Does training improve the detection of deception? A meta-analysis. Communication Research.

Article *in* Communication Research · May 2016

DOI: 10.1177/0093650214534974

4 authors:

Valerie Hauch
Universitäre Psychiatrische Kliniken Basel
6 PUBLICATIONS **235** CITATIONS
SEE PROFILE

Siegfried L. Sporer
Justus-Liebig-Universität Gießen
99 PUBLICATIONS **4,253** CITATIONS
SEE PROFILE

Stephen Michael
Whitman College
15 PUBLICATIONS **369** CITATIONS
SEE PROFILE

Christian A. Meissner
Iowa State University
111 PUBLICATIONS **6,441** CITATIONS
SEE PROFILE

Communication Research
2016, Vol. 43(3) 283–343
© The Author(s) 2014
Reprints and permissions:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0093650214534974
crx.sagepub.com
SAGE

Does Training Improve the Detection of Deception? A Meta-Analysis

Valerie Hauch[1], Siegfried L. Sporer[1], Stephen W. Michael[2], and Christian A. Meissner[3]

Abstract

This meta-analysis examined whether training improves detection of deception. Overall, 30 studies (22 published and 8 unpublished; control-group design) resulted in a small to medium training effect for detection accuracy (k = 30, g_u = 0.331) and for lie accuracy (k = 11, g_u = 0.422), but not for truth accuracy (k = 11, g_u = 0.060). If participants were guided by cues to detect the truth, rather than to detect deception, only truth accuracy was increased. Moderator analyses revealed larger training effects if the training was based on verbal content cues, whereas feedback, nonverbal and paraverbal, or multichannel cue training had only small effects. Type of training, duration, mode of instruction, and publication status were also important moderators. Recommendations for designing, conducting, and reporting training studies are discussed.

Keywords

meta-analysis, detection of deception, training, feedback, nonverbal behavior, verbal content cues

Detecting deception can be a difficult task and neither lay persons nor professionals show an impressive ability to correctly differentiate deceptive and true statements in strangers (Bond & DePaulo, 2006; Vrij, 2008). Several researchers and practitioners tried to improve this ability through different training approaches (Frank & Feeley,

[1]Justus Liebig University Giessen, Germany
[2]Mercer University, Macon, GA, USA
[3]Iowa State University, Ames, IA, USA

Corresponding Author:
Valerie Hauch, Department of Psychology and Sports Science, Justus Liebig University Giessen, Otto-Behaghel-Strasse 10F, Giessen 35394, Germany.
Email: Valerie.Hauch@psychol.uni-giessen.de

2003). The aim of the following meta-analysis was to (a) quantitatively assess the extent to which training improves the ability to detect deception and (b) determine the characteristics of the training protocol that may be most effective in improving detection accuracy. To this end, the role of several moderator variables on training effects will be investigated. Guidelines for creating new training methods and for improving already existing training programs are derived from the results. Finally, standards for designing and reporting experimental training studies are recommended.

Human Judges' Deception Detection Accuracy

Previous findings show that the ability of lay persons to correctly distinguish between deceptive and true stories is only slightly better than flipping a coin. The meta-analysis by Aamodt and Custer (2006) yielded an average detection accuracy of 54.22% (k = 156). Bond and DePaulo's (2006) large-scale meta-analysis with 206 studies reported a weighted average detection accuracy of 53.46%, which was slightly above chance level.

Regardless of overall accuracy, Bond and DePaulo found that judges rated more accounts as truthful (55.23%) than as lies (44.77%), confirming the well-known "truth bias" (Zuckerman, Koestner, Colella, & Alton, 1984). By implication, a truth bias leads to higher accuracy for detecting true stories (truth accuracy) than lies (lie accuracy), given an equal number of lies and truths to be judged.

Unweighted analyses by Bond and DePaulo (2006) supported this relation by finding a truth accuracy of 61.34% compared with a lower lie accuracy of 47.55%. This relation is referred to as the "veracity effect" (Levine, Park, & McCornack, 1999). A response bias shift may occur if a training program directs judges to look for lie or truth criteria, respectively, thus inducing a lie or truth bias.

While a truth bias is prevalent among lay judges, a lie bias may also occur under some conditions. For example, Meissner and Kassin (2002) showed that training or experience as police/parole officers or social workers was associated with a lie bias ("investigator bias"). Using signal detection theory, they showed that neither experience (k = 4) nor training (k = 2) led to better discrimination ability, but to a lie bias. However, other authors found no evidence for such a lie bias (e.g., McCornack & Levine, 1990; see the overview in Burgoon & Levine, 2009). This inconsistency in findings could be due to the use of different experimental designs, research paradigms, or participant samples (e.g., police vs. students).

Furthermore, detection accuracy is not better for professionals expected to have lie detection experience (e.g., police investigators, detectives, psychologists, or judges). Aamodt and Custer's (2006) meta-analysis suggested no relationships between detection accuracy and experience (k = 13, r = −.08, corresponding to d = −0.16) or education (k = 4, r = .03, d = 0.06). Average detection accuracy for professional lie catchers (55.51%) did not significantly differ from that of lay persons (54.22%). In the meta-analysis by Bond and DePaulo (2006), experts did not significantly outperform lay persons (k = 20, d = −0.03).

Given such discouraging findings, researchers and practitioners have tried to develop training programs to improve the ability to detect deception.

Overview of Training Studies and Their Theoretical Underpinnings

This review focuses on training approaches that involved (a) feedback on participants' judgments of truth and deception, (b) nonverbal and paraverbal cues, (c) verbal content cues, and (d) combinations of (a) to (c).

Feedback Training

Several authors attempted to improve detection of deception by providing feedback after judgments (e.g., Elaad, 2003; Porter, McCabe, Woodworth, & Peace, 2007). Why would the feedback approach work? From a theoretical perspective, the most relevant answer comes from the "law of effect" proposed by Thorndike (1913, 1927): Positive feedback is equated with reinforcement and negative feedback with punishment. Both types of feedback should have positive effects on performance because positive feedback reinforces correct behavior and negative feedback punishes incorrect behavior. Applied to the detection of deception training context, the law of effect would predict greater detection accuracy if a correct judgment is followed by positive feedback (e.g., "Your judgment was correct"), or if an incorrect judgment is followed by negative feedback (e.g., "Your judgment was incorrect").

From an empirical perspective, Kluger and DeNisi's (1996) large-scale meta-analysis on training feedback on different kinds of performance found that, on average, feedback has a moderate positive effect on performance (k = 607, d = 0.41), though effect sizes were quite heterogeneous.

Porter, Woodworth, and Birt (2000) proposed two possible mechanisms why feedback could lead to improved detection accuracy. First, feedback may lead participants "to detect (consciously or unconsciously) valid cues to deception and modify their decision-making accordingly" (p. 655). Second, feedback implies a social demand factor for making "more careful judgments" (Porter et al., 2000, p. 655), in that participants may be motivated due to increased pressure to perform better. To discover which mechanism is more likely to work, Porter et al. (2007) compared a bogus (inaccurate) feedback with an accurate feedback condition (see also Zuckerman, Koestner, & Alton, 1984). Unfortunately, neither accurate nor inaccurate feedback improved detection accuracy.

An extension of the mere feedback approach is to link information about and use of specific deception cues (see next section) with feedback about the accuracy of a given judgment (e.g., Fiedler & Walka, 1993).

In sum, we hypothesized different types of feedback to improve performance. In the following sections, we discuss training judges to use different types of cues that are thought to be associated with deception.

Nonverbal and Paraverbal Cues Training

All training studies including nonverbal or paraverbal cues share the assumption that senders show systematic differences when lying or telling the truth with respect to these behaviors. While some authors subsume vocal expressions under nonverbal behaviors, we use the term *nonverbal* referring only to visual cues, subsuming vocal expressions under "paraverbal" behavior (also called "paralanguage," "prosodic," or "vocalics").

Three meta-analyses showed only a few reliable differences of these cues in deceptive versus true stories, all small in magnitude (DePaulo, Lindsay, Malone, Muhlenbruck, Charlton, & Cooper, 2003; Sporer & Schwandt, 2006, 2007). For example, Sporer and Schwandt (2007) found a decrease in nodding ($k = 9$, $d = -0.18$), in hand and finger movements ($k = 5$, $d = -0.38$), and in leg and foot movements ($k = 15$, $d = -0.14$) for liars, whereas DePaulo et al. (2003) observed an increase in adaptors ($k = 14$, $d = 0.16$) and a decrease of illustrators ($k = 16$, $d = -0.14$) for liars. Concerning paraverbal behaviors, two meta-analyses found a significant positive effect size for liars' voice pitch (DePaulo et al., 2003: $k = 12$, $d = 0.21$; Sporer & Schwandt, 2006: $k = 7$, $d = -0.18$). Furthermore, DePaulo et al. found an increase in repetitions ($k = 4$, $d = 0.21$), and Sporer and Schwandt (2006) observed an increase in response latency for liars ($k = 18$, $d = 0.21$). Despite their significance, these effect sizes were small and varied widely across studies. Some of the differences between these meta-analyses are due to the operationalizations used and the inclusion/exclusion of different studies.

In addition, there are several theoretical approaches (e.g., Zuckerman, DePaulo, & Rosenthal, 1981) leading to different, and at times, contradictory predictions for particular behaviors (see DePaulo et al., 2003; Sporer & Schwandt, 2006, 2007, for an overview). For example, the arousal approach as well as the emotion-fear approach predicts an increase of head movements with deception, whereas the emotion-guilt approach, the attempted control approach, and the cognitive load/working memory model assume a decrease with deception (Sporer & Schwandt, 2007). Individual training studies justify their choice of cues trained with these different theoretical backgrounds or with an idiosyncratic selection of previous findings. Consequently, different training programs taught either nonverbal cues only (e.g., Vrij, 1994), a combination of nonverbal and paraverbal cues (e.g., DePaulo, Lassiter, & Stone, 1982), or compared these two (e.g., DePaulo et al., 1982). More problematically, some training programs actually instructed participants to look for *increases* in certain behaviors that were actually *negatively* related to deception according to these later meta-analyses mentioned.

Because of the small effect sizes for the validity of nonverbal and paraverbal cues, effectiveness of training approaches using such cues is predicted to be low overall.

Verbal Content Cues Training

Few training approaches have focused solely on the verbal content of statements. Three approaches provide support for the hypothesis that senders' speech content would systematically differ when telling the truth than when lying. First, Undeutsch (1967) stated that statements based on memory of real experiences differ in quality and quantity from invented and false statements. Steller and Köhnken (1989) developed a list of 19 reality criteria, referred to as criteria-based content analysis (CBCA), which integrated criteria described by Arntzen (1970, 1983), Dettenborn, Froehlich, and Szewczyk (1984), Sporer (1983), Szewczyk (1973), and Undeutsch (1967). True statements are believed to contain more of these criteria than false statements. For example, if a statement is logically structured, includes many details, for example, of conversations, the statement is more likely to be true. CBCA is only a part of statement validity analysis (SVA), which is a comprehensive approach including different methods of collecting and analyzing data to assess the credibility of statements (Steller & Köhnken, 1989). Although the validity of various CBCA criteria has been experimentally tested in numerous studies (see Vrij, 2005), there are only a few training studies yet (e.g., Akehurst, Bull, Vrij, & Köhnken, 2004; Landry & Brigham, 1992).

Empirical evidence for the validity of the CBCA criteria comes from Vrij's (2005) vote-counting review and from DePaulo et al.'s (2003) meta-analysis. Vote-counting refers to a simple tallying of significant positive, significant negative, and null findings. Vote-counting has been criticized as being an inadequate method of meta-analysis because it neither takes sample sizes nor the magnitude of observed effects (i.e., their effect sizes) into account (Hedges & Olkin, 1985; see also Sporer & Cohn, 2011). In their meta-analysis, DePaulo et al. found support for some of the CBCA criteria. Truth-tellers' accounts include more details ($k = 24$, $d = -0.30$) and more spontaneous corrections ($k = 5$, $d = -0.29$), are more logically structured ($k = 6$, $d = -0.25$), and participants admitted a lack of memory more frequently ($k = 5$, $d = -0.42$). Note, however, that DePaulo et al. used only a very small portion of the literature ($k = 5$ or 6).

Second, Johnson and Raye (1981) developed the reality monitoring (RM) approach which assumes that people rely on qualitative characteristics, such as sensory, contextual, semantic, and emotional information, when deciding whether one's *own* memory is based upon an actual event (external) or not (internal). This assumption has been extended to interpersonal RM, that is, judging the reality of other people's memories (Mitchell & Johnson, 2000; Sporer, 1997; Sporer & Sharman, 2006) including the detection of deception (for reviews, see Masip, Sporer, Garrido, & Herrero, 2005; Sporer, 2004). DePaulo et al.'s (2003) meta-analysis reported a nonsignificant tendency that sensory information was more frequently present in true accounts compared with lies ($k = 4$, $d = -0.17$). In addition, in summarizing results from RM studies from Vrij and colleagues, Sporer (2004) reported positive effect sizes for visual details, sound details, and spatial, temporal, and affective information ranging from $d = 0.43$ to $d = 1.46$, and both a positive ($d = 0.85$, in Vrij, Edward, Roberts, & Bull, 2000) and a negative ($d = -0.41$, in Vrij, Akehurst, Soukara, & Bull, 2004) effect size for cognitive operations (e.g., associations, reflections, decision processes).

Third, several studies used combinations of selected CBCA and RM criteria (e.g., Sporer & Bursch, 1996; Sporer & McCrimmon, 1997; Sporer & McFadyen, 2001). Sporer (1998, 2004) theoretically and empirically combined the CBCA and the RM approach on the basis of factor analyses and laid a theoretical foundation from research on autobiographical memory, impression management, and attribution theory resulting in a comprehensive set of truth criteria referred to as the Aberdeen Report Judgment Scales (ARJS; Sporer, 1998, 2004). A few training studies by Sporer and his colleagues using the ARJS have been conducted (e.g., Sporer, Samweber, & Stucke, 2000).

Other researchers trained their participants with different methods involving different types of verbal content analysis (see Colwell et al., 2009; deTurck, Feeley, & Roman, 1997; Santarcangelo, Cribbie, & Ebesu Hubbard, 2004). Finally, some researchers applied a mixture of nonverbal, paraverbal, and verbal content cues, for example, using the Reid Technique (Blair, 2009; Kassin & Fong, 1999) and other techniques (Hendershot, 1981).

We did not include studies that used specific computer programs, such as the Linguistic Inquiry and Word Count (Newman, Pennebaker, Berry, & Richards, 2003; Zhou, Burgoon, Nunamaker, & Twitchell, 2004) to find linguistic cues to deception because they did not involve training human raters (for a recent meta-analysis on these cues, see Hauch, Blandón-Gitlin, Masip, & Sporer, 2013).

We predict that training programs using verbal content cues yield the largest training effects compared with multichannel cues or feedback due to the larger effect sizes of the cues trained.

Previous Meta-Analyses of Training Studies

Although two previous meta-analyses on training to detect deception have been published, we identified important methodological issues that lead us to call into question the reliability of their findings. Frank and Feeley (2003) summarized 11 published studies with 20 hypothesis tests, missing several relevant studies already available at the time. In addition, the authors did not consider an important statistical problem of dependent effect sizes (Gleser & Olkin, 1994, 2009; Lipsey & Wilson, 2001): Studies with multiple training groups and only one control group were treated as if they were independent. This led to an overrepresentation of control groups, which apparently were used repeatedly for comparison. Thus, the weighted average effect size reported ($r = .20$, $d = 0.41$), with a heterogeneous effect size distribution, is likely an overestimate of the population parameter and an underestimate of its variability.

While Driskell's (2012)[1] attempt to update Frank and Feeley's (2003) meta-analysis is more comprehensive, including 16 published studies, it did not cover 13 relevant published and unpublished studies, nor 8 studies using other experimental designs. Consequently, our synthesis not only covers a larger set of studies and experimental designs but also reduces a potential publication bias by making a special attempt to include unpublished studies (most of which are conference presentations).

Driskell's meta-analysis contains similar methodological problems as Frank and Feeley's although the author seems to be aware of them. In his synthesis, Driskell found a weighted average training effect of $d = 0.50$ in 16 published training studies (from 1984 to 2006) with 30 hypothesis tests. Our synthesis included 30 studies with a total of 55 hypothesis tests.[2] While Driskell did note the problem of dependent effect sizes in a footnote (p. 728, Note 3), calculating an average effect size for 16 studies by pooling across the different comparisons does not solve this problem. Using the same control group repeatedly is likely to have led to an overestimation of the mean training effect size (see our *Discussion*). Last but not least, Driskell did not analyze lie and truth accuracy separately. This differentiation is important because an overall improvement in detection accuracy does not necessarily mean that both abilities—correctly classifying lies and truths—are improved. Therefore, a meta-analysis on training to detect deception should consider at least three dependent variables, namely, overall detection accuracy, lie accuracy, and truth accuracy.

Here we present a new meta-analysis with a substantially larger number of studies that addresses the methodological issues noted and updates the current state of knowledge on deception detection training. We also address the issue of publication bias by using newer statistical methods borrowed from the medical literature (Rothstein, Sutton, & Borenstein, 2005; Sutton, 2009), which are explained in the *Method* and *Results* section.

Designs of Training Studies to Be Included

Training studies involve several phases. The first phase is to obtain true or false statements from senders. In one paradigm, senders are asked to tell their true or false opinions, attitudes, or feelings about a particular theme, a film, or a person. Alternatively, they are instructed to tell a true or false story about a self-experienced event or about a mock crime they either did or did not commit. In the second phase, other participants, referred to as judges or receivers, are either randomly assigned to the experimental (training) and control condition (true experiment) or nonrandomly (quasi-experiment, see Campbell & Stanley, 1963). Then, a set of statements of the senders is either presented audiovisually, visually, or as a written transcript, and judged regarding their truthfulness.

Moreover, a training study can be assessed in three different designs (Campbell & Stanley, 1963; see Table 1). The first is referred to as *posttest only with control* (POWC; see Carlson & Schmidt, 1999) design and implies a training and a control group, each measured only once. The second is referred to as *one-group pretest-posttest* (OGPP) design and consists of at least one training group measured before and after training. The third is referred to as *pretest-posttest with control* (PPWC; see Carlson & Schmidt, 1999) design and includes an experimental and a control group both tested before and after training. Lipsey and Wilson (2001) suggested that studies with these different experimental designs should not be aggregated into a single meta-analysis, because different effect size measures are used that should be interpreted separately (POWC: comparison of control vs. training group; OGPP: comparison of pretest vs. posttest; PPWC: comparison of pre- and posttest *changes* in trained vs. control group). Therefore, different study designs were investigated in separate meta-analyses.

Main Hypotheses

An underlying assumption is that training people or giving feedback on any task aims to improve a particular ability (Patrick, 1992). Therefore, we expected to find an overall positive training effect regarding detection accuracy, as well as for lie and truth accuracy.

Table 1. Different Designs of Training Studies.

Design	Group	Pretest	Training	Posttest
Posttest only with control (POWC)	EG	—	T	O_1
	CG	—	—	O_2
One-group pretest-posttest (OGPP)	EG	O_1	T	O_2
	—	—	—	—
Pretest-posttest with control (PPWC)	EG	O_1	T	O_2
	CG	O_3	—	O_4

Note. EG = experimental group; CG = control group; O = observation; T = training.

Hypotheses for Potential Moderator Variables

In a meta-analysis, a moderator may account for systematic variability between studies (Hedges & Olkin, 1985; Lipsey & Wilson, 2001). Studies differ with respect to a range of independent variables, some of which could have an indirect relationship with training effects. A priori hypotheses with theoretical or empirical background rather than post hoc tests were developed for moderator variables to produce a higher level of certainty for the interpretation of the results (Wood & Eagly, 2009).

Training Category

The *training category* (nonverbal and paraverbal cues, verbal content cues, and feedback) was assumed to moderate effect sizes. Thus, training with verbal content cues was hypothesized to have stronger training effects on detection accuracy compared with the other training categories, because verbal content cues are more strongly related to deception/truthfulness compared with nonverbal or paraverbal cues (DePaulo et al., 2003; Sporer, 2004; Sporer & Schwandt, 2006, 2007; Vrij, 2005). Studies using a feedback paradigm were also expected to lead to a positive (but small) training effect due to Thorndike's (1913, 1927) "law of effect." In addition, we expected a negative training effect when studies utilized a bogus feedback paradigm.

Purpose of the Training

As discussed above, a truth bias in judgment has been revealed for lay persons (e.g., Bond & DePaulo, 2006), whereas an "investigator bias" or lie bias was found for professional lie catchers (Meissner & Kassin, 2002). In relation to training, Masip, Alonso, Garrido, and Herrero (2009) demonstrated that the particular *purpose* of a specific training, that is, focusing on either cues to deception or cues to truthfulness biased participants' responses toward deception or truth, respectively. Therefore, it was expected that training programs using cues to deception would lead to higher lie accuracy for trained compared with untrained persons. In contrast, training programs with the aim to detect the truth would lead to higher truth accuracy.

Intensity of Training

It was predicted that the *intensity of the training* is positively associated with a training effect. We defined training intensity as a conglomerate of five individual components analyzed as separate moderators: *duration*, presentation medium of cues to be learned (referred to as *training medium*), *number of practice examples, group size*, and *trainer presence*. Training intensity was expected to increase the longer the training session, and the higher the number of different media the training content was presented with (e.g., a combination of video-lecture, lecture, handwritten instructions). Providing practice examples, as opposed to no practice, smaller group sizes, and the presence of a trainer in person should also enhance training effects.

Senders' Motivation

As a general hypothesis, from a self-presentational perspective (DePaulo, 1992), one would expect that more highly motivated liars and truth-tellers will make a stronger attempt to tell more compelling stories, be more cooperative, provide more details, and so forth (DePaulo et al., 2003; Sporer, 2004). When training focuses on verbal content cues, highly motivated story-tellers who actually experienced an event should provide more details, which are used as truth criteria in the CBCA and RM approach. Consequently, when training focuses on verbal content cues, discrimination should be better and training more effective.

On the other hand, DePaulo and Kirkendol (1989) proposed the *motivational impairment effect*, which predicts the opposite for *nonverbal* cues: If senders are more highly motivated to lie successfully, they try too hard to control their behavior, but, unable to do so, display *more nonverbal* cues to deception (DePaulo et al., 2003; Sporer & Schwandt, 2006, 2007). Thus, liars should actually be *easier* to detect by judges trained to look for these *nonverbal* cues. Motivation across studies varied as a function of monetary incentives or participation in a (mock) crime.

Story Content

Originally we had coded a large number of categories regarding the content of lies/ truths that we recoded into three categories: (a) lies about attitudes (e.g., liking or disliking somebody or something), (b) lies about a personally experienced (significant) autobiographical life event (e.g., an operation), and (c) lies about an observed or staged event (e.g., a mock crime). With increasing involvement, more cues to deception may become discernible and make training more effective.

Design and Base Rate Information

First, we test whether within-participants designs are more sensitive to training effects than between-participants designs. Senders might show intraindividual differences when lying or telling the truth (Bond & DePaulo, 2008; DePaulo & Morris, 2004; Köhnken, 1989). If judges have the opportunity to evaluate both deceptive and true stories of the same sender, they should be better in their discrimination performance, because they have two behavioral excerpts of a person. Thus, we expected a higher training effect for within-participants than between-participants designs.

Second, in some studies, trainers or researchers informed their participants on the actual lie/truth ratio (the base rate) beforehand. If judges knew this base rate (usually 50%; with the exception of DePaulo et al., 1982, who used a base rate of 67%), the training effect was expected to be higher than if they did not know it due to the fact that they might not be inclined toward a truth bias or a lie bias.

Research Group

Different groups of researchers may differ regarding the effectiveness of training for reasons not documented in their reports. Different research groups may have used different types of stimulus material. Two issues need to be distinguished: (a) If studies differ in difficulty level of stimulus material, this should only affect training effects as main effects (except in case of floor or ceiling effects). One could analyze for this by using overall (or control group) accuracy as a continuous predictor in a meta-regression, but we have opted against this for space reasons. (b) A more serious problem arises if the choice of stimulus material interacts with training effectiveness (by leading to an improvement for one type of training over another).

No specific hypotheses are possible, but in case of differences, the type of training program or stimuli used by each laboratory should be scrutinized.

Publication Bias

Publication bias is related to the tendency of researchers to submit and for journals to be more likely to publish studies reporting significant results than those with nonsignificant results (Begg, 1994; Cooper, 2010; Rothstein et al., 2005; Sporer & Cohn, 2011). There is strong evidence for a publication bias in psychological treatment research (Lipsey & Wilson, 1993). It is hypothesized that published studies show stronger effects than unpublished ones. Furthermore, higher precision of estimates (i.e., smaller standard errors due to larger samples) should be negatively associated with the size of effects.

Method

Research Question and Dependent Variables

To study training effects on the ability to detect deception, three dependent variables were used: Overall *detection accuracy* was operationalized by the total number of correct judgments irrespective of truth status, divided by the total number of judgments made, multiplied by 100. *Truth* or *lie accuracy* was calculated by the number of correctly classified true/false statements, divided by the total number of true/false statements, multiplied by 100.

Inclusion and Exclusion Criteria

First, to be included, studies needed to be designed to investigate the effects of training or feedback on detection accuracy. Second, studies must have used one of the aforementioned designs: (a) POWC, (b) OGPP, or (c) PPWC. Third, studies had to report statistical data from which an effect size for detection accuracy could be derived. Fourth, participants must have judged both deceptive and true stories, whereby the actual truth status of the statement remained unknown to the participants (at least until the judgment was given). Studies in which any kind of technical tool or physiological measure (e.g., a polygraph) was used or taught to participants were excluded. In cases where the results of a specific data set were re-used or otherwise duplicated in more than one publication, we chose the publication that contained most information or with the highest peer-review journal status. A complete list of all excluded studies with the respective reasons for exclusion can be found in Appendix A.

Moreover, training studies could be constructed in one of two research paradigms. The first paradigm requires an approximately equal number of participants in the control and experimental conditions and a sample size larger than 10 participants in the OGPP design. Studies designed with the second paradigm include a relatively small number of trained participants (e.g., one to five "experts") compared with a much larger number of untrained participants. Another difference between these designs is the much larger number of judgments made in this second paradigm. Only studies that utilized the first paradigm were included in the meta-analysis.

Literature Search

The first step to locate relevant studies was to search through the reference lists of relevant review articles (Bond & DePaulo, 2006; Bull, 1989, 2004; Frank & Feeley, 2003; Vrij, 2008). The first and third author read the abstracts or methods sections to evaluate the suitability according to the aforementioned criteria. Reference sections of these potential training or feedback studies were examined for further studies.

In a second step, computer-based searches of the *Social Sciences Citation Index* with the cited reference search procedure were conducted. In addition, *PsycInfo, WorldCat*, and *Psyndex* searches were conducted using combinations (with the Boolean connector *AND*) of the three keyword categories: *training/feedback/improv*, detect**/credibility judg*, and *deceit/decept*/truth*. Repeated searches were conducted, searching for articles since 1980 until March 2009. A final search was conducted in February 2011, which located five further studies.

The third step was to execute a search with the internet searching tool *Google Scholar* using different combinations of all listed keywords. The first 20 sites of results were examined for relevant studies. The final step included sending emails to the authors of all potential training or feedback studies to request further unpublished or published articles or conference papers.

A total of 39 studies met the inclusion criteria: 31 POWC, 2 OGPP, and 6 PPWC studies with sufficient statistical data. Some studies included more than one hypothesis test, comparing more than one training group with a control group, which will be explained later.

Coding Scheme

Besides effect sizes, five groups of variables were coded: (a) general study characteristics, and information about (b) the judges, (c) the senders, (d) the training, and (e) the judgment procedure.

General study characteristics were year published, publication status (unpublished or published), and type of publication. We subdivided studies into six research groups by authors (deTurck/Feeley/Levine; Sporer et al.; Vrij et al.; Zuckerman et al.; and "other" deception researchers who only conducted one training study). Information about the judges included total sample size and *n*s for experimental and control groups), age, gender, and occupation. In addition, assignment to conditions (random vs. nonrandom) and the motivation to detect lies were coded (none; low: $1-$5, or short written instruction; medium: $6-$10 or long written instruction). Regarding information about senders, sample size, randomization, number, duration and type of stories (attitude/liking, personal autobiographical event, observed/staged event/mock crime), motivation to lie successfully (none; low to medium: $1-$50, or written instruction; high: crime), and design (between- or within-participants) were coded.

Information about the training were training category (feedback, multichannel, verbal content, combination), purpose (to detect lies or the truth), duration, training medium (written instruction or lecture or combination), number of examples, group size, trainer presence, and base rate information. Information about the judgment procedure were the medium in which stories of senders were presented and the number of judgments made.

Categories were later collapsed due to empty cells or too few studies in particular subcategories. Some continuous variables (e.g., training duration, examples, and group size) were recoded into categorical variables in order to conduct moderator analyses.

To code the dependent variables detection accuracy, lie accuracy, and truth accuracy, appropriate statistical values were coded (e.g., means and standard deviations) for each investigated experimental and control group, and/or ANOVA results (*F, df*s, and *p* values) and/or *t*-values for pairwise comparisons between two groups.

Coding Procedure and Intercoder Reliability

Two independent coders (first and third author) coded all variables listed above for each study. The Coding Manual and Coding Protocol were first established in collaboration with the second author and iteratively refined. In order to train coders and establish reliability of the coding scheme, the coders first worked simultaneously through two studies. Then, the coders rated the POWC design studies in the aforementioned manner into an Excel spreadsheet. An agreement was defined as coding exactly the same value for a particular variable. All disagreements were resolved by the concordant decision of both coders.

Cohen's *kappa* (Cohen, 1960) for categorical moderator variables that ranged from .71 to .95, and Pearson's *r* for continuous variables from .73 to .90, were highly satisfactory (Orwin & Vevea, 2009).

Effect Size Estimates for POWC

An appropriate effect size for the POWC design is the standardized mean difference (usually referred to as Cohen's *d*), where the mean of the control group is first subtracted from the mean of the experimental group and then divided by the pooled standard deviation. As this estimate slightly overestimates the population effect size for small sample sizes, *d* was adjusted with a correction factor, resulting in the unbiased estimate g_u (Borenstein, 2009; Lipsey & Wilson, 2001).

Whenever possible, cell *M*s, *SD*s, and *n*s were used for calculation. If studies reported other statistical measures, such as *t*-values, *F* values, *p* values, *Z* values, or *F* values with more than one degree of freedom (mean-square error method, Lipsey & Wilson, 2001), appropriate formulae were applied to calculate effect sizes (Borenstein, 2009). In cases where the comparison was reported simply as "nonsignificant," $g_u = 0$ was assumed.

Effect Size Estimates for OGPP

For studies using a training group tested both before and after training, we used the same formula as above for between-participants designs to calculate the standardized mean difference from means and standard deviations of pre- and posttest (d_{OGPP}; Dunlap, Cortina, Vaslow, & Burke, 1996, Formula 1; Lipsey & Wilson, 2001). If means and standard deviations were not provided, no effect size could be calculated because the formula for repeated measures designs requires the correlation between pre- and posttest (Borenstein, 2009; Dunlap et al., 1996), which was not reported in any study.

Effect Size Estimates for PPWC

Because PPWC designed studies provide more data than POWC or OGPP designs, the "standardized mean change" was computed (see Morris, 2008, Formula 12). In this formula, the difference (change) between pre- and posttest scores of the control group is subtracted from the difference between the pre- and posttest scores of the training group. The result is divided by the pooled pre- and posttest standard deviations for both control and training group (Formula 13). In other words, the effect size d_{PPWC} estimates the standardized difference between the pretest versus posttest *changes* of the training and the control group, respectively. Because this effect size is not directly comparable with the g_u for the POWC or OGPP designs, results are reported separately.

Following the recommendations by Lipsey and Wilson (2001), the different effect size metrics from the three study designs were not combined in a single meta-analysis, but analyzed separately.

Statistically Dependent Effect Sizes

An important issue of this meta-analysis is that more than half of the included studies had conducted different training approaches with more than one trained group and only one control group. For each training group versus control group comparison, a separate effect size was computed. Because these comparisons (hypothesis tests) were always between several training groups and a single identical control group, these effect sizes are statistically dependent.

Meta-analyses are based on the requirement of independent data points as the unit of analysis (Lipsey & Wilson, 2001). Inclusion of dependent effect sizes incurs problems of inflated sample size, underestimation of standard error, and overrepresentation of studies with multiple effect sizes. Therefore, the average of these dependent effect sizes of a given study and the adjusted inverse variance weights were computed. Our first meta-analysis integrated these averaged effect sizes to test the overall training effect across all studies.

To investigate the more interesting question of the effectiveness of *different types* of training, separate (and thus independent) meta-analyses were subsequently conducted for eight different types of training, with specific training type versus control group comparisons (hypothesis tests) derived from all studies that involved the respective comparison: bogus feedback or training, feedback, nonverbal cues, paraverbal cues, nonverbal and paraverbal cues, nonverbal and paraverbal cues and feedback, verbal content cues, verbal content and nonverbal and paraverbal cues.

Meta-Analytic Procedures

Before integration of effect sizes, we tested for outliers by visual inspection of the distributions of individual effect sizes and their confidence intervals, as well as by a more sophisticated method that tests standardized residuals and homogeneity after removing any particular study as recommended by Hedges and Olkin (1985). According to this method, removal of an outlier significantly reduces the heterogeneity within a set of studies.

If these techniques revealed the same effect sizes as outliers, sensitivity analyses were conducted with and without these effect sizes (Greenhouse & Iyengar, 2009). The reason for conducting outlier analyses is that outliers in meta-analyses would make the calculation of a “mean effect size” meaningless (just as outliers distort correlation coefficients or multivariate analyses).

The weighted average effect size was calculated by weighting each individual effect size (g_u) by the inverse of its variance (Lipsey & Wilson, 2001). The fixed effects model was applied, which assumes that all individual effect sizes estimate the same fixed population parameter.[3] Heterogeneity tests were calculated yielding the Q statistic, which approximates a chi-square distribution with $k - 1$ df (Lipsey & Wilson, 2001). As an additional indicator of heterogeneity, the descriptive statistic I^2 was used to indicate the proportion of total variation of effect sizes that is due to heterogeneity (Higgins & Thompson, 2002; Shadish & Haddock, 2009). As a rule of thumb, an I^2

value of 25% is considered to indicate small heterogeneity, 50% medium heterogeneity, and 75% large heterogeneity.

Meta-Analytic Procedure for OGPP and PPWC Studies

To compute the OGPP and PPWC studies' variance, the correlation between pretest and posttest measures (see Dunlap et al., 1996; Morris, 2008) is needed, which none of the studies reported. Hence, no mean effect size weighted by inverse variance weights could be computed. Instead, we calculated the unweighted mean and a mean effect size weighted by sample sizes as a tentative estimate.

Publication Bias

Publication bias is addressed both via graphical and statistical methods (Sutton, 2009). A funnel plot is presented to show an overview of the distribution of effect sizes plotted against the inverse of the standard error (Sterne, Becker, & Egger, 2005). It is assumed that results from studies with smaller sample sizes are more widely spread around the mean effect size because of larger random error (Sutton, Duval, Tweedie, Abrams, & Jones, 2000). Thus, the shape of the distribution should look like a symmetric funnel if no publication bias is present. As an additional test of publication bias, we compared results of published and unpublished studies, using publication status as a moderator.

Computer Software

For computing individual effect sizes, variances, weights, and standard errors, all formulae were programmed in Excel spreadsheets programmed in Microsoft Office Excel (2003) by the second author and cross-checked by the first author. Calculations of meta-analyses were conducted using both Excel spreadsheets and SPSS 20 for Mac, using the macros provided by Wilson (2010).

Results

Study Characteristics

The frequencies and descriptive statistics of continuous variables are displayed in Table 2. A total of 30 POWC[4] designed studies were located, of which 8 were unpublished (including 2 master's theses, a doctoral dissertation, an unpublished manuscript, and 4 conference presentations) and 22 published articles. They were conducted between 1981 and 2011. Judges were randomly assigned to experimental conditions in 20 studies; 10 studies did not report the mode of assignment. All but one study provided information about the occupation of the judges, 86.2% being students, 3.4% trainees, and 10.3% police or parole officers. In four studies, participants received some incentives to successfully detect deception and the truth.

Table 2. Frequencies and Descriptive Statistics of Continuous Variables.

Variable	*k*	*M*	*SD*	*Median*	Minimum	Maximum
N	30	121.27	98.35	100.50	20	390
N_{CG}	30	51.23	42.52	40.50	10	195
N_{EG}	30	70.03	63.80	51.00	10	281
M age	10	25.09	6.44	21.26	19.83	37
SD age	6	3.30	2.36	2.75	1.26	7
Male judges	20	62.10	70.09	54.50	0	331
Female judges	20	63.00	44.49	58.50	0	174
Number of senders	29	20.90	19.98	12.00	2	82
Male sender	22	8.77	9.09	5.50	0	36
Female sender	22	11.41	13.43	6.00	0	55
Stories per sender	29	3.52	3.52	2.00	1	16
Duration of true story	18	118.11	164.02	69.30	20	720
Duration of deceptive story	18	114.59	162.68	56.30	20	720
Duration training	14	54.29	60.89	30.00	5	180
Number of examples	27	2.19	3.48	0.00	0	15
Judgments per person	30	20.37	17.86	16.00	1	72

Note. k = number of hypothesis tests; *N* = sample size; CG = control group; EG = experimental group.

Of all studies, 70% used a within- and 30% a between-participants design for telling lies and truths. The average number of words did not differ between true (M = 118.11, SD = 164.02, Mdn = 69.30, k = 18) and deceptive statements (M = 114.59, SD = 162.68, Mdn = 56.30, k = 18, g_u = 0.02). Participants were asked to judge M = 20.37 stories per study, via an audiovisual medium (82.1%), via transcript (14.3%), or via a combination of both (3.6%). All variables coded are listed in Appendix B (Tables B1, B2 and B3).

Meta-Analytic Syntheses of Effect Sizes

This section deals with the overall effect of any type of training on detection accuracy, lie accuracy, and truth accuracy. Thus, multiple training groups were averaged resulting in one effect size per study as the unit of analysis. All groups involving bogus feedback were excluded from the analysis, because they did not have the aim to improve detection accuracy. Following Cohen's (1988) recommendation, g_u = 0.20 is considered a small, g_u = 0.50 a medium, and g_u = 0.80 a large effect size.

Overall detection accuracy. A total of 30 hypothesis tests involving n = 3,614 participants resulted in a small to medium training effect of g_u = 0.331 [0.262, 0.400]. The results were highly heterogeneous, $Q(29) = 141.44$, $p < .001$, $I^2 = 79.50$, with g_us ranging from $g_u = -0.672$ to g_u = 1.424. Of these 30 effect sizes, 2 had a significant negative, 8 a nonsignificant negative, 17 a significant positive, and 3 a nonsignificant

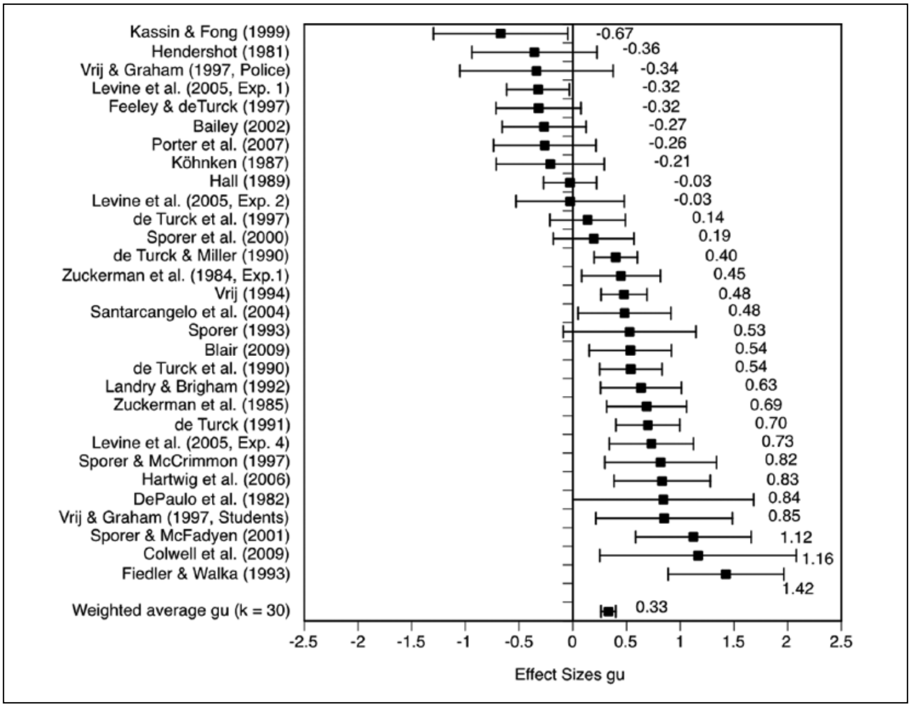

Figure 1. Effect size distribution of mean effect sizes (and 95% CIs) for overall detection accuracy.
Note. CI = confidence interval.

positive effect. Figure 1 reflects this heterogeneity, also indicating graphically that some studies on either side of the distribution may be considered outliers.

Lie accuracy. Only 11 out of 30 studies reported detection accuracy separately for lies and true accounts. These 11 hypothesis tests involving n = 1,274 judges revealed a significant training effect of g_u = 0.422 [0.299, 0.544] for lie accuracy (Figure 2). The distribution was heterogeneous, $Q(10) = 22.26$, p = .014, I^2 = 55.32. The outlier analysis identified the study by Levine, Feeley, McCornack, Hughes, and Harms (2005, Exp. 4), as an outlier. After removing that study, $Q(9)$ = 13.49, p = .142 shrank to a nonsignificant value, and I^2 = 33.27 also indicated that most of the variation was due to sampling error. The weighted average effect size slightly decreased to g_u = 0.362 [0.233, 0.491], still a small to medium training effect.

Truth accuracy. Three out of 11 studies (n = 1,274) showed significant negative, while 6 studies showed significant positive effects for truth accuracy; the remaining 2 were not significantly different from 0 (Figure 3). The analysis resulted in a nonsignificant

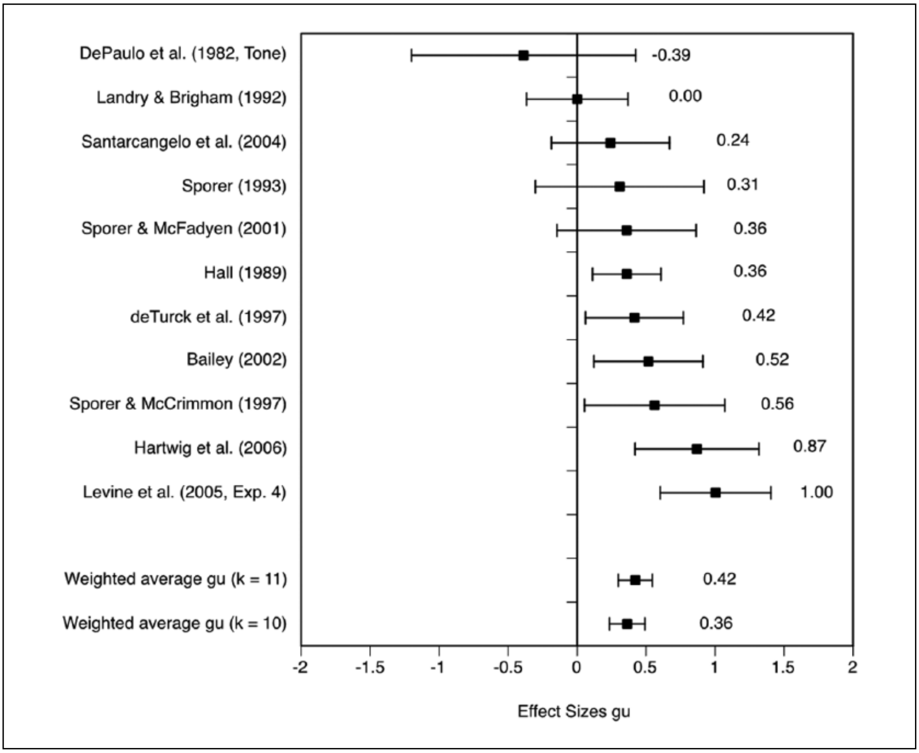

Figure 2. Effect size distribution of mean effect sizes (and 95% CIs) for lie accuracy.
Note. CI = confidence interval.

weighted average effect size of g_u = 0.060 [−0.063, 0.184], p = .337, with a highly heterogeneous effect size distribution, $Q(10)$ = 97.95, p < .001, I^2 = 89.79. Although some of the studies on either side of the distribution could formally be considered as outliers, none of them was excluded.

Moderator Analyses

This section deals with the analyses of previously selected independent variables to moderate the relationship between training and detection accuracy. The pairwise associations between all independent variables, which follow an ordinal relationship, are displayed in Table 3.

Training category. We classified all training programs into four major categories according to training content: (a) accurate *feedback* about truth status (k = 4); (b) "multichannel" category (k = 10): information about specific nonverbal and/or paraverbal cues to deception; (c) *verbal content cues* (such as CBCA, RM, or ARJS; k = 7); (d)

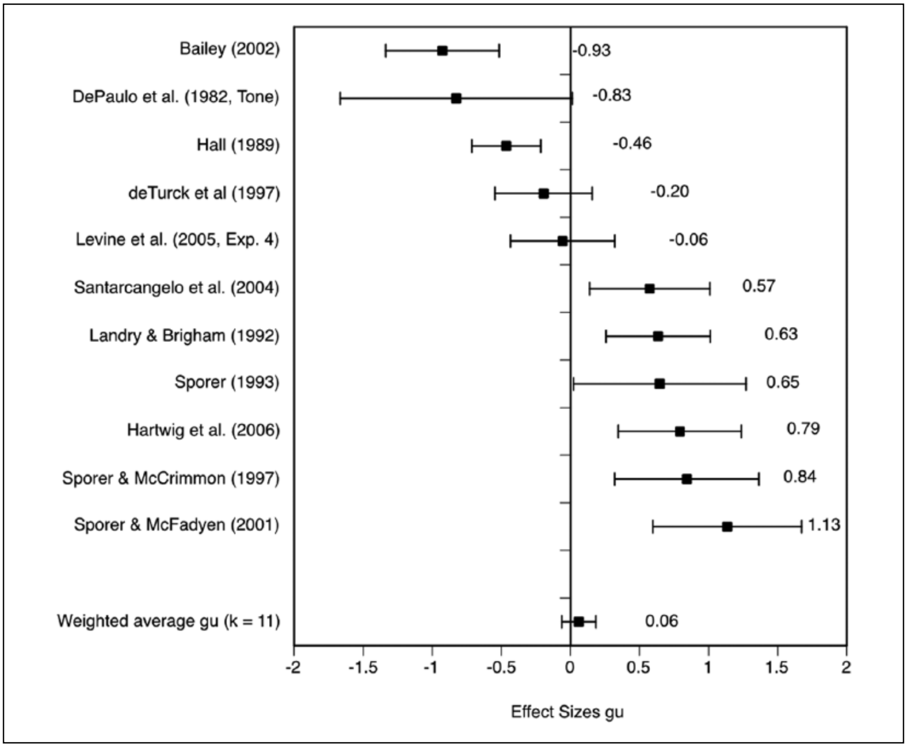

Figure 3. Effect size distribution of mean effect sizes (and 95% CIs) for truth accuracy.
Note. CI = confidence interval.

combination of at least two of the aforementioned categories (k = 9). A significant homogeneity test statistic, $Q_B(3)$ = 15.79, $p < .001$, suggested reliable differences between these categories (Figure 4), although some heterogeneity remained within each training category, $Q_W(26)$ = 134.08, $p < .001$. Studies giving feedback (k = 4, n = 693, g_u = 0.189 [0.022, 0.357]), as well as programs teaching multichannel cues (k = 10, n = 1,351, g_u = 0.276 [0.170, 0.382]), or a combination of the above paradigms (k = 9, n = 887, g_u = 0.336 [0.201, 0.470]), revealed small effect sizes, while verbal content cue training provided a medium training effect of g_u = 0.653 ([0.471, 0.835], k = 7, n = 683).

It should be noted that the variable training category is highly associated with the variable purpose in that only verbal content training studies (but no other training category) had the purpose to detect the truth (k = 5), and only two verbal content training studies had the purpose to detect lies.

Purpose of the training. The predictor variable purpose—whether training had the aim to detect lies or the truth—was assumed to moderate effect sizes for lie and truth

Table 3. Correlation Matrix (Phi or Cramer's V) of Moderator Variables.

Moderator (categories)	Examples	Medium	Group size	Trainer	Purpose	Design	Motivation	Base rate	PubStat
Duration (1-3)	.568[a]	.890[a]**	.333[a]	.661[a]*	.279[a]	.509[a]	.859[a]**	.430[a]	.417[a]
Examples (0/1)		.604[a]**	.220	.402*	−.397[a]	−.369	.853[a]**	.328	−.122
Medium (1-3)			.665[a]	.515*	.408[a]	.544*	.409[a]*	.371[a]	.189
Group Size (0/1)				.000	.192	.320[a]	.373[a]	—	.000
Trainer (0/1)					.181	−.173	.322[a]	−.225	−.070
Purpose (0/1)						.296[a]	.325[a]	.289[a]	.664**
Design (0/1)							.570*	−.348	−.230
Motivation (1-3)								.308	.235
Base rate (0/1)									.296[a]

Note. Coding categories are explained in the text. Correlations for cross tables for 2 × 2 tables are *phi* coefficients; all others Cramer's *V*. PubStat = publication status.
[a]Cross table of categories of moderators contains cell sizes with $k = 0$.
*$p < .05$. **$p < .01$.

accuracy. A total of 26 studies ($N = 3{,}070$) reported the purpose of their training, either to detect lies ($k = 21$, $n = 2{,}568$) or the truth ($k = 5$, $n = 502$). From those 11 studies reporting lie and truth accuracies, 6 had the aim to detect lies, 4 had the aim to detect the truth, and the study by Hall (1989) did not report this information. Because all studies with the aim to detect the truth implemented verbal content cue training, purpose is entirely confounded with training category (verbal content).

The moderator analysis for lie accuracy yielded a significant effect for purpose, $Q_B(1) = 4.09, p = .043$ (Figure 5). Training programs with the aim to detect lies resulted in a larger training effect for lie accuracy ($g_u = 0.550$ [0.374, 0.725]) than programs with the aim to detect the truth ($g_u = 0.246$ [0.010, 0.483]). This result was no longer significant if the outlier ($g_u = 1.003$ [0.602, 1.405], Levine et al., 2005, Exp. 4) was removed, $Q_B(1) = 1.58, p = .209$.

The moderator analysis for truth accuracy suggested a significant main effect for purpose, $Q_B(1) = 29.64$, $p < .001$, though heterogeneity within groups was still large, $Q_W(8) = 45.59, p < .001$. A large training effect for truth accuracy could only be found if trainings aimed to detect the truth ($g_u = 0.784$ [0.540, 1.029]) but not if they aimed to detect lies ($g_u = -0.050$ [−0.225, 0.124]).

Intensity of the training

a. *Duration*. The duration of the training had a mean of 54.29 ($SD = 60.89$, $k = 14$, $n = 1{,}744$), and a $Mdn = 30.00$ minutes per training. The short training category (5-20 minutes) included four studies ($n = 384$), medium training (21-60 minutes) seven studies ($n = 1{,}159$), and long training (61-180 minutes) included three studies ($n = 201$). A moderator analysis showed a significant effect, $Q_B(2) = 15.45$, $p < .004$, but heterogeneity remained within groups, $Q_W(11) = 57.70$, $p < .001$. The short training had a nonsignificant effect of $g_u = -0.030$ [−0.217, 0.157], whereas medium and long training yielded medium effects of $g_u = 0.391$ [0.271, 0.511] and $g_u = 0.491$ [0.160, 0.822], respectively.

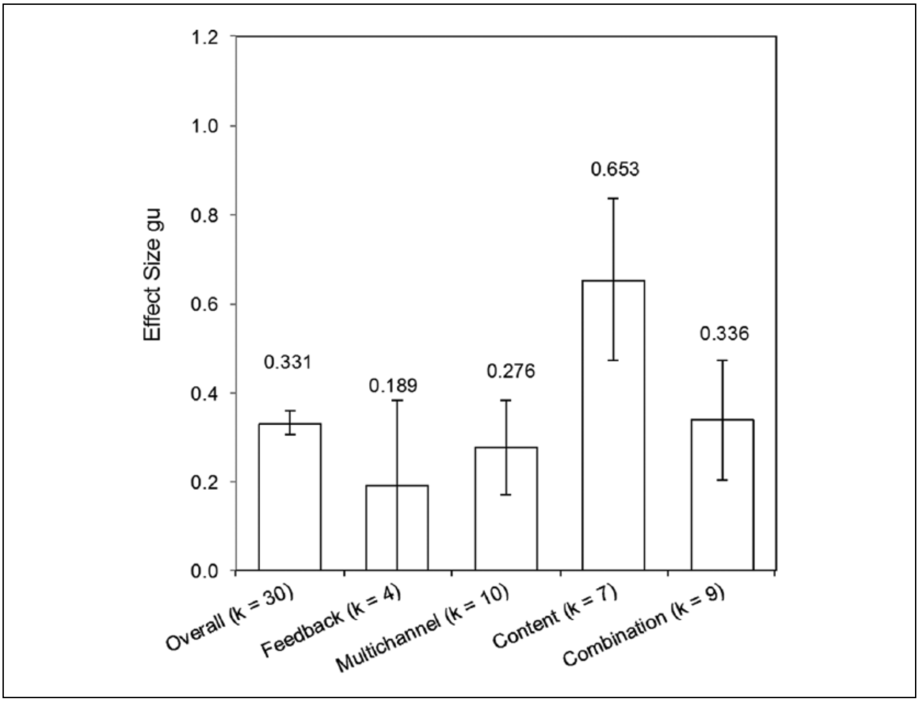

Figure 4. Moderator analysis for training category on overall accuracy.

b. *Training medium.* A moderator analysis resulted in a significant difference between groups, $Q_B(1) = 39.97$, $p < .001$, showing that training programs using written instructions ($k = 10$, $n = 940$, $g_u = 0.470$ [0.334, 0.605]), or using a combination of written instruction and lecture or video ($k = 11$, $n = 1{,}477$, $g_u = 0.443$ [0.337, 0.549]) had larger training effects than training programs using only a lecture or video format ($k = 7$, $n = 848$, $g_u = -0.067$ [−0.205, 0.071]).

c. *Number of examples*. A nonsignificant $Q_B(1) = 0.03$, $p = .860$, indicated that training effectiveness did not differ as a function of practicing examples ($k = 11$, $n = 1{,}715$, $g_u = 0.341$ [0.225, 0.456]) or no examples ($k = 16$, $n = 1{,}330$, $g_u = 0.354$ [0.256, 0.453]).

d. *Group size*. Training programs were either assessed in small groups of 1 to 6 trainees ($k = 9$, $n = 820$, $g_u = 0.308$ [0.158, 0.457]), or in larger groups of 7 to 30 trainees ($k = 6$, $n = 1{,}005$, $g_u = 0.285$ [0.157, 0.412]). A moderator analysis yielded no difference between these groups, $Q_B(1) = 0.05$, $p = .812$.

e. *Trainer presence*. Trainer presence yielded a nonsignificant $Q_B(1) = 1.55$, $p = .312$, indicating that effectiveness did not differ whether training was conducted by a live person ($k = 19$, $n = 2{,}298$, $g_u = 0.360$ [0.275, 0.445]), or without any trainer present ($k = 10$, $n = 1{,}216$, $g_u = 0.267$ [0.148, 0.386]), for example, by a computer program or only by written instructions.

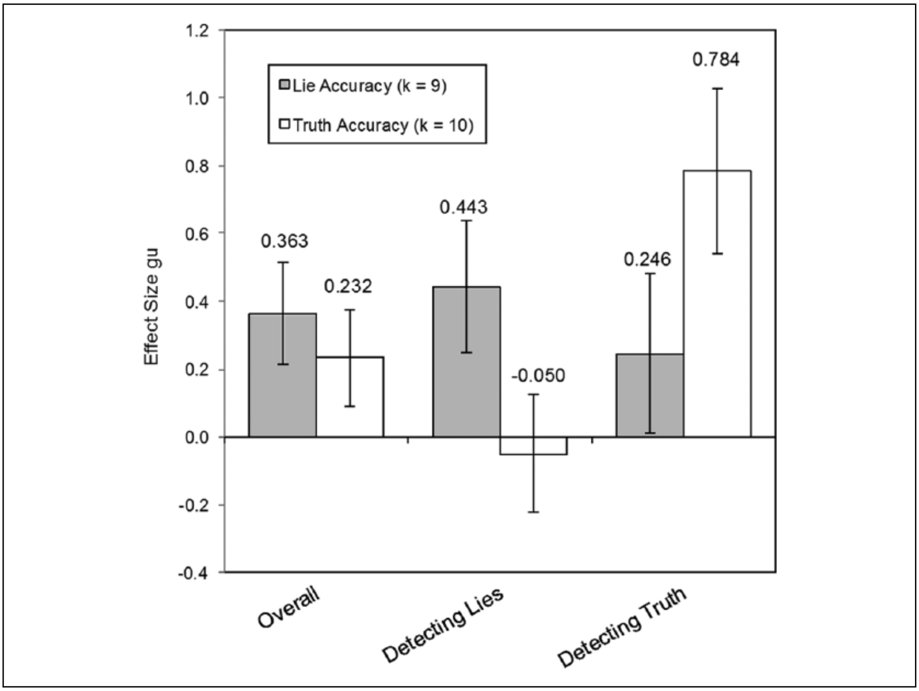

Figure 5. Moderator analyses of purpose for lie and truth accuracy.

Senders' motivation. Senders were not specifically motivated in 19 cases (n = 1,877, g_u = 0.354 [0.259, 0.449]), received low to medium motivation ($1-$50) in seven studies (n = 1,496, g_u = 0.266 [0.156, 0.375]), and were assumed to be highly motivated in four studies (n = 241, g_u = 0.510 [0.260, 0.760]). A moderator analysis resulted in a nonsignificant $Q_B(2) = 3.56$, $p = .169$, leading to the conclusion that senders' incentives did not moderate the training effect.

To test for a possible motivational impairment effect, we separately analyzed studies that used either only multichannel cues (nonverbal or paraverbal) or only verbal content cues. Training with multichannel cues was more effective under medium motivation of senders than studies where senders were not motivated, $Q_B(1) = 14.62$, $p <$ .001. When senders were not explicitly motivated, there was no training effect, g_u = 0.011 [−0.171, 0.193], k = 5, n = 406. For medium motivation stories, there was a significant training effect, g_u = 0.451 [0.318, 0.584], k = 4, n = 925. The study by Hendershot (1981), which was the only one classified as a high motivation study, showed a negative training effect (g_u = −0.358, n = 20).

When training was conducted with verbal content cues only, the difference in training effectiveness was not significant, $Q_B(1) = 1.76$, $p = .185$. When senders were not explicitly motivated, there was a medium size significant training effect, g_u = 0.590 [0.386, 0.795], k = 5, n = 502. For high motivation stories, there was a strong training effect, g_u = 0.895 [0.494, 1.297], k = 2, n = 181, but this was based on only two studies.

Story content. There were no significant differences in training effects as a function of story content, $Q_B(2) = 0.96$, p = .620: attitudes (k = 9, n = 1,520, g_u = 0.301 [0.201, 0.410]); personal autobiographical events (k = 8, n = 664; g_u = 0.395, [0.232, 0.558]); observed or staged events (k = 9, n = 1,116, g_u = 0.303 [0.179, 0.428]).

Design. Nine studies (n = 961) used a between- (senders telling the truth *or* lying), and 21 a within-participants design (n = 2,653; senders telling the truth *and* lying). The experimental design did not moderate the training effect, $Q_B(1) = 0.52$, p = .472.

Base rate information. The significant homogeneity test statistic, $Q_B(1) = 4.53$, p = .033, suggested that the training effect was larger if participants were aware of the lie/truth ratio beforehand (k = 5, n = 527, g_u = 0.446 [0.261, 0.630]) than if they were not (k = 19, n = 1,997, g_u = 0.221 [0.126, 0.316]).

Research group. There was a significant difference among the six different research groups, $Q_B(5) = 32.12$, p < .001 (see Appendix B: Table B1). The largest effect sizes were obtained in two studies by Zuckerman and colleagues (n = 249, g_u = 0.566 [0.305, 0.827]), and four studies by Sporer and colleagues (n = 388, g_u = 0.572 [0.329, 0.816]), although these were not significantly different from three studies by Vrij and colleagues (n = 429, g_u = 0.450 [0.256, 0.645]), nor from the 10 studies from other deception researchers who had conducted only a single training study (n = 728, g_u = 0.453 [0.297, 0.608]). Eight training studies from deTurck, Feeley, and/or Levine showed a small weighted average effect size of g_u = 0.285 [0.177, 0.394], n = 1,368, which was significantly smaller than the effect sizes obtained from the labs by Zuckerman or Sporer, but not significantly different from effect sizes found from Vrij's or other deception researchers' laboratories. Only one group (that we referred to as "others" who reported three studies) resulted in a nonsignificant, slightly negative training effect (g_u = −0.126 [−0.322, 0.071], n = 452) that differed significantly from all other groups.

It should be noted that this moderator variable was highly correlated with other moderator variables showing that research groups systematically differ with respect to study characteristics. For example, all four studies conducted by Sporer and colleagues trained criteria to detect the truth and not lies (only the study by Landry & Brigham, 1992, also used truth criteria), all studies by Zuckerman applied an attitude/liking paradigm, and studies by deTurck/Feeley/Levine did not ask senders to lie about a personal event.

Publication status. The 22 published studies (n = 2,734) differed from the eight unpublished studies (n = 880), $Q_B(1) = 4.12$, p = .042, suggesting a publication bias. The training effect for published studies was significantly higher (g_u = 0.371 [0.292, 0.450]) than for unpublished studies (g_u = 0.202 [0.060, 0.344]). It should be noted that

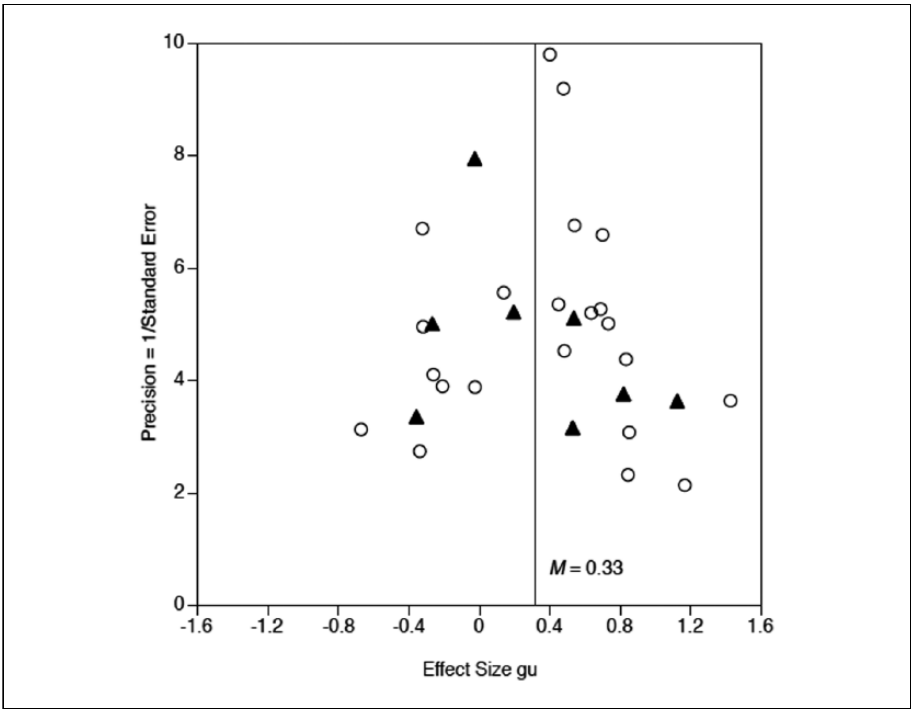

Figure 6. Funnel plot of effect sizes of published (open circles) and unpublished (black triangles) studies for overall detection accuracy and the inverse of the standard error.

publication status was confounded with purpose. Unpublished studies tended to train people to detect the truth, while only one published study did (Landry & Brigham, 1992).

Figure 6 displays the funnel plot of the effect sizes of published and unpublished studies and the inverse of the standard error (precision = 1/*SE*). Although it is difficult to ascertain asymmetry of funnel plots by visual inspection, there appear to be fewer published studies with lower precision and a negative effect size or an effect size close to zero, indicating the possibility of publication bias.

However, more formal tests to address publication bias, such as Begg and Mazumdar's (1994) rank correlation test, or Egger's regression test (Egger, Davey Smith, Schneider, & Minder, 1997; see Sutton, 2009), yielded significant results that would have suggested a publication bias. Duval and Tweedie's (2000a, 2000b) trim and fill method, which estimates and adjusts for the numbers and outcomes of missing studies by an iterative method, suggested only a slight downward adjustment of the mean overall effect size from $g_u = 0.331$ to $g_u = 0.312$. Note also that in Figure 6, there are as many unpublished studies above the mean weighted effect size as below, which would be an argument against a publication bias.

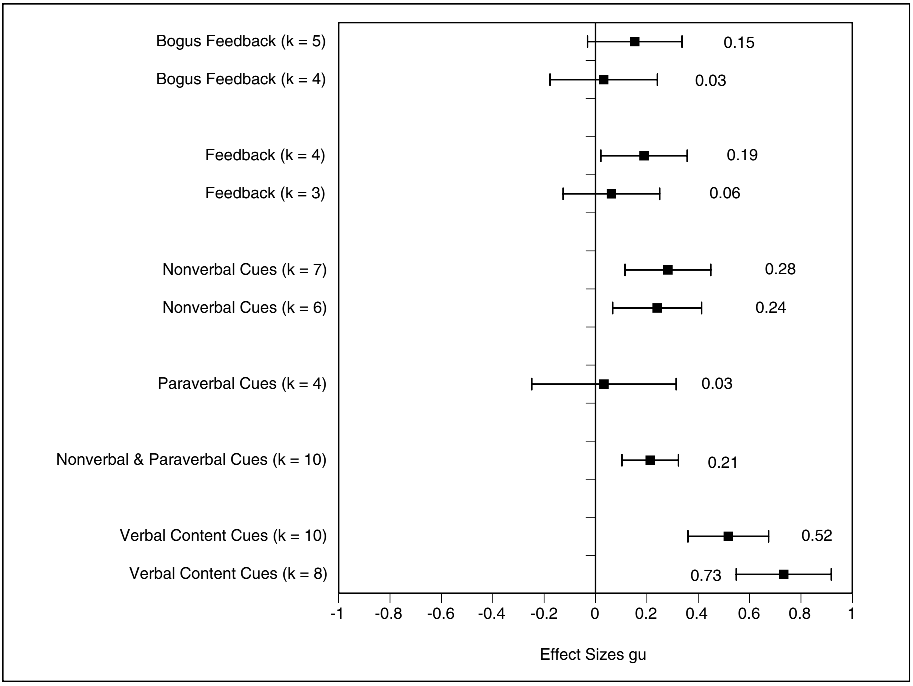

Figure 7. Overview of meta-analyses for different types of training.

Effect Size Analyses for Different Training Types

To evaluate differences between the contents of training, all training programs were classified into eight different types: bogus feedback, feedback, nonverbal cues, paraverbal cues, nonverbal and paraverbal cues, nonverbal and paraverbal cues and feedback, verbal content cues, and verbal content and nonverbal and paraverbal cues. This approach involved synthesizing all studies using a particular training separately as tests for the efficacy of these specific training procedures versus a control group (Appendix C). If any training study contained two or more training contents of the same type, the effect sizes were averaged to avoid dependence of effect sizes using the same control groups. Figure 7 displays the weighted average effect sizes and CIs sorted by their effect sizes.

Bogus feedback or training. Two studies (Porter et al., 2007; Zuckerman, Koestner, & Alton, 1984, Exp. 2) implemented bogus feedback, and three studies (Levine et al., 2005, Exp. 1, 2, and 4) conducted a bogus training. The weighted average effect size of these five studies (n = 486 judges) was g_u = 0.153 [−0.030, 0.337], p = .102, with quite a heterogeneous distribution, $Q(4)$ = 13.01, p = .011, I^2 = 69.24, and individual effect sizes ranging from −0.373 to 0.565. If the outlier from Levine et al. (2005, Exp. 4; g_u = 0.565) was removed, the weighted average training effect was g_u = 0.032

[−0.176, 0.241], $p = .760$, and the homogeneity test was no longer significant, $Q(3) = 7.35$, $p = .061$, $I^2 = 59.20$. These results suggest that bogus feedback did not have an effect on detection accuracy.

Accurate feedback. A total of four accurate feedback studies, involving $n = 712$ judges, provided a small weighted average effect size of $g_u = 0.189$ [0.022, 0.357], $p = .027$, indicating that judges who were given feedback were slightly better than untrained judges. But the results were quite heterogeneous, $Q(3) = 15.21$, $p = .002$, $I^2 = 80.27$, primarily due to an outlier by Zuckerman, Koestner, and Colella (1985; $g_u = 0.692$). Removing this outlier, the weighted average effect size became nonsignificant, $g_u = 0.062$ [−0.126, 0.250], $p = .518$, indicating that feedback had no effect.

Nonverbal cues. Seven hypothesis tests ($n = 559$) resulted in effect sizes ranging from −0.339 (Vrij & Graham, 1997, Exp. 2) to 0.849 (Vrij & Graham, 1997, Exp. 1) for studies training on nonverbal cues only. The weighted average effect size was $g_u = 0.282$ [0.115, 0.449], $p = .001$, but rather heterogeneous, $Q(6) = 13.73$, $p = .033$, $I^2 = 56.29$. Removing the outlier (Vrij & Graham, 1997, Exp. 1) resulted in $g_u = 0.240$ [0.067, 0.413], $p = .007$, and a nonsignificant homogeneity test statistic. Thus, nonverbal cue training had a small positive effect on detection accuracy—with and without the outlier.

Paraverbal cues. A nonsignificant weighted average effect size of 0.033 [−0.247, 0.314], $p = .815$, occurred for four paraverbal cue training studies ($n = 194$). Although effect sizes ranged from a minimum of −0.397 (deTurck et al., 1997) to a maximum of 0.842 (DePaulo et al., 1982), the homogeneity test statistic yielded a nonsignificant value, $Q(3) = 6.60$, $p = .086$, $I^2 = 54.54$. Thus, on average, training with paraverbal cues had no effect on detection accuracy.

Combination of nonverbal and paraverbal cues. Ten studies with a total of $n = 1,308$ judges evaluated a training with a combination of nonverbal and paraverbal cues, yielding a significant training effect of $g_u = 0.213$ [0.103, 0.323], $p < .001$. The minimum effect size was $g_u = -0.480$ (Blair, 2009) and the maximum was $g_u = 1.360$ (Fiedler & Walka, 1993), yielding a quite heterogeneous distribution, $Q(9) = 69.37$, $p < .001$, $I^2 = 87.03$, with several outliers on either side of the distribution (standardized residuals larger than |2.5|).

Combination of nonverbal and paraverbal cues with feedback. Only three studies involving a total of $n = 488$ judges conducted training with a combination of nonverbal cues and feedback (Vrij, 1994; $g_u = 0.485$), or a combination of nonverbal and paraverbal cues and feedback (deTurck, Harszlak, Bodhorn, & Texter, 1990: $g_u = 0.541$; Fiedler & Walka, 1993: $g_u = 1.495$) reporting medium to very large positive effect sizes. Due to a quite heterogeneous effect size distribution, $Q(2) = 8.32$, $p = .016$, $I^2 = 74.06$, no weighted average effect size was calculated.

Verbal content cues. Ten hypothesis tests ($n = 645$) yielded effect sizes ranging from $g_u = -0.429$ (Feeley & deTurck, 1997) to $g_u = 1.165$ (Colwell et al., 2009), resulting in a quite heterogeneous effect size distribution, $Q(9) = 31.39$, $p < .001$, $I^2 = 71.33$. Meta-analysis resulted in a medium size training effect of 0.517 [0.359, 0.674], $p < .001$. Analysis of outliers suggested two studies (Feeley & deTurck, 1997; Sporer et al., 2000) as outliers. When they were removed, the weighted average training effect turned out to be even larger ($g_u = 0.733$ [0.547, 0.918], $p < .001$), with a homogeneous effect size distribution, $Q(7) = 8.78$, $p < .269$, $I^2 = 20.25$.

Combination of nonverbal, paraverbal, and verbal content cues. Three studies trained judges ($n = 190$) with a combination of nonverbal, paraverbal, and verbal content cues. The results were quite contradictory, with a significant negative effect size of $g_u = -0.672$ ([−1.297, −0.047], Kassin & Fong, 1999), a nonsignificant effect size of $g_u = -0.358$ ([−0.941, 0.224], Hendershot, 1981), and a large positive effect size of $g_u = 1.261$ ([0.785, 1.737]; Blair, 2009). Due to the large heterogeneity, $Q(2) = 29.85$, $p < .001$, $I^2 = 93.30$, as well as the small number of studies, no synthesis was attempted.

Results for OGPP Designs

The two OGPP studies used a multimedia training system called Agent99 Trainer (see Table 4). Because the studies by Crews, Cao, Lin, Nunamaker, and Burgoon (2007) and George, Biros, Adkins, Burgoon, and Nunamaker (2004) did not report the correlation between pretest and posttest outcomes, no meta-analysis in the same metric as the previously reported effect sizes could be conducted. As reported in Table 4, consistent medium to large positive effect sizes ranging from $d_{\text{OGPP}} = 0.474$ to $d_{\text{OGPP}} = 1.566$ were found, with quite a large unweighted mean effect size ($d_{\text{OGPP}} = 0.973$). If effect sizes were weighted by sample size, the average effect size was $d_{\text{OGPP}} = 0.693$. Thus, a large standardized pre- to posttest change effect size was observed as detection accuracy was higher after training than before.

Results for PPWC Designs

PPWC studies used different forms of training (see Table 5). None of the six PPWC designed studies reported information about the correlation between pre- and posttest measures, so that no meta-analysis could be conducted. However, the standardized mean change effect size was calculated for each training group (Table 5). Effect sizes ranged from $d_{\text{PPWC}} = -1.112$ (Porter et al., 2000) to $d_{\text{PPWC}} = 1.161$ (Blair, 2006), revealing quite a heterogeneous distribution. The unweighted average effect size was $d_{\text{PPWC}} = 0.203$ ($n = 647$); weighted by sample size, it was $d_{\text{PPWC}} = 0.180$. Therefore, on average, the training group might have a small advantage in pretest-posttest change compared with the control group regarding their detection accuracy.

Discussion

This meta-analysis showed that training improved the overall ability to detect deception with a small to medium effect size. This finding is especially encouraging if we think about the disillusioning 54% detection accuracy found in Aamodt and Custer's

Table 4. Summary of Effect Sizes for OGPP Studies.

Authors (year)	Trained group	*n*	d_{OGPP}
Crews, Cao, Lin, Nunamaker, and Burgoon (2007)	1. Agent 99 Trainer	15	1.566
	2. Lecture-Training	14	0.992
	Combined	29	1.362
George, Biros, Adkins, Burgoon, and Nunamaker (2004)	1. Video-Training ("Control")	28	0.697
	2. Agent 99 Trainer	40	0.476
	3. Agent 99 T. + Ask questions (Qs)	29	0.430
	4. Agent 99 T. + Ask Qs + More Content	42	0.474
	5. Agent 99 T. + Ask Qs + More Content + Quiz	38	0.708
	Combined	177	0.583
Unweighted *M*		206	0.973
Weighted *M*[a]		206	0.693

Note. OGPP = one-group pretest-posttest; n = sample size; d_{OGPP} = effect size *d* for OGPP studies; T = Trainer.
[a]Effect sizes weighted by total sample size.

(2006), as well as Bond and DePaulo's (2006), meta-analyses. However, the mean training effects observed in our meta-analysis were not as strong as those in Driskell's (2012) meta-analysis, and many of the cautions spelled out in Frank and Feeley's (2003) summary still apply for our updated set of studies. Lie accuracy increased with training while there was no significant effect on truth accuracy.

As training effects varied widely, we took a closer look at subgroups of studies differing in training content and other variables to identify the most promising approaches.

Which Trainings Appear Most Promising for Overall Detection Accuracy?

Training on verbal content cues. As hypothesized, training with verbal content cues had the largest training effect on detection accuracy. This could be due to theoretically more differentiated and empirically tested assumptions (e.g., CBCA, RM, and ARJS criteria; Köhnken, 2004; Masip et al., 2005; Sporer, 1998, 2004) of this approach. Also, DePaulo et al. (2003) found higher effect sizes for CBCA and RM verbal content cues than for nonverbal and paraverbal cues in their meta-analysis (but a large-scale meta-analysis of individual verbal content cues is still wanting).

Focusing on verbal content rather than on heuristic cues like nonverbal behavior has also been demonstrated to result in higher detection rates in a series of recent studies based on dual process theories of credibility attribution (Reinhard, Sporer, & Scharmach, 2013; Reinhard, Sporer, Scharmach, & Marksteiner, 2011).

Table 5. Summary of Effect Sizes for PPWC Studies.

Authors (Year)	Trained group vs. control group	N_{CG}	N_{EG}	d_{PPWC}
Costanzo (1992)	1. Lecture	35	35	−0.005
	2. Practice	35	35	0.715
	Combined	35	70	0.363
Porter, Woodworth, and Birt (2000)	1. Feedback	32	32	−0.341
	2. Feedback and cues	32	31	−1.112
	3. Training of parole officers	32	32	−0.041
	Combined	32	95	−0.437
Blair and McCamey (2002)	Behavior Analysis Interview-Training	25	27	0.488
Dziubinski (2003)	Different trainings	28	89	0.410
George, Marett, Burgoon, Crews, Cao, Lin, and Biros (2004)	1. Agent99 Trainer	29	26	−0.663
	2. Lecture and combination	29	59	−0.035
	Combined	29	85	−0.489
Blair (2006)	1. Training	40	49	1.161
	2. Training and response bias information	40	43	0.573
	Combined	40	92	0.882
Unweighted *M*		189	458	0.203
Weighted[a] *M*		189	458	0.180

Note. PPWC = pretest-posttest with control; N_{CG} = sample size of control group; N_{EG} = sample size of experimental group; d_{PPWC} = effect size *d* for PPWC studies.
[a]Effect sizes weighted by total sample size.

We found additional support for verbal content training within the second meta-analytic approach, where these programs showed a medium size training effect, with the exception of studies by Feeley and deTurck (1997) and Köhnken (1987) who obtained negative training effects.

Multichannel studies. Studies with the use of multichannel training programs showed only a small training effect. The second meta-analytic approach supported this finding: Training programs using only paraverbal cues yielded no training effect, whereas training programs using only nonverbal cues, or a combination of nonverbal and paraverbal cues, showed a marginal training effect. Considering that recent meta-analyses found either no or only faint relations for most nonverbal and paraverbal cues to deception (DePaulo et al., 2003; Sporer & Schwandt, 2006, 2007), people trained to focus on these cues, which were presumably not present in the stimulus material and therefore may simply not be diagnostic for differentiating between truth and deception, are likely to fail.

Although subjective ratings of nonverbal behaviors may be more likely to be associated with deception than more objective frequency counts (DePaulo et al., 2003; DePaulo & Morris, 2004, Hauch et al., 2013), these cues have not been incorporated into the training programs reviewed here (for an exception, see Fiedler & Walka, 1993, who used subjective ratings of *channel discrepancies*).

Feedback. Feedback studies resulted in a small effect for detection accuracy, as it was expected from the "law of effect" (Thorndike, 1913, 1927). In contrast to the medium effect size (d = 0.41) from Kluger and DeNisi's (1996) meta-analysis on all kinds of feedback interventions, we found a markedly smaller effect of g_u = 0.19. This difference may be explained by the fact that participants in the feedback studies reviewed here only learned about the outcome (truth or lie) and not upon which cues they should have based their judgment (Fiedler & Walka, 1993). In other words, if trainees only learn about the outcome of their judgments, but not what they may have done right or wrong in evaluating signs of deceit or truths, and how to weight these signals (the process of lie detection), we cannot really expect large effects from feedback in this domain.

This is not to say that feedback could not be *more* beneficial than in the studies reviewed here. For example, training studies using Agent99 Trainer (comprehensive computer training program using a combination of nonverbal, paraverbal, and verbal content cues) evaluated not only the final outcome but also an increase in knowledge, which was tested via pop-up quizzes (Biros, Sakamoto, et al., 2005; George, Marett, Burgoon, et al., 2004). Similarly, more interactive approaches (e.g., Agent99 Trainer) where trainees navigate through the materials taught may be promising, as discussed in Lin, Crews, Cao, Nunamaker, and Burgoon (2003). Unfortunately, reports of these studies (Crews et al., 2007; George, Marett, Burgoon, et al., 2004) did not provide enough statistical details necessary for meta-analytic synthesis or information about the precise content of the training itself.

Finally, neither bogus feedback nor bogus training had any positive or deteriorating effects.

Combinations of approaches. Compared with the mere feedback approach, combining information about nonverbal and paraverbal cues and providing feedback may be more promising. The second meta-analytic approach found that three studies implementing this technique yielded quite large training effects (especially the study by Fiedler & Walka, 1993). Here, participants seemed to have learned to detect particular cues they were searching for and applied them appropriately to make a lie-truth judgment (Fiedler & Walka, 1993). This was demonstrated by conducting a Brunswikian lens model analysis that tests whether people actually use ecologically valid cues for their judgments (Fiedler & Walka, 1993; Hartwig & Bond, 2011; Reinhard et al., 2011; Sporer, Masip, & Cramer, 2014).

Combinations of cue training without feedback (nonverbal, paraverbal, and verbal content) were adapted in three studies indicating quite contradictory effects. For instance, training with the Reid Technique, which is very popular in the United States, resulted in a large positive training effect in Blair's (2009) study, while it showed a detrimental effect in the study by Kassin and Fong (1999). Methodological differences in participant samples or in stimuli (suspect interviews in actual theft cases in the former, interviews after a mock crime in the latter study) might explain the divergent outcomes.

Is Training Equally Effective for Lie and Truth Accuracy?

Surprisingly, trainings improved only lie accuracy but not truth accuracy (except for verbal content trainings, which were more successful with classifying true stories correctly). To understand this finding, the purpose of training has to be taken into account, which turned out to be an important moderator variable.

When judges were trained to focus on the truth (e.g., using credibility criteria, such as CBCA, which are usually positively associated with veracity; see Landry & Brigham, 1992), truth accuracy was increased; when trained with cues to deception (e.g., speech errors, or adaptors; see deTurck et al., 1997), there was no training effect for truth accuracy. As one cannot infer causal relationships from meta-analytic findings, more direct evidence for a potential response bias shift as a function of training purpose comes from Masip et al.'s (2009) fake training study. They trained participants in two experiments (PPWC design) either to detect deception (with nondiagnostic deception cues), or to detect the truth (with nondiagnostic truthfulness cues), or did not train them at all. Regardless of accuracy, a strong shift in response bias toward the respective direction of the trained cues was found, whereas no response bias shift was observed for untrained control participants. Consequently, the truth bias invoked by teaching verbal content *truth* criteria (as in the studies in this meta-analysis) usually is likely to result in a truth bias, and hence in a veracity effect (Levine et al., 1999). Unfortunately, it was not possible to test for response bias shifts in this meta-analysis due to missing information (about the truth-lie judgments regardless of accuracy) in most studies. Due to the fact that all content training studies reviewed here utilized truth criteria, in future studies and in training attempts, should be made to avoid such a truth bias shift.

In contrast, no moderating effect of purpose was found for lie accuracy, except for an outlier effect size by Levine et al. (2005, Exp. 4) that was excluded.

Comparison With Driskell's Meta-Analytic Findings

In the introduction, we addressed several methodological issues in Driskell's (2012) meta-analysis that may have affected his findings to be different from ours. To begin with, Driskell found a medium training effect in detection accuracy of $d = 0.50$ compared with a smaller $d = 0.33$ in this meta-analysis. This difference is probably due to three facts: First, Driskell calculated the weighted average of dependent training groups and treated them as if they were independent. This contradicts the assumption of independent data points in a meta-analysis (Lipsey & Wilson, 2001). We avoided this problem by averaging the individual effect sizes if more than one training group was applied, and by sorting these training groups into one of eight categories to calculate further sub-meta-analyses.

Secondly, our meta-analysis included more studies, both published and unpublished. Thirteen relevant studies (posttest only with control group design) conducted between 1981 and 2009 were not contained in Driskell's meta-analysis. A separate meta-analysis of these 13 omitted studies yielded a nonsignificant training effect of $d = 0.08$ ($p = .202$). Because Driskell's meta-analysis did not contain these studies, his weighted average effect size of $d = 0.50$ appears to overestimate the training effect. Furthermore, we analyzed eight additional training studies implementing other experimental designs.

Third, Driskell included only published studies, which is likely to lead to a publication bias, especially when we think of our results that unpublished studies revealed a smaller training effect than published studies.

Despite these differences, there is an interesting feature of Driskell's review our analyses could not address. He sorted the trained cues according to DePaulo et al.'s (2003) analytical approach. This analysis showed that training programs might be more effective if certain deception cues were included (e.g., cues reflecting more tension, discrepancy, fewer details, fewer illustrators, or phrase repetitions).

Methodological Implications

Reporting standards. There were various shortcomings regarding the reporting of important independent and dependent variables. To evaluate the differential effectiveness of specific training characteristics, their detailed documentation is necessary, which many studies failed to do. To facilitate planning, analyzing, replicating, and comparing training studies in the future, we outline several methodological recommendations.

Most importantly, training studies should report not only overall detection accuracy but also separately lie and truth accuracy rates. In order to investigate the relation between response bias, training, purpose of the training (see Masip et al., 2009), and training effects, means and standard deviations for lie and truth judgments in addition to detection accuracy, lie and truth accuracy are necessary. Signal detection theory (Green & Swets, 1966), as suggested by Meissner and Kassin (2002), Sporer (2004), and Masip et al. (2009), should be utilized to differentiate training effects from response bias.

Researchers should draw on a theoretical framework to specify hypotheses and to design training interventions and their components. To understand and evaluate the effectiveness of these components, the very content of a training program and the specific cues should always be described, especially for multichannel programs or combinations of different training strategies. Without these details, the chances to replicate training success are practically impossible. Information about the training procedure (intensity, duration, group size, trainer presence, etc.) should always be described.

Experimental design issues. With regard to experimental designs (see Table 1), we make several methodological suggestions. Although most training studies were designed with the POWC, Campbell and Stanley (1963) cautioned researchers about the fact that the experimental and control groups may not be equal before treatment. They further recommended always randomly assigning participants to the conditions. Using a pretest could establish group equivalence before interventions. With the OGPP, the pretest-posttest changes may have been due to a training effect, but could also have been produced by other change-producing events (*history*), physiological or psychological processes such as fatigue or boredom (*maturation*), or to the effect of *testing* itself.

The most extensive and time-consuming PPWC was utilized in six studies. This design controls for many alternative explanations and allows for a better understanding of the underlying mechanisms (for an excellent discussion of program evaluation research, in particular, process and outcome evaluation, see Rossi, Lipsey, & Freeman, 2009).

Whenever researchers decide to train professionals, they should include a control group of lay persons. If only police officers were trained without a pretest (e.g., Köhnken, 1987; Vrij, 1994), or without a control group of lay persons, we do not know how accurate these officers were without training, or whether the professional groups' detection accuracy differs compared with the ability of lay persons. When only professional groups are compared before and after training, we do not know whether the pretest itself sensitized them to become better, or if lay persons may have performed better (see Shadish, Cook, & Campbell, 2002).

A specific issue in deception research is the question whether a sender provides both a truthful and a deceptive account (within-participants design), or only one account (between-participants design). Surprisingly, our results showed no differences regarding training effectiveness.

Another deception specific question is whether researchers should or should not inform their participants about the lie/truth base rates of senders' stories. Detection accuracy was higher for trained compared with control judges in studies with base rate information. Because base rates are rarely known in real life, using 50-50 base rates as in most studies may jeopardize ecological validity of findings when different base rates are likely for certain types of lies.

Problems of ecological validity: sender motivation. Last but not least, a prevailing problem of all deception studies is the lack of negative consequences to senders in cases of detection (Miller & Stiff, 1993). The question arises whether money or the awareness that a mock crime was staged motivates participants in an experiment sufficiently to be comparable with suspects in police interrogations, defendants in the courtroom, or other high stake situations such as business or political negotiations, or in personal relationship contexts.

Although the moderator of sender motivation across all studies did not show any differences, there were simply too few studies with high sender motivation to draw any firm conclusions. When looking at studies focusing on nonverbal cues only, effect sizes were larger when sender motivation was medium (g_u = 0.451) than with no motivation (g_u = 0.011). For verbal content cues trainings, the two studies that used high motivation material yielded nonsignificantly larger effect sizes (g_u = 0.895) than the five no motivation studies (g_u = 0.590). This could be interpreted as evidence for a motivational impairment effect, but other differences between the small sets of studies could also be responsible for the (lack of) differences.

Perhaps, researchers should attempt to apply paradigms in which senders are motivated by the opportunity to escape mild punishment (within ethical limits) in case of detection, rather than being motivated by money or through a mock crime. For example, researchers could give participants money first for taking part in the experiment, but tell them that they have to pay back large portions if their lies are detected. For applications in criminal justice contexts, researchers might also use corroborated cases of perjury, or of true versus false alibis.

Practical Implications

On the basis of the present meta-analytic results, we venture some recommendations for conducting training programs to maximize training effects.

Content of the training. Because training programs including verbal content cues (such as CBCA, RM, or ARJS) led to highest training effects, we recommend the use of these cues in future training programs. Cue selection should be based on effect sizes from past studies, not on vote-counting (as in Masip et al., 2005; Vrij, 2005, 2008). Because some of the training effects and the cues used may be domain specific, trainings should take that into account when selecting cues.

As most verbal content training studies utilized truth criteria, attempts should be made to avoid a response bias toward the truth. Perhaps, using feedback in addition to verbal content cues might be worth exploring.

Presentation format. A training program should include written instructions about the training content—either on its own or in combination with a (video) lecture session. Surprisingly, if participants were trained via a (video) lecture session only, the training was not effective at all. This could be due to the opportunity for participants to reread instructions and internalize the training content at their own pace and return to any section for clarification.

Use of examples. Trainees should practice their abilities with examples from different senders, although we did not find an advantage for using examples per se. Trainees should learn to become more familiar with the cues and their coding/rating with different types of accounts from different contexts. Also, practice has been demonstrated to improve reliability of coding (Küpper & Sporer, 1995).

Duration and number of training sessions. The tendency that longer training sessions are more effective than shorter ones leads us to recommend a minimum length of 60 minutes or more, depending on content and use of examples.

Even when short-term training effects could be demonstrated, long-lasting training effects should be investigated using follow-up posttests with different delay intervals to capture long-lasting training effects. We recommend multiple training sessions (e.g., Akehurst et al., 2004) for professionals such as forensic psychologists, police officers, or judges to ensure that the training content will be retained and refreshed (for general guidelines on how to conceptualize training programs, see Docan-Morgan, 2007).

Counteracting biases of professional groups. As outlined earlier, some professional groups tend to have a lie bias (e.g., Meissner & Kassin, 2002). Therefore, training these professionals with verbal content cues related to the truth may counteract their lie bias.

Furthermore, professionals who have been involved in the practice of detection of deception for years might be somewhat reluctant to accept training contents offered by psychologists, particularly when the stimulus materials appear to lack face validity. For example, police officers might be "out of practice at being taught" (Akehurst et al., 2004, p. 888) and hence show difficulties learning a comprehensive credibility assessment method, which may contradict some of their personal on the job experience.

In general, we suggest that a training program should be customized to the specific needs, level of knowledge, and expertise of law enforcement personnel (see also Docan-Morgan, 2007).

Limitations

In the following, we address four potential limitations that should be kept in mind when interpreting the results of this meta-analysis.

First, when looking at mean effect sizes, one should never do this without taking into account their variance. For any meta-analysis, different subclasses of persons, training programs, outcomes, settings, or times can lead to large heterogeneity (Matt & Cook, 2009). As we observed large heterogeneity, especially for overall detection accuracy and truth accuracy, we addressed this issue by conducting several moderator analyses. (We also calculated random effects models that yielded comparable conclusions, available from the authors.)

Second, conducting several moderator variable analyses by blocking studies into different subgroups (as is done in many meta-analyses) may lead to a confounding of moderator variables, which poses a threat to generalized inferences (Matt & Cook, 2009; Pigott, 2012). In the present meta-analysis, many moderator variables were at least partially confounded with each other. Therefore, we had inspected cross tables of moderators and their intercorrelations to assure a sufficient number of studies in each subgroup for tests to approach independence (analogous to orthogonality of contrasts in ANOVA). Meta-regression analyses may have been a better solution to this problem (Pigott, 2012), but the limited number of studies made us decide against this solution.

Third, our moderator analysis yielded a publication bias (Lipsey & Wilson, 1993; Sporer & Cohn, 2011) despite our attempts to avoid such a bias by including almost 30% unpublished studies. All authors were contacted and asked for further (unpublished or submitted) experimental training studies in order to counteract publication bias before meta-analytic syntheses were conducted.

Fourth, all training studies were laboratory experiments in which independent variables were varied and manipulated. Because the ground truth in real world settings cannot be established with certainty, it is a challenge for researchers to create and evaluate training programs with real life events (e.g., witness or suspect statements).

Conclusions

Although training studies were quite heterogeneous with respect to their effect size, content, and operationalization, we found a small to medium training effect for overall detection accuracy and lie accuracy, but not for truth accuracy. Truth accuracy was only improved if verbal content cues to detect the truth were utilized, although this result should be interpreted with caution, because it could be due to a shift in response bias toward correctly detecting the truth. Training with verbal content cues yielded the highest training effect, whereas training with nonverbal cues, paraverbal cues, or feedback resulted in quite small or nonsignificant training effects. Therefore, researchers and practioners should not base their trainings on these unreliable cues but focus on verbal content training.

Appendix A

Summary of Excluded Studies and Reason for Exclusion

Authors	Reason for exclusion
Akehurst, Bull, Vrij, and Köhnken (2004)	Lack of statistical data for computing an effect size
Biros (2004)	Review of Biros et al. (2002) and Cao et al. (2003)
Biros, George, and Zmud (2002)	Task for judges was to find error in complex working situations instead of statements
Biros, George, and Zmud (2005)	Summary of Biros et al. (2002) with implications
Biros, Hass, Wiers, Twitchell, Adkins, Burgoon, and Nunamaker (2005)	No deception detection training study
Biros, Sakamoto, Geroge, Adkins, Kruse, Burgoon, and Nunamaker (2005)	Cue knowledge score investigated (pop-up quizzes; no detection accuracy)
Blair, Levine, and Shaw (2010)	Participants received additional information about the context/situation, but were not trained
Burgoon, Nunamaker, George, Adkins, Kruse, and Biros (2007)	Grant report that includes two training studies (George, Marett, et al., 2004, and George, Biros, et al., 2004, which are both included)
Cao, Crews, Lin, Burgoon, and Nunamaker (2003)	Same data set as Crews, Cao, Lin, Nunamaker, and Burgoon (2007)

(continued)

Appendix A (continued)

Authors	Reason for exclusion
Cao, Lin, Deokar, Burgoon, Crews, and Adkins (2004)	Usability study (no deception detection training study)
Cao, Crews, Nunamaker, Burgoon, and Lin (2004)	Usability study (no deception detection training study)
Clark (1983)	Not retrievable
Dando and Bull (2011)	Training study, but lack of untrained control group or pre-test; only five trainees.
Elaad (2003)	Lack of statistical data for computing an effect size
Enos, Shriberg, Graciarena, Hirschberg, and Stolcke (2007)	No training study, computer program used—no human judges
Ford (2004)	Detection of five specific cue categories (emotional, arousal, memory, cognitive effort, communication tactics) measured between pre- and posttest, no data provided for detection accuracy overall
Geiselman, Elmgren, Green, and Rystad (2011)	Training program is based upon cues derived from the same stimulus material (Exp. 2) that the experimental (training) group also rated later (Exp. 3)
George, Biros, Burgoon, and Nunamaker (2003)	Same data set as George, Marett, Burgoon et al. (2004)
George, Biros, Burgoon, Nunamaker et al., (2008)	Summary of two training studies (George, Marett, et al., 2004, and George, Biros, et al., 2004)
George, Marett, and Tilley (2004)	No deception detection training study
Hill and Craig (2004)	Detection of pain in facial expressions
Horvath, Jayne, and Buckley (1994)	No between- or within-participants design
Y. C. Lin (1999)	Not retrievable
M. Lin, Crews, Cao, Nunamaker, and Burgoon (2003)	Article reports results from Cao et al. (2003)
Mann, Vrij, and Bull (2006)	No training or feedback
Marett, Biros, and Knode (2004)	Relationship between training and accuracy not investigated
Masip, Alonso, Garrido, and Herrero (2009)	No purpose of improving detection accuracy
McKenzie, Scerbo, and Catanzaro (2003)	No deception detection training study
Parker and Brown (2000)	Training of only two individuals was not clearly described; no usable results of means/detection accuracy
Porter, Juodis, ten Brinke, Klein, and Wilson (2010)	Lack of statistical data for computing an effect size
Seager (2001)	No specific detection deception training
Warren, Schertler, and Bull (2009)	Training with facial (micro-)expression tools; lack of control group
Yang (1996)	Not retrievable

Appendix B

Coding Decisions for All Variables

Table B1. Coding of General Study Characteristics and Characteristics From Judges.

Authors (Year)	Publ. Status	Type of Publ.	Research Group	Occupation	*M* Age	*SD* Age	Males	Females	Motivation	Rating medium	# Judgments	Randomization
DePaulo, Lassiter, and Stone (1982)	publ.	PR	I Study	Students	na	na	22	22	no	audiov.	72	randomized
Zuckerman, Koestner, and Alton (1984; Exp. 1)	publ.	PR	Zuckerman	Students	na	na	69	63	no	audiov.	8	na
Zuckerman, Koestner, and Colella (1985)	publ.	PR	Zuckerman	Students	na	na	60	47	no	na	64	randomized
Köhnken (1987)	publ.	PR	I Study	Police officers	28.70	7.10	80	0	no	audiov.	4	randomized
Hall (1989)	unpubl.	Diss.	Others	Students	na	na	93	162	low	audiov.	28	randomized
deTurck and Miller (1990)	publ.	PR	deTurck/Feeley/ Levine	Students	na	na	na	na	no	audiov.	16	randomized
deTurck, Harszlak, Bodhorn, and Texter (1990)	publ.	PR	deTurck/Feeley/ Levine	Students	na	na	na	na	no	audiov.	8	na
deTurck (1991)	publ.	PR	deTurck/Feeley/ Levine	Students	na	na	na	na	no	audiov.	16	na
Landry and Brigham (1992)	publ.	PR	I Study	Students	na	na	na	na	no	na	12	randomized
Fiedler and Walka (1993)	publ.	PR	I Study	Students	na	na	na	na	no	audiov.	40	randomized
Vrij (1994)	publ.	PR	Vrij	Police Officers	37.00	na	331	29	no	audiov.	20	randomized
deTurck, Feeley, and Roman (1997)	publ.	PR	deTurck/Feeley/ Levine	Students	na	na	na	na	no	audiov.	8	randomized
Feeley and deTurck (1997)	publ.	PR	deTurck/Feeley/ Levine	Students	na	na	65	57	no	audiov.	4	na
Vrij and Graham (1997; Exp. 1)	publ.	PR	Vrij	Students	21.00	na	na	na	no	audiov.	20	na
Vrij and Graham (1997; Exp. 2)	publ.	PR	Vrij	Police Officers	34.00	na	na	na	no	audiov.	20	na
Kassin and Fong (1999)	publ.	PR	I Study	Students	na	na	11	29	no	audiov.	8	randomized
Sporer, Samweber, and Stucke (2000)	unpubl.	PR	Sporer	Students	na	na	54	54	no	transcr.	16	na
Santarcangelo, Cribbie, and Ebesu Hubbard (2004)	publ.	PR	I Study	Students	20.80	na	16	81	no	audiov.	60	randomized

(continued)

Table B1. (continued)

	Publ.	Type of								Rating		
Authors (Year)	Status	Publ.	Research Group	Occupation	*M* Age	*SD* Age	Males	Females	Motivation	medium	# Judgments	Randomization
Levine, Feeley, McCornack, Hughes, and Harms (2005; Exp. 1)	publ.	PR	deTurck/Feeley/ Levine	Students	19.91	1.36	82	174	no	audiov.	16	randomized
Levine et al. (2005; Exp. 2)	publ.	PR	deTurck/Feeley/ Levine	Students	21.51	1.49	26	64	no	audiov.	16	randomized
Levine et al. (2005; Exp. 4)	publ.	PR	deTurck/Feeley/ Levine	Students	19.83	1.26	86	72	no	audiov.	16	randomized
Hartwig, Granhag, Strömwall, and Kronkvist (2006)	publ.	PR	1 Study	Trainees	28.20	4.00	55	27	low	in pers.	1	randomized
Porter, McCabe, Woodworth, and Peace (2007)	publ.	PR	1 Study	Students	19.95	4.59	32	119	low	audiov.	12	na
Colwell et al. (2009)	publ.	PR	1 Study	Students	na	na	na	na	low	transcr.	30	na
Hendershot (1981)	unpubl.	Thesis	Others	na	na	na	na	na	no	audiov.	32	na
Bailey (2002)	unpubl.	Thesis	Others	Students	na	na	28	72	no	audiov.	30	randomized
Sporer (1993)	unpubl.	PP	Sporer	Students	na	na	20	20	no	transcr.	8	randomized
Blair (2009)	unpubl.	UM	1 Study	Students	na	na	96	64	no	audiov.	10	randomized
Sporer and McCrimmon (1997)	unpubl.	PP	Sporer	Students	na	na	0	60	no	audiov.	8	randomized
Sporer and McFadyen (2001)	unpubl.	PP	Sporer	Students	na	na	16	44	no	transcr.	8	randomized

Note. Publ. = Publication; publ. = published; unpubl. = unpublished; PR = peer review; Diss. = dissertation; PP = paper presented; UM = unpublished manuscript; 1 Study = single publication from deception researcher; na = not available; audiov. = audiovisual; transcr. = transcript; pers. = person;. # = number.

Table B2. Coding of Study Characteristics From Senders.

Authors (Year)	Randomization	Design	# Senders	Males	Females	Stories per sender	Story content	Senders' motivation	Duration truth (s)	Duration lie (s)
DePaulo, Lassiter, and Stone (1982)	randomized	Within	12	6	6	6	Attitude/Liking	None	20.00	20.00
Zuckerman, Koestner, and Alton (1984; Exp. 1)	na	Within	8	4	4	8	Attitude/Liking	None	25.00	25.00
Zuckerman, Koestner, and Colella (1985)	randomized	Within	8	4	4	8	Attitude/Liking	None	25.00	25.00
Köhnken (1987)	randomized	Between	4	na	na	1	Observed Event	Low	272.00	234.50
Hall (1989)	not rand.	Within	14	na	na	4	Attitude/Liking	Low	105.00	105.00
deTurck and Miller (1990)	not rand.	Within	32	16	16	16	Attitude/Liking	Low	na	na
deTurck, Harszlak, Bodhorn, and Texter (1990)	na	Between	32	13	19	1	Staged Live Event	Medium	na	na
deTurck (1991)	na	Within	16	8	8	4	Attitude/Liking	Medium	na	na
Landry and Brigham (1992)	na	Within	12	6	6	2	Sign. Pos./Neg. Event	None	105.00	105.00
Fiedler and Walka (1993)	not rand.	Within	10	5	5	4	Attitude/Liking, Mock Crime	None	150.00	150.00
Vrij (1994)	randomized	Within	20	14	6	2	Staged Live Event	None	44.00	44.00
deTurck, Feeley, and Roman (1997)	randomized	Between	32	na	na	1	Staged Live Event	Medium	na	na
Feeley and deTurck (1997)	randomized	Between	8	4	4	1	Staged Live Event	Medium	176.45	176.45
Vrij and Graham (1997; Exp. 1)	not rand.	Within	10	5	5	2	Staged Live Event	None	30.00	30.00
Vrij and Graham (1997; Exp. 2)	not rand.	Within	10	5	5	2	Staged Live Event	None	30.00	30.00
Kassin and Fong (1999)	not rand.	Within	12	6	6	2	Mock Crime	None	90.00	90.00
Sporer, Samweber, and Stucke (2000)	na	Within	72	36	36	2	Personal Event	None	na	na
Santarcangelo, Cribbie, and Ebesu Hubbard (2004)	na	Between	60	na	na	1	na	None	45.00	45.00
Levine, Feeley, McCornack, Hughes, and Harms (2005; Exp. 1)	na	Within	2	1	1	8	Attitude/Liking	None	na	na
Levine et al. (2005; Exp. 2)	na	Within	2	1	1	8	Attitude/Liking	None	na	na

(continued)

Table B2. (continued)

Authors (Year)	Randomization	Design	# Senders	Males	Females	Stories per sender	Story content	Senders' motivation	Duration truth (s)	Duration lie (s)
Levine et al. (2005; Exp. 4)	na	Within	2	1	1	8	Attitude/Liking	None	na	na
Hartwig, Granhag, Strömwall, and Kronkvist (2006)	randomized	Between	82	27	55	1	Mock Crime	High	720.00	720.00
Porter, McCabe, Woodworth, and Peace (2007)	na	Between	12	na	na	1	Sign. Neg. Event	None	120.00	120.00
Colwell et al. (2009)	na	Between	30	na	na	1	Observed Event/ Mock Crime	High	na	na
Hendershot (1981)	not rand.	Within	16	16	0	2	Mock Crime	High	na	na
Bailey (2002)	na	Within	30	15	15	1	Attitude/Liking, Mock Crime	None	30.00	30.00
Sporer (1993)	na	Within	na	na	na	na	Personal Event	None	na	na
Blair (2009)	randomized	Between	10	na	na	1	Sign. Neg. Event	High	na	na
Sporer and McCrimmon (1997)	na	Within	24	0	24	2	Personal Event	None	69.30	56.30
Sporer and McFadyen (2001)	na	Within	24	0	24	2	Personal Event	None	69.30	56.30

Note. rand. = randomized; # = number; na = not available; Sign. = significant; Pos. = positive; Neg. = negative; s = seconds.

Table B3. Coding of Training Characteristics.

Authors (Year)	N_{CG}	N_{EG}	Training category	Purpose	Duration in minutes	Medium	Examples	Group size[a]	Trainer presence	Base rate info
DePaulo, Lassiter, and Stone (1982)	11	11	Multichannel	Lies	na	Written	na	na	Present	no
Zuckerman, Koestner, and Alton (1984; Exp. 1)	43	89	Feedback	Lies	na	na	na	2	Present	no
Zuckerman, Koestner, and Colella (1985)	63	54	Feedback	Lies	na	na	8	2	Present	na
Köhnken (1987)	20	60	Combination	na	45	Written and Lecture	na	2	Present	no
Hall (1989)	81	281	Feedback	na	na	Lecture Video	4	5	Present	no
deTurck and Miller (1990)	195	195	Multichannel	Lies	30	Written and Lecture	5	3	Present	no
deTurck, Harszlak, Bodhorn, and Texter (1990)	94	94	Multichannel	Lies	30	Demo-Video and Lecture	5	na	Present	yes
deTurck (1991)	91	92	Multichannel	Lies	30	Written, Demo and Lecture	5	na	Present	no
Landry and Brigham (1992)	64	50	Verbal Content	Truth	45	Written and Lecture	0	5	Present	no
Fiedler and Walka (1993)	24	48	Combination	Lies	na	Written	0	2	Present	no
Vrij (1994)	144	216	Combination	Lies	na	Written	na	na	Absent	na
deTurck, Feeley, and Roman (1997)	41	123	Multichannel	Lies	30	Written, Demo, and Lecture	5	na	Absent	yes
Feeley and deTurck (1997)	33	96	Combination	Lies	na	Written	na	2	Absent	no
Vrij and Graham (1997; Exp. 1)	20	20	Combination	Lies	na	Written	0	na	Absent	yes
Vrij and Graham (1997; Exp. 2)	14	15	Combination	Lies	na	Written	0	na	Absent	yes
Kassin and Fong (1999)	20	20	Combination	Lies	50	Written and Lecture Video	0	2	Present	no
Sporer, Samweber, and Stucke (2000)	54	54	Verbal Content	Truth	na	Written	0	na	Present	na
Santarcangelo, Cribbie, and Ebesu Hubbard (2004)	30	67	Combination	Lies	na	Written and Lecture	0	na	Present	no

(continued)

Table B3. (continued)

Authors (Year)	N_{CG}	N_{EG}	Training category	Purpose	Duration in minutes	Medium	Examples	Group size[a]	Trainer presence	Base rate info
Levine, Feeley, McCornack, Hughes, and Harms (2005; Exp. 1)	124	71	Multichannel	Lies	5	Lecture Video	0	na	Absent	no
Levine et al. (2005; Exp. 2)	31	28	Multichannel	Lies	5	Lecture Video	0	3	Absent	no
Levine et al. (2005; Exp. 4)	54	52	Multichannel	Lies	5	Lecture Video	0	3	Absent	no
Hartwig, Granhag, Strömwall, and Kronkvist (2006)	41	41	Verbal Content	Lies	180	Demo-Video and Lecture	4	na	Present	yes
Porter, McCabe, Woodworth, and Peace (2007)	50	51	Feedback	na	na	Lecture	0	na	Present	no
Colwell et al. (2009)	10	10	Verbal Content	Lies	180	Written and Lecture	3	na	Present	na
Hendershot (1981)	14	14	Multichannel	na	120	Lecture	15	4	Present	no
Bailey (2002)	50	50	Multichannel	Lies	5	Lecture	3	2	Present	no
Sporer (1993)	20	20	Verbal Content	Truth	na	Written	0	na	na	no
Blair (2009)	40	120	Combination	Lies	na	Demo-Video and Lecture	2	na	Present	na
Sporer and McCrimmon (1997)	30	30	Verbal Content	Truth	na	Written	0	1	Absent	no
Sporer and McFadyen (2001)	30	30	Verbal Content	Truth	na	Written	0	1	Absent	no

Note. N = sample size; CG = control group; EG = experimental group; na = not available.

[a] Coding for group size (in persons): 1 = 1-2, 2 = 3-6, 3 = 7-10, 4 = 11-20, 5 = 20-30.

Appendix C

Summary of Individual Training Groups, Sample Size, Training Content, and Coding for Type of Training

Authors (Year)	Training group (#)	N_{EG}	Training content	Coding: Type of training
DePaulo, Lassiter, and Stone (1982)	Attend to Tone (1)	11	Attention to voice	Paraverbal Cues
	Attend to Word (2)	11	Attention to spoken message	Verbal Content Cues[a]
	Attend to Visual (3)	11	Attention to nonverbal signs	Nonverbal Cues[a]
Zuckerman, Koestner, and Alton (1984; Exp. 1)	(4 After) Feedback (1)	22	FB given after first 4 judgments for each sender	Feedback
	(8 After) Feedback (2)	21	FB given after each of 8 judgments	Feedback
	(4 Before) Feedback (3)	22	FB given before first 4 judgments	Feedback
	Mixed (4)	24	FB given before first 4 judgments and after last 4	Feedback
Zuckerman, Koestner, and Alton (1984; Exp. 2)	(4 Before) Feedback (1)	20	FB given before first 4 judgments	Feedback[a]
	Bogus (after 8) Feedback (2)	19	Bogus FB given after all 8 judgments (half correct, half false)	Bogus Feedback
Zuckerman, Koestner, and Colella (1985)	Feedback (1)	54	Feedback	Feedback
Köhnken (1987)	Nonverbal Training (1)	20	Head movements, eye blink, gaze, illustrators, adaptors, body movements, and leg and foot movements	Nonverbal Cues
	Speech Training (2)	20	Speech rate, filled pauses, word fragments, stuttering, repetitions, self-reflections, parenthetic remarks, corrections, false starts, diversity of vocabulary, syntax complexity	Paraverbal Cues
	Verbal Content Training (CBCA; 3)	20	Logical consistency, amount of detail, space-time interrelationships, accounts of unusual details, spontaneous details	Verbal Content Cues
Hall (1989)	Mixed Feedback (1)	99	Feedback 2 before and 2 after statement (in training session)	Feedback
	Before Feedback (2)	94	Feedback before statement (in training session)	Feedback
	After Feedback (3)	88	Feedback after statement (in training session)	Feedback

(continued)

Appendix C (continued)

Authors (Year)	Training group (#)	N_{EG}	Training content	Coding: Type of training
deTurck and Miller (1990)	Nonverbal and Paraverbal Training (1)	195	4 Paraverbal cues: response latency, message duration, pauses, speech errors 2 Nonverbal cues: adaptors, hand gestures	Nonverbal and Paraverbal Cues
deTurck, Harszlak, Bodhorn, and Texter (1990)	Nonverbal and Paraverbal Training and Feedback (1)	94	4 Paraverbal cues: response latency, message duration, pauses, speech errors 2 Nonverbal cues: adaptors, hand gestures	Nonverbal and Paraverbal Cues and Feedback
deTurck (1991)	Nonverbal and Paraverbal Training (1)	91	4 Paraverbal cues: response latency, message duration, pauses, speech errors 2 Nonverbal cues: adaptors, hand gestures	Nonverbal and Paraverbal Cues
Landry and Brigham (1992)	CBCA Training (1)	50	14 CBCA-Criteria (logical structure, quantity of details, contextual embedding, descriptions of interactions, reproduction of conversation, unexpected complications during the incident, unusual details, superfluous details, accounts of subjective mental state, attribution of perpetrator's mental state, spontaneous corrections, admitting lack of memory, raising doubts about one's own testimony, self-deprecation)	Verbal Content Cues

(continued)

Appendix C (continued)

Authors (Year)	Training group (#)	N_{EG}	Training content	Coding: Type of training
Fiedler and Walka (1993)	Nonverbal Training (1)	24	7 Cues: disguised smiling, lack of head movements, self-adaptors, increased pitch, reduced speech rate and pauses, channel discrepancies	Nonverbal and Paraverbal Cues
	Nonverbal and Paraverbal Training and Feedback (2)	24	7 Cues: disguised smiling, lack of head movements, self-adaptors, increased pitch, reduced speech rate and pauses, channel discrepancies	Nonverbal and Paraverbal Cues and Feedback
Vrij (1994)	Information and Feedback (1)	108	Hand and finger movements and Feedback	Nonverbal Cues and Feedback
	Information (2)	108	Hand and finger movements	Nonverbal Cues
deTurck, Feeley, and Roman (1997)	Visual Training (1)	41	Adaptors, hand gestures, head movements, hand shrugs	Nonverbal Cues
	Vocal Training (2)	41	Speech errors, pauses, response latency, message duration	Paraverbal Cues
	Visual and Vocal Training (3)	41	Speech errors, adaptors, hand gestures	Nonverbal and Paraverbal Cues
Feeley and deTurck (1997)	Plausibility (1)	32	Attention to the verbal or spoken message	Verbal Content Cues
	Nervousness (2)	32	Attention to nervousness	Nonverbal and Paraverbal Cues
	Nonverbal (3)	32	Attention to the communicator's nonverbal behavior	Nonverbal Cues
Vrij and Graham (1997; Exp. 1)	Information (1)	20	Hand and Finger movements (and personality traits)	Nonverbal Cues

(continued)

Appendix C (continued)

Authors (Year)	Training group (#)	N_{EG}	Training content	Coding: Type of training
Vrij and Graham (1997; Exp. 2)	Information (1)	15	Hand and Finger movements (and personality traits)	Nonverbal Cues
Kassin and Fong (1999)	Reid Technique (1)	20	Verbal Behavior: *Truthful*: direct, spontaneous, helpful, concerned; denials are broad, sweeping and unequivocal; first-person pronouns, descriptive verbs, unqualified language. *Deceptive*: guarded, unhelpful, unconcerned, they hesitate, shake their hands or mumble, responses are general or evasive, omit details, weak, narrowly defined, or qualified phrases; Nonverbal behavior: *Truthful*: sit upright, face the interrogator, lean forward, use hands and arms, maintain appropriate eye contact. *Deceptive*: rigid body posture, slouch backward, align nonfrontally, cross arms or legs, exhibit various grooming gestures, cover eyes and mouth, either stare or avoid eye contact	Verbal Content and Nonverbal and Paraverbal Cues
Sporer, Samweber, and Stucke (2000)	ARJS Guidance (1)	54	9 Criteria: realism and coherence, spatial information, time information, sensory impressions, emotions and feelings, verbal and nonverbal interactions, complications/extraordinary details, corrections/memory failure, lack of social desirability	Verbal Content Cues

(continued)

Appendix C (continued)

Authors (Year)	Training group (#)	N_{EG}	Training content	Coding: Type of training
Santarcangelo, Cribbie, and Ebesu Hubbard (2004)	Visual Training (1)	21	Self-adaptors, hand gestures, foot and leg movements, postural shifts	Nonverbal Cues
	Vocal Training (2)	20	Pauses, speech errors, response latency, hesitation	Paraverbal Cues
	Verbal Training (3)	26	Plausibility, concreteness, consistency and clarity	Verbal Content Cues
Levine, Feeley, McCornack, Hughes, and Harms (2005; Exp. 1)	Nonverbal and Paraverbal Training (1)	71	Response latencies, adaptors, speech errors, and pauses	Nonverbal and Paraverbal Cues
	Bogus Training (2)	61	Eye contact, speech speed, posture, foot movements	Bogus Training
Levine et al. (2005; Exp. 2)	Nonverbal and Paraverbal Training (1)	28	Response latencies, adaptors, speech errors, and pauses	Nonverbal and Paraverbal Cues
	Bogus Training (2)	31	Eye contact, speech speed, posture, foot movements	Bogus Training
Levine et al. (2005; Exp. 4)	Nonverbal and Paraverbal Training (1)	52	Response latencies, foot movements, speech errors, and pauses	Nonverbal and Paraverbal Cues
	Bogus Training (2)	52	Eye contact, speech speed, posture, adaptors	Bogus Training
Hartwig, Granhag, Strömwall, and Kronkvist (2006)	Strategic Use of Evidence Technique (1)	41	Time of evidence-disclosure in interview (= evidence-statement consistency)	Verbal Content Cues
Porter, McCabe, Woodworth, and Peace (2007)	Accurate Feedback (1)	50	Feedback	Feedback
	Bogus Feedback (2)	50	Feedback	Bogus Feedback

(continued)

Appendix C (continued)

Authors (Year)	Training group (#)	N_{EG}	Training content	Coding: Type of training
Colwell et al. (2009)	Asessment Criteria Indicative of Deception Training (1)	10	Honest stories: longer responses, addition of new details during latter segments of the interview, and more admissions of potential error over entire interview	Verbal Content Cues
Hendershot (1981)	Verbal and Nonverbal Training (1)	14	Verbal and nonverbal cues (e.g., eye movements, speech content)	Verbal Content and Nonverbal and Paraverbal Cues
Bailey (2002)	Nonverbal and Paraverbal Training (1)	50	5 Paraverbal: high pitched voice, more speech hesitations, more speech errors, higher speech rate, longer pause durations 3 Nonverbal: fewer illustrators, fewer hand and finger movements, fewer leg and foot	Nonverbal and Paraverbal Cues
Sporer (1993)	CBCA Training (1)	20	5 Verbal Content Cues: logical consistency, quantity of details, description of unusual details, description of emotion, lack of social desirability	Verbal Content Cues
Blair (2009)	DePaulo-Meta-Analysis Training (1)	40	6 Paraverbal: response length, response latency, rate of speech, non-ah speech disturbances, silent pauses, filled pauses; 4 nonverbal: foot or leg movements, nervous/ tense, self-fidgeting, fidgeting	Nonverbal and Paraverbal Cues
	Reid Technique (2)	40	Inbau-Training (paraverbal, nonverbal, and verbal content)	Verbal Content and Nonverbal and Paraverbal Cues
	DePaulo and Reid Training (3)	40	DePaulo and Reid Training combined	Verbal Content and Nonverbal and Paraverbal Cues

(continued)

Appendix C (continued)

Authors (Year)	Training group (#)	N_{EG}	Training content	Coding: Type of training
Sporer and McCrimmon (1997)	CBCA/RM Training (1)	30	9 ARJS Criteria: logical structure, spatial details, time details, sensory impressions, emotions and feelings, nonverbal and verbal interactions, complications and/or unusual and/or superfluous details, spontaneous corrections or admission of memory failure, negative statements about the self	Verbal Content Cues
Sporer and McFadyen (2001)	CBCA/RM Training (1)	30	9 ARJS Criteria: logical structure, spatial details, time details, sensory impressions, emotions and feelings, nonverbal and verbal interactions, complications and/or unusual and/or superfluous details, spontaneous corrections or admission of memory failure, negative statements about the self	Verbal Content Cues

Note. # = number; N_{EG} = sample size of experimental group; FB = feedback; CBCA = criteria-based content analysis; ARJS = Aberdeen Report Judgment Scales; RM = reality monitoring.

[a]No effect size could be computed. Numbers in parentheses indicate the number of different training groups in each study.

Acknowledgments

Grateful thanks are due to those researchers of primary studies who took the effort to respond to our inquiries as well as to reviewers of previous versions of this article for their fruitful and inspiring comments and suggestions. The authors also wish to thank Dr. Iris Blandón-Gitlin for her insightful comments on a previous version of the article.

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

Notes

1. Unfortunately, we were not aware of Driskell's meta-analysis while we prepared ours (main work between January 2009 and March 2011) until our first version of this article.
2. We also found some discrepancies between Driskell's and our calculations of individual effect sizes for the same studies. While some of these differences may be accounted for by our calculations using somewhat more conservative formulae (using *M*s, *SD*s, and cell *n*s rather than *F* values with the respective *df*s), all our values were coded by two independent coders and cross-validated.
3. Although we have also calculated the random effects model (available on request), we only present the results of the fixed effect model here, which provided clearer results (smaller confidence intervals) and does not require as many studies as the random effects model (Cooper, Hedges, & Valentine, 2009; Hedges & Vevea, 1998).
4. The study by Zuckerman, Koestner, & Alton (1984, Exp. 2) was excluded for the first meta-analysis, because the average effect size included both a feedback and a bogus feedback group. For the latter, the requirement to aim at an increase in detection accuracy was not fulfilled.

References

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

Aamodt, M. G., & Custer, H. (2006). Who can best catch a liar? A meta-analysis of individual differences in detecting deception. *The Forensic Examiner*, *15*, 7-11. Retrieved from https://www.ncjrs.gov/App/publications/Abstract.aspx?id=236906

Akehurst, L., Bull, R., Vrij, A., & Köhnken, G. (2004). The effects of training professional groups and lay persons to use criteria-based content analysis to detect deception. *Applied Cognitive Psychology*, *18*, 877-891. doi:10.1002/acp.1057

Arntzen, F. (1970). *Psychologie der Zeugenaussage. Einführung in die forensische Aussagepsychologie* [Psychology of eyewitness testimony. Introduction to forensic psychology of statement analysis]. Göttingen, Germany: Hogrefe.

Arntzen, F. (1983). *Psychologie der Zeugenaussage. Systematik der Glaubwürdigkeitsmerkmale* [Psychology of eyewitness testimony. A system of credibility criteria]. München, Germany: C.H. Beck.

*Bailey, J. T. (2002). *Detecting deception when motivated: The effects of accountability and training on veracity judgments* (Unpublished master's thesis). Ohio University, Athens. (OCLC: 52189763)

Begg, C. B. (1994). Publication bias. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 399-409). New York, NY: Russell Sage Foundation.

Begg, C. B., & Mazumdar, M. (1994). Operating characteristics of a rank correlation test for publication bias. *Biometrics*, *50*, 1088-1101. doi:10.2307/2533446

Biros, D. P. (2004, October). *Scenario-based training for deception detection*. Proceedings of the 1st Annual Conference on Information Security Curriculum Development, ACM, New York, NY. doi:10.1145/1059524.1059531

Biros, D. P., George, J. F., & Zmud, R. (2002). Inducing sensitivity to deception in order to improve decision-making performance: A field study. *MIS Quarterly*, *26*, 119-144. doi:10.2307/4132323

Biros, D. P., George, J. F., & Zmud, R. (2005). Inside the fence: Sensitizing decision makers to the possibility of deception in the data they use. *MIS Quarterly Executive*, *4*, 261-267. Retrieved from http://misqe.org/ojs2/index.php/misqe/article/view/74

Biros, D. P., Hass, M. C., Wiers, K., Twitchell, D., Adkins, M., Burgoon, J. K., & Nunamaker, J. F. (2005, January). *Task performance under deceptive conditions: Using military scenarios in deception detection research*. Proceedings of the 38th Annual Hawaii International Conference on System Sciences, Big Island, Hawaii. doi:10.1109/HICSS.2005.578

Biros, D. P., Sakamoto, J., George, J. F., Adkins, M., Kruse, J., Burgoon, J. K., & Nunamaker, J. F., Jr. (2005, January). *A quasi-experiment to determine the impact of a computer based deception detection training system: The use of Agent99 Trainer in the U.S. military*. Proceedings of the 38th Annual Hawaii International Conference on System Sciences, Big Island, Hawaii. doi:10.1109/HICSS.2005.42

*Blair, J. P. (2006). Can detection of deception response bias be manipulated? *Journal of Crime & Justice*, *29*, 141-152. doi:10.1080/0735648X.2006.9721652

*Blair, J. P. (2009). *Deception detection: Do laboratory cues generalize to the field?* Unpublished manuscript.

Blair, J. P., Levine, T. R., & Shaw, A. S. (2010). Content in context improves deception detection accuracy. *Human Communication Research*, *36*, 423-442. doi:10.1111/j.1468–2958.2010.01382.x

*Blair, J. P., & McCamey, W. P. (2002). Detection of deception: An analysis of the behavioral analysis interview technique. *Illinois Law Enforcement Executive Forum*, *2*, 165-169. Retrieved from http://www.reid.com/pdfs/Blair2002Detection%20of.pdf

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, C. F., & DePaulo, B. M. (2008). Individual differences in judging deception: Accuracy and bias. *Psychological Bulletin*, *134*, 477-492. doi:10.1037/0033-2909.134.4.477

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* (2nd ed., pp. 221-235). New York, NY: Russell Sage Foundation.

Bull, R. (1989). Can training enhance the detection of deception? In J. Yuille (Ed.), *Credibility assessment* (pp. 83-100). Dordrecht, The Netherlands: Kluwer. doi:10.1007/978-94-015-7856-1_5

Bull, R. (2004). Training to detect deception from behavioural cues: Attempts and problems. In P. A. Granhag & L. A. Strömwall (Eds.), *Deception detection in forensic contexts* (pp. 251-268). Cambridge, UK: Cambridge University Press. doi:10.1017/CBO9780511490071.011

Burgoon, J. K., & Levine, T. R. (2009). Advances in deception detection. In S. W. Smith & S. R. Wilson (Eds.), *New directions in interpersonal communication research* (pp. 201-220). Thousand Oaks, CA: Sage.

Burgoon, J. K., Nunamaker, J. F., George, J. F., Adkins, M., Kruse, J., & Biros, D. (2007). Detecting deception in the military infosphere: Improving and integrating human detection capabilities with automated tools. In C. Wang, S. King, R. Wachter, R. Herklotz, C. Arney, G. Toth, . . . T. Combs (Eds.), *Information security research: New methods for protecting against cyber threats* (pp. 606-627). Indianapolis, IN: Wiley.

Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research*. Chicago, IL: Rand McNally.

Cao, J., Crews, J. M., Lin, M., Burgoon, J., & Nunamaker, J. F. (2003). Designing Agent99 Trainer: A learner-centered, web-based training system for deception detection. In H. Chen (Ed.), *Lecture Notes in Computer Sciences: Vol. 2665. Intelligence and security informatics* (pp. 358-365). Berlin, Germany: Springer-Verlag. Retrieved from http://link.springer.com/chapter/10.1007%2F3-540-44853-5_30

Cao, J., Crews, J. M., Nunamaker, J. F., Burgoon, J. K., & Lin, M. (2004, January). *User experience with Agent99 Trainer: A usability study*. Proceedings of the 37th Hawaii International Conference on System Sciences, Big Island, Hawaii. doi:10.1109/HICSS.2004.1265153

Cao, J., Lin, M., Deokar, A., Burgoon, J. K., Crews, J. M., & Adkins, M. (2004). Computer-based training for deception detection: What users want? In H. Chen (Ed.), *Lecture Notes in Computer Sciences: Vol. 3073. Intelligence and security informatics* (pp. 163-175). Berlin, Germany: Springer-Verlag. doi:10.1007/978-3-540-25952-7_12

Carlson, K. D., & Schmidt, F. L. (1999). Impact of experimental design on effect size: Findings from the research literature on training. *Journal of Applied Psychology*, *84*, 851-862. doi:10.1037/0021-9010.84.6.851

Clark, L. M. (1983). *Training humans to become better decoders of deception* (Unpublished master's thesis). University of Georgia, Athens. (OCLC:10040606)

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: Lawrence Erlbaum.

*Colwell, K., Hiscock-Anisman, C., Memon, A., Colwell, L. H., Taylor, L., & Woods, D. (2009). Training in assessment criteria indicative of deception to improve credibility judgments. *Journal of Forensic Psychology Practice*, *9*, 199-207. doi:10.1080/15228930902810078

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

Cooper, H., Hedges, L. V., & Valentine, J. C. (Eds.). (2009). *The handbook of research synthesis and meta-analysis* (2nd ed.). New York, NY: Russell Sage Foundation.

*Costanzo, M. (1992). Training students to decode verbal and nonverbal cues: Effects on confidence and performance. *Journal of Educational Psychology*, *84*, 308-313. doi:10.1037//0022-0663.84.3.308

*Crews, J. M., Cao, J., Lin, M., Nunamaker, J. F., & Burgoon, J. K. (2007). A comparison of instructor-led vs. web-based training for detecting deception. *Journal of Science, Technology, Engineering and Math Education*, *8*, 31-39. Retrieved from http://jstem.org/ojs/index.php?journal=JSTEM&page=article&op=viewFile&path[]=1350&path[]=1185

Dando, C. J., & Bull, R. (2011). Maximising opportunities to detect verbal deception: Training police officers to interview tactically. *Journal of Investigative Psychology and Offender Profiling*, *8*, 189-202. doi:10.1002/jip.145

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., Lassiter, G. D., & Stone, J. I. (1982). Attentional determinants of success at detecting deception and truth. *Personality and Social Psychology Bulletin*, *8*, 273-279. doi:10.1177/0146167282082014

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, UK: Cambridge University Press. doi:10.1017/CBO9780511490071.002

Dettenborn, H., Froehlich, H., & Szewczyk, H. (1984). *Forensische Psychologie* [Forensic psychology]. Berlin, Germany: Deutscher Verlag der Wissenschaften.

*deTurck, M. A. (1991). Training observers to detect spontaneous deception: The effects of gender. *Communication Reports*, *4*, 79-89. doi:10.1080/08934219109367528

*deTurck, M. A., Feeley, T. H., & Roman, L. (1997). Vocal and visual cue training in behavioral lie detection. *Communication Research Reports*, *14*, 249-259. doi:10.1080/08824099709388668

*deTurck, M. A., Harszlak, J. J., Bodhorn, D., & Texter, L. (1990). The effects of training social perceivers to detect deception from behavioral cues. *Communication Quarterly*, *38*, 1-11. doi:10.1080/01463379009369753

*deTurck, M. A., & Miller, G. R. (1990). Training observers to detect deception: Effects of self-monitoring and rehearsal. *Human Communication Research*, *16*, 603-620. doi:10.1111/j.1468-2958.1990.tb00224.x

Docan-Morgan, T. (2007). Training law enforcement officers to detect deception: A critique of previous research and framework for the future. *Applied Psychology in Criminal Justice*, *3*, 143-171. Retrieved from http://www.relationalturningpoints.org/uploads/2007_-_Training_Law.pdf

Driskell, J. E. (2012). Effectiveness of deception detection training: A meta-analysis. *Psychology, Crime & Law*, *18*, 713-731. doi:10.1080/1068316X.2010.535820

Dunlap, W. P., Cortina, J., Vaslow, J. B., & Burke, M. J. (1996). Meta-analysis of experiments with matched groups or repeated measures designs. *Psychological Methods*, *1*, 170-177. doi:10.1037/1082-989X.1.2.170

Duval, S. J., & Tweedie, R. L. (2000a). A nonparametric “trim and fill” method of accounting for publication bias in meta-analysis. *Journal of the American Statistical Association*, *95*, 89-98. doi:10.1080/01621459.2000.10473905

Duval, S. J., & Tweedie, R. L. (2000b). Trim and fill: A simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. *Biometrics*, *56*, 455-463. doi:10.1111/j.0006-341X.2000.00455.x

*Dziubinski, M. A. (2003). *Deception detection in a computer-mediated environment: Gender, trust, and training issues* (Doctoral dissertation). Air Force Institute of Technology, Wright- Patterson Air Force Base, OH. Retrieved from http://www.dtic.mil/cgi-bin/GetTRDoc?AD=ADA420817

Egger, M., Davey Smith, G., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *British Medical Journal*, *315*, 629-634. Retrieved from http://jpkc.hrbmu.edu.cn/lxbx/cankao/Bias%20in%20meta-analysis%20detected%20by%20a%20simple,%20graphical%20test.pdf

Elaad, E. (2003). Effects of feedback on the overestimated capacity to detect lies and the underestimated ability to detect lies. *Applied Cognitive Psychology*, *17*, 349-363. doi:10.1002/acp.871

Enos, F., Shriberg, E., Graciarena, M., Hirschberg, J., & Stolcke, A. (2007, August). *Detecting deception using critical statements*. Proceedings of the 10th European Conference on Speech Communication and Technology - Interspeech, Antwerp, Belgium. Retrieved from http://www-speech.sri.com/papers/IS07-enos-p1085.pdf

*Feeley, T. H., & deTurck, M. A. (1997). Case-relevant vs. case-irrelevant questioning in experimental lie detection. *Communication Reports*, *10*, 35-45. doi:10.1080/08934219709367657

*Fiedler, K., & Walka, I. (1993). Training lie detectors to use nonverbal cues instead of global heuristics. *Human Communication Research*, *20*, 199-223. doi:10.1111/j.1468-2958.1993.tb00321.x

Ford, C. L. (2004). *Determination of the trainability of deception detection cues* (Unpublished thesis). Air Force Institute of Technology, Wright- Patterson Air Force Base, OH. Retrieved from http://www.dtic.mil/dtic/tr/fulltext/u2/a423153.pdf

Frank, M. G., & Feeley, T. H. (2003). To catch a liar: Challenges for research in lie detection training. *Journal of Applied Communication Research*, *31*, 58-75. doi:10.1080/00909880305377

Geiselman, R. E., Elmgren, S., Green, C., & Rystad, I. (2011). Training laypersons to detect deception in oral narratives and exchanges. *American Journal of Forensic Psychology*, *32*, 1-22.

*George, J. F., Biros, D. P., Adkins, M., Burgoon, J. K., & Nunamaker, J. F. (2004). Testing various modes of computer-based training for deception detection. In H. Chen (Ed.), *Lecture Notes in Computer Sciences: Vol. 3073. Intelligence and security informatics* (pp. 411-417). Berlin, Germany: Springer-Verlag. doi:10.1007/978-3-540-25952-7_31

George, J. F., Biros, D. P., Burgoon, J. K., & Nunamaker, J. F., Jr. (2003, June). *Training professionals to detect deception*. Proceedings of the First NSF/NIJ Symposium on Intelligence and Security Informatics, Tucson, AZ. doi:10.1007/3-540-44853-5_31

George, J. F., Biros, D. P., Burgoon, J. K., Nunamaker, J. F., Crews, J. M., Cao, J., Marret, K., Adkins, M., Kruse, J., & Lin, M. (2008). The role of e-training in protecting information assets against deception attacks. *MIS Quarterly Executive*, *7*, 57-69. Retrieved from http://misqe.org/ojs2/index.php/misqe/article/view/188

*George, J. F., Marett, K., Burgoon, J. K., Crews, J., Cao, J., Lin, M., & Biros, D. P. (2004, January). *Training to detect deception: An experimental investigation*. Proceedings of the 37th Hawaii International Conference on System Sciences, Big Island, Hawaii. doi:10.1109/HICSS.2004.1265082

George, J. F., Marett, K., & Tilley, P. (2004, January). *Deception detection under varying electronic media and warning conditions*. Proceedings of the 37th Hawaii International Conference on System Sciences, Big Island, Hawaii. doi:10.1109/HICSS.2004.1265080

Gleser, L. J., & Olkin, I. (1994). Stochastically dependent effect sizes. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 339-355). New York, NY: Russell Sage Foundation.

Gleser, L. J., & Olkin, I. (2009). Stochastically dependent effect sizes. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2nd ed., pp. 357-376). New York, NY: Russell Sage Foundation.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York, NY: Wiley.

Greenhouse, J. B., & Iyengar, S. (2009). Sensitivity analysis and diagnostics. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *Handbook of research synthesis and meta-analysis* (2nd ed., pp. 417-433). New York, NY: Russell Sage Foundation.

*Hall, S. (1989). *The generalizability of learning to detect deception in effective and ineffective deceivers* (Unpublished doctoral dissertation). Auburn University, AL. doi:oclc/20840284

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

*Hartwig, M., Granhag, P. A., Strömwall, L. A., & Kronkvist, O. (2006). Strategic use of evidence during police interviews: When training to detect deception works. *Law and Human Behavior*, *30*, 603-619. doi:10.1007/s10979-006-9053-9

Hauch, V., Blandón-Gitlin, I., Masip, J., & Sporer, S. L. (2013). *Are computers effective lie detectors? A meta-analysis of linguistic cues to deception*. Manuscript submitted for publication.

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

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

*Hendershot, J. (1981). *Detection of deception in low and high socialization subjects with trained and untrained judges* (Unpublished master's thesis). Auburn University, AL. doi:oclc/8096203

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

Hill, M. L., & Craig, K. D. (2004). Detecting deception in facial expressions of pain: Accuracy and training. *The Clinical Journal of Pain*, *20*, 415-422. doi:10.1097/00002508-200411000-00006

Horvath, F., Jayne, B., & Buckley, J. (1994). Differentiation of truthful and deceptive criminal suspects in behavior analysis interviews. *Journal of Forensic Sciences*, 39, 793-807. Retrieved from https://www.ncjrs.gov/App/publications/Abstract.aspx?id=148725

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

*Kassin, S. M., & Fong, C. T. (1999). "I'm innocent!": Effects of training on judgments of truth and deception in the interrogation room. *Law and Human Behavior*, *23*, 499-516. doi:10.1023/A:1022330011811

Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin*, *119*, 254-284. doi:10.1037//0033-2909.119.2.254

*Köhnken, G. (1987). Training police officers to detect deceptive eyewitness statements: Does it work? *Social Behaviour*, *2*, 1-17.

Köhnken, G. (1989). Behavioral correlates of statement credibility: Theories, paradigms and results. In H. Wegener, F. Lösel, & J. Haisch (Eds.), *Criminal behavior and the justice system: Psychological perspectives* (pp. 271-289). New York, NY: Springer-Verlag. doi:10.1007/978-3-642-86017-1_18

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 agreement for credibility criteria: An empirical study]. In G. Bierbrauer, W. Gottwald, & B. Birnbreier-Stahlberger (Eds.), *Verfahrensgerechtigkeit-Rechtspsychologische Forschungsbeiträge für die Justizpraxis* (pp. 187-213). Köln, Germany: Otto Schmidt Verlag.

*Landry, K., & Brigham, J. C. (1992). The effect of training in criteria-based content analysis on the ability to detect deception in adults. *Law and Human Behavior*, *16*, 663-675. doi:10.1007/bf01884022

*Levine, T. R., Feeley, T. H., McCornack, S. A., Hughes, M., & Harms, C. M. (2005). Testing the effects of nonverbal behavior training on accuracy in deception detection with the inclusion of a bogus training control group. *Western Journal of Communication*, *69*, 203-217. doi:10.1080/10570310500202355

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

Lin, M., Crews, J. M., Cao, J., Nunamaker, J. F., Jr., & Burgoon, J. K. (2003, August). *Agent99 trainer: Designing a web-based multimedia training system for deception detection knowledge transfer*. Proceedings of the Ninth Americas Conference on Information Systems (AMCIS 2003), Tampa, FL. Retrieved from http://aisel.aisnet.org/amcis2003/334/

Lin, Y. C. (1999). *A study of training on deception detection: The effects of the specific six cues versus heuristics on deception detection accuracy* (Unpublished master's thesis). State University of New York, Buffalo.

Lipsey, M. W., & Wilson, D. B. (1993). The efficacy of psychological, educational, and behavioral treatment. *American Psychologist*, *48*, 1181-1209. doi:10.1037//0003-066X.48.12.1181

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

Mann, S., Vrij, A., & Bull, R. (2006). Looking through the eyes of an accurate lie detector. *The Journal of Credibility Assessment and Witness Psychology*, *7*, 1-16. Retrieved from http://truth.charleshontsphd.com/JCAAWP/2006_1_16/2006_1_16.pdf

Marett, K., Biros, D. P., & Knode, M. L. (2004). Self-efficacy, training effectiveness, and deception detection: A longitudinal study of lie detection training. In H. Chen (Ed.), *Lecture Notes in Computer Sciences: Vol. 3073. Intelligence and security informatics* (pp. 187-200). Berlin, Germany: Springer-Verlag. doi:10.1007/978-3-540-25952-7_14

Masip, J., Alonso, H., Garrido, E., & Herrero, C. (2009). Training to detect what? The biasing effects of training on veracity judgments. *Applied Cognitive Psychology*, *23*, 1282-1296. doi:10.1002/acp.1535

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

Matt, G. E., & Cook, T. D. (2009). Threats to the validity of generalized inferences. In H. Cooper, L. V. Harris, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2nd ed., pp. 537-560). New York, NY: Russell Sage Foundation.

McCornack, S. A., & Levine, T. R. (1990). When lovers become leery: The relationship between suspicion and accuracy in detecting deception. *Communication Monographs*, *57*, 219-230. doi:10.1080/03637759009376197

McKenzie, F. R., Scerbo, M., & Catanzaro, J. (2003). Generating nonverbal indicators of deception in virtual reality training. *Journal of WSCG*, *11*(1), 314-321. Retrieved from http://www.researchgate.net/publication/2474474_Generating_Nonverbal_Indicators_of_Deception_in_Virtual_Reality_Training

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

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

Mitchell, K. J., & Johnson, M. K. (2000). Source monitoring: Attributing mental experiences. In E. Tulving & F. I. M. Craik (Eds.), *The oxford handbook of memory* (pp. 179-195). New York, NY: Oxford University Press.

Morris, S. B. (2008). Estimating effect sizes from pretest-posttest-control group designs. *Organizational Research Methods*, *11*, 364-386. doi:10.1177/1094428106291059

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

Orwin, R. G., & Vevea, J. L. (2009). Evaluating coding decisions. In H. Cooper, L. V. Harris, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2nd ed., pp. 177-205). New York, NY: Russell Sage Foundation.

Parker, A. D., & Brown, J. (2000). Detection of deception: Statement validity analysis as a means of determining truthfulness or falsity of rape allegations. *Legal and Criminological Psychology*, *5*, 237-259. doi:10.1348/135532500168119

Patrick, J. (1992). *Training: Research and practice*. Padstow, UK: Academic Press.

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

Porter, S., Juodis, M., ten Brinke, L. M., Klein, R., & Wilson, K. (2010). Evaluation of the effectiveness of a brief deception detection training program. *The Journal of Forensic Psychiatry & Psychology*, *21*, 66-76. doi:10.1080/14789940903174246

*Porter, S., McCabe, S., Woodworth, M., & Peace, K. A. (2007). "Genius is 1% inspiration and 99% perspiration": Or is it? An investigation of the impact of motivation and feedback on deception detection. *Legal and Criminological Psychology*, *12*, 297-309. doi:10.1348/135532506X143958

*Porter, S., Woodworth, M., & Birt, A. R. (2000). Truth, lies, and videotape: An investigation of the ability of federal parole officers to detect deception. *Law and Human Behavior*, *24*, 643-658. doi:10.1023/A:1005500219657

Reinhard, M.-A., Sporer, S. L., & Scharmach, M. (2013). Perceived familiarity with a judgmental situation improves lie detection ability. *Swiss Journal of Psychology*, *72*, 53-61. doi:10.1024/1421-0185/a000098

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

Rossi, P. H., Lipsey, M. W., & Freeman, H. E. (Eds.). (2009). *Evaluation: A systematic approach*. Thousand Oaks, CA: Sage.

Rothstein, H. R., Sutton, A. J., & Borenstein, M. (Eds.). (2005). *Publication bias in meta-analysis: Prevention, assessment and adjustments*. Chichester, UK: John Wiley.

*Santarcangelo, M., Cribbie, R. A., & Ebesu Hubbard, A. S. (2004). Improving accuracy of veracity judgment through cue training. *Perceptual & Motor Skills*, *98*, 1039-1048. doi:10.2466/pms.98.3.1039-1048

Seager, P. B. (2001). *Improving the ability of people to detect lies* (Unpublished doctoral dissertation). University of Hertfordshire, Hatfield, UK.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston, MA: Houghton Mifflin.

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* (2nd ed., pp. 257-277). New York, NY: Russell Sage Foundation.

Sporer, S. L. (1983, August). *Content criteria of credibility: The German approach to eyewitness testimony*. Paper presented in G.S. Goodman (Chair), The child witness: Psychological and legal issues. Symposium presented at the 91st Annual Convention of the American Psychological Association in Anaheim, CA.

*Sporer, S. L. (1993, April). *Münchhausen's Zopf: Zur Diskrimination wahrer von erfundenen Geschichten* [Baron Muenchhausen's pony tail: Discriminating true from invented stories]. Paper presented at the 35th Tagung experimentell arbeitender Psychologen in Münster, Germany.

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, UK: Cambridge University Press. doi:http://dx.doi.org/10.1017/CBO9780511490071.004

Sporer, S. L., & Bursch, S. E. (1996, April). *Detection of deception by verbal means: Before and after training*. Paper presented at the 38th Tagung experimentell arbeitender Psychologen in Eichstätt, Germany.

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

Sporer, S. L., Masip, J., & Cramer, M. (2014). Guidance to detect deception with the Aberdeen Report Judgment Scales: Are verbal content cues useful to detect false accusations? *American Journal of Psychology*, *127*, 43-61. doi:10.5406/amerjpsyc.127.1.0043

*Sporer, S. L., & McCrimmon, S. (1997, July). *A pleasant—or not so pleasant—dinner evening? Guiding people to detect what really happened*. Paper presented at the Tagung der Fachgruppe Sozialpsychologie der Deutschen Gesellschaft für Psychologie in Konstanz, Germany.

*Sporer, S. L., & McFadyen, C. J. C. (2001, June). *The medium is the message? Detecting deception from videotapes and transcripts with the Aberdeen Report Judgments Scales*. Paper presented at the 11th European Conference of Psychology and Law in Lisbon, Portugal.

*Sporer, S. L., Samweber, M. C., & Stucke, T. S. (2000, March). *Twisting the outcome: Discriminating truths from factually experiences events*. Paper presented at the American Psychology-Law Society Conference in New Orleans, LA.

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:10.1037/1076-8971.1.13.1.1

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*, 837-854. doi:10.1002/acp.1234

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, NY: Springer-Verlag.

Sterne, J. A. C., Becker, B. J., & Egger, M. (2005). The funnel plot. In H. R. Rothstein, A. J. Sutton, & M. Borenstein (Eds.), *Publication bias in meta-analysis: Prevention, assessment and adjustments* (pp. 75-98). West Sussex, UK: Wiley.

Sutton, A. J. (2009). Publication bias. In H. Cooper, L. V. Harris, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2nd ed., pp. 435-452). New York, NY: Russell Sage Foundation.

Sutton, A. J., Duval, S. J., Tweedie, R. L., Abrams, K. R., & Jones, D. R. (2000). Empirical assessment of effect of publication bias on meta-analyses. *British Medical Journal*, *320*, 1574-1577. doi:10.1136/bmj.320.7249.1574

Szewczyk, H. (1973). Kriterien der Beurteilung kindlicher Zeugenaussagen [Criteria for the evaluation of child witnesses]. *Probleme und Ergebnisse der Psychologie*, *46*, 47-66.

Thorndike, E. L. (1913). *Educational psychology, Volume I: The original nature of man*. New York: Teachers College, Columbia University.

Thorndike, E. L. (1927). The law of effect. *American Journal of Psychology*, *39*, 212-222.

Undeutsch, U. (1967). Beurteilung der Glaubhaftigkeit von Aussagen [Evaluation of the credibility of statements]. In U. Undeutsch (Ed.), *Handbuch der Psychologie Vol. 11: Forensische Psychologie* (pp. 26-181). Göttingen, Germany: Hogrefe.

*Vrij, A. (1994). The impact of information and setting on detection of deception by police detectives. *Journal of Nonverbal Behavior*, *18*, 117-136. doi:10.1007/BF02170074

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, 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., & Graham, S. (1997). Individual differences between liars and the ability to detect lies. *Expert Evidence: The International Digest of Human Behaviour Science and Law*, *5*, 144-148. doi:10.1023/A:1008835204584

Warren, G., Schertler, E., & Bull, P. (2009). Detecting deception from emotional and unemotional cues. *Journal of Nonverbal Behavior*, *33*, 59-69. doi:10.1007/s109190080057-7

Wilson, D. B. (2010). *Meta-analysis macros for SAS, SPSS, and Stata*. Retrieved from http://mason.gmu.edu/~dwilsonb/ma.html

Wood, W., & Eagly, A. H. (2009). Advantages of certainty and uncertainty. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *Handbook of research synthesis and meta-analysis* (2nd ed., pp. 455-472). New York, NY: Russell Sage Foundation.

Yang, C. C. (1996). *The effects of training, rehearsal, and consequences for lying on deception detection accuracy* (Unpublished master's thesis). State University of New York, Buffalo.

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

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. 1-59). New York, NY: Academic Press.

*Zuckerman, M., Koestner, R. E., & Alton, A. O. (1984). Learning to detect deception. *Journal of Personality and Social Psychology*, *46*, 519-528. doi:10.1037/0022-3514.46.3.519

*Zuckerman, M., Koestner, R. E., & Colella, M. J. (1985). Learning to detect deception from three communication channels. *Journal of Nonverbal Behavior*, *9*, 188-194. doi:10.1007/BF01000739

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

Author Biographies

Valerie Hauch (Diploma, University of Giessen, 2010) is a doctoral student at the Department of Social Psychology and Psychology and Law at the University of Giessen. Her research focuses on meta-analyses in the field of detection of deception and her dissertation deals with meta-analyses on linguistic and verbal content cues to deception.

Siegfried L. Sporer (PhD, University of New Hampshire, 1980) is Professor for Social Psychology and Psychology and Law at the University of Giessen, Germany. His research has focused on eyewitness testimony, facial recognition and person identificantion, and eyewitness meta-memory as well as nonverbal, paraverbal, linguistic and content cues to deception and the detection of deception. In recent years, he has specialized on meta-analyses of various aspects of eyewitness testimony and deception. His email address is Sporer@psychol.uni-giessen.de.

Stephen W. Michael (PhD, University of Texas at El Paso, 2013) is currently a Visiting Assistant Professor in the Psychology Department at Mercer University. His research interests include deception and investigative interviewing. His email address is Michael_SW@mercer.edu.

Christian A. Meissner (PhD, Florida State University, 2001) is Professor in the Cognitive Psychology program at Iowa State University. His research focuses on applied cognition, including: the role of memory, attention, perception, and decision processes in real world tasks; areas of application include face recognition, forensic interviewing, deception detection, and legal decision making. His email address is cameissn@iastate.edu.