indices

As assumed, inter-rater reliability as measured with *percentage agreement* was found to be (very) high. These findings can be attributed to the fact that *PCA* does not take agreement by chance into account (Cohen, 1960; Frick & Semmel, 1978), leading to an overestimation of actual agreement. This problem is more pronounced when the number of rating categories is low (e.g., for dichotomous ratings, chance agreement is 50%), as was the case in most studies (see *rating scales* in Table 2). Extremely high PCA values may also be an artefact when base rates are very low or very high (Sporer, 2012). We found some evidence for this U-shaped relationship when plotting average base rates against weighted average *PCA*s (Figure 1).

In line with our assumption, all other indices showed lower inter-rater reliabilities. Weighted meta-analyses across all criteria on Pearson's *r* (*M* = .70, *SD*

= .10), unweighted meta-analyses on Maxwell's *RE* (*M* = .62, *SD* = .12), and *kappa* or *ICC* (M_{Mdn} = .55, SD_{Mdn} = .14) revealed adequate to good--but not high or perfect inter-rater reliabilities. As expected, unweighted meta-analyses of Cohen's *kappa* (*M* = .45, *SD* = .08) revealed somewhat lower reliability values than the other indices. However, according to Landis and Koch (1977), *kappa* values ranging from .30 to .57 still reflect fair to moderate consistency between evaluators. Yet, for forensic applications, higher values are desirable.

Base rates. With base rates substantially below or above .50, *kappa* generally leads to low reliabilities (Shrout et al., 1987). Indeed, in our data, with low (< .30) or high (> .65) base rates, mean *kappa*s tended to have low values indicating an inverted U-shaped relationship (Figure 3). A similar pattern was found for Pearson's *r* (Figure 2). In our meta-regressions on Pearson's *r*, eleven (out of 15) criteria had either positive linear or curvilinear, quadratic relationships with base rates. This is another indication that inter-rater reliability of CBCA criteria (measured with *PCA*, *kappa*, or Pearson's *r*) is *not* independent of their base rates, which should always be considered when interpreting these coefficients.

Attempts to Account for Heterogeneity

We found reliability values to be highly heterogeneous for almost all criteria. In order to explain this variability, we tested specific predictions concerning how the research paradigm, the amount of training received, the rating scale used, and the base rates (*see above*) would be associated with reliability. These predictions were tested for Pearson's *r*.

Research paradigm. We argued that because CBCA was designed for autobiographical events (like sexual abuse) and because of the increased ecological validity of field studies and quasi-experiments, inter-rater reliability would be higher

for these paradigms than in laboratory experiments. This hypothesis was strongly supported for 15 out of 19 criteria in the blocking analyses, and in 9 of 16 meta-regression analyses. Presumably, field studies and quasi-experiments yield much longer, more detailed but also more variable accounts compared to laboratory simulations, which in turn may allow for higher reliability correlations.

Rater training. We expected that the amount of training would increase the raters' expertise, thus leading to higher inter-rater reliability. In the blocking analyses, this hypothesis was supported for ten criteria, with no associations for five criteria, and significant associations in the opposite direction emerged for the remaining four. However, some of these associations disappeared or even became negative when rating scale used and research paradigm were controlled for. It might be the case that some criteria are easier to train and apply than others. For example, training seemed to be particularly useful with the criteria *external associations*, *logical consistency*, *contextual embedding*, *subjective mental state*, and *spontaneous corrections* whereas it was negatively associated with *quantity of details*, *descriptions of interactions* and *reproduction of conversation* in the meta-regressions.

It is possible that the way we measured the intensity of training (number of differentiated training components) was not sensitive enough. An alternative estimate of training intensity is the duration of training, but this information was provided for only 30 studies (and Pearson's *r* was not used in all of these studies). Somewhat surprisingly, the number of training components did not significantly correlate with training duration, *rho*(30) = .087, *p* = .647). Apparently, short training programs might include all training components (e.g., Dunbar et al., 2012), while longer training programs may contain a few components only (e.g., Granhag,

Strömwall, & Landström, 2006). In any case, it is remarkable that for the 30 studies reporting training duration, the average length of training was quite extensive (*M* = 23.09 hs; *SD* = 40.07; *Mdn* = 8.75). If these 30 studies are representative of all the ones included, then the length of training may explain why reliability was generally rather high. Alternatively, only studies that used extensive training also reported in detail on the reliability of individual criteria.

To our knowledge, only Köhnken (2004) and his students (Höfer, 1995; Krause, 1997; Petersen, 1997) systematically examined the impact of training intensity, length and type on reliability in a direct manner. They developed an extensive three-week training program that included several different components (e.g., Köhnken, 2004; see also Gödert et al., 2005, for a shorter version). In one of the studies, Petersen (1997) demonstrated that reliabilities for all criteria and statements were higher for the most extensively trained group, than for less extensively trained groups or the control group. To further investigate the relationship between training intensity/duration and inter-rater reliability, research should be conducted where the duration and intensity of training are manipulated systematically.

Rating scale. When looking at the definitions of CBCA criteria (Table 1) readers may wonder why so many researchers have reduced rating scales to presence ratings. For most criteria, the underlying construct to be rated is clearly continuous (e.g., *quantity of details*, *description of interactions*, *superfluous details*, *offense specific elements*). Perhaps, this (in our view inappropropriate) reduction of information may have come about by the desire to apply *PCA* as a simple measure of agreement. Although simple presence ratings can easily be added up to a summary score, the German originators of SVA/CBCA have strongly advised against

this (e.g., Steller, 2013). Volbert (2008) summarized this most aptly in the title of a review article: "Credibility assessment--more than Criteria-based Content Analysis". Whether or not it does or does not make sense to create summary scores in laboratory simulation studies, is beyond the scope of this paper (see Sporer, 2012). Within the context of forensic SVA, it is clearly not appropriate.

Research Implications and Forensic Applications

This meta-analysis demonstrated that inter-rater reliabilities for most CBCA criteria were fairly adequate, especially in field studies and quasi-experiments. This is important because reliability is a central prerequisite for the validity of any assessment procedure. However, these conclusions depend on the type of coefficient investigated. Conclusions are not as optimistic when Cohen's *kappa* was used. Because *kappa* is generally lower for criteria with low or high base rates (Shrout et al., 1987), dichotomizing criteria into "present" vs. "absent" may exacerbate this problem.

Although the problems with using *PCA* as a measure of inter-coder agreement have been known for over 50 years (Cohen, 1960; Fleiss et al., 2003; Shrout et al., 1987; Uebersax, 1987), its frequent use in the CBCA literature is quite disconcerting. Hence, we recommend that PCA should not be used as a single indicator of reliability (if at all) but should always be supplemented by Cohen's (weighted) *kappa* (for categorical data), or intra-class correlation coefficients (*ICC*s) for continuous data. *ICC*s have the advantage that they take systematic differences between coders into account. However, the type of *ICC* used must be specified depending on the design. We also recommend that reliabilities always be reported for *all* criteria individually because they may vary depending on the domain and type of stimulus materials (due to respective differences in base rates).

Many studies have investigated only portions of the accounts to study inter-coder reliability. We recommend having all accounts coded by two (or even more) raters, which would improve reliability estimates for the study at hand (see Rosenthal, 1995). As our data have shown, training coders does pay off at least for some criteria, and by increasing reliability, validity may also be increased. With an eye on validity, we also recommend more fine-grained differentiations (e.g., 0 to 4 scales) rather than simple presence ratings which should increase the discriminatory power of the respective items. However, using more fine-grained rating scales may be more difficult to apply and may require additional training (cf. Köhnken, 2004). This might explain why presence ratings tended to result in higher reliabilities than rating scales in this review.

For other forms of reliability, in particular test-retest reliability and internal consistency, not enough studies were available for a systematic review. Nonetheless, these are important aspects of reliability future research needs to address. Space limitations prevent us to report results for other related criteria, or for refinements in operationalizations, which some researchers have used to supplement the 19 CBCA criteria listed in Table 1 (e.g., Niehaus, 2001). For example, reliability of some CBCA criteria has been improved by treating them as a scale made up of several items (e.g., Küpper & Sporer, 1997; Sporer, 2004). Usually, scale reliabilities are higher than reliabilities of individual items, provided the scale has satisfactory internal consistency and positive corrected item-total correlations (at least > .20; Cortina, 1993). Reliability can also be improved by using more than two raters for all observations, employing the Spearman-Brown formula (Anastasi, 1990; Rosenthal, 1995). Unfortunately, most studies have excluded *offense specific elements* as inappropriate when a topic other than sexual abuse was

investigated. However, criterion 19 can be creatively adapted as "event-specific elements" (which can be operationalized by systematic surveys of experts for a specific type of event).

As noted before, in forensic applications CBCA is never used on its own but only as an element of a more comprehensive hypothesis testing strategy known as SVA. While SVA expert testimony is common in central Europe within the context of inquisitorial legal systems, it is usually not admissible in adversary court systems. However, CBCA is now discussed (and trained) as a "credibility assessment tool" worldwide (although it was never meant to be used as such in isolation). As a research instrument, this procedure has to meet traditional psychometric quality standards like objectivity, reliability and validity.

The extent to which CBCA ratings, along with many other factors, contribute to positive or negative evaluations of a witness's credibility by experts is still largely unknown. If assessments differ between experts, courts will be ill-served and miscarriages of justice may ensue. As recent court cases (BGH, 1999; see Jansen, 2012; Steller, 2013; Volbert & Steller, 2014) have made clear, a thorough training in diagnostic and clinical psychology is necessary as a basis for SVA and CBCA in individual cases. We are concerned that through the increased availability of various "criteria lists" in publications, in workshops and on the Internet, more and more people may feel "qualified" to conduct credibility assessments. Without proper training, this is a bad idea.

Due to the complex nature of credibility assessment, the information integration inherent in SVA does not lend itself to simple reliability analyses (although some authors have at least addressed this problem; e.g., Gumpert & Lindblad, 1999). Only when the subjective nature of these evaluations is appreciated

by decision makers can responsible decisions ensue, whether in the courtroom or any other lie-truth discrimination setting.

However, the numerous studies we were able to localize and integrate in this report do demonstrate that quite a few criteria of CBCA, the core component of SVA, can be assessed reliably. We have also demonstrated how reliability can and should be measured for *any* content-oriented approach and hope that future researchers and practitioners will heed to our advice.

References

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

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

*Akehurst, L., Manton, S., & Quandte, S. (2011). Careful calculation or a leap of faith? A field study of the translation of CBCA ratings to final credibility judgements. *Applied Cognitive Psychology, 25*, 236-243. doi:10.1002/acp.1669

Anastasi, A. (1990). *Psychological testing.* New York, NY: Macmillan Publishing Company.

*Anson, D. A., Golding, S. L., & Gully, K. J. (1993). Child sexual abuse allegations: Reliability of Criteria-based Content Analysis. *Law and Human Behavior, 17*, 331-341. doi:10.1007/bf01044512

Ariely, D. (2012). *The (honest) truth about dishonesty. How we lie to everyone—especially ourselves*. New York, NY: Harper.

Arntzen, F. (1970). *Psychologie der Zeugenaussage* [Psychology of testimony]. Göttingen, Germany: Verlag für Psychologie, Hogrefe.

Arntzen, F. (1983). *Psychologie der Zeugenaussage: System der Glaubwürdigkeits-merkmale* [Psychology of testimony: System of credibility criteria]. München, Germany: Beck.

Arntzen, F. (1992). Die Situation der forensischen Aussagepsychologie in der BRD. [Current state of psychology of testimony in the Federal Republic of Germany]. In A. Trankell (Ed.), *Reconstructing the past: The role of psychologists in criminal trials* (pp. 107-120). Deventer, The Netherlands: Kluwer.

Bartko, J. J. (1966). Intraclass correlation coefficient as a measure of reliability. *Psychological Reports, 19*, 3-11.

Bayerl, P. S., & Paul, K. I. (2011). What determines inter-coder agreement in manual annotations? A meta-analytic investigation. *Computational Linguistics, 37*, 699-725. Retrieved from: http://www.mitpressjournals.org/doi/pdf/10.1162/COLI_a_00074

BGH (1954). *BGH Urteil vom 14. Dezember 1954 - 5StR 416/54* - Gr. Strafk. b. d. AG Celle. Zuziehung eines Sachverständigen bei Kinderaussagen [Admission of experts for testimonies of children].

BGH (1999). BGH Urteil vom 30.Juli 1999 - 1StR 618/98 - LG Ansbach. StPO § 244 Abs. 4 Satz 2 Wissenschaftliche Anforderungen an Aussagepsychologische Begutachtugen [German Code of Criminal Procedure § 244 Paragraph 4 Sentence 2: Scientific requirements on credibility assessments]. Reprinted in *Praxis der Rechtspsychologie, 9*, 113-125.

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

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

*Blandón-Gitlin, I., Pezdek, K., Rogers, M., & Brodie, L. (2005). Detecting deception in children: An experimental study of the effect of event familiarity on CBCA ratings. *Law and Human Behavior, 29*, 187-197. doi:10.1007/s10979-005-2417-8

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

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

*Boychuk, T. D. (1991). *Criteria-based Content Analysis of children's statements about sexual abuse: A field-based validation study* (Unpublished doctoral dissertation). Arizona State University, Phoenix, AZ.

*Bradford, D. (2006). *Detection of deception in the confessional context* (Doctoral dissertation, University of New South Wales, Australia). Retrieved from http://unsworks.unsw.edu.au/fapi/datastream/unsworks:1567/SOURCE02

*Buck, J. A., Warren, A. R., Betman, S. I., & Brigham, J. C. (2002). Age differences in Criteria-based Content Analysis scores in typical child sexual abuse interviews. *Journal of Applied Developmental Psychology, 23*, 267-283. doi:10.1016/s0193-3973(02)00107-7

*Caso, L., Vrij, A., Mann, S., & De Leo, G. (2006). Deceptive responses: The impact of verbal and non-verbal countermeasures. *Legal and Criminological Psychology, 11*, 99-111. doi:10.1348/135532505X49936

*Chang, G. H.-Y. (2008). *Effectiveness of content analysis in assessing suspect credibility: Counterterrorism implications* (Unpublished doctoral dissertation). University of Nebraska, Lincoln, NE.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*, 37-46. doi:10.1177/001316446002000104

Cohen, J. (1968). Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin, 70*, 213-220. doi:10.1037/h0026256

*Connolly, D., & Lavoie, A. A. (2009). *The efficacy of CBCA and RM in discriminating between reports of single, repeated, and fabricated events* (Simon Fraser University, Burnaby, Canada). Retrieved from summit.sfu.ca/system/files/iritems1/8675/b51756961.pdf

Cooper, H., Hedges, L. V., & Valentine, J. C. (2009). (Eds.) *The handbook of research synthesis and meta-analysis.* New York, NY: Russell Sage Foundation.

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and application. *Journal of Applied Psychology, 78*, 98-104. doi.org/10.1037/0021-9010.78.1.98

*Craig, R. A., Scheibe, R., Raskin, D. C., Kircher, J. C., & Dodd, D. H. (1999). Interviewer questions and content analysis of children's statements of sexual abuse. *Applied Developmental Science, 3*, 77-85. doi:10.1207/s1532480xads0302_2

Cronbach, L. J. (1990). *Essentials of psychological testing.* Grand Rapids, MI: Harper & Row.

*Dana-Kirby, L. (1997). *Discerning truth from deception: Is Criteria-based Content Analysis effective with adult statements?* (Unpublished doctoral dissertation). University of Oregon, Eugene, OR.

Dettenborn, H., Fröhlich, H. H., & Szewczyk, H. (1984). *Forensische Psychologie: Lehrbuch der gerichtlichen Psychologie für Juristen, Kriminalisten, Psychologen, Pädagogen und Mediziner* [Forensic psychology: Textbook of

psychology and law for legal professionals, criminologists, psychologists, educators, and physicians]. Berlin, Germany: VEB Deutscher Verlag der Wissenschaften.

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

Donner, A., & Klar, N. (1996). The statistical analysis of kappa statistics in multiple samples. *Journal of Clinical Epidemiology, 49*, 1053-1058. doi:0.1016/0895-4356(96)00057-1

*Dukala, K., Sporer, S. L., & Polczyk, R. (in prep.). *Can cognitive interview impair lie detection using CBCA criteria in elderly witnesses' testimonies?* Manuscript in preparation, University of Krakow, Poland.

*Dunbar, N. E., Harvell, L., Jensen, M. L., Burgoon, J. K., & Kelley, K. (2012). The viability of using rapid judgments as a method of deception detection. In M. Jensen, J. Meservy, J. Burgoon, & J. Nunamaker (Eds.), *Proceedings of the Rapid Screening Technologies, Deception Detection and Credibility Assessment Symposium*, Grand Wailea, Maui, HI. Retrieved from: http://www.hicss.hawaii.edu/Reports/HICSS45_RapidScreeningTechnologiesDeceptionDetectionandCredibilityAssessmentSymposium.pdf

*Eggers, J. (2002). *Glaubwürdigkeit von Zeugenaussagen: Eine Evaluation des Kieler Trainingsprogramms zur Beurteilung der Glaubwürdigkeit von Zeugenaussagen* [Credibility of testimonies: An evaluation of the "Kieler training program"] (Unpublished diploma thesis). Christian-Albrechts Universität Kiel, Kiel, Germany.

Ekman, P. (2009). *Telling lies. Clues to deceit in the marketplace, politics, and marriage*. New York, NY: W. W. Norton & Company.

Farah, M. J., Hutchinson, J. B., Phelps, E. A., & Wagner, A. D. (2014). Functional MRI-based lie detection: scientific and societal challenges. *Nature Reviews Neuroscience, 15*, 123-131. doi:10.1038/nrn3665

Fleiss, J. L. (1981). *Statistical methods for rates and proportions.* New York, NY: Wiley.

Fleiss, J. L., & Cohen, J. (1973). Equivalence of weighted kappa and intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement, 33*, 613-619. doi:10.1177/001316447303300309

Fleiss, J. L., Levin, B., & Paik, M. C. (2003). *Statistical methods for rates and proportions.* Hoboken, NJ: Wiley.

Frick, T., & Semmel, M. I. (1978). Observer agreement and reliabilities of classroom observational measures. *Review of Educational Research, 48*, 157-184. doi:10.3102/00346543048001157

*Gödert, H. W., Gamer, M., Rill, H. G., & Vossel, G. (2005). Statement validity assessment: Inter-rater reliability of Criteria-based Content Analysis in the mock-crime paradigm. *Legal and Criminological Psychology, 10*, 225-245. doi:10.1348/135532505x52680

*Granhag, P. A., & Strömwall, L. A. (2002). Repeated interrogations: Verbal and non-verbal cues to deception. *Applied Cognitive Psychology, 16*, 243-257. doi:10.1002/acp.784

Granhag, P. A., & Strömwall, L. A. (Eds.) (2004). *The detection of deception in forensic contexts.* Cambridge, UK: Cambridge University Press.

*Granhag, P. A., Strömwall, L. A., & Landström, S. (2006). Children recalling an event repeatedly: Effects on RM and CBCA scores. *Legal and Criminological Psychology, 11*, 81-98. doi:10.1348/135532505x49620

Gumpert, C. H., & Lindblad, F. (2000). Expert testimony on child sexual abuse: A qualitative study of the Swedish approach to statement analysis. *Expert Evidence, 7*, 279-314. doi:10.1023/a:1016657130623

Gwet, K. L. (2012). *Inter-rater reliability: The definite guide to measuring the extent of agreement among multiple raters.* Gaithersburg, MD: Advanced Analytics.

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

*Hänert, P. (2007). *Die Validität inhaltlicher Glaubhaftigkeitsmerkmale unter suggestiven Bedingungen: Eine emprirische Untersuchung an Vorschulkindern* [Validity of Criteria-based Content Analysis under suggestive conditions: An empirical study with preschool children] (Doctoral dissertation, Christian-Albrechts-Universität Kiel, Kiel, Germany). Retrieved from http://d-nb.info/1019813482/34

Hauch, V., Blandón-Gitlin, I., Masip, J., & Sporer, S. L. (2015). Are computers effective lie detectors? A meta-analysis of linguistic cues to deception. *Personality and Social Psychology Review, 19*, 307-342. doi: 10.1177/1088868314556539

Hauch, V., Sporer, S. L., Michael, S., & Meissner, C. A. (2014). Does training improve the detection of deception? A meta-analysis. *Communication Research*. Advance online publication. doi:10.1177/0093650214534974

Hedges, L. V. (1982). Fitting categorical models to effect sizes from a serious of experiments. *Journal of Educational Statistics, 7*, 119-137. doi:10.2307/1164961

*Heinze, Y. (1996). *Inhaltliche Realkennzeichen in Aussagen Jugendlicher: Eine Simulationsstudie zur wissenschaftlichen Evaluation der inhaltsorientierten Aussageanalyse* [Credibility criteria in statements of adolescents: An experimental investigation of Criteria-based Content Analysis] (Unpublished diploma thesis). Westfälische Wilhelms-Universität Münster, Münster, Germany.

*Herrmann, M., & Jena, S. (1995). *Realkennzeichen im Einzellfall* [Content credibility criteria in individual cases] (Unpublished diploma thesis). Freie Universität Berlin, Berlin, Germany.

*Hettler, S. (2005). *Evaluation eines erweiterten Kanons inhaltlicher Kennzeichen wahrer und falscher Zeugenaussagen* [Evaluation of an extension of Criteria-based Content Analysis in self-experienced and fabricated testimonies] (Unpublished diploma thesis). Universität Konstanz, Konstanz, Germany.

Higgins, J. P. T., & Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine, 21*, 1539-1558. doi:10.1002/sim.1186

*Höfer, E. (1995). *Glaubwürdigkeitsdiagnostik unter differentiellen Beanspruchungsbedingungen* [Credibility assessment under varying stress conditions] (Unpublished doctoral dissertation). Christian-Albrechts-Universität Kiel, Kiel, Germany.

*Honts, C. R., & Devitt, M. K. (1993). *Credibility analysis of verbatim statements (CAVS).* University of North Dakota, Psychology Department. Grand Forks, ND. Retrieved from:

http://oai.dtic.mil/oai/oai?verb=getRecord&metadataPrefix=html&identifier=ADA271575

*Horowitz, S. W., Lamb, M. E., Esplin, P. W., Boychuk, T. D., Krispin, O., & Reiter-Lavery, L. (1997). Reliability of Criteria-based Content Analysis of child witness statements. *Legal and Criminological Psychology, 2*, 11-21. doi:10.1111/j.2044-8333.1997.tb00329.x

*Horstmann, S. (1996). *Inhaltliche Realkennzeichen in Aussagen Jugendlicher: Eine Simulationsstudie zur wissenschaftlichen Evaluation der inhaltsorientierten Aussageanalyse* [Credibility criteria in statements of adolescents: An experimental study on the evaluation of Criteria-based Content Analysis] (Unpublished diploma thesis). Westfälische Wilhelms-Universität Münster, Münster, Germany.

Hunter, J., & Schmidt, F. (2004). *Methods of meta-analysis: Correcting error and bias in research findings.* Thousand Oaks, CA: Sage.

*Janka, C. (2003). *Der Einfluss des Zeitintervalls zwischen Ereignis und Aussage auf die inhaltliche Qualität erlebnisbasierter und intentional falscher Aussagen* [The effect of the event-statement retention interval on the quality of self-experienced vs. fabricated statements] (Unpublished diploma thesis). Technische Universität Berlin, Berlin, Germany.

Jansen, G. (2012). *Zeuge und Aussagepsychologie* [Witness and psychology of testimony]. Heidelberg: C. F. Müller.

*Joffe, R. D. (1992). *Criteria-based Content Analysis: An experimental investigation with children* (Doctoral dissertation, University of British Columbia, Canada). Retrieved from: https://circle.ubc.ca/handle/2429/1723

Köhnken, G. (1982). *Sprechverhalten und Glaubwürdigkeit: Eine experimentelle Studie zur extralinguistischen und textstilistischen Aussageanalyse* [Speech behavior and credibility: An experimental study on extra-linguistic and text-stylistic statement analysis.] Unpublished doctoral dissertation, University of Kiel, Kiel, West Germany.

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

Köhnken, G. (2004). Statement Validity Analysis and the "detection of the truth". In P. A. Granhag & L. A. Strömwall (Eds.), *The detection of deception in forensic contexts* (pp. 41-63). Cambridge, UK: Cambridge University Press.

*Krahé, B., & Kundrotas, S. (1992). Glaubwürdigkeitsbeurteilung bei Vergewaltigungsprozessen: Ein aussagenanalytisches Feldexperiment [Credibility assessment in rape trails: A field study]. *Zeitschrift für experimentelle und angewandte Psychologie, 4*, 588-620.

*Krahé, B., Reimer, T., & Scheinberger, R. (1995, October). *Kriterienorientierte Aussagenanalyse als Instrument zur Glaubhaftigkeitsbegutachtung: Eine methodenkritische Untersuchung* [Criteria-based Content Analysis as an instrument for credibility assessment: A critical-methodological investigation]. Paper presented at the 6th Arbeitstagung der Fachgruppe Rechtspsychologie, Bremen, Germany.

Krause, S. (1997). *Konzeption und Evaluation der Validität des Kieler Trainings-programmes zur Beurteilung der Glaubwürdigkeit von Zeugenaussagen* [Development and validation of the Kieler training program for credibility

assessment of witness' testimonies] (Unpublished diploma thesis). Christian-Albrechts-Universität Kiel, Kiel, Germany.

Krippendorff, K. (1970). Estimating reliability, systematic error and random error of interval data. *Educational and Psychological Measurement, 30*, 61-70. doi:10.1177/001316447003000105

Küpper, B., & Sporer, S. L. (1995). Beurteilerübereinstimmung bei Glaubwürdigkeitsmerkmalen: Eine empirische Studie [Inter-rater reliability of content credibility criteria: An empirical study]. In G. Bierbrauer, W. Gottwald, & B. Birnbreier-Stahlberger (Eds.), *Verfahrensgerechtigkeit* (pp. 187-213). Köln, Germany: Otto Schmidt.

Lamb, M. E., Sternberg, K. J., & Esplin, P. W. (1994). Factors influencing the reliability and validity of statements made by young victims of sexual maltreatment. *Journal of Applied Developmental Psychology, 15*, 255-280. doi: 10.1016/0193-3973(94)90016-7

Lamb, M. E., Sternberg, K. J., Esplin, P. W., I. Hershkowitz, & Y. Orbach (1997a). Assessing the credibility of children's allegations of sexual abuse: A survey of recent research. *Learning and Individual Differences, 9*, 175-194. doi: 10.1016/S1041-6080(97)90005-4

*Lamb, M. E., Sternberg, K. J., Esplin, P. W., Hershkowitz, I., Orbach, Y., & Hovav, M. (1997b). Criterion-based Content Analysis: A field validation study. *Child Abuse & Neglect, 21*, 255-264. doi:10.1016/s0145-2134(96)00170-6

*Lamers-Winkelman, F., Buffing, F., & van der Zanden, A. P. (1992, June-July). *What children can or will tell about sexual abuse: Preliminary results.* Paper presented at the NATO Advanced Study Institute: The child witness in context: Cognitive, social and legal perspectives, Lucca, Italy.

Landis, J. R., & Koch, G. G. (1977). Measurement of observer agreement for categorical data. *Biometrics, 33*, 159-174. doi:10.2307/2529310

*Leal, S., Vrij, A., Warmelink, L., Vernham, Z., & Fisher, R. P. (2013). You cannot hide your telephone lies: Providing a model statement as an aid to detect deception in insurance telephone calls. *Legal and Criminological Psychology*. Advance online publication. doi:10.1111/lcrp.12017

*Lee, Z., Klaver, J. R., & Hart, S. D. (2008). Psychopathy and verbal indicators of deception in offenders. *Psychology, Crime, & Law, 14*, 73-84. doi:10.1080/10683160701423738

Levine, T. R. (2014). Truth-Default Theory (TDT): A theory of human deception and deception detection. *Journal of Language and Social Psychology, 33*, 378-392. doi:10.1177/0261927X14535916

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis.* Thousand Oaks, CA: Sage.

*Maier, B. (2007). *Glaubhaftigkeitsdiagnostik von Zeugenaussagen: Eine diskriminanzanalystische Untersuchung* [Credibility assessment of testimonies: An investigation with discriminant analysis]. Saarbrücken, Germany: VDM Verlag.

*Manzanero, A. L., Recio, M., Alemany, A., Vallet, R., Aróztegui, J., & Sporer, S. L. (2014). *Credibility assessment of true and simulated victims with intellectual disability.* Unpublished manuscript, Universidad Complutense de Madrid, Spain.

Masip, J., & Garrido, E. (2006). *La evaluación del abuso sexual infantil. Análisis de la validez de las declaraciones del niño* [The assessment of child sexual abuse. Validity assessment of children's statements]. Alcalá de Guadaíra, Spain: Editorial MAD.

Masip, J., Sporer, S. L., Garrido, E., & Herrero, C. (2005). The detection of deception with the reality monitoring approach: A review of the empirical evidence. *Psychology, Crime, & Law, 11,* 99-122. doi:10.1080/10683160410001726356

Maxwell, A. E. (1977). Coefficients of agreement between observers and their interpretation. *British Journal of Psychiatry, 130*, 79-83. doi:10.1192/bjp.130.1.79

*Mazzoni, G., & Ambrosio, K. (2002). *An analysis of eyewitness report in children: Using the CBCA with 7-year-old children*. Unpublished manuscript, Seton Hall University, NJ.

McGraw, K. O., & Wong, S. P. (1996a). Forming inferences about some intraclass correlation coefficients. *Psychological Methods, 1*, 30-46. doi:10.1037/1082-989x.1.4.390

McGraw, K. O., & Wong, S. P. (1996b). Forming inferences about some intraclass correlations coefficients: Correction. *Psychological Methods, 1*, 390. doi:10.1037//1082-989x.1.4.390

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

*Metzger, G. (1996). *Inhaltsgestützte Beurteilung der Glaubwürdigkeit von Zeugenaussagen* [Criteria-based Content Analysis for credibility assessment of testimonies] *(Unpublished diploma thesis).* Christian-Albrechts-Universität Kiel, Kiel, Germany.

Miller, G. R., & Stiff, J. B. (1993*). Deceptive communication.* Newbury Park, CA: Sage.

National Research Council. Committee to Review the Scientific Evidence on the Polygraph. Division of Behavioral and Social Sciences and Education (2003). *The polygraph and lie detection.* Washington, DC: The National Academies Press.

*Naumann, T. (2005). *Zur Anwendbarkeit der kriterienorientierten Inhaltsanalyse bei nicht-erlebnisbegründeten Aussagen nach Vorabinformation unterschiedlichen Ausmaßes* [Applicability of Criteria-based Content Analysis after varying amount of advance information] (Unpublished diploma thesis). Technische Universität Carolo-Wilhelmina zu Braunschweig, Braunschweig, Germany.

*Niehaus, S. (2001). *Zur Anwendbarkeit inhaltlicher Glaubhaftigkeitsmerkmale bei Zeugenaussagen unterschiedlichen Wahrheitsgehaltes* [Applicability of content criteria to testimonies with different truth status]. Frankfurt am Main, Germany: Europäische Hochschulschriften.

*Peace, K. A., & Porter, S. (2011). Remembrance of lies past: A comparison of the features and consistency of truthful and fabricated trauma narratives. *Applied Cognitive Psychology, 25*, 414-423. doi:10.1002/acp.1708

*Petersen, R. (1997). *Konzeption und Evaluation der Validität des Kieler Trainingsprogrammes zur Beurteilung der Glaubwürdigkeit von Zeugenaussagen* [Development and evaluation of the Kiel training program for credibility assessment of witness' testimonies] (Unpublished diploma thesis). Christian-Albrecht-Universität Kiel, Kiel, Germany.

Pigott, T. D. (2012). *Advances in meta-analysis.* New York, NY: Springer. doi:10.1007/978-1-4614-2278-5

*Porter, S., Peace, K. A., & Emmett, K. A. (2007). You protest too much, methinks: Investigating the features of truthful and fabricated reports of traumatic

experiences. *Canadian Journal of Behavioural Science, 39*, 79-91. doi:10.1037/cjbs2007007

*Porter, S., & Yuille, J. C. (1996). The language of deceit: An investigation of the verbal clues to deception in the interrogation context. *Law and Human Behavior, 20*, 443-458. doi:10.1007/bf01498980

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

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

Reinhard, M.-A., Sporer, S. L., Scharmach, M., & Marksteiner, T. (2011). Listening, not watching: Situational familiarity and the ability to detect deception. *Journal of Personality and Social Psychology, 101*, 467–484. doi:10.1037/a0023726

*Roma, P., Martini, P. S., Sabatello, U., Tatarelli, R., & Ferracuti, S. (2011). Validity of Criteria-based Content Analysis (CBCA) at trial in free-narrative interviews. *Child Abuse & Neglect, 35*, 613-620. doi:10.1016/j.chiabu.2011.04.004

Rosenthal, R. (1995). Methodology. In A. Tesser (Ed.), *Advanced social psychology* (pp. 17-49). New York, NY: McGraw-Hill.

*Ruby, C. L., & Brigham, J. C. (1998). Can Criteria-based Content Analysis distinguish between true and false statements of African-American speakers? *Law and Human Behavior, 22*, 369-388. doi:10.1023/a:1025766825429

*Rutta, Y. (2001). *Der Effekt von Hintergrundwissen über aussagepsychologische Methodik auf die inhaltliche Qualität von intentionalen Falschaussagen* [The

effect of background knowledge on the methods of credibility assessment on the quality of false statements] (Unpublished diploma thesis). Freie Universität Berlin, Berlin, Germany.

*Saacke, I. (1995). *Aussagequalität und kognitiver Entwicklungsstand: Eine Studie zur Überprüfung des Zusammenhangs zwischen der Auftretenshäufigkeit der Realkennzeichen in kindlichen Erlebnisberichten und der kognitiven Entwicklung im Kindesalter* [Quality of statements and cognitive mental age: An investigative study on the relationship between the frequency of occurrence of reality criteria in children's self-experienced accounts and cognitive development in childhood] (Unpublished diploma thesis). Freie Universität Berlin, Berlin, Germany.

*Sallmon, Y., & Volbert, R. (2013, September). *Zu gut, um wahr zu sein: Angaben zur Erinnerungsqualität als Mittel zur Differenzierung zwischen wahren und erfundenen Aussagen* [Too good to be true: Self-reports on memory qualities as a means to differentiate between true and false testimonies]. Paper presented at the 15. Tagung der Fachgruppe Rechtspsychologie der Deutschen Gesellschaft für Psychologie, Bonn, Germany.

*Santtila, P., Roppola, H., Runtti, M., & Niemi, P. (2000). Assessment of child witness statements using Criteria-based Content Analysis (CBCA): The effects of age, verbal ability, and interviewer's emotional style. *Psychology, Crime, & Law, 6*, 159-179. doi:10.1080/10683160008409802

*Saykaly, C., Talwar, V., Lindsay, R. C. L., Bala, N. C., & Lee, K. (2013). The influence of multiple interviews on the verbal markers of children’s deception. *Law and Human Behavior, 37*, 187-196. doi:10.1037/lhb0000023

*Scheinberger, R. (1993). *Inhaltliche Realkennzeichen in Aussagen von Erwachsenen* [Content credibility criteria in testimonies of adults] (Unpublished diploma thesis). Ferie Universität Berlin, Berlin, Germany.

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

Shadish, W. R., & Haddock, C. K. (2009). Combining estimates of effect size. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (pp. 257-277). New York, NY: Russel Sage Foundation.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86*, 420-428. doi:10.1037//0033-2909.86.2.420

Shrout, P. E., Spitzer, R. L., & Fleiss, J. L. (1987). Quantification of agreement in psychiatric diagnosis revisited. *Archives of General Psychiatry, 44*, 172-177. doi:10.1001/archpsyc.1987.01800140084013

Spitznagel, E. L., & Helzer, J. E. (1985). A proposed solution to the base rate problem in the kappa-statistic. *Archives of General Psychiatry, 42*, 725-728.

Sporer, S. L. (1983, August). *Content criteria of credibility: The German approach to eyewitness testimony*. Paper presented at the 91st Annual Meeting of the American Psychological Association, Anaheim, CA.

*Sporer, S. L. (1997a). The less travelled road to truth: Verbal cues in deception detection in accounts of fabricated and self-experienced events. *Applied Cognitive Psychology, 11*, 373-397. doi:10.1002/(sici)1099-0720(199710)11:5<373::aid-acp461>3.3.co;2-s

*Sporer, S. L. (1997b). Realitätsüberwachungskriterien und forensische Glaubwürdigkeitskriterien im Vergleich: Validitätsprüfung anhand selbsterlebter und erfundener Geschichten [Comparing reality monitoring criteria and forensic credibility criteria: Validity experiments with self-experienced and invented accounts]. In L. Greuel, T. Fabian, & M. Stadler (Eds.), *Psychologie der Zeugenaussage* (pp. 71-85). München, Germany: Psychologie Verlags Union.

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

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

*Sporer, S. L., & Bursch, S. E. (2003). *Training to detect deception by verbal means: Learning to discriminate or change in response bias?* Unpublished manuscript, University of Giessen, Giessen, Germany.

Sporer, S. L., Hauch, V., Blandón-Gitlin, I., & Masip, J. (2014). *Validity of CBCA criteria: A meta-analysis.* Manuscript in preparation.

Sporer, S. L., & Schwandt, B. (2006). Paraverbal correlates of deception: A meta-analytic synthesis. *Applied Cognitive Psychology, 20*, 421-446. doi:10.1002/acp.1190

Sporer, S. L., & Schwandt, B. (2007). Moderators of nonverbal indicators of deception. A meta-analytic synthesis. *Psychology, Public Policy, and Law, 13*, 1-34. doi:10.1037/1076-8971.13.1.1

Steller, M. (1989). Recent developments in statement analysis. In J. C. Yuille (Ed.), *Credibility assessment* (pp. 135-154). Dordrecht, Netherlands: Kluwer Academic Publishers.

Steller, M. (2013). Vier Jahrzehnte forensische Aussagepsychologie: Eine nicht nur persönliche Geschichte [Four decades of forensic psychology of testimony: Not only a personal story]. *Praxis der Rechtspsychologie, 23*, 11-32.

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

Steller, M., Volbert, R., & Wellershaus, P. (1993). *Zur Beurteilung von Kinderaussagen: Aussagepsychologische Konstrukte und methodische Strategien* [On the assessment of children's testimonies: Psychological constructs and methodological strategies]. In L. Montada (Ed.), *Bericht über den 38. Kongreß der Deutschen Gesellschaft für Psychologie in Trier 1992* (Volume 2, pp. 367-376). Göttingen: Hogrefe.

*Steller, M., Wellershaus, P., & Wolf, P. (1992). Realkennzeichen in Kinderaussagen: Empirische Grundlagen der kriterienorientierten Aussagenanalyse [Reality criteria in children's testimonies: Empirical principles of Criteria-based Content Analysis]. *Zeitschrift für Experimentelle und Angewandte Psychologie, 39*, 151-170.

*Strömwall, L. A., Bengtsson, L., Leander, L., & Granhag, P. A. (2004). Assessing children's statements: The impact of a repeated experience on CBCA and RM ratings. *Applied Cognitive Psychology, 18*, 653-668. doi:10.1002/acp.1021

Szewczyk, H. (1973). Kriterien der Beurteilung kindlicher Zeugenaussagen [Content criteria for credibility assessment of children's testimonies]. *Probleme und Ergebnisse der Psychologie, 46*, 47-66.

Trankell, A. (1972). *Reconstructing the past: The role of psychologists in criminal trials*. Deventer, The Netherlands: Kluwer.

Uebersax, J. S. (1987). Diversity of decision-making models and the measurement of interrater agreement. *Psychological Bulletin, 101*, 140-146. doi:10.1037/0033-2909.101.1.140

Undeutsch, U. (1967). Beurteilung der Glaubhaftigkeit von Zeugenaussagen [Credibility assessments of eyewitness testimonies]. In U. Undeutsch (Ed.), *Handbuch der Psychologie, Band 11: Forensische Psychologie* (pp. 26-181). Göttingen, Germany: Hogrefe.

Undeutsch, U. (1982). Statement reality analysis. In A. Trankell (Ed.), *Reconstructing the past: The role of psychologists in criminal trials* (pp. 27-56). Deventer, Netherlands: Kluwer.

Vacha-Haase, T. (1998). Reliability generalization: Exploring variance in measurement error affecting score reliability across studies. *Educational and Psychological Measurement, 58*, 6-20. doi:10.1177/0013164498058001002

Vacha-Haase, T., & Thompson, B. (2011). Score reliability: A retrospective look back at 12 years of reliability generalization studies. *Measurement and Evaluation in Counseling and Development, 44*, 159-168. doi:10.1177/0748175611409845

Van Koppen, P., & Saks, M. (2003). Preventing bad psychological scientific evidence in the Netherlands and the United States. In P. J. van Koppen, & S. D. Penrod (Eds.), *Adversarial vs. inquisitorial justice: Psychological perspectives on criminal justice systems* (pp. 283-307). New York, NY: Plenum.

Volbert, R. (2008). Glaubhaftigkeitsdiagnostik--mehr als Merkmalsorientierte Inhaltsanalyse [Credibility assessment--more than Criteria-based Content Analysis]. *Forensische Psychiatrie, Psychologie und Kriminologie, 2*, 12-19. doi: 0.1007/s11757-008-0055-y

*Volbert, R., & Lau, S. (2013). Unspezifität des autobiographischen Gedächtnisses bei Depressiven: Konsequenzen für die Aussagequalität [Unspecificity of autobiographical memory of depressive patients: Consequences for statement quality]. *Praxis der Rechtspsychologie, 23*, 54-71.

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

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

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

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

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

*Vrij, A., Edward, K., & Bull, R. (2001). People's insight into their own behaviour and speech content while lying. *British Journal of Psychology, 92*, 373-389. doi:10.1348/000712601162248

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

*Vrij, A., & Heaven, S. (1999). Vocal and verbal indicators of deception as a function of lie complexity. *Psychology, Crime, & Law, 5*, 203-215. doi:10.1080/10683169908401767

*Vrij, A., Kneller, W., & Mann, S. (2000). The effect of informing liars about Criteria-based Content Analysis on their ability to deceive CBCA-raters. *Legal and Criminological Psychology, 51*, 57-70. doi:10.1348/135532500167976

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

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

Wegener, H. (1989). The present state of statement analysis. In J. C. Yuille (Ed.), *Credibility assessment* (pp. 121-133). Dordrecht, The Netherlands: Kluwer Academic Publishers.

*Wehner, I. (2006). *Erhebung und Beurteilung von Tatverdächtigenaussagen* [Collection and assessment of suspect statements]. Frankfurt, Germany: Verlag für Polizeiwissenschaft.

Wells, G. L., & Loftus, E. F. (1991). Commentary: Is this child fabricating? Reactions to a new assessment technique. In J. Doris (Ed.), *The suggestibility of children's recollections* (pp. 168-171). Washington, DC: American Psychological Association.

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

Wilson, D.B. (2002). *Meta-analysis macros for SAS, SPSS, and Stata.* Retrieved from: http://mason.gmu.edu/~dwilsonb/ma.html

*Wrege, J. (2004). *Der Einfluss von Hintergrundinformation auf spezielle Glaubwürdigkeitsmerkmale* [The impact of background information on special credibility criteria] (Unpublished diploma thesis). Freie Universität Berlin, Berlin, Germany.

*Zaparniuk, J., Yuille, J. C., & Taylor, S. (1995). Assessing the credibility of true and false statements. *International Journal of Law and Psychiatry, 18*, 343-352. doi:10.1016/0160-2527(95)00016-b

Footnotes

[1] Originally, a search on both CBCA and Reality Monitoring (see Masip et al., 2005; Sporer, 2004) studies was performed, because the current authors are conducting meta-analyses on both sets of content criteria. This might have led to the initially high number of references.

[2] We excluded those studies for which we could not obtain reliability data on separate CBCA criteria. Also, a table with reported reliability values of the average (e.g., sum score) or a range (i.e., highest and lowest reliability) can be requested from the first author.

[3] Although most indices indicated high reliability for Criteria 10, 16 and 18, some exceptions occurred: *kappa* values were only .42 for Criterion 10 and .43 for Criterion 16. Also, Criteria 16 (.48) and 18 (.48) revealed somewhat lower median *weighted kappa/ICC* values.

Table 1

Content Criteria for Statement Analysis (adapted from Steller & Köhnken, 1989)

Criterion Number	Criterion Name and Brief Description
General Characteristics	
01	*Logical structure* Logical coherence or contextual homogeneity of a statement without logical inconsistencies.
02	*Unstructured production* The described sequence of events is not reported in a chronological order and contains digressions, but can nevertheless be joined together into a logical account.
03	*Quantity of details* Detailed description of persons, events, environments, or circumstances.
Specific Contents	
04	*Contextual embedding* The described event is intertwined with other events happening at the time, with daily routines, etc.
05	*Descriptions of interactions* Report of a chain of actions and reactions between the narrator and the perpetrator.
06	*Reproduction of conversation* Literal reproduction of a dialogue or utterances of at least one person.
07	*Unexpected complications during the incident* Description of unexpected difficulties that interrupted the normal progress of the event.
Peculiarities of Contents	
08	*Unusual details* Description of odd or unexpected, but not impossible, details.
09	*Superfluous details* Report of details that are irrelevant or do not contribute to the accusation (such as peripheral details).

10 *Accurately reported details misunderstood*
Accurate descriptions of acts or events that the narrator does not understand (e.g., a child accurately describes an ejaculation misrepresenting the semen as urine).

11 *Related external associations*
Description of an event involving the narrator and the alleged perpetrator that is different from the target event but is related to it (e.g., sexual comments made by the perpetrator to the presumed victim several days before the alleged sexual abuse).

12 *Accounts of subjective mental state*
Descriptions of the narrator's own emotions or cognitions at the time of the event.

13 *Attribution of perpetrator's mental state*
Descriptions of emotions, motivations, or cognitions attributed to the perpetrator at the time of the event.

Motivation-Related Contents

14 *Spontaneous corrections*
The narrator corrects herself or himself or adds more differentiating details.

15 *Admitting lack of memory*
The narrator says s/he does not know or cannot remember.

16 *Raising doubts about one's own testimony*
The narrator expresses concern that his testimony may look implausible or unbelievable.

17 *Self-deprecation*
The narrator provides self-incriminating or self-accusing details.

18 *Pardoning the perpetrator*
The narrator provides information that exonerates the accused perpetrator, or refrains from incriminating him/her further.

Offense-Specific Elements

19 *Details characteristic of the offense*
The narrator provides correct crime-specific details that are not common knowledge.

Table 2

Report, Study and Rating Characteristics

Report Characteristics

Publication Year (*k* = 77)	Range	*Mdn*	*Mode*
	1991 - 2013	2000	1997

Publication Status (*k* = 81) [a]	Number	%
Journal articles	42	51.85
Diploma, Bachelor, or Master theses	14	17.28
Dissertations	11	13.58
Conference presentations	5	6.17
Unpublished manuscripts	3	3.70
Book chapters	2	2.47
Manuscripts in press	2	2.47
Published research reports	1	1.23
Manuscript under review	1	1.23

Language of Report (*k* = 82)	Number	%
English	58	70.73
German	24	29.27

Study Characteristics

Research Paradigm (*k* = 82)	*k*	%
Experiment	44	53.66
Quasi-experiment	25	30.49
Field study	13	15.85

Design (*k* = 80)	*k*	%
Between-participants	49	61.25
Within-participants	31	38.75

Senders	*M*	*SD*	Range
Number of senders per study (*k* = 80)	65.18	48.35	1 – 291
Number of female senders per study (*k* = 67)	40.40	35.14	0 – 233
Number of male senders per study (*k* = 67)	24.79	22.75	0 – 104
Age (*k* = 58)	20.29	11.49	5.40 – 71.21

Language of Statements (*k* = 81) [a]	*k*	%
English	40	49.38
German	26	32.10
Swedish	5	6.17
Italian	4	4.94
Dutch	3	3.70

Hebrew	1	1.23
Polish	1	1.23
Spanish	1	1.23

Mode of Production (*k* = 79)	*k*	%
Spoken	70	88.61
Handwritten	7	8.86
Typed	2	2.53

Status of the Liar (*k* = 81)	*k*	%
Victim	25	30.86
Actor	22	27.16
Witness	10	12.35
Perpetrator	8	9.88
Several roles	16	19.75

Type of Event (*k* = 82) [a]	*k*	%
Significant life event	30	36.59
Participate	13	15.85
Sexual abuse	11	13.41
Watch video	8	9.76
Trivial life event	6	7.32
Mock crime	5	6.10
Other real crime (not sexual abuse)	5	6.10
Observed staged event	2	2.44
Several events	2	2.44

Emotional Valence (*k* = 82)	*k*	%
Negative	57	69.51
Neutral	15	18.29
Both negative and positive	8	9.76
Positive	2	2.44

Motivation of Senders (*k* = 78) [a]	*k*	%
None	34	43.59
Low	27	34.62
Medium	4	5.13
High	13	16.67

Interview Style (*k* = 75)	*k*	%
Free report only	26	34.67
Subsequent semi-structured interview	28	37.33
Subsequent structured interview	17	22.67
Cognitive Interview	4	5.33

Raters and Rating Process Characteristics

Number of raters per study (*k* = 81)	*k*	%
Two	52	64.20
Three	13	16.05
Four	7	8.64

Between six and 119	9	11.11

Occupation of raters (*k* = 54) [a]	*k*	%
Graduate students	22	40.74
Undergraduate students	18	33.33
Psychologists	8	14.81
Several groups	3	5.56
Ph.D.	2	3.70
Police officers	1	1.85

Training components (*k* = 81)	*k*	%
(1) Background literature	50	61.73
(2) Operational definitions (*k* = 82)	77	93.90
(3) Examples (*k* = 82)	46	56.10
(4) Lecture(s)	57	70.37
(5a) No practice (0/2)	19	23.46
(5b) Practice without discussion/feedback/homework (1/2)	13	16.05
(5c) Practice with discussion/feedback/homework (2/2)	49	60.49

Rating scales (*k* = 82) [a]	*k*	%
Several rating scales	24	29.27
0 (absent); 1 (present); 2 (strongly present)	16	19.51
Dichotomous	14	17.07
Other Likert scale	10	12.20
Frequency count	7	8.54
0 to 4	5	6.10
1 to 7	5	6.10
1 to 5	1	1.22

Percentage of statements rated by two or more raters (*k* = 82)	*k*	%
100%	60	73.17
12.5% to 83.0%	22	26.83

Number of accounts per study after adjustment (*k* = 82)	*M*	*SD*	*Mdn*	*Mode*
	54.00	42.84	42.52	40.01

Inter-rater reliability indices reported as the single measure (*k* = 82)	*k*	%
Several	31	37.80
One	40	48.78
Pearson's *r*	17	20.73
Percentage agreement (*PCA*)	7	8.54
ICC	7	8.54
Cohen's *kappa*	6	7.32
Maxwell's *RE*	2	2.44
Spearman's *rho*	1	1.22
Weighted kappa	0	0.00
Range or average reported	11	13.41

Notes. [a] = The sum of percentages do not add up exactly to 100.00% due to rounding issues.

Table 3

Meta-Analyses of Inter-Rater Reliability Measured with Pearson's r

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	Sum score
k	31	30	35	34	30	32	27	31	29	15	19	33	31	31	31	22	14	13	7	6
N	1768	1717	1843	1760	1660	1635	1514	1655	1481	926	1026	1677	1675	1319	1319	1007	406	509	176	277
Min	-.29	-.18	-.65	-.07	-.29	.46	-.13	.03	-.28	-.04	-.09	.35	.03	-.13	.02	-.06	-.22	-.06	-.15	.44
Max	1.00	.92	.99	1.00	1.00	1.00	1.00	1.00	.93	1.00	1.00	1.00	1.00	.93	1.00	1.00	1.00	1.00	1.00	.98
$r_{unweighted}$	**.69**	**.46**	**.71**	**.69**	**.66**	**.86**	**.64**	**.62**	**.50**	**.80**	**.65**	**.79**	**.77**	**.58**	**.79**	**.73**	**.80**	**.71**	**.68**	**.90**
$r_{weighted}$	**.69**	**.46**	**.73**	**.71**	**.65**	**.86**	**.64**	**.62**	**.52**	**.81**	**.67**	**.79**	**.76**	**.60**	**.78**	**.73**	**.79**	**.72**	**.71**	**.90**
Lower *CI*	.49	.37	.62	.59	.55	.79	.51	.49	.42	.60	.47	.72	.67	.50	.67	.56	.59	.15	-.55	.66
Upper *CI*	.82	.55	.82	.80	.73	.90	.75	.73	.62	.92	.80	.85	.83	.69	.85	.84	.89	.93	.98	.97
SE	.16	.06	.10	.11	.08	.10	.10	.10	.07	.22	.15	.09	.10	.07	.11	.15	.19	.37	.64	.33
Z	**5.4**	**8.5**	**9.0**	**8.2**	**9.8**	**12.2**	**7.4**	**7.3**	**8.3**	**5.0**	**5.4**	**12.2**	**10.2**	**9.3**	**9.4**	**6.2**	**5.5**	**2.3**[a]	1.2	**4.3**
Q	**1124**	**136**	**567**	**576**	**238**	**449**	**338**	**406**	**155**	**533**	**331**	**327**	**382**	**177**	**394**	**396**	**140**	**776**	**406**	**146**
I^2	97.3	78.7	94.0	94.3	87.8	93.1	92.3	92.6	81.9	97.4	94.6	90.2	92.1	83.0	92.4	94.7	90.7	98.5	98.5	96.6

tau^2	0.69	0.07	0.32	0.35	0.14	0.29	0.23	0.26	0.10	0.68	0.37	0.20	0.24	0.16	0.31	0.43	0.40	1.84	3.78	0.69
1st Qu.	.27	.23	.51	.34	.44	.69	.32	.29	.37	.04	.27	.58	.53	.32	.60	.14	.30	.15	-.05	.73
Mdn	.51	.35	.68	.56	.53	.83	.49	.48	.48	.46	.48	.72	.64	.58	.74	.61	.64	.33	.35	.91
3rd Qu.	.69	.61	.81	.84	.74	.89	.71	.75	.62	.87	.74	.84	.81	.75	.84	.74	.87	.70	.73	.97
IQR	.42	.37	.30	.50	.30	.20	.39	.46	.25	.84	.47	.26	.29	.43	.23	.60	.57	.55	.78	.24
$k_{Base\ rate}$	20	19	19	16	18	17	13	17	16	6	10	16	14	15	12	9	5	3	3	4
$M_{Base\ rate}$	.79	.31	.64	.37	.32	.33	.27	.23	.25	.02	.13	.37	.21	.30	.38	.21	.11	.16	.38	.41

Notes. k = number of hypothesis tests; N = adjusted number of statements that were coded by at least two independent raters; Min = Minimum; Max = Maximum; r = Correlation Coefficient (Pearson's r or Spearman's r); *CI* = 95% confidence interval; *SE* = Standard Error; Z = z test; Q = homogeneity test statistic; I^2 = measure of heterogeneity; Qu. = Quartile; IQR = Interquartile range; *Mdn* = Median; values in **bold** indicate significance at $p < .001$; [a] = $p = .019$.

Table 4

Meta-Analyses of Inter-Rater Reliability Measured with Percentage Agreement (PCA) without PCA Values of ≥ .999

	1	2*	3	4	5	6	7	8	9	10	11*	12	13	14*	15	16	17	18	19	Sum score*
k	25	29	28	23	26	24	24	24	24	16	22	24	21	26	22	14	13	10	6	8
N	1466	1688	1605	1402	1563	1511	1477	1518	1507	1080	1325	1427	1439	1359	1043	602	595	623	252	574
Min	.33	.30	.24	.27	.18	.21	.25	.33	.20	.28	.37	.26	.27	.16	.11	.16	.34	.58	.39	.32
Max	.98	.94	.99	.94	.98	.95	.98	.96	.96	.99	.97	.95	.95	.92	.97	.98	.98	.95	.88	.90
$PCA_{unweighted}$	**.74**	**.67**	**.67**	**.65**	**.71**	**.72**	**.72**	**.68**	**.66**	**.85**	**.75**	**.70**	**.75**	**.68**	**.65**	**.78**	**.72**	**.75**	**.66**	**.75**
$PCA_{weighted}$	**.79**	**.70**	**.70**	**.68**	**.77**	**.77**	**.79**	**.73**	**.71**	**.93**	**.83**	**.73**	**.81**	**.71**	**.70**	**.90**	**.86**	**.80**	**.75**	**.76**
Lower *CI*	.68	.59	.57	.55	.66	.66	.68	.61	.59	.85	.73	.60	.70	.59	.57	.78	.74	.65	.50	.55
Upper *CI*	.90	.81	.83	.81	.88	.88	.89	.85	.84	1.00	.93	.85	.91	.84	.84	1.01	.99	.96	.99	.97
SE	.06	.06	.07	.07	.05	.06	.05	.06	.06	.04	.05	.06	.05	.06	.07	.06	.06	.08	.13	.11
Z	**14.1**	**12.5**	**10.7**	**10.1**	**14.0**	**13.6**	**14.8**	**12.1**	**11.2**	**21.4**	**16.2**	**11.70**	**15.0**	**11.5**	**10.0**	**15.5**	**13.4**	**10.0**	**6.0**	**7.1**
Q	**390**	**305**	**752**	**354**	**374**	**325**	**212**	**312**	**448**	**71.5**	**104**	**309**	**175**	**323**	**395**	**84**	**73**	**70**	**25**	**158**
I^2	93.8	90.8	96.4	94.0	93.3	92.9	89.2	92.6	94.9	79.0	79.9	92.6	88.6	92.3	94.7	84.4	83.5	87.2	79.6	95.6

tau^2	0.02	0.03	0.06	0.04	0.02	0.02	0.02	0.03	0.04	0.01	0.01	0.03	0.13	0.04	0.04	0.01	0.01	0.13	0.02	0.04
1st Qu.	.63	.55	.50	.51	.56	.61	.61	.55	.57	.88	.61	.56	.67	.62	.53	.75	.48	.63	.50	.74
Mdn	.75	.68	.70	.67	.79	.80	.76	.73	.68	.94	.80	.78	.83	.74	.68	.90	.80	.75	.68	.79
3rd Qu.	.88	.79	.84	.81	.88	.89	.83	.85	.81	.95	.90	.82	.91	.83	.84	.95	.94	.86	.83	.88
IQR	.26	.24	.34	.30	.32	.27	.22	.30	.24	.07	.28	.26	.24	.21	.32	.20	.46	.23	.33	.17
$k_{\text{Base rate}}$	18	21	19	15	19	17	16	17	18	10	15	17	17	17	13	9	7	4	4	7
$M_{\text{Base rate}}$	.76	.36	.64	.45	.33	.32	.22	.20	.26	.04	.20	.45	.12	.30	.35	.19	.11	.46	.73	.40

Notes. *k* = number of hypothesis tests; *N* = adjusted number of statements that were coded by at least two independent raters; Min = Minimum; Max = Maximum; *PCA* = Percentage Agreement; *CI* = 95% confidence interval; *SE* =Standard Error; *Z* = *Z*-test; *Q* = homogeneity test statistic; I^2 = measure of heterogeneity; Qu. = Quartile; IQR = Interquartile Range; *Mdn* = Median; values in **bold** indicate significance at $p < .001$; * = No values of .999 occurred for this criterion.

Table 5

Unweighted Meta-Analyses of Inter-Rater Reliability Measured with Cohen's Kappa

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	Sum score
k	22	20	22	20	23	21	18	21	20	15	19	21	20	21	15	10	12	9	8	6
N	1420	1410	1422	1386	1508	1450	1284	1417	1388	1048	1262	1334	1356	1188	778	461	612	492	357	445
Min	-.03	-.15	.00	-.07	-.08	.03	.00	-.16	.03	-.07	-.05	.07	.04	-.01	.16	-.03	-.02	-.09	-.03	.16
Max	1.00	.95	.95	1.00	.97	.97	1.00	1.00	1.00	1.00	1.00	1.00	1.00	.93	1.00	1.00	1.00	1.00	.70	.80
$kappa_{unweighted}$	.35	.32	.45	.45	.45	.57	.45	.45	.44	.42	.41	.52	.51	.39	.54	.43	.57	.51	.30	.54
1st Q.	.01	.07	.24	.27	.17	.40	.27	.17	.15	.02	.10	.25	.25	.17	.36	.01	.33	.22	.08	.30
Mdn	.38	.32	.50	.43	.41	.61	.39	.52	.48	.47	.27	.58	.55	.47	.50	.32	.61	.56	.23	.65
3rd Q.	.50	.46	.61	.67	.73	.74	.74	.68	.63	.80	.71	.71	.72	.55	.75	.88	.90	.79	.53	.75
IQR	.49	.39	.37	.40	.56	.34	.47	.51	.48	.79	.61	.46	.47	.38	.39	.87	.57	.57	.44	.45
$k_{\text{Base rate}}$	19	16	18	16	18	16	13	16	15	10	14	17	15	15	10	7	6	4	6	4
$M_{\text{Base rate}}$	.77	.47	.67	.50	.35	.28	.23	.19	.27	.04	.20	.40	.18	.30	.36	.22	.09	.45	.50	.35
$SD_{\text{Base rate}}$	.19	.25	.18	.21	.21	.17	.16	.14	.19	.04	.17	.21	.23	.24	.20	.28	.06	.43	.38	.11

Notes. k = number of hypothesis tests; *N* = adjusted number of statements that were coded by at least two independent raters; Min = Minimum; Max = Maximum; *Mdn* = Median; Qu. = Quartile; IQR = Interquartile Range*; M* = mean; *SD* = standard deviation.

Table 6

Unweighted Meta-Analyses of Inter-Rater Reliability Measured with Weighted Kappa and ICC

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	Sum score
k_{total}	10	10	13	10	10	11	9	9	10	5	10	10	9	11	9	6	6	4	2	2
$k_{weighted\ kappa}$	4	4	4	4	4	4	4	4	4	2	4	4	3	4	4	2	4	3	1	0
k_{ICC}	6	6	9	6	6	7	5	5	6	3	6	6	6	7	5	4	2	1	1	2
N	573	573	880	758	573	811	533	533	573	397	573	573	533	651	373	210	240	185	24	123
Min	-.03	.17	.37	.04	.08	.32	.02	.20	.15	-.02	-.03	.40	.50	-.10	.26	-.02	.00	.03	-.01	.91
Max	.77	.91	.88	.92	1.00	1.00	.75	1.00	.90	1.00	.95	.96	.92	.89	.88	1.00	1.00	1.00	.66	.91
1st Quartile	.01	.29	.53	.32	.31	.46	.38	.32	.29	.11	.00	.54	.53	.33	.42	.06	.46	.22	[a]	[a]
Mdn	.18	.55	.75	.49	.53	.68	.45	.67	.46	.67	.55	.62	.58	.59	.60	.48	.79	.46	[a]	[a]
3rd Quartile	.36	.76	.77	.67	.77	.87	.65	.88	.73	.92	.75	.88	.64	.75	.82	.73	.92	.72	[a]	[a]
Interquartile Range	.35	.47	.24	.36	.47	.41	.27	.56	.44	.81	.75	.34	.11	.42	.40	.67	.46	.51	[a]	[a]
$k_{\text{Base rate}}$	6	6	8	7	5	6	4	4	4	3	5	5	4	6	4	3	3	1	1	2
$M_{\text{Base rate}}$	.71	.30	.60	.28	.26	.29	.15	.13	.30	.02	.07	.32	.10	.39	.34	.10	.06	.09	.28	.25
$SD_{\text{Base rate}}$	.17	.24	.09	.17	.18	.23	.18	.08	.14	.01	.06	.10	.05	.38	.28	.10	.05	[a]	[a]	.01

Notes. *k* = number of hypothesis tests; *N* = adjusted number of statements that were coded by at least two independent raters; Min = Minimum; Max = Maximum; *Mdn* = Median; [a] = Value is not meaningful due to small *k*.

Table 7

Unweighted Meta-Analyses of Inter-Rater Reliability Measured with Maxwell's Random Error

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	Sum score
k	13	13	13	13	13	13	13	12	13	12	12	13	13	11	7	7	6	6	6	5
N	1022	1022	1022	1022	1022	1022	1022	926	1022	942	942	1022	1022	742	335	335	295	295	295	381
Min	.57	.06	.14	.05	.07	.38	-.01	.35	.18	.03	.00	.12	.03	.24	.22	.57	.37	.67	-.22	.21
Max	.95	.84	.90	.93	.92	1.00	.92	.93	.94	1.00	.84	.90	.93	.83	.73	1.00	.88	1.00	1.00	.74
1st Quartile	.66	.30	.28	.48	.20	.52	.56	.40	.37	.80	.51	.46	.50	.39	.35	.72	.73	.83	.61	.49
Mdn	.77	.40	.40	.67	.62	.73	.67	.49	.52	.91	.59	.61	.64	.50	.53	.91	.81	.85	.73	.51
3rd Quartile	.82	.76	.58	.75	.69	.83	.79	.64	.75	.95	.76	.65	.83	.76	.54	.95	.88	.88	.81	.73
IQR	.16	.46	.30	.27	.49	.31	.23	.24	.38	.16	.25	.19	.33	.36	.19	.24	.15	.05	.21	.24
$k_{\text{Base rate}}$	10	10	10	10	10	10	10	9	10	9	9	10	10	11	4	4	3	3	3	4
$M_{\text{Base rate}}$	.74	.43	.65	.58	.36	34	.27	.20	.28	.06	.25	.35	.12	.31	.45	.01	.08	.35	.68	.42
$SD_{\text{Base rate}}$	.21	.30	.18	.29	.21	.19	.18	.16	.23	.09	.17	.18	.05	.23	.14	.01	.06	.51	.27	.16

Notes. *k* = number of hypothesis tests; *N* = adjusted number of statements that were coded by at least two independent raters; Min = Minimum; Max = Maximum; IQR = Interquartile Range; *M* = Mean (Maxwell's *RE*); *Mdn* = Median.

Table 8

Results for Moderator-Analyses for Research Paradigm as a Predictor of Pearson's r

Research Paradigm		1	2	3	4	5	6	7	8	9	10
Laboratory Experiment	k_1	25	23	26	24	22	21	18	22	19	11
	$r\ [CI_{low};CI_{high}]$	.55 [.51;.58]	.40 [.36;.45]	.69 [.66;.72]	.58 [.54;.61]	.61 [.57;.64]	.76 [.74;.78]	.59 [.55;.63]	.49 [.45;.53]	.45 [.40;.50]	.34 [.27;.40]
Field- and Quasi-Experiment	k_2	6	7	9	10	8	11	9	9	10	4
	$r\ [CI_{low};CI_{high}]$	.67 [.59;.73]	.44 [.32;.55]	.79 [.74;.83]	.84 [.80;.87]	.60 [.51;.68]	.93 [.91;.94]	.73 [.66;.78]	.68 [.60;.75]	.55 [.46;.63]	.93 [.90;.95]
Q_b		7.47**	0.39	12.10***	67.86***	0.03	85.71***	11.72***	15.59***	3.64*	173.96***

(*Table 8 continues*)

Table 8 continued

Research Paradigm		11	12	13	14	15	16	17	18	19
Laboratory Experiment	k_1	13	23	21	22	22	18	8	8	5
	$r\ [CI_{low};CI_{high}]$	.35 [.29;.41]	.66 [.63;.69]	.60 [.57;.64]	.55 [.50;.59]	.70 [.67;.73]	.56 [.52;.61]	.62 [.54;.70]	.79 [.75;.83]	.95 [.93;.97]
Field- and Quasi-Experiment	k_2	6	10	10	9	9	4	6	5	2
	$r\ [CI_{low};CI_{high}]$	.78 [.72;.82]	.85 [.82;.88]	.77 [.72;.82]	.71 [.64;.76]	.80 [.75;.84]	.72 [.58;.82]	.82 [.76;.87]	.66 [.56;.74]	.48 [.06;.75]
Q_b		75.01***	48.24***	22.28***	13.79***	10.12**	4.56*	16.67***	8.62**	#

Notes. Moderator analyses were not computed with subgroup *k*s < 3. *k* = number of studies in each group; *r* = weighted average effect size (Pearson's *r*); Q_b= homogeneity test statistic: variability between group means explained by the categorical variable; # = moderator analysis not computed.
***p < .001, **p < .01, *p < .05.

Table 9

Results for Moderator-Analyses for Training Intensity as a Predictor of Pearson's r

Number of Training Components		1	2	3	4	5	6	7	8	9	10
2-4 components	k_1	9	7	10	11	9	11	7	9	10	4
	$r\ [CI_{low};CI_{high}]$	.38 [.29;.47]	.41 [.31;.50]	.80 [.76;.83]	.45 [.37;.53]	.68 [.61;.73]	.80 [.76;.83]	.57 [.48;.64]	.43 [.34;.52]	.45 [.36;.53]	.28 [.15;.40]
5-6 components	k_2	22	23	25	23	21	21	20	22	19	11
	$r\ [CI_{low};CI_{high}]$	.61 [.57;.64]	.41 [.36;.45]	.68 [.65;.71]	.68 [.65;.71]	.59 [.55;.62]	.81 [.79;.83]	.63 [.59;.66]	.55 [.50;.58]	.48 [.43;.52]	.54 [.49;.59]
Q_b		25.13***	0.01	20.32***	33.03***	6.04*	0.13	1.98	5.73*	0.35	16.29***

(*Table 9 continues*)

Table 9 continued

Number of Training Components		11	12	13	14	15	16	17	18	19
2-4 components	k_1	5	11	8	9	8	9	7	6	3
	$r\ [CI_{low};CI_{high}]$	.06 [-.06;.19]	.61 [.54;.68]	.54 [.45;.61]	.38 [.29;.47]	.83 [.77;.88]	.38 [.29;.47]	.79 [.66;.87]	.99 [.98;.99]	-.12 [-.67;.52]
5-6 components	k_2	14	22	23	22	23	22	7	7	4
	$r\ [CI_{low};CI_{high}]$	.58 [.53;.63]	.72 [.70;.75]	.66 [.62;.69]	.61 [.57;.64]	.71 [.68;.74]	.61 [.57;.64]	.71 [.65;.76]	.41 [.32;.50]	.95 [.93;.96]
Q_b		66.50***	10.74**	8.24***	25.13***	11.14***	25.13***	1.46	399.79***	27.99***

Notes. Moderator analyses were not computed with subgroup *k*s < 3. *k* = number of studies in each group; *r* = weighted average effect size (Pearson's *r*); Q_b= homogeneity test statistic: variability between group means explained by the categorical variable.
***p < .001, **p < .01, *p < .05.

Table 10

Results for Moderator-Analyses for Rating Scale as a Predictor of Pearson's r

Rating Scale		1	2	3	4	5	6	7	8	9	10
Presence	k_1	12	12	11	8	10	8	4	7	7	3
	$r\ [CI_{low};CI_{high}]$	.66 [.61;.70]	.45 [.38;.51]	.55 [.49;.61]	.62 [.56;.67]	.71 [.67;.75]	.80 [.77;.83]	.63 [.57;.69]	.61 [.55;.66]	.47 [.40;.53]	.64 [.56;.71]
Likert or weighting	k_2	18	17	19	20	17	18	19	18	18	10
	$r\ [CI_{low};CI_{high}]$	.49 [.44;.54]	.37 [.31;.43]	.68 [.65;.72]	.56 [.51;.61]	.51 [.46;.56]	.73 [.69;.76]	.61 [.56;.65]	.45 [.39;.50]	.46 [.40;.51]	.23 [.15;.31]
Frequency counts	k_3			5	6		6	4	6	4	2
	$r\ [CI_{low};CI_{high}]$			.90 [.88;.92]	.80 [.76;.84]		.93 [.91;.94]	.63 [.56;.71]	.56 [.47;.63]	.53 [.42;.62]	.94 [.91;.96]
Q_b		24.07***	3.22	146.66***	47.59***	35.83***	97.76***	0.86	17.46**	1.25	#

(*Table 10 continues*)

Table 10 continued

Rating Scale		11	12	13	14	15	16	17	18	19
Presence	k_1	5	8	9	8	8	7	1	1	1
	$r\ [CI_{low};CI_{high}]$	.51 [.43;.59]	.68 [.63;.73]	.65 [.60;.69]	.63 [.57;.70]	.86 [.83;.88]	.81 [.76;.84]	.78 [.61;.88]	1.00 [1.00;1.00]	.76 [.45;.91]
Likert or weighting	k_2	11	17	17	18	18	13	11	10	6
	$r\ [CI_{low};CI_{high}]$	.38 [.29;.46]	.68 [.64;.72]	.59 [.54;.63]	.53 [.47;.58]	.58 [.52;.62]	.37 [.29;.44]	.63 [.54;.71]	.31 [20;.41]	.94 [.92;.96]
Frequency counts	k_3		7	5	5	5		2	2	0
	$r\ [CI_{low};CI_{high}]$		.78 [.73;.82]	.73 [.67;.78]	.65 [.57;.71]	.80 [.76;.84]		.81 [.74;.86]	.67 [.56;.75]	
Q_b		5.17**	10.93**	12.29**	8.34*	99.71***	100.16***	#	#	#

Notes. Frequency counts were not included in moderator analyses. Moderator analyses were not computed with subgroup *k*s < 3. *k* = number of studies in each group; *r* = weighted average effect size (Pearson's *r*); Q_b= homogeneity test statistic: variability between group means explained by the categorical variable; # = moderator analysis not computed.
***p < .001, **p < .01, *p < .05.

Table 11

Results for Meta-Regressions with Base Rate and Squared Base Rates (as Indicators of Linear and Curvilinear Relationships)

	1	2	3	4	5	6	7	8	9
k	20	19	19	16	18	17	13	17	16
Q_{Model}	1.38	16.59***	28.04***	14.79***	2.78	2.49	10.33**	31.75***	14.34***
$Q_{Residual}$	278.89***	58.82***	69.84***	207.87***	118.64***	251.16***	93.52***	181.43***	38.51***
$B_{Base\ Rate}$	0.27	0.64***	1.71***	0.64***	0.24	0.21	0.38*	1.07***	0.19
$beta_{Base\ Rate}$	.07	.50***	.52***	.24***	-.16	.06	.25*	.43***	.10
$B_{Base\ Rate^2}$	0.19	1.45*	0.73	1.85*	0.07	0.88	0.62	4.57***	5.95***
$beta_{Base\ Rate^2}$	.01	.29*	.04	.13*	.01	.06	.09	.53***	.53***
R^2	.00	.22	.29	.07	.02	.01	.10	.15	.27
$R^2_{95\%\ CI}$	[-.02; .02]	[-.06; .50]	[-.01; .58]	[-.13; .27]	[-.09; .13]	[-.07; .09]	[-.14; .34]	[-.11; .41]	[-.04; .58]

(*Table 11 continues*)

Table 11 continued

	11	12	13	14	15	16
k	10	16	14	15	12	9
Q_{Model}	80.45***	51.03***	6.39*	5.47	19.30***	21.85***
$Q_{Residual}$	144.70***	169.20***	139.97***	58.19***	152.76***	123.20***
$B_{Base\ Rate}$	4.14***	1.69***	1.01*	0.54*	0.42*	1.66***
$beta_{Base\ Rate}$	.57***	.66***	.65*	.42*	.15*	.36***
$B_{Base\ Rate}^2$	27.49***	6.43***	1.21	1.25	3.21***	4.65*
$beta_{Base\ Rate}^2$	.47***	.46***	.52	.31	.27***	.17*
R^2	.36	.23	.04	.09	.11	.15
$R^2_{95\%\ CI}$	[.01; .71]	[-.07; .53]	[-.12; .20]	[-.13; .31]	[-.15; .37]	[-.16; .46]

Notes. Q_M = homogeneity test statistic for the regression model: variability between group means explained by the regression model; Q_R = homogeneity test statistic for the residual: unexplained variability between group means expected by chance; *B* = unstandardized regression coefficient; *beta* = standardized *regression weights*; R^2 = determination coefficient: amount of variability explained by predictor variable(s); meta-regression for Criteria 10, 17, 18, and 19 is not meaningful due to small k.
***$p < .001$, **$p < .01$, *$p < .05$.

Table 12

Results for Multiple Meta-Regression for Rating Scale, Research Paradigm, and Training Intensity as Predictors of Pearson's r

	1	2	3	4	5	6	7	8	9
k	30	29	30	28	27	26	23	25	25
Q_{Model}	54.71***	4.11	59.16***	80.15***	50.68***	65.73***	2.48	40.55***	1.44
$Q_{Residual}$	1064.99***	131.34***	315.22***	365.30***	183.99***	247.48***	284.71***	316.09***	149.94***
$B_{Paradigm}$	0.26***	0.07	0.38***	0.65***	-0.01	0.62***	0.07	0.50***	0.09
$beta_{Paradigm}$ (1=Laboratory; 2=Field)	.11***	.07	.29***	.36***	-.01	.40***	.05	.25***	.08
$B_{Training}$	0.26***	-0.04	-0.26***	0.29***	-0.25***	-0.18**	0.09	0.11	0.04
$beta_{Training}$ (1=short; 2=long)	.13***	-.05	-.20***	.22***	-.26***	-.16**	.07	.09	.04
$B_{Rating\ Scale}$	-0.23***	-0.10	0.11	-0.05	-0.38***	-0.24***	-0.02	-0.28***	-0.01
$beta_{Rating\ Scale}$ (1=Presence; 2 = Likert)	-.14***	-.16	.10	-.04	-.45***	-.24***	-.02	-.26***	-.02
R^2	0.49	0.03	0.16	0.18	0.22	0.21	0.01	0.11	0.01
$R^2_{95\%\ CI}$	[.27; .71]	[-.08; .14]	[-.05; .37]	[-.04; .40]	[-.02; .46]	[-.03; .45]	[-.06; .08]	[-.09; .31]	[-.06; .08]

(*Table 12 continues*)

Table 12 continued

	11	12	13	14	15	16
k	16	26	26	26	26	20
Q_{Model}	74.55***	23.43***	15.00**	10.68**	96.04***	122.09***
$Q_{Residual}$	163.05***	241.78***	322.83***	124.08***	279.30***	271.63***
$B_{Paradigm}$	0.20*	0.36***	0.25**	.12	0.17	0.61***
$beta_{Paradigm}$ (1=Laboratory; 2=Field)	.14*	.25***	.16**	.12	.10	.25***
$B_{Training}$	0.75***	0.15*	0.12	0.19*	-0.16	0.03
$beta_{Training}$ (1=short; 2=long)	.62***	.15*	.10	.18*	-.09	.01
$B_{Rating\ Scale}$	0.25**	0.00	-0.08	-0.19**	-0.61***	-0.83***
$beta_{Rating\ Scale}$ (1=Presence; 2 = Likert)	.22**	.00	-.08	-.24**	-.47***	-.57***
R^2	0.32	0.09	0.04	0.08	0.26	0.31
$R^2_{95\%\ CI}$	[.03; .61]	[-.09; .27]	[-.09; .16]	[-.09; .25]	[.01; .51]	[.04; .58]

Notes. Q_M= homogeneity test statistic for the regression model: variability between group means explained by the regression model; Q_R = homogeneity test statistic for the residual: unexplained variability between group means expected by chance; *B* = unstandardized regression coefficient; *beta* = standardized *regression weights*; R^2 = determination coefficient: amount of variability explained by predictor variable(s); meta-regression for Criteria 10, 17, 18, and 19 is not meaningful due to small k. ***p < .001, **p < .01, *p < .05.

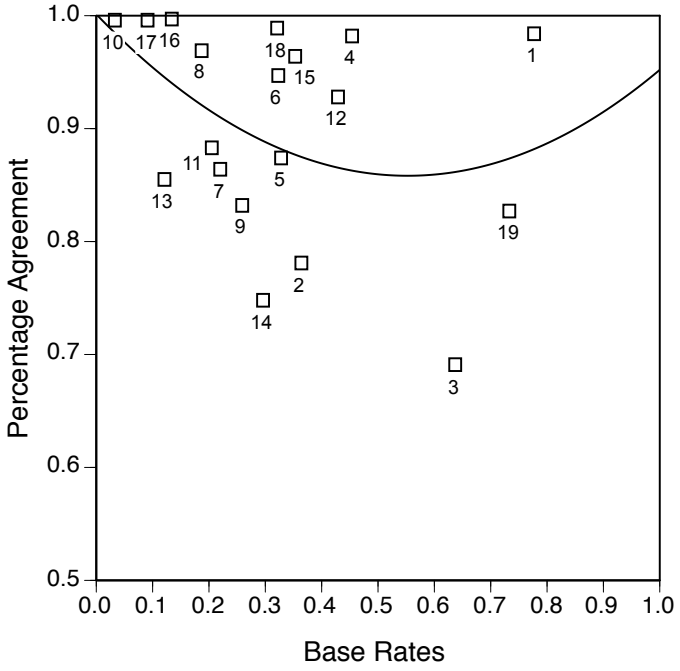

Figure 1. Base rates plotted against weighted average *percentage agreement* for 19 CBCA criteria.

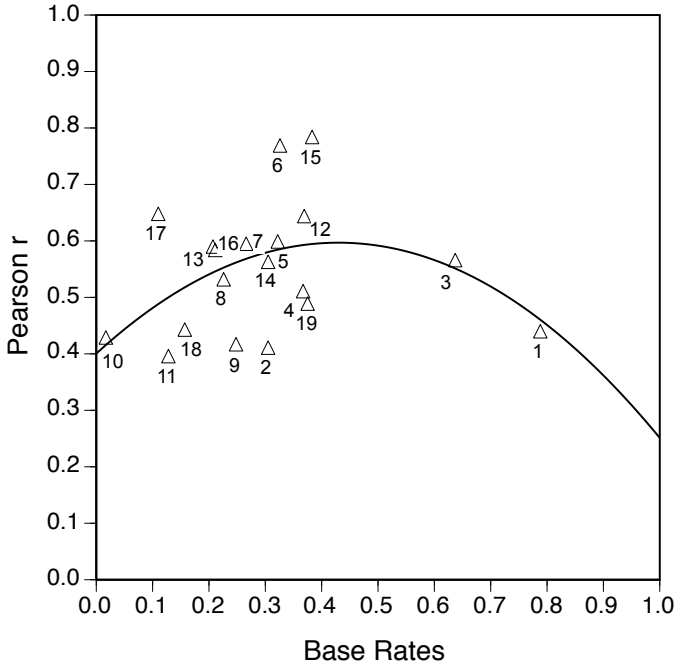

Figure 2. Base rates plotted against weighted average *Pearson r* for 19 CBCA criteria.

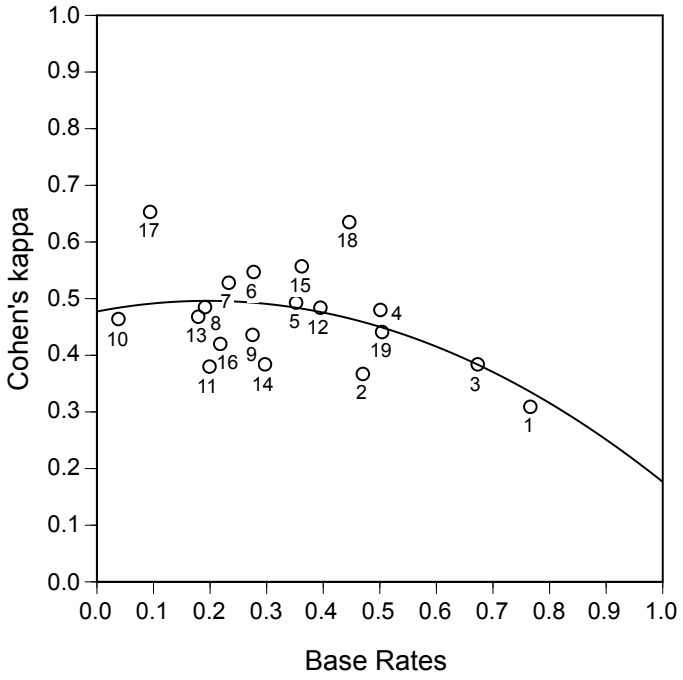

Figure 3. Base rates plotted against weighted average *kappa* for 19 CBCA criteria.

Appendix A

Coding Decisions for Study and Sender Characteristics

Authors	Source	Paradigm	Design	Language	Mode	$N_{acc.}$	N_{female}	N_{male}	N_{total}	*M* Age	Liars' Status	Event Type	Val-ence	Moti-vation	Interview
Akehurst, Köhnken, & Höfer (2001)	journal	exp.	betw.	E	spoken	1	na	na	66	na	actor	part.	neutral	none	semi-struct.
Akehurst, Manton, & Quandte (2011)	journal	field	betw.	E	spoken	1	26	5	31	10.87	victim	sex. abuse	neg.	high	semi-struct.
Anson, Golding, & Gully (1993)	journal	field	betw.	E	spoken	1	13	10	23	8.00	victim	sex. abuse	neg.	high	na
Bensi, Gambetti, Nori, & Giusberti (2009)	journal	exp.	betw.	I	spoken	1	19	21	40	26.52	witness	watch video	neg.	low	free report
Blandón-Gitlin, Pezdek, Lindsay, & Hagen (2009; Exp. 2)	journal	quasi	within	E	spoken	2	37	14	51	19.20	actor	part.	neutral	none	free report
Blandón-Gitlin, Pezdek, Rogers, Brodie (2005)	journal	exp.	betw.	E	spoken	1	43	51	94	10.50	actor	part.	neutral	none	struct.
Bogaard, Meijer, & Vrij (2014)	journal	quasi	within	D	written	2	36	28	64	21.09	wit./act./vic./p.	sign. LE	neg.	low	free report
Boychuk (1991)	Diss.	field	betw.	E	spoken	1	60	15	75	na	victim	sex. abuse	neg.	high	semi-struct.
Bradford (2006)	Diss.	exp.	betw.	E	spoken	1	16	4	20	19.90	actor	trivial LE	neg.	none	free report
Buck, Warren, Betman, & Brigham (2002)	journal	field	betw.	E	spoken	1	80	24	104	6.45	victim	sex. abuse	neg.	high	semi-struct.
Caso, Vrij, Mann, & De Leo (2006)	journal	exp.	within	I	spoken	2	na	na	64	na	actor	part.	neutral	none	struct.
Chang (2008)	Diss.	field	betw.	E	written	1	31	38	69	na	na	crime	neg.	high	free report

Connolly & Lavoie (2009)	ms. subm.	exp.	betw.	E	spoken	1	22	18	40	7.43	actor	part.	neutral	low	semi-struct.
Craig, Scheibe, Raskin, Kircher, & Dodd (1999)	journal	field	betw.	E	spoken	1	37	11	48	8.90	victim	sex. abuse	neg.	high	struct.
Dana-Kirby (1997; Exp. 1)	Diss.	exp.	within	E	spoken	4	na	na	81	na	actor	trivial LE	neg./ pos.	low	struct.
Dana-Kirby (1997; Exp. 2)	Diss.	exp.	within	E	spoken	2	na	na	126	na	actor	trivial LE	neutral	low	semi-struct.
Dana-Kirby (1997; Exp. 3)	Diss.	field	betw.	E	spoken	1	na	na	22	na	perpetr.	crime	neg.	high	na
Dukala, Sporer & Polczyk (in prep.)	unpubl. ms.	exp.	betw.	P	spoken	1	70	10	80	71.21	witness	watch video	neg.	low	struct. & CI
Dunbar, Harvell, Jensen, Burgoon, & Kelley (2012; Exp.1)	present.	exp.	within	E	spoken	2	90	104	194	28.51	actor	sex. ab./ sign. LE	neg./ pos.	med.	struct.
Dunbar, Harvell, Jensen, Burgoon, & Kelley (2012; Exp.2)	present.	exp.	betw.	E	spoken	1	52	32	84	22.76	perpetr.	part.	neg.	med.	struct.
Eggers (2002)	Thesis	exp.	betw.	G	spoken	1	34	26	60	na	wit./ vic./p.	crime	neg.	na	semi-struct.
Gödert, Gamer, Rill, & Vossel (2005)	journal	exp.	betw.	G	spoken	1	68	0	68	26.00	perpetr.	mock crime	neg.	low	semi-struct.
Granhag & Strömwall (2002)	journal	exp.	betw.	S	spoken	1	14	10	24	na	witness	observe event	neg.	low	free report
Granhag, Strömwall, & Landström (2006)	journal	exp.	betw.	S	spoken	1	na	na	38	12.40	actor	part.	neutral	na	CI
Hänert (2007)	Diss.	exp.	betw.	G	spoken	1	31	43	74	5.40	witness	observe event	neutral	none	struct.
Heinze (1996) & Horstmann (1996)	Thesis	exp.	within	G	spoken	2	22	20	42	14.11	victim	sign. LE	neg.	none	semi-struct.

Herrmann & Jena (1995; Sender 1)	Thesis	quasi	within	G	spoken	30	0	1	1	13.00	wit./act. /vic.	sign. LE	neg./ pos.	none	semi-struct.
Herrmann & Jena (1995; Sender 2)	Thesis	quasi	within	G	spoken	30	1	0	1	28.00	wit./act. /vic.	sign. LE	neg./ pos.	none	semi-struct.
Hettler (2005) & Maier (2007)	Thesis	quasi	betw.	G	spoken	1	14	26	40	27.75	wit./act.	sign. LE	neg.	none	semi-struct.
Höfer (1995)	Diss.	exp.	within	G	spoken	2	27	29	56	28.11	witness	watch video	neg.	none	free report
Honts & Devitt (1993)	report	quasi	within	E	spoken	2	13	13	26	na	actor	trivial LE	neg.	med.	semi-struct.
Horowitz, Lamb, Esplin, Boychuk, Krispin, & Reiter-Lavery (1997)	journal	field	betw.	E	spoken	1	79	21	100	8.58	victim	sex. abuse	neg.	high	semi-struct.
Janka (2003)	Thesis	quasi	within	G	na	4	8	6	14	na	wit./vic.	sign. LE	neg.	low	semi-struct.
Joffe (1992; 2nd grade children)	Diss.	exp.	betw.	E	spoken	1	23	20	43	na	actor	part.	neutral	none	semi-struct.
Joffe (1992; 4th grade children)	Diss.	exp.	betw.	E	spoken	1	27	23	50	na	actor	part.	neutral	none	semi-struct.
Krahe & Kundrotas (1992)	journal	field	betw.	G	spoken	1	30	0	30	na	victim	sex. abuse	neg.	high	na
Krahe, Reimer, & Scheinberger (1995)	present.	field	betw.	G	spoken	1	30	0	30	na	victim	sex. abuse	neg.	high	na
Lamb, Sternberg, Esplin, Hershkowitz, Orbach, & Hovav (1997b)	journal	field	betw.	H	spoken	1	70	28	98	8.72	victim	sex. abuse	neg.	high	na
Lamers-Winkelmann, Buffing, & van der Zanden (1992)	present.	field	betw.	D	spoken	1	75	28	103	5.90	victim	sex. abuse	neg.	high	na
Leal, Vrij, Warmelink, Vernham, & Fisher (2013; Exp. 1)	online first	quasi	betw.	E	spoken	1	14	26	40	35.61	victim	trivial LE	neg.	none	free report

Leal, Vrij, Warmelink, Vernham, & Fisher (2013; Exp. 2)	online first	quasi	betw.	E	spoken	1	61	22	83	25.31	victim	trivial LE	neg.	none	free report
Lee, Klaver, & Hart (2008)	journal	exp.	within	E	spoken	2	0	45	45	32.91	perpetr.	crime	neg.	none	free report
Manzanero, Recio, Alemany, Vallet, Aróztegui, & Sporer (2014)	unpubl. ms.	quasi	betw.	Sp.	spoken	1	13	16	29	32.17	victim	sign. LE	neg.	low	struct.
Mazzoni & Ambrosio (2002)	na	exp.	within	I	spoken	2	na	na	30	7.00	victim	sign. LE	neg.	none	semi-struct.
Merckelbach (2004; Exp.2)	journal	quasi	within	na	written	2	38	0	38	19.50	victim	sign. LE	neg.	none	free report
Metzger (1996)	Thesis	exp.	betw.	E	spoken	1	na	na	66	na	actor	part.	neutral	none	struct.
Naumann (2005)	Thesis	quasi	betw.	G	spoken	1	10	0	10	26.30	actor	sign. LE	neg./ pos.	low	semi-struct.
Niehaus (2001)	Diss.	quasi	betw.	G	spoken	1	41	39	80	8.75	victim	sign. LE	neg.	low	struct.
Peace & Porter (2011)	journal	quasi	within	E	written	2	233	58	291	19.64	victim	sign. LE	neg.	low	free report
Petersen (1997; Seminar Group)	Thesis	exp.	na	G	na	na	na	na	na	na	victim	sign. LE	neg.	na	free report
Petersen (1997; Training Group)	Thesis	exp.	na	G	na	na	na	na	na	na	victim	sign. LE	neg.	na	free report
Porter, Peace, & Emmett (2007)	journal	exp.	within	E	written	2	94	32	126	19.86	victim	sign. LE	neg.	none	free report
Porter & Yuille (1996)	journal	exp.	betw.	E	spoken	1	44	16	60	na	perpetr.	mock crime	neg.	low	semi-struct.
Porter, Yuille, & Lehman (1999)	journal	exp.	within	E	spoken	2	na	na	75	19.20	victim	sign. LE	neg.	low	struct.

Roma, Martini, Sabatello, Tatrelli, & Ferracuti (2011)	journal	field	betw.	I	spoken	1	86	23	109	8.58	victim	sex. abuse	neg.	high	semi-struct.
Ruby & Brigham (1998; Black speakers)	journal	quasi	within	E	spoken	2	3	3	6	na	wit./act. /vic./p.	sign. LE	neg.	none	free report
Ruby & Brigham (1998; White speakers)	journal	quasi	within	E	spoken	2	3	3	6	na	wit./act. /vic./p.	sign. LE	neg.	none	free report
Rutta (2001)	Thesis	exp.	within	G	typed	2	na	na	16	na	wit./act. /vic./p.	sign. LE	neg.	low	semi-struct.
Saacke (1995)	Thesis	quasi	within	G	spoken	1	47	41	88	8.50	wit./act. /vic.	sign. LE	neg.	low	semi-struct.
Sallmon & Volbert (2013)	present.	quasi	betw.	G	spoken	1	36	33	69	25.80	act./vic.	sign. LE	neg.	low	free report
Santtila, Roppola, Runtti, & Niemi (2000)	journal	quasi	within	S	spoken	2	35	33	68	10.50	victim	sign. LE	neg.	none	semi-struct.
Saykaly, Talwar, Lindsay, Bala, & Lee (2013)	journal	exp.	within	E	spoken	2	38	40	78	7.58	actor	part.	pos.	none	struct.
Scheinberger (1993)	Thesis	quasi	betw.	G	spoken	1	30	0	30	30.00	actor	sign. LE	pos.	low	free report
Schellemann-Offermanns & Merckelbach (2010)	journal	quasi	within	D	typed	2	30	30	60	na	victim	sign. LE	neg.	none	free report
Sporer (1997a)	journal	exp.	betw.	G	spoken	1	40	40	80	25.00	actor	sign. LE	neutral	none	free report
Sporer (1997b)	chapter	exp.	betw.	G	written	1	100	100	200	21.00	actor	sign. LE	neg./ pos.	none	free report
Sporer & Bursch (2003)	unpubl. ms.	exp.	betw.	G	written	2	100	100	200	21.00	actor	sign. LE	neg./ pos.	none	free report
Steller, Wellershaus, & Wolf (1992)	journal	quasi	within	G	spoken	2	47	41	88	8.50	wit./act. /vic.	sign. LE	neg.	low	semi-struct.

Strömwall, Bengtsson, Leander, & Granhag (2004)	journal	exp.	betw.	S	spoken	1	na	na	41	11.85	actor	part.	neutral	none	CI
Volbert & Lau (2013; Exp.2)	journal	quasi	betw.	G	spoken	1	36	0	36	46.75	actor	sign. LE	neg./ pos.	none	struct.
Vrij, Akehurst, Soukara, & Bull (2002)	journal	exp.	betw.	E	spoken	1	na	na	130	14.09	wit./act.	part.	neutral	low	struct.
Vrij, Akehurst, Soukara, & Bull (2004)	journal	exp.	betw.	E	spoken	1	na	na	91	na	wit./act.	observe event /part.	neutral	low	free report
Vrij, Edward, & Bull (2001)	journal	exp.	within	E	spoken	2	76	10	86	25.32	witness	watch video	neg.	low	struct.
Vrij, Edward, Roberts, & Bull (2000)	journal	exp.	betw.	E	spoken	1	53	20	73	28.89	witness	watch video	neg.	low	struct.
Vrij & Heaven (1999)	journal	exp.	within	E	spoken	2	16	24	40	23.00	witness	watch video	neg.	none	struct.
Vrij, Kneller, & Mann (2000)	journal	exp.	betw.	E	spoken	1	22	8	30	26.30	witness	watch video	neg.	none	free report
Vrij, Mann, Fisher, Leal, Milne, & Bull (2008)	journal	exp.	betw.	E	spoken	1	40	40	80	20.88	perpetr.	mock crime	neg.	low	free report
Vrij, Mann, Kristen, & Fisher (2007)	journal	exp.	betw.	E	spoken	1	50	70	120	22.07	perpetr.	mock crime	neg.	med.	na
Wehner (2006)	chapter	exp.	betw.	G	spoken	1	30	26	56	24.60	act./p.	mock crime	neg.	low	CI
Willén & Strömwall (2012)	journal	quasi	within	S	spoken	2	9	21	30	34.20	perpetr.	crime	neg.	none	semi-struct.
Wrege (2004)	Thesis	quasi	within	G	spoken	2	10	6	16	27.00	wit./act. /vic.	sign. LE	neg.	low	semi-struct.
Zaparniuk, Yuille, & Taylor (1995)	journal	exp.	betw.	E	spoken	1	24	16	40	19.80	witness	watch video	neg.	none	semi-struct.

Notes. Exp. = Experiment; journal = article published in journal; Diss. = Dissertation; ms. subm. = manuscript submitted for review; present. = paper or poster presented at conference, or summary of presented paper in edited conference proceedings; unpubl. ms. = unpublished manuscript; report = published research report; online first = online first version of to be printed journal article; chapter = book chapter; quasi = quasi-experiment; betw. = between-participants design (lie or truth); within = within-participants design (lie and truth); E = English, G = German, D = Dutch, H = Hebrew, I = Italian, P = Polish, S = Swedish; Sp = Spanish; na = not available; written = handwritten; $N_{acc.}$ = number of accounts per person; *M* = Mean; wit. = witness; act. = actor; vic. = victim; p./perpetr. = perpetrator; sex. ab. = sexual abuse; sign. = significant; LE = life event; part. = participate; neg. = negative; pos. = positive; med. = medium; struct. = structured; free report = free report only; CI = Cognitive Interview.

Appendix B

Coding Decisions for Rater, Rating Process, and Training Characteristics

Authors	N_{rater}	Occupation	Mode	Scale	Dura-tion	Litera-ture	Defini-tions	Exam-ples	Lec-ture	Practice	Sum	% rated	$N_{adj.}$
Akehurst et al. (2001)	3	na	transcript	other Likert	na	no	yes	yes	yes	2 comp.	5	100.0	66
Akehurst et al. (2011)	2	psych.	transcript	1 to 5	na	no	yes	yes	yes	2 comp.	5	100.0	31
Anson et al. (1993)	4	na	audiovisual	0/1	na	no	yes	no	yes	2 comp.	4	100.0	23
Bensi et al. (2009)	2	Ph.D.	transcript	0-1-2	6.00	yes	yes	no	yes	2 comp.	5	100.0	40
Blandón-Gitlin et al. (2009; Exp. 2)	2	grad. students	transcript	0-1-2	na	yes	yes	no	yes	2 comp.	5	100.0	51
Blandón-Gitlin et al. (2005)	2	psych.	transcript	0/1	na	yes	yes	yes	yes	2 comp.	6	100.0	48
Bogaard et al. (2014)	2	na	transcript	0-1-2	1.50	yes	yes	yes	no	2 comp.	5	100.0	64
Boychuk (1991)	3	psych./ grad. students	transcript	0-1-2 to 0/1	201.33	yes	yes	no	yes	2 comp.	5	100.0	75
Bradford (2006)	2	students	transcript	1 to 7	na	no	yes	no	no	2 comp.	3	100.0	20
Buck et al. (2002)	2	grad. students	transcript	0/1	na	yes	yes	no	no	2 comp.	4	100.0	104
Caso et al. (2006)	2	na	transcript	frequencies	na	yes	yes	yes	yes	2 comp.	6	20.0	10
Chang (2008)	4	grad. students	transcript	0/1	35.00	no	yes	no	yes	2 comp.	4	100.0	69

Connolly & Lavoie (2009)	2	grad. students	transcript	0-1-2	16.00	yes	yes	yes	yes	2 comp.	6	100.0	40
Craig et al. (1999)	4	students	transcript	0/1	2.00	no	yes	no	yes	2 comp.	4	100.0	48
Dana-Kirby (1997; Exp. 1)	2	students	audiovisual	0/1	1.00	no	yes	yes	yes	1 comp.	4	100.0	150
Dana-Kirby (1997; Exp. 2)	4	students	audiovisual	0-1-2	1.00	no	yes	yes	yes	1 comp.	4	100.0	126
Dana-Kirby (1997; Exp. 3)	2	students	transcript	0/1	na	no	yes	yes	yes	1 comp.	4	100.0	22
Dukala et al. (in prep.)	4	grad. students	transcript	0-1-2	10.00	yes	yes	yes	yes	2 comp.	6	50.0	40
Dunbar et al. (2012; Exp.1)	8	students	audiovisual	1 to 7	1.50	yes	yes	yes	yes	2 comp.	6	100.0	194
Dunbar et al. (2012; Exp.1)	8	students	audio-visual	1 to 7	1.50	yes	yes	yes	yes	2 comp.	6	100.0	84
Eggers (2002)	6	grad. students	transcript	0 to 4	na	yes	yes	yes	yes	2 comp.	6	100.0	60
Gödert et al. (2005)	3	students	transcript	other Likert	18.00	yes	yes	yes	yes	2 comp.	6	100.0	68
Granhag & Strömwall (2002)	2	na	na	frequencies	na	na	yes	yes	na	na	2	20.0	14
Granhag et al. (2006)	2	na	transcript	2 types	40.00	yes	no	no	no	2 comp.	3	20.0	16
Hänert (2007)	2	grad. students	transcript	other Likert	na	yes	yes	yes	yes	2 comp.	6	16.2	12
Heinze (1996) & Horstmann (1996)	3	psych./ grad. students	transcript	2 types	na	yes	yes	yes	yes	2 comp.	6	100.0	42
Herrmann & Jena (1995; Sender 1)	3	na	transcript	3 types	na	yes	yes	yes	yes	2 comp.	6	100.0	15

Herrmann & Jena (1995; Sender 2)	2	na	transcript	3 types	na	yes	yes	yes	yes	2 comp.	6	100.0	15
Hettler (2005) & Maier (2007)	3	grad. students	transcript	2 types	na	yes	yes	yes	yes	2 comp.	6	28.3	80
Höfer (1995)	3	students	audiovisual	4 or more types	80.00	yes	yes	yes	yes	2 comp.	6	100.0	56
Honts & Devitt (1993)	2	Ph.D.	transcript	0-1-2	na	yes	yes	yes	no	no	3	100.0	26
Horowitz et al. (1997)	3	grad. students	transcript	0/1	na	yes	yes	yes	yes	2 comp.	6	100.0	100
Janka (2003)	2	grad. students	transcript	3 types	na	yes	yes	yes	no	no	3	100.0	28
Joffe (1992; 2nd grade children)	4	students	transcript	other Likert	20.00	no	yes	yes	yes	2 comp.	5	100.0	43
Joffe (1992; 4th grade children)	4	students	transcript	other Likert	20.00	no	yes	yes	yes	2 comp.	5	100.0	50
Krahe & Kundrotas (1992)	31	police	transcript	0-1-2	na	no	yes	no	no	no	1	100.0	30
Krahe et al. (1995)	na	psych./ (grad.) students	transcript	other Likert	na	yes	no	no	yes	no	2	100.0	30
Lamb et al. (1997b)	3	na	transcript	0/1	na	yes	yes	no	no	2 comp.	4	100.0	89
Lamers-Winkelmann et al. (1992)	2	na	transcript	0/1	16.00	yes	yes	no	yes	2 comp.	5	100.0	103
Leal et al. (2013; Exp. 1)	2	na	transcript	2 types	na	no	yes	no	yes	no	2	100.0	40
Leal et al. (2013; Exp. 2)	2	na	transcript	2 types	na	yes	yes	no	yes	2 comp.	5	100.0	83
Lee et al. (2008)	2	na	transcript	0/1	na	no	yes	no	yes	no	2	83.0	75

Manzanero et al. (2014)	2	psych.	transcript	frequencies	na	no	yes	no	yes	no	2	100.0	29
Mazzoni & Ambrosio (2002)	2	na	transcript	0-1-2	na	no	no	no	no	no	0	100.0	30
Merckelbach (2004; Exp.2)	2	psych.	transcript	other Likert	na	yes	yes	no	no	no	2	100.0	38
Metzger (1996)	3	students	transcript	other Likert	na	yes	yes	yes	yes	2 comp.	6	100.0	93
Naumann (2005)	2	psych.	transcript	2 types	3.00	yes	yes	yes	yes	no	4	100.0	10
Niehaus (2001)	9	grad. students	transcript	2 types	na	yes	yes	yes	yes	2 comp.	6	100.0	80
Peace & Porter (2011)	2	students	transcript	2 types	20.00	no	yes	no	yes	no	2	20.0	58
Petersen (1997; Seminar Group)	6	grad. students	transcript	0 to 4	7.50	no	yes	yes	yes	no	3	100.0	4
Petersen (1997; Training Group)	6	grad. students	transcript	0 to 4	7.50	no	yes	yes	yes	no	3	100.0	4
Porter et al. (2007)	2	na	transcript	1 to 7	na	no	no	no	no	no	0	20.0	48
Porter & Yuille (1996)	2	na	transcript	2 types	24.00	no	yes	no	yes	no	2	13.3	8
Porter et al. (1999)	3	na	transcript	2 types	na	no	yes	no	yes	no	2	15.0	25
Roma (2011)	2	psych.	transcript	0/1	7.50	yes	yes	no	yes	1 comp.	4	100.0	109
Ruby & Brigham (1998; Black speakers)	119	students	transcript	0-1-2	0.75	no	yes	yes	no	1 comp.	3	100.0	6
Ruby & Brigham (1998; White speakers)	119	students	transcript	0-1-2	0.75	no	yes	yes	no	1 comp.	3	100.0	6

Rutta (2001)	2	psych.	transcript	2 types	na	yes	yes	yes	no	1 comp.	4	15.6	5
Saacke (1995)	2	grad. students	transcript	3 types	na	yes	yes	no	yes	1 comp.	4	100.0	88
Sallmon & Volbert (2013)	2	grad. students	transcript	frequencies	na	yes	yes	yes	yes	2 comp.	6	42.0	29
Santtila et al. (2000)	2	na	transcript	2 types	na	yes	yes	no	no	2 comp.	4	29.4	20
Saykaly et al. (2013)	2	grad. students	transcript	frequencies	na	no	yes	no	yes	1 comp.	3	25.0	59
Scheinberger (1993)	2	grad. students	transcript	0-1-2	na	no	yes	yes	yes	2 comp.	5	100.0	30
Schellemann-Offermanns & Merckelbach (2010)	2	students	transcript	other Likert	na	yes	yes	no	no	no	2	100.0	60
Sporer (1997a)	2	students	transcript	0-1-2	5.00	yes	yes	yes	yes	2 comp.	6	100.0	80
Sporer (1997b)	2	grad. students	transcript	0-1-2	na	yes	yes	yes	no	2 comp.	5	100.0	200
Sporer & Bursch (2003)	2	grad. students	transcript	1 to 7	10.00	yes	yes	no	no	2 comp.	4	100.0	200
Steller et al. (1992)	3	grad. students	transcript	other Likert	90.00	no	yes	yes	yes	1 comp.	4	100.0	88
Strömwall et al. (2004)	2	na	transcript	0-1-2	na	yes	yes	no	yes	2 comp.	5	20.0	18
Volbert & Lau (2013; Exp. 2)	2	psych.	transcript	3 types	na	yes	yes	yes	no	1 comp.	4	16.7	9
Vrij et al. (2002)	2	na	transcript	2 types	na	yes	yes	yes	yes	2 comp.	6	100.0	95
Vrij et al. (2004)	2	na	transcript	4 or more types	na	yes	yes	yes	yes	2 comp.	6	100.0	91

Vrij et al. (2001)	2	na	transcript	2 types	na	yes	yes	no	yes	2 comp.	5	100.0	86
Vrij; Edward, et al. (2000)	2	na	transcript	2 types	na	yes	yes	no	no	2 comp.	4	100.0	73
Vrij & Heaven (1999)	2	na	transcript	2 types	na	no	no	no	no	no	0	100.0	40
Vrij, Kneller, et al. (2000)	2	na	transcript	0/1	na	yes	yes	no	no	2 comp.	4	100.0	30
Vrij et al. (2008)	2	na	transcript	frequencies	na	no	yes	no	yes	no	2	50.0	40
Vrij et al. (2007)	2	na	transcript	frequencies	na	yes	yes	yes	yes	2 comp.	6	50.0	60
Wehner (2006)	2	grad. students	audiotape	0 to 4	40.00	yes	yes	yes	yes	2 comp.	6	100.0	56
Willén & Strömwall (2012)	2	students	transcript	0 to 4	6.00	yes	yes	yes	no	2 comp.	5	25.0	15
Wrege (2004)	2	grad. students	transcript	3 types	na	no	yes	yes	no	1 comp.	3	12.5	8
Zaparniuk (1995)	3	na	audiotape	0/1	na	no	yes	no	yes	1 comp.	3	50.0	20

Notes. Exp. = Experiment; N_{rater} = number of raters; na = not available; Mode = Mode of presentation; other Likert = other Likert scale (e.g., 1 to 4); Duration = average training duration per rater (in hours); comp. = number of the following components that are fulfilled: a) practice, b) feedback/discussion, c) homework tasks; Sum = sum of five training variables (literature (0/1), definitions (0/1), examples (0/1), lecture (0/1), and practice (0-1-2)): minimum = 0, maximum = 6; % rated = percentage of accounts that were independently coded by at least two raters; $N_{adj.}$ = number of adjusted statements--adjustment: total number of statements multiplied by the percentage of statements that were coded by at least two raters.

Appendix C

List of Studies Reporting on Different Reliability Indices for CBCA Criteria

Authors	Pearson's *r* / Spearman's *rho* / *phi*	(Absolute) *PCA*	Cohen's *kappa*	Weighted *kappa* / *ICC*	Maxwell's *RE*
Akehurst et al. (2001)	CBCA01-10, 12-16	CBCA01-10, 12-16	#	#	#
Akehurst et al. (2011)	CBCA01-19, sum score	#	#	#	#
Anson et al. (1993)	#	CBCA01-19, sum score	CBCA01-15, 17-19, sum score	#	CBCA01-15, 17-19, sum score
Bensi et al. (2009)	CBCA01, 02, 05, 08, 09, 13, 14, 15, 16	#	#	#	#
Blandón-Gitlin et al. (2009; Exp. 2)	CBCA01-07, 09, 12-16	CBCA01-07, 09, 12-16	CBCA01-07, 09, 12-16	#	#
Blandón-Gitlin et al. (2005)	#	#	CBCA01-15	#	#
Bogaard et al. (2014)	#	#	CBCA01, 03-06, 08, 09, sum score	#	#
Boychuk (1991)	#	#	range (CBCA01-19)	#	#
Bradford (2006)	CBCA01-09, 11, 12, 14, 15, 17, 18, sum score	#	#	#	#
Buck et al. (2002)	#	#	#	#	CBCA01-19
Caso et al. (2006)	#	CBCA03, 05, 06, 07, 12, 14, 15	#	#	#

Chang (2008)	#	#	#	*ICC*: range (CBCA06)[h]	#
Connolly & Lavoie (2009)	CBCA01-07, 09, 11, 12, 14, 15, 17	CBCA01-07, 09-17	CBCA01-07, 09-17	w. *kappa*: CBCA01-07, 09-17	CBCA01-07, 09-17
Craig et al. (1999)	#	range (CBCA01-15), sum score	range (CBCA01-15)	#	#
Dana-Kirby (1997; Exp. 1)	#	CBCA01-14, sum score	CBCA01-14, sum score	#	CBCA01-14, sum score
Dana-Kirby (1997; Exp. 2)	#	CBCA01-14, sum score	CBCA01-14, sum score	#	CBCA01-14, sum score
Dana-Kirby (1997; Exp. 3)	#	CBCA01-14, sum score	CBCA01-14, sum score	#	CBCA01-14, sum score
Dukala et al. (in prep.)	CBCA01-05, 08-17; *rho*: CBCA01-05, 08-17	CBCA01-05, 08-17	CBCA01-05, 08-17	*ICC*: CBCA01-05, 08-17[e]	#
Dunbar et al. (2012; Exp.1)	#	#	#	*ICC*: CBCA03, 04, 06, 14[f]	#
Dunbar et al. (2012; Exp.1)	#	#	#	*ICC*: CBCA03, 04, 06, 14[f]	#
Eggers (2002)	CBCA01-09, 11-19	CBCA01-19	CBCA01-18	#	CBCA01-19
Gödert et al. (2005)	CBCA01-10, 12, 13, 14, 15, 16, 18	CBCA01-18[a]	#	w. *kappa* and *ICC*: CBCA01-18[dg]	#
Granhag & Strömwall (2002)	CBCA03	CBCA03	#	#	#
Granhag et al. (2006)	CBCA01, 02, 03, 05, 06, 08, 12, 14, 15, 16	CBCA01, 02, 03, 05, 06, 08, 12, 14, 15, 16	CBCA01, 02, 03, 05, 06, 08, 12, 14, 15, 16	#	#
Hänert (2007)	CBCA01-09, 11, 13-16	#	#	#	#

Heinze (1996) & Horstmann (1996)	CBCA04-18; *phi*: CBCA01, 02, 03	#	#	#	#
Herrmann & Jena (1995; Sender 1)	CBCA02, 03, 04, 06-10, 12-15	#	#	#	#
Herrmann & Jena (1995; Sender 2)	CBCA02, 03, 04, 06, 08, 12-15	#	#	#	#
Hettler (2005) & Maier (2007)	sum score (CBCA01-09, 11-18)	#	#	#	#
Höfer (1995)	CBCA02, 03, 07, 12, 13, 14; *phi*: CBCA01, 10, 15, 16	#	#	#	#
Honts & Devitt (1993)	range (CBCA01-09, 12-17, sum score)	#	#	#	#
Horowitz et al. (1997)	#	CBCA01-19	CBCA01-19	#	CBCA01-19
Janka (2003)	CBCA04-09, 11-18	CBCA01-15, 17, 18	#	w. *kappa*: CBCA01-18	#
Joffe (1992; 2nd grade children)	#	#	#	*ICC*: CBCA01-03, 05-09, 11-15, sum score[g]	#
Joffe (1992; 4th grade children)	#	#	#	*ICC*: CBCA01-03, 05-09, 11-16, sum score[g]	#
Krahe & Kundrotas (1992)	#	#	CBCA01-19	#	#
Krahe et al. (1995)	#	#	#	#	#
Lamers-Winkelmann et al. (1992)	#	CBCA01-19	CBCA01-19	#	#
Lamb et al. (1997b)	#	sum score (CBCA01-14)	#	#	#

Leal et al. (2013; Exp. 1)	#	#	#	*ICC*: sum score (CBCA01-09, 11-12, 14-17)[h]	#
Leal et al. (2013; Exp. 2)	#	#	#	*ICC*: sum score (CBCA01-09, 11-12, 14-17)[h]	#
Lee et al. (2008)	#	#	CBCA01-17, 19	#	#
Manzanero et al. (2014)	#	CBCA01, 02 (range: 04-16, 19)	#	#	#
Mazzoni & Ambrosio (2002)	*rho*: sum score (CBCA01-19)	#	#	#	#
Merckelbach (2004; Exp.2)	range (CBCA01, 03-06, 08, 09, 12, 13)	#	#	#	#
Metzger (1996)	CBCA01-09, 11-16	#	#	#	#
Naumann (2005)	#	#	sum score (CBCA01-15)	#	#
Niehaus (2001)	CBCA01-15, 17, 18	CBCA01-15, 17, 18, sum score	CBCA01-15, 17, 18, sum score	w. *kappa*: CBCA01-15, 17, 18, sum score	#
Peace & Porter (2011)	CBCA01, 03	#	#	#	#
Petersen (1997; Seminar Group)	CBCA01-19	CBCA01-19[a]	CBCA01-19	#	CBCA01-19
Petersen (1997; Training Group)	CBCA01-19	CBCA01-19[a]	CBCA01-19	#	CBCA01-19
Porter et al. (2007)	range (CBCA01, 12, 13)	#	#	#	#
Porter & Yuille (1996)	CBCA12 (CBCA01, 02, 03, 07, 08, 09, 11, 14, 15: min. rel.)	#	#	#	#

Porter et al. (1999)	CBCA04 (CBCA01, 03, 15: min. rel.)	#	#	#	#
Roma (2011)	#	#	#	#	CBCA01-14
Ruby & Brigham (1998; Black speakers)	#	CBCA01-09, 11, 12, 14-17	#	#	#
Ruby & Brigham (1998; White speakers)	#	CBCA01-09, 11, 12, 14-17	#	#	#
Rutta (2001)	CBCA01, 03, 04, 06, 09, 12, 14, 15, 17	#	#	#	#
Saacke (1995)	#	CBCA01-18	CBCA02-18	#	#
Sallmon & Volbert (2013)	#	#	#	*ICC*: CBCA03[i]	#
Santtila et al. (2000)	*rho*: CBCA06, 09 (range: CBCA01-05, 07, 08, 10-14)	#	#	#	#
Saykaly et al. (2013)	#	#	CBCA14	#	#
Scheinberger (1993)	CBCA03-09, 11-15, 18	#	#	#	#
Schellemann-Offermanns & Merckelbach (2010)	range (CBCA01, 03-06, 08, 09, 12, 13)	#	#	#	#
Sporer (1997a)	Pearson's *r* and *phi*: CBCA01-09, 12, 13[b]	CBCA01-09, 12, 13[b]	CBCA01-09, 12, 13	#	CBCA01-09,12,13
Sporer (1997b)	Pearson's *r* and *phi*: CBCA01-13	CBCA01-13[b]	CBCA01-13	#	CBCA01-13
Sporer & Bursch (2003)	Pearson's *r* and *rho*: CBCA01-13	#	#	*ICC*: CBCA01-13[e]	#

Steller et al. (1992)	#	CBCA01-18	#	#	#
Strömwall et al. (2004)	CBCA01, 03-08, 10-16, 19	CBCA01-19	CBCA01, 03-08, 10-16, 19	#	#
Volbert & Lau (2013; Exp. 2)	CBCA05-10, 12-19	#	#	w. *kappa*: CBCA01-19	#
Vrij et al. (2002)	CBCA01-08, 12-16	#	#	#	#
Vrij et al. (2004)	CBCA01-09, 11-16	#	#	#	#
Vrij et al. (2001)	CBCA01-06, 08, 13-16	CBCA01-06, 08, 13-16[c]	CBCA01-06, 08, 13-16	#	#
Vrij; Edward, et al. (2000)	CBCA01-06, 08, 09, 12-16, 18	#	#	#	#
Vrij & Heaven (1999)	#	CBCA03, 16	#	#	#
Vrij, Kneller, et al. (2000)	#	CBCA02 (range: CBCA01, 03, 04, 05, 08, 13-16)	#	#	#
Vrij et al. (2008)	CBCA04	#	#	#	#
Vrij et al. (2007)	CBCA01-09, 11-18	#	#	#	#
Wehner (2006)	Pearson's *r* and *rho*: CBCA01-04, 07, 08, 09, 12, 14, 15, sum score	CBCA01-04, 07, 08, 09, 12, 14, 15, sumcores	#	#	#
Willén & Strömwall (2012)	CBCA01-09, 11-17, 19	#	#	w. *kappa*: CBCA01-09, 11-17, 19	#
Wrege (2004)	CBCA06, 12, 13, 15; *rho*: CBCA01-08, 13,	#	#	#	#

	14, 15, 17				
Zaparniuk (1995)	#	CBCA01-03, 05-18	#	#	#

Notes. min. rel. = minimum reliability; # = no index reported; whenever several correlation coefficients (Pearson's *r*, Spearman's *rho*, *phi*) were reported, Pearson's *r* was chosen for analysis when possible; [a] = in addition to absolute *PCA*, extended *PCA* was reported; [b] = two values were reported for each index: Pearson's *r* and *PCA* for 0-1-2 rating, and *phi* and *PCA* for dichotomous rating (for analysis, *PCA* for 0-1-2 rating was chosen); [c] = two values were reported: before and after data transformation (for analysis, *PCA* after data transformation was chosen); w. *kappa* = weighted *kappa*; [d] = when weighted *kappa* and *ICC* were reported, *ICC* was chosen for analysis; [e] = when $ICC_{average}$ and ICC_{single} were reported, $ICC_{average}$ was chosen for analysis; [f] = one-way random version of ICC; [g] = generalizability coefficient; [h] = type of *ICC* not specified; [i] = two values reported: consistency and absolute agreement (absolute agreement chosen for analysis).

DISCUSSION

This dissertation reported two meta-analyses on the detection of deception with linguistic and verbal content cues. Aside from other approaches to detect deception (e.g., nonverbal, paraverbal, psychophysiological), former research (e.g., Amado, Arce, & Fariña, 2015; Bond & DePaulo, 2006; DePaulo, Lindsay, Malone, Muhlenbruck, Charlton, & Cooper, 2003; Hauch, Sporer, Michael, & Meissner, 2014) strongly suggested that the approach of analyzing the content of a statement is probably most promising to outperform the average 54% discrimination accuracy (Aamodt & Custer, 2006; Bond & DePaulo, 2006). Therefore, the aim of this dissertation was to further the knowledge on linguistic and verbal content cues to deception by means of meta-analyses.

To this end, the *first meta-analysis* examined if and to what extent linguistic cues assessed by computer programs can distinguish deceptive from true statements. In other words, the focus of this enterprise was the validity of linguistic cues to deception. The *second meta-analysis* deals with *Criteria-based Content Analysis* criteria (CBCA, Steller & Köhnken, 1989), or credibility criteria as a special kind of verbal content cues rated by human raters. More specifically, the amount of agreement that can be reached by several independent evaluators--that is, their inter-rater reliability--was systematically investigated. In the following, main and most important findings of both meta-analyses are discussed. A more detailed discussion can be found in the corresponding discussion sections of each meta-analysis.

Linguistic cues to deception assessed by computer programs

In the first meta-analysis, after an exhaustive literature search in various interdisciplinary research areas, 44 hypothesis tests were included. Furthermore,

from a wealth of different linguistic cues (> 200), 40 were selected, operationally defined and allocated to six principal research questions. Under each question, a direction of effect was hypothesized for every single linguistic marker from different theoretical perspectives. Moreover, several theoretically or methodologically important independent variables were supposed to be associated with effect sizes (e.g., event type and personal involvement, emotional valence, intensity of interaction, motivation of the liar, production mode, program type, publication status, experimental design).

The first research question “Do liars experience greater cognitive load?” was mainly supported in that compared to true statements, lies seem to be shorter and less elaborate than true statements. These findings provided support for the theoretical assumption of a working memory model to deception in that constructing a lie is cognitively more demanding for working memory capacity than telling the truth (Sporer, 2015; Sporer & Schwandt, 2006, 2007, based on Baddeley, 2000, 2006). More specifically, differences between liars and truth-tellers in word quantity were largest when negative emotional events were told to an interaction partner, but seem to be reversed in computer-mediated communication, or when analyzed from specific programs designed to detect deception.

The second research question was based on an impression management approach in that a liar’s profile or self-presentation is less convincing than a truthful one (DePaulo et. al, 2003). Thus, it was hypothesized that compared to truth-tellers, liars language is less certain and more vague as indexed with the use of more modal verbs or tentative constructions. In general, weak support was found for the opposite: Truth-tellers seem to slightly use more tentative words than liars. This effect occurred only in a within-experimental design, but is unrelated to other

independent variables. However, this finding could actually be seen as support for the self-presentational perspective, because truth-tellers do not require any special effort to appear credible or avoid insecurities (see Volbert & Steller, 2014), and therefore might express more words related to uncertainty.

The third complex of research questions dealt with emotional processes: Due to Ekman's emotional approach (1988, 2001) liars feel more negative emotions than truth-tellers, a transfer of these emotions on language was assumed. Indeed, liars expressed more negative emotional words, negations, and more emotional words overall than truth-tellers. Differences in negative emotions between liars and truth-tellers were most pronounced in settings when storytellers were highly motivated in providing negative emotional, self-experienced events to an interaction partner, and only occurred in studies using a within-participants design. No differences were found in the use of positive emotional words although from an autobiographical "fading affect bias" it was assumed that liars use fewer positive emotional words (Walker & Skowronski, 2009). These results suggested that it is necessary to differentiate the type of emotional valence (i.e., negative, positive, neutral) expressed in language.

The fourth research question focused on the psychological distance, or immediacy (e.g., Wiener & Mehrabian, 1968) expressed by manipulating the use of personal pronouns and active versus passive speech. The hypothesis that liars are less immediate and distance themselves more from events was partially supported. In general, liars used slightly fewer self-references and more other references than truth-tellers. This difference in self-references was most stressed in negative events told to an interacting person, and in published studies implementing a within-participants design. On the other side, liars expressed more other references in

unpublished studies with a between-participants design, where senders were asked to provide statements on neutral events or attitudes without an interaction partner. These results impressively showed a varying association of independent variables with specific linguistic cues to deception.

The fifth research question investigated the amount of details in statements. From a reality monitoring approach (Johnson & Raye, 1981; Sporer, 1997, 2004) it was assumed that liars would use less sensory and perceptual details than truth-tellers because the critical event is internally generated and not based on memory of an actual experience. Some support for this assumption was found for words expressing sensory-perceptual processes, especially hearing and quantifiers, but not for other indicators. These effects were strongest when senders were highly motivated, handwrite their accounts, the study was published, and implemented a within-participants design. Contrary to the hypothesis, liars tended to use more motion verbs, especially when statements were spoken or handwritten, or when the study was published. This result could be explained from a working memory perspective, in that liars might use more simple and less complex words, like motion verbs due to greater cognitive demands (see also Newman, Pennebaker, Berry, & Richards, 2003).

The sixth research question, “Do liars refer less often to cognitive processes?”, was derived from autobiographical memory research. Indeed, liars expressed fewer words related to their cognitions. Put differently, truth-tellers provided more cognitive words--possibly to generate cues for memory retrieval (e.g., Conway, 1990). Again, this effect was only present in studies implementing a within-participants design.

Besides these research questions, the results of miscellaneous linguistic cues to deception that could not be allocated to a specific research question or could not be predicted from a theory, were presented in the meta-analysis. However, the above findings altogether strongly suggested that a theoretical foundation and integration with derived hypotheses is of utmost relevance when investigating the outcome of linguistic cues to deception.

Setting the fact aside that most theories received at least some support, it is important to discuss some factors that limit the generalizability of these findings. It should be mentioned first that a meta-analysis always constitutes a weighted mean of mean results and left a large amount of variance within and between studies in the dark (e.g., Lipsey & Wilson, 2001). Therefore, moderator analyses could bring a more complex association between effect sizes and independent variables to light. Here, one or several moderator variables--such as the emotional valence of the event, the personal investment or motivation of the sender, or methodological factors (i.e., experimental design, publication status)--moderated almost all main effects. In other words, the specific context of a statement, or the research setup can already make a difference in the linguistic profile of liars and truth-tellers (see also Hancock and Woodworth, 2013). Another limitation is the general small effect size (according to Cohen, 1988), although it can be compared to effect sizes found in previous meta-analyses (e.g., DePaulo et al., 2003; Sporer & Schwandt, 2006, 2007). Put differently, the difference between liars' and truth-tellers' linguistic profile on average--and even when considering context variables--is quite small.

Furthermore, two probably important variables could not be investigated in this meta-analysis: Language and age of senders. As Newman et al. (2003) already pointed out, different languages and their specific grammatical rules (e.g., the use of

personal pronouns in Spanish vs. English) can lead to different linguistic profiles--regardless of the truth status. Furthermore, the language proficiency also plays an important role in linguistic differences between liars and truth-tellers (for an example of credibility criteria rated by human evaluators, see Evans & Michael, 2014). Additionally, the cultural or ethnic background may also play a role in speech and the use of specific words (e.g., Matsumoto, Hwang, & Sandoval, 2014). Furthermore, the storyteller's age matters in that children's language is generally more simple and less complex because their cognitive ability, regulation of language and memory retrieval strategies develop throughout childhood (e.g., Volbert & Steller, 2014). For example, in a recent study, Williams, Talwar, Lindsay, Bala, and Lee (2014), directly compared children's with adult's linguistic profile. Results revealed that younger children (4-5 years) and older children (6-7 years) differ from adults (18-25 years) in the use of personal pronouns, emotional and cognitive process words, and exclusive terms when lying or telling the truth. Together these issues suggest that by analyzing the language of truth-tellers and liars, it is imperative to take the age of the storyteller and different languages into account.

The results and limitations of the first meta-analysis lead to the following practical and empirical implications: Computer programs may be useful to find some preliminary linguistic differences between truth-tellers and liars. Therefore, predictions regarding the outcome of linguistic cues should be based a priori on suitable theories. Also, the context of a statement, specific characteristics of the storyteller, and the research paradigm should always be taken into account when interpreting these effects. However, at this point, computer programs should not be used as a (single) tool to decide whether a person is lying or telling the truth, especially not in a criminal justice context. Therefore, future research on linguistic

analyses with computer programs should focus on its potential in further uncovering linguistic profiles between liars and truth-tellers. This could be implemented for example by refining word categories and linguistic cues, or by complementing words to dictionaries with the help of theoretical assumptions. Moreover, other relevant personal factors that could lead to different narrative styles, like age (e.g., Williams et al., 2014), language (e.g., Masip, Bethencourt, Lucas, Sánchez-San Segundo, & Herrero, 2012), or fantasy proneness (e.g., Schelleman-Offermans & Merckelbach, 2010), should be investigated more closely. These and further situational variables should be subject to future research before applying a computer-based linguistic lie detection assessment to individual forensic cases (if at all).

Inter-rater reliability of CBCA criteria

The second meta-analysis investigated the inter-rater reliability of CBCA criteria. CBCA is an important component of Statement Validity Assessment (SVA, Köhnken, 2004; Steller & Köhnken, 1989), a clinical credibility assessment procedure used by psychological forensic expert witnesses in court in many countries, especially when no external and explicit evidence is at hand (Köhnken, 2004; Sporer, 1983; Steller & Köhnken, 1989). Aside from its forensic application, CBCA criteria have been object of many research investigations as a means to detect deception (e.g., Vrij, 2008). Within this procedure, the critical statement is analyzed in view of 19 criteria, which are subsumed under five broader categories: *general characteristics* of the statement (i.e., 01-logical structure, 02-unstructured production, 03-quantity of details), *specific contents* (i.e., 04-contextual embedding, 05-descriptions of interactions, 06-reproduction of conversation, 07-unexpected complications), *peculiarities of the content* (i.e., 08-unusual details, 09-superfluous details, 10-accurately reported details misunderstood, 11-related external

associations, 12-acounts of subjective mental state, 13-attribution of perpetrator's mental state), *motivation-related contents* (i.e., 14-spontaneous correction, 15-admitting lack of memory, 16-raising doubts about ones own memory, 17-self-deprecation, 18-pardoning the perpetrator), and *offense-specific elements* (i.e., 19-details characteristic of the offense). All criteria are assumed to appear more often in truthful than in deceptive statements (e.g., Undeutsch, 1967; Steller & Köhnken, 1989). Aside from Undeutsch' working hypothesis (Köhnken, 1990), several authors attempted to develop a more comprehensive theoretical background from cognitive and motivational perspectives (Köhnken, 2004; Niehaus, 2001; Sporer, 1997; Volbert & Steller, 2014). Empirical support for this assumption for most criteria came from two recent meta-analyses on the validity of CBCA criteria (Amado, Arce, & Fariña, 2015; Sporer, Hauch, Blandón-Gitlin, & Masip, 2015). As an important prerequisite of its validity, the inter-rater reliability is of utmost importance (e.g., Köhnken, 2004; Küpper & Sporer, 1995; Steller & Köhnken, 1989).

Therefore, the first meta-analysis on the inter-rater reliability of CBCA criteria was conducted. To this end, after an exhaustive literature search, 82 hypothesis tests that fulfilled eligibility criteria were included. Besides six reliability indices that constitute the dependent variables, four independent variables were investigated: Research paradigm, amount of rater's training, rating scale, and base rate, or frequency of occurrence of the individual CBCA criteria.

As expected, inter-rater reliabilities as measured with Pearson's *r* were good to excellent for almost all criteria. Lowest but still good reliability values (according to Fleiss, 1981) were found for unstructured production (02) and superfluous details (09). For *percentage agreement*, all criteria except contextual embedding (04) resulted in good agreement rates. Unweighted analyses of Cohen's *kappa* displayed

moderate to fair agreement (according to Landis & Koch, 1977) for all criteria with lowest values for two criteria (unstructured production (02), details characteristic of the offense (19)). Combined analyses of weighted *kappa* and *ICC* showed good reliabilities except for logical structure (01). Unweighted analyses on Maxwell's *RE* resulted in good reliabilities except for two criteria with marginal reliability (unstructured production (02), quantity of details (03)). Moreover, further analyses revealed that independent variables significantly moderate inter-rater reliabilities in the expected direction (at least for the first of the following three variables). First, the base rate was highly associated with reliability in that low and high base rates accompany lower Pearson's *r* and lower *kappa*, but higher *percentage agreement*. Second, Pearson's *r* was found to be higher in field studies and quasi-experiments than in laboratory experiments for most CBCA criteria. Third, less fine-grained rating scales were associated with higher Pearson's *r* for some CBCA criteria. Fourth, the amount of training was positively, negatively or not associated with Pearson's *r*--depending on the specific CBCA criterion.

Taken together, these meta-analytic findings suggest that on average, the inter-rater reliability is sufficient to good for most CBCA criteria. Compared to inter-rater reliabilities of psychiatric diagnoses in initial DSM-5 field trials (Freedman, Lewis, Michels, Pine, Schultz, Tamminga, Gabbard, Shur-Fen Gau, Javitt, Oquendo, Shrout, Vieta, & Yager, 2013, Figure 1), the inter-rater reliability of CBCA raters lies in the middle of this range. However, a few criteria with less straightforward definitions like unstructured production (02; see also Anson, Golding, & Gully, 1993) resulted in a somewhat lower inter-rater reliability. In interpreting the amount of agreement, it is important to take moderator variables--especially its base rate and research paradigm--into account.

Despite these findings, some limiting aspects of this meta-analysis and implications for future research need to be discussed. Almost all meta-analyses resulted in highly heterogeneous and/or skewed distribution of effect sizes. Although relevant moderator variables were detected, a large amount of variance still left unexplained for most criteria. Therefore, future research should attempt to more precisely quantify the link between variations of specific context variables with its inter-rater reliability. To this attempt, it would also be desirable to maintain the power of future meta-analyses by including more studies overall, and by reducing the number of excluded studies due to missing data (e.g., Cohn & Becker, 2003). Thus, it would be helpful if future studies could report all data on independent and dependent variables--regardless of their significance--to avoid publication bias (e.g., Sporer & Cohn, 2011; Sutton, 2009). As more and more scientific journals provide the opportunity to upload supplemental material online, this would probably not be an unrealistic suggestion. Also, a (much) higher percentage of multiple rated statements (i.e., 100%), would also lead to a higher number of included rated statements for future syntheses. Furthermore, as only a few CBCA field studies with high-stake, real life forensic cases meeting high quality research standards were conducted (e.g., Akehurst, Manton, & Quante, 2011; Roma, Martini, Sabatello, Tatarelli, & Ferracuti, 2011), future research should try to focus on these type of investigations. By doing so, conclusions with higher ecological validity and generalizability can be made. Finally, future studies should refrain from using *percentage agreement* as the only measure due to its shortcomings (e.g., Cohen, 1960). Instead, the following reliability indices should be employed: *ICC*s or Pearson's *r* for continuous data, and *kappa* or *weighted kappa* for categorical data. As a more sophisticated and unbiased estimator of *kappa*, the *prevalence-adjusted, bias-adjusted kappa* (*PABAK*, Birt,

Bishop, & Carlin, 1993) might also be used to control for its base rate (i.e., prevalence) dependence.

Moreover, three research directions that would probably further our knowledge in using CBCA to detect deception are the following: (a) to examine the relationship between inter-rater reliability and validity (as is planned in Sporer et al., 2015), (b) to more deeply investigate the application of the entire credibility assessment procedure SVA in terms of its validity and reliability (see also Volbert & Steller, 2014, for this suggestion), and (c) to establish a link between human CBCA ratings and linguistic computer analyses, for example by editing dictionaries from a theoretical view or adding specific words or sentences expected to be found as a CBCA criterion (e.g., Sim & Lamb, 2013).

Setting research implications aside, how do these meta-analytic findings support the practical application of CBCA? As mentioned earlier, the inter-rater reliability is the most important prerequisite of its validity. In other words, although CBCA ratings have subjective or clinical components, this meta-analysis showed that reliability is maintained. In other words, the fact that two independent psychological expert witnesses come to the same rating of CBCA criteria is above chance level. However, the inter-rater reliability of the entire SVA procedure and its final credibility assessment in court are in need of more empirical investigations (Volbert & Steller, 2014). This issue is impressively demonstrated in the introducing example, the “Kachelmann-trial”: Here, two psychological expert witnesses came to the same evaluation that critical statements of the alleged victim Claudia D. are probably not based on real experience (Friedrichsen, 2011, May). Although the overall conclusion coincided, it may be the case that they were at least partially

derived from different ratings of CBCA criteria--besides considering and weighting other important factors from SVA.

Conclusion

These meta-analyses integrated previous research on the detection of deception from linguistic and verbal content cues. The first meta-analysis demonstrated that theoretically based linguistic cues to deception might exist. However, their validity is limited to small effect sizes and to their association with a number of situational variables and research settings. Although some linguistic cues were detected, in 2015, computer programs should not be used for forensic or any other application to judge whether a person is lying or telling the truth. The second meta-analysis revealed that the inter-rater reliability of CBCA criteria was generally sufficient to good for most criteria. However, results were also highly associated with situational and methodological variables. Finally, with these meta-analyses, this dissertation added empirical knowledge on deception detection research from a verbal perspective, and provided practical implications and orientations for future research.

References

Aamodt, M. G. & Custer, H. (2006). Who can best catch a liar? A meta-analysis of individual differences in detecting deception. *Forensic Examiner, 15*, 6-11.

Akehurst, L., Manton, S., & Quandte, S. (2011). Careful calculation or a leap of faith? A field study of the translation of CBCA ratings to final credibility judgements. *Applied Cognitive Psychology, 25*, 236-243. doi:10.1002/acp.1669

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

Anson, D. A., Golding, S. L., & Gully, K. J. (1993). Child sexual abuse allegations: Reliability of Criteria-based Content Analysis. *Law and Human Behavior, 17*, 331-341. doi:10.1007/bf01044512

Baddeley, A. (2000). The episodic buffer: A new component of working memory? *Trends in Cognitive Sciences, 4*, 417–423. doi:10.1016/S1364-6613(00)01538-2

Baddeley, A. (2006) Working memory: An overview. In Pickering S. (Ed.), *Working memory and education* (pp. 1–31). New York: Academic Press.

Birt, T., Bishop, J., & Carlin, J. B. (1993). Bias, prevalence, and kappa. *Journal of Clinical Epidemiology, 46*, 423-429. doi:10.1016/0895-4356(93)90018-V

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

Cohen, J. (1960). A coefficient of agreement for nominal scales. Educational and *Psychological Measurement, 20*, 37-46. doi:10.1177/001316446002000104

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

Cohn, L. D., & Becker, B. J. (2003). How meta-analysis increases statistical power. *Psychological Methods, 8*, 243-253. doi:10.1037/1082-989X.8.3.243

Conway, M. A. (1990). *Autobiographical memory. An introduction*. Buckingham: Open University Press.

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

Ekman, P. (1988). Lying and nonverbal behavior: Theoretical issues and new findings. *Journal of Nonverbal Behavior, 12*, 163–176. doi:10.1007/BF00987486

Ekman, P. (2001). *Telling Lies: Clues to Deceit in the Marketplace, Politics, and Marriage*. New York: W. W. Norton.

Evans, J. R. & Michael, S. W. (2014). Detecting deception in non-native English speakers. *Applied Cognitive Psychology, 28*, 226-237. doi:10.1002/acp.2990

Fleiss, J. L. (1981). *Statistical methods for rates and proportions.* New York: Wiley.

Freedman, R., Lewis, D. A., Michels, R., Pine, D. S., Schultz, S. K., Tamminga, C. A., Gabbard, G. O., Shur-Fen Gau, S., Javitt, D. C., Oquendo, M. A., Shrout, P. E., Vieta, E., & Yager, J. (2013, Editors' Office). The Initial Field Trials of DSM-5: New Blooms and Old Thorns. *American Journal of Psychiatry, 170*, 1, 1-5. doi:10.1176/appi.ajp.2012.12091189)

Friedrichsen, G. (2011, May). *Gutachter im Kachelmann-Prozess: "Vielleicht hat sie das Messer nur gefühlt?" [Expert witnesses in Kachelmann-trial: "Perhaps she has just felt the knife?"]* Retrieved from http://www.spiegel.de/panorama/justiz/

gutachter-in-kachelmann-prozess-vielleicht-hat-sie-das-messer-nur-gefuehlt-a-761541.html on November 14th 2015.

Hancock, J., & Woodworth, M. (2013). An “eye” for an “I”: The challenges and opportunities for spotting credibility in a digital world. In B. S. Cooper. D. Griesel, & M. Ternes (Eds.), *Applied issues in investigative interviewing, eyewitness memory, and credibility assessment* (pp. 325-340). New York: Springer.

Hauch, V., Sporer, S. L., Michael, S. W., & Meissner, C. A. (2014). *Does training improve the detection of deception? A meta-analysis.* Communication Research. Advance online publication. doi:10.1177/0093650214534974

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

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

Köhnken, G. (2004). Statement Validity Analysis and the "detection of the truth". In P. A. Granhag & L. A. Strömwall (Eds*.), The detection of deception in forensic contexts* (pp. 41-63). Cambridge, UK: Cambridge University Press.

Küpper, B., & Sporer, S. L. (1995). Beurteilerübereinstimmung bei Glaubwürdigkeitsmerkmalen: Eine empirische Studie [Inter-rater reliability of content credibility criteria: An empirical study]. In G. Bierbrauer, W. Gottwald, & B. Birnbreier-Stahlberger (Eds.), *Verfahrensgerechtigkeit* (pp. 187-213). Köln, Germany: Otto Schmidt.

Landis, J. R., & Koch, G. G. (1977). Measurement of observer agreement for categorical data. Biometrics, 33, 159-174. doi:10.2307/2529310

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis.* Thousand Oaks: Sage Publications.

Masip, J., Bethencourt, M., Lucas, G., Sánchez-San Segundo, M., & Herrero, C. (2012). Deception detection from written accounts. *Scandinavian Journal of Psychology, 53*, 103-111. doi:10.1111/j.1467-9450.2011.00931.x

Matsumoto, D., Hwang, H. C., & Sandoval, V. A. (2014). Ethnic similarities and differences in linguistic indicators of veracity and lying in a moderately high stakes scenario. *Journal of Police and Criminal Psychology, 30*, 15-26. doi:10.1007/s11896-013-9137-7

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

Niehaus, S. (2001). *Zur Anwendbarkeit inhaltlicher Glaubhaftigkeitsmerkmale bei Zeugenaussagen unterschiedlichen Wahrheitsgehaltes* [Applicability of content criteria to testimonies with different truth status]. Frankfurt am Main, Germany: Europäische Hochschulschriften.

Roma, P., Martini, P. S., Sabatello, U., Tatarelli, R., & Ferracuti, S. (2011). Validity of Criteria-based Content Analysis (CBCA) at trial in free-narrative interviews. *Child Abuse & Neglect, 35*, 613-620. doi:10.1016/j.chiabu.2011.04.004

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

Sim, M. P. Y., & Lamb, M. E. (2013). Children's disclosure of child sexual abuse: How motivational factors affect linguistic categories related to deception

detection. *Psychology, Crime, & Law, 19, 8*, 649-660. doi:10.1080/1068316X.2012.719621

Sporer, S. L. (1983, August). *Content criteria of credibility: The German approach to eyewitness testimony*. Paper presented at the 91st Annual Meeting of the American Psychological Association, Anaheim, CA.

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

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

Sporer, S. L. (2015). *Deception and cognitive load: Expanding our horizon with a working memory model*. Manuscript submitted for publication.

Sporer, S. L., & Cohn, L. D. (2011). Meta-analysis. In B. D. Rosenfeld, & S. D. Penrod (Eds.), *Research methods in forensic psychology* (pp. 43-62). New York: Wiley.

Sporer, S. L., Hauch, V., Blandón-Gitlin, I., & Masip, J. (2015, August). *Content cues to veracity: A meta-analysis of the validity of Criteria-based Content Analysis.* Paper presented at the European Association of Psychology and Law Conference in Nuremberg, Germany.

Sporer, S. L., & Schwandt, B. (2006). Paraverbal correlates of deception: A meta-analysis. *Applied Cognitive Psychology, 20,* 421-446. doi:10.1002/acp.1190

Sporer, S. L., & Schwandt, B. (2007). Moderators of nonverbal indicators of deception: A meta-analytic synthesis. *Psychology, Public Policy, and Law, 13,* 1-34. doi:10.1037/1076-8971.13.1.1

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

Sutton, A. J. (2009). Publication bias. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2nd ed., pp. 435-452). New York, NY: Russell Sage Foundation.

Undeutsch, U. (1967). Beurteilung der Glaubhaftigkeit von Zeugenaussagen. In U. Undeutsch (Ed.), *Handbuch der Psychologie, Band 11: Forensische Psychologie* (pp. 26-181). Göttingen, Germany: Hogrefe.

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

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

Walker, W. R. & Skowronski, J. J. (2009). The fading affect bias: But what the hell is it for? *Applied Cognitive Psychology, 23,* 1122–1136. doi:10.1002/acp.1614

Wiener, M., & Mehrabian, A. (1968). Language within language: Immediacy, a channel in verbal communication. New York: Appleton-Century-Crofts.

Williams, S. M., Talwar, V., Lindsay, R. C. L., Bala, N., & Lee, K. (2014). Is the truth in your words? Distinguishing children's deceptive from truthful statements. *Journal of Criminology, 2014*, 1-9. doi:10.1155/2014/547519

DEUTSCHE ZUSAMMENFASSUNG

In verschiedenen sozialen Situationen des alltäglichen Lebens ist ein jeder gelegentlich mit dem Verdacht konfrontiert, dass ein Gesprächspartner oder eine Gesprächspartnerin nicht die Wahrheit erzählt. Die Erkenntnis, dass Menschen im Durchschnitt mindestens zweimal pro Tag lügen (DePaulo, Kashy, Kirkendol, Wyer, & Epstein, 1996), stützt diesen Verdacht. Ebenso kommt es in anderen Lebensbereichen, wie zum Beispiel im Arbeitskontext in Bewerbungsgesprächen, oder im forensischen Kontext zu Lügen, also absichtlichen Falschaussagen (Vrij, 2008). Besonders bei letztgenanntem können die negativen Konsequenzen von nicht erkannten Falschaussagen oder für unwahr gehaltene wahre Aussagen sowohl für die einzelne Person, als auch für das Allgemeinwohl, erheblich sein. Daher ist die Frage, wie Lügen von wahren Aussagen unterschieden werden können, von großer Relevanz.

Da Menschen bei dieser Entscheidung im Durchschnitt mit einer Urteilsgüte von 54% nur leicht oberhalb des Zufallsniveaus liegen (Aamodt & Custer, 2006; Bond & DePaulo, 2006), versuchen Forscher und Praktiker seit jeher diese Fähigkeit zu verbessern, indem sie nach Indikatoren (als Synonyme werden Merkmale, Kriterien oder Hinweise verwendet) suchen, die zur validen Unterscheidung von Lüge und Wahrheit herangezogen werden können. Diese Indikatoren können sich auf psychophysiologische Reaktionen (z. B. Hautleitfähigkeit, Atemfrequenz), nonverbales Verhalten (Gestik, Mimik, Körperbewegungen) oder paraverbales Verhalten (Stimmhöhe, Antwortlatenz), oder auf den Inhalt einer Aussage beziehen (siehe Überblick in Sporer & Köhnken, 2008).

Forschungsergebnisse der letzten Dekade zeigen, dass inhaltliche Merkmale dabei im Vergleich zu anderen Methoden die vielversprechendste Validität

aufweisen (Amado, Arce, & Fariña, 2015; Bond & DePaulo, 2006; DePaulo, Lindsay, Malone, Muhlenbruck, Charlton, & Cooper, 2003; Hauch, Sporer, Michael, & Meissner, 2014). Daher beschäftigt sich die Dissertation mit der Frage, ob bestimmte inhaltliche Merkmale, die zur Unterscheidung von Lüge und Wahrheit herangezogen werden, wissenschaftlichen Gütekriterien der Validität und Reliabilität genügen. Hierfür wurde die Methodik der Metaanalyse angewendet, welche sich dadurch auszeichnet empirische Primärstudien, die dieselbe Forschungsfrage untersuchen, zu integrieren (z. B. Lipsey & Wilson, 2001). Im Vergleich zu einer qualitativen Literaturzusammenfassung (*Review*), werden Studienergebnisse mithilfe von methodologischen Standards, insbesondere Effektstärken, selektiert, quantifiziert und analysiert (APA 2008; Cooper, Hedges, & Valentine, 2009; Sporer & Cohn, 2011). Während die erste Metaanalyse die Validität linguistischer Kriterien, die mithilfe von Computerprogrammen analysiert wurden, untersucht, ergründet die zweite Metaanalyse die Beurteilerübereinstimmung (Interrater-Reliabilität) von Glaubhaftigkeitsmerkmalen (Steller & Köhnken, 1989). Das Ziel der Dissertation ist es somit, einen empirisch-systematischen Überblick über zwei bedeutsame Zweige der Forschung zur Entdeckung von Täuschung mit inhaltlichen Merkmalen zu liefern.

Metaanalyse I: Validität linguistischer Lügenkriterien

Als eine Möglichkeit inhaltliche Lügenmerkmale zu erforschen, untersuchten Wissenschaftler und Wissenschaftlerinnen aus verschieden Forschungsbereichen wie Psychologie, Kommunikationswissenschaften, Computerlinguistik oder Computerwissenschaften, mithilfe von verschiedenen Computerprogrammen Texte wahrer und falscher Aussagen. Hierbei wurden meist a priori grammatikalische Kategorien (z. B. Wortanzahl, Personalpronomen, Verben), oder inhaltliche Kategorien (z. B. emotional-getönte Wörter, oder Wörter, die

wahrnehmungsgebundene oder kognitive Prozesse abbilden), definiert, die linguistische Kriterien darstellen (als Beispiel siehe *Linguistic Inquiry and Word Count*, LIWC, Pennebaker, Francis, & Booth, 2001). Die Anzahl der zu einer Kategorie zugehörigen auftretenden Wörter wird dann zwischen wahren und erfundenen (transkribierten) Aussagen verglichen. Erstaunlicherweise wurde bereits 1974 die erste empirische Untersuchung mit einer linguistischen Computeranalyse von Knapp, Hart und Dennis durchgeführt.

Die in der vorliegenden Metaanalyse durchgeführte umfassende Literatursuche in renommierten wissenschaftlichen Datenbaken ergab, dass fast 40 Jahre später circa 400 Studien durchgeführt worden waren. Nach Abgleich mit den Ein- und Ausschlusskriterien (wie zum Beispiel die Angabe unabhängiger statistischer Daten über die Häufigkeit linguistischer Merkmale in wahren und erfundenen Aussagen), wurden 44 Hypothesentests in die Metaanalyse eingeschlossen. Aus über 200 linguistischen Kriterien wurden 40 relevante Kriterien sechs Forschungsfragen zugeordnet und deren Richtung im Rahmen theoretischer Einbettung vorhergesagt. Ebenso wurden neben Studiencharakteristika (Publikationsjahr, Anzahl der Stichprobe, Alter, Anzahl der Aussagen, etc.) weitere relevante unabhängige Variablen, wie die Art (selbsterlebte Erfahrung, Einstellung, Verschiedenes) und die emotionale Valenz (neutral, negativ, positiv) des Ereignisses, die Art der Interaktion zwischen Aussagendem und Zuhörer (keine Interaktion, computervermittelt, Interview, Gespräch), die Motivation des Aussagenden (keine, wenig bis mittel, viel), die Art der Berichterstattung (hand- oder maschinenschriftlich, mündlich), der Programmtyp (LIWC, allgemein, spezifisch), das experimentelle Design (Zwischensubjekt- und Innersubjektfaktoren) und der

Publikationsstatus (publiziert, unpubliziert), von zwei trainierten, unabhängigen Beurteilern und Beurteilerinnen kodiert.

Mithilfe der beschriebenen Methodik sollten folgende Ziele der Metaanalyse erreicht werden: Systematische Integration interdisziplinärer Forschungsergebnisse zur Abbildung linguistischer Unterschiede zwischen wahren und erfundenen Aussagen, operationalisierende Definitionen linguistischer Kriterien, Vorhersage der Richtung des Effektes mithilfe theoretischer Einbettung und Quantifizierung des Zusammenhangs zwischen Kontextvariablen und Effektstärken. Im Folgenden werden die Hauptergebnisse der Forschungsfragen und der linguistischen Kriterien, welche jeweils in Richtung der Lüge im Vergleich zur Wahrheit formuliert sind, dargestellt. Dabei bedeuten negative Effektstärken, dass Lügner ein Merkmal häufiger und positive Effektstärken, dass Wahrheitssagende ein Merkmal häufiger (im Vergleich zu dem jeweiligen Pendant) benutzten. Zudem werden die Ergebnisse der Kontextvariablen und Implikationen für Forschung und Praxis diskutiert. Da bis zur Durchführung der Metaanalyse keine deutschsprachige Studie vorlag, werden im folgenden die Übersetzungen des LIWC-Wörterbuches von Wolf, Horn, Mehl, Haug, Pennebaker und Kordy (2008) teilweise übernommen.

Die erste Forschungsfrage, ob Lügner kognitiv stärker belastet sind, wurde anhand des Arbeitsgedächtnismodell des Lügens (Sporer, 2015; Sporer & Schwandt, 2006, 2007, basierend auf Baddeley, 2000, 2006; Walczyk, Igou, Dixon, & Tcholakian, 2013) gestellt. Es wurde angenommen, dass das Arbeitsgedächtnis beim Lügen aufgrund mehrerer paralleler Aufgaben insgesamt stärker beansprucht wird, wodurch weniger Kapazität für die Sprachproduktion verbleibt. Die Vorhersage, dass Lügen im Durchschnitt kürzer sind, also weniger Wörter insgesamt beinhalten, wurde bestätigt. Zudem benutzten Lügner hypothesenkonform weniger

unterschiedliche Wörter und sich ausschließende Konjunktionen (z. B. außer, ohne, jedoch), wodurch sich eine falsche Aussage eher einfacher und weniger komplex als eine wahre Aussage gestaltet. Der positive Haupteffekt der Wortanzahl war am größten, wenn negativ-emotionale Ereignisse einem direkten Interaktionspartner gegenüber geschildert wurden. Dahingegen zeigte sich ein invertierter Effekt bei computerbasierter Kommunikation oder bei der Auswertung mit speziellen Computerprogrammen. Demnach waren unter diesen Umständen Lügen im Durchschnitt länger als wahre Aussagen. Weitere linguistische Merkmale, wie zum Beispiel Wort- oder Satzlänge, Verbanzahl, Schreibfehler, oder weitere kausale Konjunktionen (z. B. weil, daher, aufgrund) zeigten keine signifikanten Unterschiede. Als hypothesenkonträre Ausnahme zeigte sich bei Lügnern eine erhöhte Satzanzahl.

Der Selbstpräsentationsansatz von DePaulo und Kollegen (DePaulo et al., 2003) stellte den theoretischen Hintergrund für die zweite Forschungsfrage, ob Lügner unsicherer sind als Wahrheitssagende, dar. Es wurde angenommen, dass eine täuschende Selbstpräsentation weniger überzeugend ist und dass Lügner sich durch moralische Bedenken und negative Gefühle mehr von ihrer Aussage distanzieren. Wenn sich diese Distanz in der Sprache niederschlägt (Wiener & Mehrabian, 1968), könnten Lügen eher vage und unsicher formuliert sein (Kuiken, 1981). Die Metaanalysen zeigten keine Unterschiede in Modalverben (z. B. sollten, könnten) und bei Wörtern, die Gewissheit ausdrücken (z. B. finite Temporaladverbien, wie nie, immer, alle). Im scheinbaren Widerspruch zur Hypothese wurde in wahren Aussagen eine erhöhte Anzahl von Wörtern gefunden, die eine Vorläufigkeit ausdrücken (z. B. vielleicht, beinahe), allerdings nur in Studien mit einem Innersubjektfaktoren-Design. Aus einer anderen Perspektive des Selbstpräsentationsansatzes konnte dieser Effekt damit erklärt werden, dass sich

Personen, die die Wahrheit sagen, nicht zusätzlich anstrengen, glaubhaft zu erscheinen oder Unsicherheiten zu vermeiden - im Gegensatz zu Lügnern, die damit vermehrt beschäftigt sind (DePaulo et al., 2003; Volbert & Steller, 2014).

Der dritte Fragenkomplex beschäftigte sich mit emotionalen Prozessen. Aufgrund Ekman's emotionalen Ansatzes (1988, 2001) wurde angenommen, dass Lügner negative Emotionen, wie Schuld, Scham, Angst oder Ärger empfinden. Daher wurde erwartet, dass sich diese Emotionen in der Sprache wiederfinden. In der Tat benutzten Lügner im Durchschnitt mehr Verneinungen (z. B. nein, nicht) und negativ-emotionale Wörter, insbesondere Wörter, die Ärger ausdrücken. Die Unterschiede in negativ-emotionalen Wörtern zeigten sich nur im Innersubjektfaktoren-Design und wenn der Aussagende hoch motiviert war einem direkten Interaktionspartner ein negativ-emotionales, selbsterlebtes Ereignis mitzuteilen. Dahingegen fanden sich keine Unterschiede für Traurigkeits- oder Angstwörter. Der aus der autobiographischen Gedächtnisforschung stammende „fading affect bias" (Walker & Skowronski, 2009), dass negative Ereignisse schneller verblassen als positive, wurde nicht bestätigt, da sich im Durchschnitt keine Unterschiede in positiv-emotionalen Wörtern darstellten.

Die vierte Forschungsfrage „Distanzieren sich Lügner stärker von ihrer Aussage?" basierte auf dem Konstrukt der Unmittelbarkeit (*Immediacy*) von Wiener und Mehrabian (1968), welcher den Grad der Direktheit und Intensität der Interaktion zwischen Aussagendem und Aussage beschreibt. Es wurde angenommen, dass sich Lügner mehr von ihrer Aussage distanzieren, indem sie weniger Selbstbezüge, mehr Fremdbezüge und passive Verben oder Generalisierungen benutzen. Tatsächlich verwendeten Lügner im Durchschnitt etwas weniger Personalpronomina der 1. Person (z. B. ich, wir) und mehr Personalpronomina der 2. und 3. Person (z.

B. du, er, ihr, sie). Es ergaben sich keine signifikanten Effekte für Generalisierungen (z. B. jeder, alle) oder passive Verben. Der Effekt der Selbstbezüge war besonders deutlich, wenn in einer publizierten Studie mit Innersubjektfaktoren-Design negative Ereignisse einem Interaktionspartner geschildert werden sollten. Demgegenüber zeigte sich ein moderierender Zusammenhang bei Fremdbezügen in die Richtung, dass die erwarteten Effektstärken vor allem in unpublizierten Studien mit Zwischensubjektfaktoren-Design, in dem neutrale Ereignisse oder Einstellungen ohne Anwesenheit eines Interaktionspartners berichtet wurden, auftraten.

Ob Lügner insgesamt weniger Details berichten, untersuchte die fünfte Forschungsfrage. Vor dem theoretischen Hintergrund des Realitätsüberwachungs-ansatzes (Johnson & Raye, 1981; Sporer, 1997, 2004) wurde angenommen, dass Lügner weniger sensorische und kontextbezogene Details berichten (können), da sie nicht zurückgreifen können auf durch Wahrnehmung eines tatsächlichen Ereignisses extern-generierte Erinnerungen, sondern auf intern-generierte Erinnerungen basierend auf Phantasien, Gedanken oder Träume,. Es zeigten sich hypothesenkonforme, kleine Effekte für sensorisch-wahrnehmungsbezogene Prozesse, insbesondere Wörter, die das Hören (z. B. zuhören, sprechen) betreffen, und für Numerale (z. B. wenig, alles, viel). Diese Effekte wurden besonders bei hochmotivierten Personen, die ihre Aussage handschriftlich verfassten, und in publizierten Studien mit einem Innersubjektfaktoren-Design deutlich. Für andere linguistische Kriterien, wie Sehen, Fühlen, Zeit, Ort, Präpositionen, Zahlen oder Modifizierer (Summe aus Adverbien und Adjektiven) zeigten sich keine Effekte. Entgegen der Hypothese traten Bewegungsverben (z. B. laufen, gehen) häufiger bei Lügnern auf, insbesondere wenn diese ihre Aussage mündlich oder handschriftlich darboten, und wenn die Studie publiziert war. Dieses Ergebnis konnte mithilfe des

Arbeitsgedächtnismodells erklärt werden, da Lügner aufgrund eingeschränkter kognitiver Kapazität weniger komplexe Wörter, wie zum Beispiel Bewegungsverben benutzen (siehe Newman, Pennebaker, Berry, & Richards, 2003).

Die sechste Forschungsfrage „Beziehen sich Lügner weniger häufig auf kognitive Prozesse?“ fand theoretische Einbettung in der autobiographischen Gedächtnisforschung (Conway, 1990). Da Erinnerungen an tatsächlich Erlebtes Hinweise zur Wiederherstellung, unterstützende Erinnerungen oder weitere kognitive Operationen bedürfen, wurde angenommen, dass Lügen weniger kognitive Wörter beinhalten, da die zugrunde liegenden Prozesse andere sind (z. B. Sporer, 2015; Walczyk et al., 2013). Für die Kriterien kognitive Prozesse (z. B. weil, denken, wissen) und Einsicht (z. B. bemerken, bewusst, entscheiden) ergaben sich hypothesenkonforme kleine Effektstärken, was bedeutet, dass Personen, die die Wahrheit sagen, durch die Wortwahl mehr Bezüge zu kognitiven Prozessen herstellten. Diese Effekte wurden im Durchschnitt nur in Studien mit einem Innersubjektfaktoren-Design gefunden.

Neben den sechs Forschungsfragen zugeordneten linguistischen Merkmalen wurden weitere 29 Kriterien einer Restkategorie zugeordnet, da diese theoretisch nicht hergeleitet werden konnten. Diese Analysen ergaben, dass wahre Aussagen mehr Wörter bezüglich Hemmung (z. B. abstreiten, unterdrücken), Menschen (z. B. Baby, Junge), Biologie, physische Zustände und Essen, beinhalteten. Lügen enthielten mehr einschließende Wörter (z. B. auch, insgesamt), soziale Prozesse, Freizeitwörter und Zukunftsverben.

Insgesamt zeigten die Ergebnisse, dass mithilfe von Theorien bestimmte linguistische Merkmale zwischen wahren und erfundenen Aussagen vorhergesagt werden können und sich im Durchschnitt unterscheiden. Ebenso demonstrierten die

Moderatoranalysen, dass die Haupteffekte der allermeisten Kriterien innerhalb verschiedener Kategorien der Moderatorvariablen zu unterschiedlichen Effektstärken führen. Daher kann zusammenfassend festgehalten werden, dass die Validität einzelner, hypothesengeleiteter linguistischer Lügenkriterien unter bestimmten Bedingungen gegeben zu sein scheint. Diese Interpretation unterliegt allerdings folgenden limitierenden Faktoren: Erstens mussten aufgrund von nicht vorhandenen oder beschaffbaren Daten mindestens 50 relevante Studien ausgeschlossen werden. Da Hinweise aus den Moderatoranalysen für einen Publikationsbias (Sutton, 2009) vorlagen, kann eine Verzerrung der Effektstärken durch den Ausschluss dieser Studien nicht quantifiziert werden. Zweitens waren durchschnittliche Effektstärken nach Cohen's (1988) Einteilung (klein: d = 0.20, mittel: d = 0.50, groß: d = 0.80) eher klein ausgeprägt. Der Median aller untersuchten gerichteten Effektstärken ergab einen Wert von |d = 0.19|, welcher allerdings im Vergleich zum Median von |d = 0.10| in DePaulo et al.'s Metaanalyse (2003) mit 158 Lügenkriterien sogar leicht höher liegt. Drittens zeigte sich eine erhebliche Heterogenität der individuellen Effektstärken innerhalb einzelner Kriterien, welche nur teilweise durch den Ausschluss von Ausreißern oder Moderatorvariablen erklärt werden konnte. Oftmals verbleibt ein erheblicher Anteil unaufgeklärter Varianz zwischen den Studieneffekten. Viertens konnte der Zusammenhang der Validität linguistischer Merkmale mit zwei wesentlichen Kontextvariablen - Sprache und Alter - mangels ausreichender Studienanzahl nicht quantifiziert werden. Insgesamt führen diese vier Limitationen dazu, dass die Aussagekraft der einzelnen Haupteffekte linguistischer Lügenmerkmale als eingeschränkt zu beurteilen ist.

Schlussendlich führen die Befunde und Limitationen der Metaanalyse zu folgenden Implikationen für zukünftige Forschung: Aufgrund der interdisziplinären

Vielfalt der Kriterien, die in dieser Arbeit definiert, kategorisiert und einer Theorie zugeordnet wurden, könnten zukünftige Untersuchungen auf diese Definitionen zurückgreifen. Ebenso empfiehlt es sich, die untersuchten Lügenkriterien anhand von spezifischen Theorien a priori vorherzusagen. Des Weiteren wäre es erstrebenswert für zukünftige Projekte, relevante Kontextvariablen zu untersuchen und in die Interpretation der Ergebnisse miteinzubeziehen (siehe Hancock & Woodworth, 2013). Des Weiteren wäre die Entwicklung von Computerprogrammen, welche über ein Wort-Auszählen hinausgehen und den situativen Kontext einer Aussage als auch Personencharakteristika berücksichtigen, wünschenswert. Abschließend wäre ein direkter Vergleich von menschlichen und computerbasierten Urteilen relevant, um Stärken und Schwächen jeder Seite nutzbar zu machen.

Insgesamt zeigte diese Metaanalyse, dass linguistische Lügenmerkmale unter bestimmten Bedingungen existieren. Allerdings darf daraus nicht geschlussfolgert werden, dass Computerprogramme als Basis für eine Einzelfallentscheidung über Lüge oder Wahrheit herangezogen werden können.

Metaanalyse 2: Interrater-Reliabilität der Glaubhaftigkeitskriterien

Vor Strafgerichten kommt es häufig zu Verhandlungen, in denen Aussage gegen Aussage steht und keine zuverlässigen und unabhängigen Beweismittel, wie zum Beispiel DNA-Spuren oder Videobandaufzeichnungen, zur Verifizierung einer Aussage herangezogen werden können (Steller & Köhnken, 1989; Undeutsch, 1982; Vrij, 2008). Daher werden in Deutschland in speziellen Fällen aufgrund eines Bundesgerichtshofurteils (BGH, 1999) und auch in einigen anderen Ländern, wie zum Beispiel der Schweiz, Niederlande oder Spanien, psychologische Gutachter und Gutachterinnen - so genannte Rechts- oder Aussagepsychologen und -psychologinnen - zu Rate gezogen. Diese untersuchen anhand eines

systematischen klinisch-psychologischen Beurteilungsprozesses, der so genannten Glaubhaftigkeitsbegutachtung (Köhnken, 1990; Steller & Volbert, 1999; Volbert, 1995), die aussagende Person, welche in den allermeisten Fällen ein Opferzeuge oder eine Opferzeugin (z. B. eines Missbrauchs oder Vergewaltigung) ist. Bei diesem umfangreichen und hypothesengeleiteten Prozedere müssen verschiedenste Faktoren, wie die Entstehung der Aussage, die Interviewtechnik, Personencharakteristika, die Aussagekonstanz und nicht zuletzt die kritische Aussage selbst untersucht werden (z. B. Steller & Volbert, 1999; Volbert, 2010).

Diese letztgenannte merkmalsorientierte Aussagen- oder Inhaltsanalyse wird mit der von Steller und Köhnken (1989) aus der bisherigen Literatur zusammengestellten Liste von 19 Glaubhaftigkeitskriterien oder Realkennzeichen durchgeführt (*Criteria-based Content Analysis*, CBCA). Diese werden fünf übergeordneten Kategorien zugeordnet: Allgemeine Merkmale (01-Logische Konsistenz, 02-Ungeordnet sprunghafte Darstellung, 03-Quantitativer Detaillreichtum), spezielle Inhalte (04-Raum-zeitliche Verknüpfungen, 05-Interaktionsschilderung, 06-Wiedergabe von Gesprächen, 07-Schilderungen von Komplikationen im Handlungsverlauf), inhaltliche Besonderheiten (08-Schilderung ausgefallener Einzelheiten, 09-Schilderung nebensächlicher Einzelheiten, 10-Phänomengemäße Schilderung unverstandener Handlungselemente, 11-Indirekt handlungsbezogene Schilderungen, 12-Schilderung eigener psychischer Vorgänge, 13-Schilderung psychischer Vorgänge des Angeschuldigten), motivationsbezogene Inhalte (14-Spontane Verbesserungen der eigenen Aussage, 15-Eingeständnis von Erinnerungslücken, 16-Einwände der Richtigkeit gegen die eigene Aussage, 17-Selbstbelastungen, 18-Entlastung des Angeschuldigten), und deliktspezifische Inhalte (19-Deliktspezifische Aussagenelemente). Die von Steller (1989) als

Undeutsch-Hypothese bezeichnete Grundannahme von Udo Undeutsch (1967), dass sich wahre von erfundenen Aussagen in ihrer Qualität unterscheiden, führt zu der Hypothese, dass das Vorliegen eines jeden Realkennzeichens in einer Aussage für den Erlebnisbezug, bzw. für die Wahrheit der Aussage, spricht. Im Gegenzug bedeutet die Abwesenheit eines Realkennzeichens jedoch nicht, dass eine Lüge vorliegt. Eine umfangreichere theoretische Untermauerung (im Vergleich zur Arbeitshypothese von Undeutsch; siehe Köhnken, 1990) im Hinblick auf kognitive Faktoren als auch motivationale Aspekte der Selbstpräsentation wurde von verschiedenen Autoren unternommen (Köhnken, 2004; Niehaus, 2001; Sporer, 1997; Volbert & Steller, 2014).

Als wichtige Voraussetzung für die forensisch-praktische Anwendbarkeit müssen psychometrische Gütekriterien, wie die Validität und Reliabilität erfüllt sein (Köhnken, 2004; Steller, 1989; Steller & Köhnken, 1989; Wells & Loftus, 1991). Inwieweit die Glaubhaftigkeitskriterien tatsächlich zwischen wahren und erfundenen Aussagen unterscheiden - das heißt die Frage nach der Validität - wurde in vielen experimentellen Studien und einigen Feldstudien untersucht (Vrij, 2008). In einer aktuellen Metaanalyse über 18 dieser publizierten Studien, in denen Kinderaussagen analysiert wurden, zeigten sich mehrheitlich mittlere Effektstärken in die erwartete Richtung (Amado et al., 2015). Mit anderen Worten: Die Realkennzeichen treten wie erwartet durchschnittlich häufiger in wahren als in erfundenen Kinderaussagen auf. Ebenfalls zeigen vorläufige Teilergebnisse einer weiteren aktuellen Metaanalyse über 58 Studien, dass für die meisten Kriterien im Durchschnitt signifikante positive Effektstärken zu finden sind (Sporer, Hauch, Blandón-Gitlin, & Masip, 2015). Diese Effekte werden von mindestens drei unabhängigen Variablen moderiert: Dem Studienparadigma (Experiment vs. Quasi-Experiment und Feldstudie), dem

Studiendesign (Zwischensubjekt- und Innersubjektfaktoren) und dem Alter der Aussagenden (Kinder bis 12 Jahre vs. Erwachsene ab 18 Jahre).

Als unabdingbare Voraussetzung für die Validität eines psychologischen Messinstrumentes ist eine hohe Beurteilerübereinstimmung, die so genannte Interrater-Reliabilität, anzusehen (Anastasi, 1990; Cronbach, 1990; Küpper & Sporer, 1995). Hierbei stellt sich die Frage, ob und inwiefern zwei Beurteiler oder Beurteilerinnen zu demselben Rating eines einzelnen Realkennzeichens gelangen. Da zu dieser Fragestellung eine Vielzahl empirischer Studien durchgeführt wurden und bisher noch keine quantitative Zusammenfassung vorlag, wurde diese Metaanalyse als notwendiges Unterfangen angesehen. Die Ziele der Metaanalyse bestanden demnach darin, sämtliche Studien über die Beurteilerübereinstimmung quantitativ zu integrieren und verschiedene Reliabilitätsindizes zu schätzen. Des Weiteren sollte der Zusammenhang zwischen Reliabilität und unabhängigen Variablen eruiert werden.

Nach einer umfangreichen Literaturrecherche in wissenschaftlich relevanten digitalen Datenbanken und analogen Bibliotheken wurden 52 englischsprachige und 22 deutschsprachige unveröffentlichte und veröffentlichte Studien nach Abgleich mit Ein- und Ausschlusskriterien inkludiert. Hierbei wurden zum Beispiel nur Studien eingeschlossen, welche die Einschätzung der Glaubhaftigkeitskriterien von mindestens zwei unabhängigen und blinden Beurteilern oder Beurteilerinnen miteinander verglichen und einen entsprechenden Reliabilitätsindex berichteten. Die abhängigen Variablen (bzw. Effektstärken) stellten gängige Reliabilitätsindizes (Pearson's *r*, *Prozentuale Übereinstimmung*, Cohen's *kappa*, weighted (gewichtetes) *kappa*, Intraklassenkorrelationskoeffizient (*ICC*, intra-class correlation coefficient) und Maxwell's random error, *RE*) dar. Neben wichtigen Studiencharakteristika

wurden die vom Wahrheitsstatus unabhängige Basisrate (Auftretenshäufigkeit) der Realkennzeichen, das Forschungsparadigma (Experimente, Quasi-Experimente und Feldstudien), die Trainingsintensität der Beurteiler und Beurteilerinnen und die Beurteilungsskalierung (dichotom, Präsenzrating (0-1-2), Likert-Skalierung, Häufigkeitszählung) als unabhängige Variablen von zwei unabhängigen Experten oder Expertinnen mit zufriedenstellenden Interrater-Reliabilitäten kodiert.

Für den Reliabilitätsindex Pearson's *r* ergaben sich in den gewichteten Metaanalysen weitestgehend gute bis exzellente Reliabilitäten für die meisten Realkennzeichen. Die geringsten Werte, obwohl diese nach Fleiss' (1981) Kategorisierung noch als gut zu bewerten sind, zeigten sich für 02-Ungeordnet sprunghafte Darstellung und 09-Schilderung nebensächlicher Einzelheiten. Die gewichteten Metaanalysen zur *Prozentualen Übereinstimmung* zeigten durchgehend zufriedenstellende Werte. Ungewichtete Metaanalysen des Reliabilitätsindex *kappa* lieferten für alle bis auf zwei Kriterien (02-Ungeordnet sprunghafte Darstellung und 19-Deliktspezifische Aussagenelemente) eine moderate bis faire Übereinstimmung (nach der Kategorisierung von Landis & Koch, 1977). Aufgrund ähnlicher psychometrischer Kennwerte wurde das gewichtete *kappa* und *ICC* gemeinsam untersucht. Die ungewichtete Metaanalysen ergaben gute Reliabilitäten für alle Realkennzeichen mit Ausnahme der 01-Logischen Konsistenz. Ebenso zeigte der Index Maxwell's *RE* in einer ungewichteten Metanalyse gute Reliabilitäten mit Ausnahme zweier Kriterien (02-Ungeordnet sprunghafte Darstellung und 03-Quantitativer Detailreichtum). Neben diesen Hauptbefunden deuteten weitere Analysen (Moderatoranalyse und Metaregression) auf Zusammenhänge zwischen unabhängigen Variablen und Reliabilität hin. Wie erwartet zeigte sich, dass eine besonders niedrige oder hohe Basisrate mit niedrigen Reliabilitäten (gemessen mit

Pearson's *r* und *kappa*), jedoch mit erhöhter *Prozentualer Übereinstimmung* einhergeht. Des Weiteren ergaben sich hypothesenkonform für die meisten Realkennzeichen höhere Reliabilitäten (Pearson's *r*) in Quasi-Experimenten und Feldstudien als in Experimenten. Bezüglich der Beurteilungsskalen wurden höhere Werte (Pearson's *r*) für weniger differenzierte Skalen (wie z. B. Präsenzratings oder dichotome Urteile) im Vergleich zu Likert-Skalen oder Häufigkeitszählungen für einige Glaubhaftigkeitskriterien gefunden. Die Annahme, dass ein intensiveres Training mit einer höheren Reliabilität (Pearson's *r*) einhergeht, wurde für fünf Realkennzeichen bestätigt, es zeigten sich jedoch ebenso Befunde in die konträre Richtung für drei weitere Kriterien.

Zusammengefasst zeigte die Metaanalyse, dass die Beurteiler-übereinstimmung für die meisten Realkennzeichen und innerhalb der meisten Reliabilitätsindizes als hinreichend bis gut zu beurteilen ist. Dabei ergaben sich wie angenommen besonders für Kriterien mit einer klaren Operationalisierung (wie zum Beispiel 06-Wiedergabe von Gesprächen, 10-Unverstandene Handlungselemente, oder 16-Einwände der Richtigkeit der eigenen Aussage) konsistent überzeugende Reliabilitäten. Besonders zwei Kriterien (02-Ungeordnet sprunghafte Darstellung, 09-Schilderung nebensächlicher Einzelheiten) mit eher uneindeutigen Definitionen wiesen vergleichsweise niedrigere Reliabilitäten auf. Als weitere wesentliche Befunde zeigten sich Zusammenhänge der Reliabilität mit der Basisrate einzelner Kriterien, der Wahl des Forschungsparadigmas, der Beurteilungsskala und der Trainingsintensität.

Neben den Hauptaussagen sind folgende Limitationen zu nennen und im Hinblick auf zukünftige Forschung kritisch zu diskutieren. Zunächst waren die Ergebnisse mehrheitlich von großer Heterogenität. Obwohl die Varianz durch die

erwähnten Variablen teilweise erklärt wurde, verblieb ein erheblicher Anteil ungeklärter Varianz. Daher wird es wichtig sein in weiterer Forschung, ungeklärte Varianz durch systematische Untersuchungen weiterer unabhängiger Variablen zu quantifizieren. Dies könnte in zukünftigen Metaanalysen ebenfalls möglich sein, wenn die Autoren der Studien alle Informationen zu unabhängigen als auch abhängigen Variablen offen darlegten (z. B. indem zusätzliche Daten auf entsprechenden Internetseiten der Zeitschriften hochgeladen würden). Da es in manchen Studien sehr schwierig oder nicht möglich gewesen ist, einige Daten zu erfassen, litten viele Ergebnisse unter dem Ausschluss relevanter Studien, womit die Power der Metaanalyse reduziert wurde (Cohn & Becker, 2003). Ebenso wurde die Aussagekraft der metaanalytischen Ergebnisse dadurch vermindert, dass häufig nur ein Anteil der gesamten Aussagen einer Studie mehrfach kodiert wurde. Es wäre daher für zukünftige Forschung wünschenswert, dass nach Möglichkeit *alle* unabhängigen Beurteiler und Beurteilerinnen *alle* Aussagen - und nicht nur einen Anteil - evaluieren könnten. Für die Auswahl des Reliabilitätsindexes wäre es erstrebenswert, wenn nicht nur auf die *Prozentuale Übereinstimmung* (aufgrund der Überschätzung durch die Zufallsübereinstimmung, siehe Cohen, 1960), sondern vielmehr auf psychometrisch höherwertige Indizes, wie Pearson's *r*, *ICC* oder *kappa* abgestellt würde. Ebenfalls sollte, wie bereits erwähnt, immer die Basisrate der einzelnen Kriterien bei der Interpretation mitberücksichtigt werden. Abschließend bleibt zu erwähnen, dass die einzelnen Realkennzeichen nur einen Teil der forensisch-praktischen Arbeit der Glaubhaftigkeitsbegutachtung ausmachen und sich zukünftige Forschung mit den Gütekriterien des gesamten Prozesses der Glaubhaftigkeitsbegutachtung beschäftigen sollte (siehe Volbert & Steller, 2014).

Trotz dieser Limitationen und Anregungen zeigte diese Metaanalyse insgesamt, dass die Beurteilerübereinstimmung der allermeisten Realkennzeichen als zufriedenstellend zu beurteilen ist. Daher ist eine wesentliche Bedingung für die Validität der Realkennzeichen und für die Anwendung in der forensischen Aussagepsychologie gegeben.

Schlussfolgerung

Das Ziel der Dissertation war es, den Forschungsbereich der Entdeckung von Täuschung aus inhaltlicher Perspektive zu ergänzen. Die erste Metaanalyse zeigte, dass einige hypothesengeleitete linguistische Lügenmerkmale existieren, wobei die Validität aufgrund kleiner Effektstärken und Zusammenhänge zu Kontextvariablen als eingeschränkt zu beurteilen ist. Daher wurde empfohlen, Computerprogramme im Jahr 2015 als Entscheidungsgrundlage über Wahrheit oder Lüge nicht anzuwenden. Die zweite Metaanalyse ergab, dass die Beurteilerübereinstimmung für die meisten der 19 Glaubhaftigkeitsmerkmale anhand verschiedener Reliabilitätsindizes als hinreichend bis gut zu bewerten ist, wobei unabhängige Variablen bei der Interpretation berücksichtigt werden müssen. Insgesamt kann geschlussfolgert werden, dass inhaltliche Kriterien zur Differenzierung von Lüge und Wahrheit - unter bestimmten Voraussetzungen - psychometrische Gütekriterien erfüllen.

Literatur

Aamodt, M. G. & Custer, H. (2006). Who can best catch a liar? A meta-analysis of individual differences in detecting deception. *Forensic Examiner, 15*, 6-11.

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

Anastasi, A. (1990). *Psychological testing*. New York: Macmillan Publishing Company.

APA Publications and Communications Board Working Group on Journal Article Reporting Standards (2008). Reporting standards for research in Psychology. Why do we need them? What might they be? *American Psychologist, 63*, 839-851. doi:10.1037/0003-066X.63.9.839

Baddeley, A. (2000). The episodic buffer: A new component of working memory? *Trends in Cognitive Sciences, 4*, 417–423. doi:10.1016/S1364-6613(00)01538-2

Baddeley, A. (2006) Working memory: An overview. In Pickering S. (Ed.), *Working memory and education* (pp. 1–31). New York: Academic Press.

BGH (1999). BGH Urteil vom 30.Juli 1999 - 1StR 618/98 - LG Ansbach. StPO § 244 Abs. 4 Satz 2 Wissenschaftliche Anforderungen an Aussagepsychologische Begutachtugen. *Praxis der Rechtspsychologie, 9*, 113-125.

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

Cohen, J. (1960). A coefficient of agreement for nominal scales. Educational and *Psychological Measurement, 20*, 37-46. doi:10.1177/001316446002000104

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

Cohn, L. D., & Becker, B. J. (2003). How meta-analysis increases statistical power. *Psychological Methods, 8*, 243-253. doi:10.1037/1082-989X.8.3.243

Conway, M. A. (1990). *Autobiographical memory. An introduction*. Buckingham: Open University Press.

Cooper, H., Hedges, L. V., & Valentine, J. C. (2009). (Eds.) *The handbook of research synthesis and meta-analysis*. New York: Russell Sage Foundation.

Cronbach, L. J. (1990). Essentials of psychological testing. Grand Rapids, MI: Harper & Row.

DePaulo, B. M., Kashy, D. A., Kirkendol, S. E., Wyer, M. M., & Epstein, J. A. (1996). Lying in everyday life. *Journal of Personality and Social Psychology, 70*, 979-995. doi:10.1037/0022-3514.70.5.979

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

Ekman, P. (1988). Lying and nonverbal behavior: Theoretical issues and new findings. *Journal of Nonverbal Behavior, 12*, 163–176. doi:10.1007/BF00987486

Ekman, P. (2001). *Telling Lies: Clues to Deceit in the Marketplace, Politics, and Marriage*. New York: W. W. Norton.

Fleiss, J. L. (1981). *Statistical methods for rates and proportions.* New York: Wiley.

Hancock, J., & Woodworth, M. (2013). An “eye” for an “I”: The challenges and opportunities for spotting credibility in a digital world. In B. S. Cooper. D. Griesel, & M. Ternes (Eds.), *Applied issues in investigative interviewing,*

eyewitness memory, and credibility assessment (pp. 325-340). New York: Springer.

Hauch, V., Sporer, S. L., Michael, S. W., & Meissner, C. A. (2014). *Does training improve the detection of deception? A meta-analysis.* Communication Research. Advance online publication. doi:10.1177/0093650214534974

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

Knapp, M. L., Hart, R. P., & Dennis H. S. (1974). An exploration of deception as a communication construct. *Communication Research, 1*, 15-29. doi:10.1111/j.1468-2958.1974.tb00250.x

Köhnken, G. (1990). *Glaubwürdigkeit: Untersuchungen zu einem psychologischen Konstrukt*. München, Germany: Psychologie-Verlags-Union.

Köhnken, G. (2004). Statement Validity Analysis and the "detection of the truth". In P. A. Granhag & L. A. Strömwall (Eds.*), The detection of deception in forensic contexts* (pp. 41-63). Cambridge, UK: Cambridge University Press.

Kuiken, D. (1981). Nonimmediate language style and inconsistency between private and expressed evaluations. *Journal of Experimental Social Psychology, 17*, 183-196. doi:10.1016/0022-1031(81)90013-5

Küpper, B., & Sporer, S. L. (1995). Beurteilerübereinstimmung bei Glaubwürdigkeitsmerkmalen: Eine empirische Studie. In G. Bierbrauer, W. Gottwald, & B. Birnbreier-Stahlberger (Eds.), *Verfahrensgerechtigkeit* (pp. 187-213). Köln, Germany: Otto Schmidt.

Landis, J. R., & Koch, G. G. (1977). Measurement of observer agreement for categorical data. *Biometrics, 33*, 159-174. doi:10.2307/2529310

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis.* Thousand Oaks: Sage Publications.

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

Niehaus, S. (2001). *Zur Anwendbarkeit inhaltlicher Glaubhaftigkeitsmerkmale bei Zeugenaussagen unterschiedlichen Wahrheitsgehaltes*. Frankfurt am Main, Germany: Europäische Hochschulschriften.

Pennebaker, J. W., Francis, M.E., Booth, R.J. (2001). *Linguistic Inquiry and Word Count (LIWC): LIWC 2001.* Mahwah, NJ: Erlbaum.

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

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

Sporer, S. L. (2015). *Deception and cognitive load: Expanding our horizon with a working memory model*. Manuscript submitted for publication.

Sporer, S. L., Hauch, V., Blandón-Gitlin, I., & Masip, J. (2015, August). *Content cues to veracity: A meta-analysis of the validity of Criteria-based Content Analysis.* Paper presented at the European Association of Psychology and Law Conference in Nuremberg, Germany.

Sporer, S. L., & Cohn, L. D. (2011). Meta-analysis. In B. D. Rosenfeld, & S. D. Penrod (Eds.), *Research methods in forensic psychology* (pp. 43-62). New York: Wiley.

Sporer, S. L., & Köhnken, G. (2008). Nonverbale und paraverbale Korrelate von Täuschung. In M. Steller, & R. Volbert (Eds.), *Handbuch der Psychologie. Rechtspsychologie* (pp. 353-363). Göttingen: Hogrefe.

Sporer, S. L., & Schwandt, B. (2006). Paraverbal correlates of deception: A meta-analysis. *Applied Cognitive Psychology, 20,* 421-446. doi:10.1002/acp.1190

Sporer, S. L., & Schwandt, B. (2007). Moderators of nonverbal indicators of deception: A meta-analytic synthesis. *Psychology, Public Policy, and Law, 13,* 1-34. doi:10.1037/1076-8971.13.1.1

Steller, M. (1989). Recent developments in statement analysis. In J. C. Yuille (Ed.), *Credibility assessment* (pp. 135-154). Dordrecht, Netherlands: Kluwer Academic Publishers.

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

Steller, M. & Volbert, R. (1999). Forensisch-aussagepsychologische Begutachtung (Glaubwürdigkeitsbegutachtung). Wissenschaftliches Gutachten für den Bundesgerichtshof. *Praxis der Rechtspsychologie, 9*, 46-112.

Sutton, A. J. (2009). Publication bias. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2nd ed., pp. 435-452). New York, NY: Russell Sage Foundation.

Undeutsch, U. (1967). Beurteilung der Glaubhaftigkeit von Zeugenaussagen. In U. Undeutsch (Ed.), *Handbuch der Psychologie, Band 11: Forensische Psychologie* (pp. 26-181). Göttingen, Germany: Hogrefe.

Undeutsch, U. (1982). Statement reality analysis. In A. Trankell (Ed.), *Reconstructing the past: The role of psychologists in criminal trials* (pp. 27-56). Deventer, Netherlands: Kluwer.

Volbert, R. (1995). Glaubwürdigkeitsbegutachtung bei Verdacht auf sexuellen Mißbrauch von Kindern. *Zeitschrift für Kinder- und Jugendpsychiatrie, 23*, 20-26.

Volbert, R. (2010). Aussagepsychologische Begutachtung. In R. Volbert & K.-Pl. Dahle (Eds.), *Forensisch-psychologische Diagnostik im Strafverfahren* (pp. 18-66), Göttingen, Germany: Hogrefe.

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

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

Walczyk, J. J., Igou, F. P., Dixon, A. P., & Tcholakian, T. (2013). Advancing lie detection by inducing cognitive load on liars: A review of relevant theories and techniques guided by lessons from polygraph-based approaches. *Frontiers in Psychology, 4*(14), 1-13. doi:10.3389/fpsyg.2013.00014

Walker, W. R. & Skowronski, J. J. (2009). The fading affect bias: But what the hell is it for? *Applied Cognitive Psychology, 23,* 1122–1136. doi:10.1002/acp.1614

Wells, G. L., & Loftus, E. F. (1991). Commentary: Is this child fabricating? Reactions to a new assessment technique. In J. Doris (Ed.), *The suggestibility of*

children's recollections (pp. 168-171). Washington, DC: American Psychological Association.

Wiener, M., & Mehrabian, A. (1968). *Language within language: Immediacy, a channel in verbal communication.* New York: Appleton-Century-Crofts.

Wolf, M., Horn, A. B., Mehl, M. R., Haug, S., Pennebaker, J. W., und Kordy, H. (2008). Computergestützte quantitative Textanalyse. Äquivalenz und Robustheit der deutschen Version des Linguistic Inquiry and Word Count. *Diagnostica, 54*, 85-98. doi:10.1026/0012-1924.54.2.85

PUBLICATION STATUS

The *first meta-analysis* on linguistic cues to deception assessed by computer programs is published in a peer-reviewed APA journal called “Personality and Social Psychology Review” (http://psr.sagepub.com/content/19/4/307) and is referenced as:

Hauch, V., Blandón-Gitlin, I., Masip, J., & Sporer, S. L. (2015). Are computers effective lie detectors? A meta-analysis of linguistic cues to deception. *Personality and Social Psychology Review, 19*, 307-342. doi: 10.1177/1088868314556539.

A revised version of the *second meta-analysis* on the inter-rater reliability of CBCA criteria is accepted for publication in a special issue on “field reliability and validity of forensic psychological assessment instruments and procedures” from a peer-reviewed APA journal named “Psychological Assessment” (http://www.apa.org/pubs/journals/ pas/call-for-papers-field-reliability.aspx) and is referred to as:

Hauch, V., Sporer, S. L., Masip, J. & Blandón-Gitlin, I. (in press). Can credibility criteria be assessed reliably? Meta-analysis of Criteria-based Content Analysis. *Psychological Assessment.*

ERKLÄRUNG

Ich erkläre: Ich habe die vorgelegte Dissertation selbständig und nur mit den Hilfen angefertigt, die ich in der Dissertation angegeben habe. Alle Textstellen, die wörtlich oder sinngemäß aus veröffentlichten oder nicht veröffentlichten Schriften entnommen sind, und alle Angaben, die auf mündlichen Auskünften beruhen, sind als solche kenntlich gemacht.

_____

Valerie Hauch