Perspectives on Psychological Science
2023, Vol. 18(6) 1368–1387
© The Author(s) 2023
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/17456916221149943
www.psychologicalscience.org/PPS

Is It the Judge, the Sender, or Just the Individual Message? Disentangling Person and Message Effects on Variation in Lie-Detection Judgments

Sarah Volz[1], Marc-André Reinhard[1], and Patrick Müller[2]
[1]Department of Psychology, University of Kassel, and [2]Faculty of Civil Engineering, Building Physics, and Business, University of Applied Sciences Stuttgart

Abstract

Research suggests that people differ more in their ability to lie than in their ability to detect lies. However, because studies have not treated senders and messages as separate entities, it is unclear whether some senders are generally more transparent than others or whether individual messages differ in their transparency of veracity regardless of senders. Variance attributable to judges, senders, and messages was estimated simultaneously using multiple messages from each sender (totaling more than 45,000 judgments). The claim that the accuracy of a veracity judgment depends on the sender was not supported. Messages differed in their detectability (21% explained variance), but senders did not. Message veracity accounted for most message variation (16.8% of the total variance), but other idiosyncratic message characteristics also contributed significantly. Consistent with the notion that a (mis)match between sender demeanor and veracity determines accuracy, lie and truth detectability differed individually within senders. Judges primarily determined variance in lie-versus-truth classifications (12%) and in confidence (46%) but played no role regarding judgment accuracy (< 0.01%). This work has substantial implications for the design and direction of future research and underscores the importance of separating senders and messages when developing theories and testing derived hypotheses.

Keywords

deception detection, sources of variation, credibility, lie-detection ability, confidence

Humankind has long tried to find ways to detect lies (see, e.g., Lykken, 1974; Trovillo, 1939). However, in ad hoc veracity judgments, the overall accuracy rate for discriminating between truth and deception is 54% (for a meta-analysis, see Bond & DePaulo, 2006), just above the level that could be expected by chance. Although many studies have investigated human lie-detection ability and influencing factors thereof (see, e.g., Aamodt & Custer, 2006; Bond & DePaulo, 2006), some research also suggests that the senders might have a strong impact on veracity judgments and the accuracy of such judgments (e.g., Bond & DePaulo, 2008). Put differently, some individuals may be better liars than others, which would affect lie-detection accuracy regardless of who is judging these individuals.

The Problem of the Sender–Message Entanglement

Several attempts have been made to discern the variability in lie detection that is due to *judges*, who are making the veracity judgments, and the variability that is due to *senders*, who are delivering the truths and lies being judged (e.g., Bond & DePaulo, 2008; Levine, 2016; Levine et al., 2011, 2022). However, we argue that the separation into these two sources of variation is insufficient because it conflates the sender with the

Corresponding Author:
Sarah Volz, Department of Psychology, University of Kassel
Email: sarah.volz@uni-kassel.de

message that the sender is conveying (i.e., the lie or truth). Previous work on the sources of variability rarely explicitly made this distinction between sender and message and often used only one message per sender (*sender–message entanglement*); therefore, there was no differentiation between a global person effect of senders across their messages and the individual situated messages these senders delivered. Because senders either delivered a truthful or a deceptive message, such studies also confounded *message veracity* with the sender (and the message). Hence, these studies neither accounted for within-sender message-to-message variability resulting from the idiosyncratic features of messages nor determined how message veracity affected variability.

Because the sender–message entanglement is evident both in theorizing and in the methodological approaches of previous work, we first discuss on a theoretical level why for some of the lie-detection variables not only a global sender effect but also an effect of the individual message seems plausible. On a methodological level, we address the problem of the sender–message entanglement by using multiple truthful and deceptive messages from each sender and account for the variation that is due to these messages. In a large data set of more than 45,000 veracity judgments, we simultaneously estimate the extent to which judges, senders, and messages explain the variation in three lie-detection variables: the variation in the type of judgment (i.e., in whether a message is classified as a lie vs. as truth), the variation in the accuracy of judgments (i.e., in whether a judgment is correct), and, for the first time, the variation in confidence with which a veracity judgment is made.

The entanglement of the concepts of senders and messages not only led to ambiguities in theoretical reasoning but potentially also to faulty conclusions because effects of the situated message and its veracity were neglected. To illustrate, a person tells two lies: One is “I walked on Mars yesterday,” and the other is “I like soccer.” For the message “I walked on Mars yesterday,” most likely everyone would be able to tell that this is a lie. However, from a high detection rate for this lie, one could not conclude that the person is generally an unsuccessful liar (compared with other senders); anyone who makes this statement would be identified as a liar reliably. Thus, a high detection rate for this message says less about the sender’s ability to lie or the lie-detection ability of the person making the judgment than it does about the message itself. For the message “I love soccer,” the detection rate would be lower than for the previous statement. This lower detection rate likely occurs regardless of who delivers the statement and regardless of who makes the judgment. Highlighting the importance of disentangling the message from the sender, if the same individual told these two lies, the detection rates for the lies would differ. Therefore, using only one of these lies in a study could lead to faulty conclusions about this individual’s ability to lie.[1]

Most of the few studies on sources of variation in lie detection (e.g., Bond & DePaulo, 2008; Levine, 2016; Levine et al., 2011; Masip et al., 2020) examined standard deviations for two variables on the judge dimension (ability and credulity) and two variables on the sender dimension (detectability and credibility). There are study designs in lie detection in which judges make judgments about multiple messages, which allows their ability and credulity to be determined across multiple messages, whereas senders deliver only one message (see Bond & DePaulo, 2006), and their detectability and credibility are determined only by that one truthful or deceptive message. Therefore, in studies that examine variation in lie detection with such a design, variables calculated for the sender dimensions did not reflect pure person effects. Moreover, the variation of senders’ detectability might have been overestimated in such studies when veracity was confounded with the sender (and the message). Because of the typical truth bias (for a meta-analysis, see Bond & DePaulo, 2006), truthful messages oftentimes have higher detectability rates than deceptive messages. When only one message per sender is used, senders randomly assigned to the truth condition will likely receive higher detectability scores than senders assigned to the lie condition (for an exception resulting from a demeanor induction that was independent of veracity, see Levine et al., 2011). This veracity effect (Levine et al., 1999) could have led to systematic detectability differences and therefore increased sender variation without being an individual sender difference. Because judges’ ability scores were calculated across truthful and deceptive messages, these scores were less affected by this effect.

In an influential meta-analysis on sources of variability in lie detection, Bond and DePaulo (2008) wrote:

> While gauging differences among individuals as judges of deceit, we also assess differences among them as liars. . . . A sender is perfectly detectable if that sender is always judged to be lying when s/he is telling a lie and always judged to be telling the truth when s/he is telling the truth. (p. 478)

Even though Bond and DePaulo stated that sender detectability is to be determined across several lies and truths from a sender, previous work on sources of variability in lie detection sometimes equated the sender with the message and used only one message per sender (e.g., Levine et al., 2011). Because the meta-analysis revealed larger variability on the stimulus side than the judge side, Bond and DePaulo (2008) concluded that "the accuracy of a deception judgment depends more on the liar than the judge" (p. 486). However, the sender–message entanglement made it impossible to tell whether the accuracy of a deception judgment actually depends on the liar (i.e., a global person effect of the senders) or on the lies themselves (i.e., an effect of the individual messages).

A study of Levine (2016) and a replication of it (Masip et al., 2020) partly addressed the problem of the sender–message entanglement by conducting a round-robin experiment in which participants acted as both judges and senders. As judges, participants judged multiple messages, and their ability and credulity were calculated across all messages they judged. As senders, participants delivered multiple messages, and their detectability and credibility were calculated across all messages they delivered. Even though sender scores were thereby less influenced by individual messages than had been the case in previous studies, variability resulting from messages was still not determined. Accordingly, to the best of our knowledge, there are no studies to date that properly separate the person effects of judges and senders and the effects of messages when examining veracity judgments, neither on a methodological nor theoretical level.[2] In the following, we describe which sources of variation seem plausible for the three variables judgment type (lie vs. truth classifications), accuracy, and confidence when judge, sender, and message are treated as separate entities.

Research Question 1: What Determines Whether a Message Is Classified as Lie or Truth?

Some previous work found a person effect of judges to be the main source of variation in lie-versus-truth classifications (Levine et al., 2022; Masip et al., 2020). A person effect of judges would suggest that some individuals generally have a higher tendency to rate messages as truthful compared with others. Because base-rate assumptions can affect truth-judgment rates (see Street & Richardson, 2015), interindividual-judge differences might reflect their individual estimates of the probability of truth versus deception occurring. Base-rate assumptions might be influenced by the "deception consensus effect" (e.g., Markowitz, 2022; Markowitz & Hancock, 2018; see also Sagarin et al., 1998); that is, individuals who lie often also expect others to lie frequently. Moreover, base-rate assumptions also show in the "investigator bias" (Meissner & Kassin, 2002); individuals experienced in lie detection (e.g., police officers) are confronted with lies more often than naive individuals. A higher generalized suspicion of these experienced individuals (see, e.g., Masip et al., 2005) is assumed to result in a higher tendency to classify messages as lies (e.g., Bond & DePaulo, 2006; Garrido et al., 2004).

Other previous work identified the sender as the main source of variation in lie-versus-truth classifications (e.g., Bond & DePaulo, 2008; Levine, 2016; Levine et al., 2011). However, for studies that involved a sender–message entanglement (e.g., Bond & DePaulo, 2008; Levine et al., 2011), it is unclear whether senders were actually the main source of variation or whether the variation was, at least partly, due to idiosyncratic messages. Theoretically, a sender effect would fit the conceptualization of sender demeanor introduced by Levine et al. (2011). They argued that a sender's credibility generalizes across both situations and judges and is largely independent of veracity. Consistent with this idea, a series of studies by Frank and Ekman (2004) found that the amount of truth judgments that senders received was stable across two situations. From an evolutionary perspective, senders should be interested in their truths being recognized as truthful and their lies remaining undetected (e.g., Buller & Burgoon, 1996; Solbu & Frank, 2019); hence, senders strive to be regarded as credible in all of their messages. If some senders are more successful at this endeavor than others, senders would be a source of variation for the classification of messages as lies versus truths. Supporting this argument, sender characteristics such as social and emotional skills, personality traits (e.g., DePaulo & Rosenthal, 1979; Riggio et al., 1987), and attractiveness (see, e.g., Patzer, 1983; Zebrowitz et al., 1996) have been identified as predictors of deception success or generally higher perceived credibility (for a review, see also Semrad et al., 2019). In addition, individuals who lie more frequently than others could be perceived as more credible across their truthful and deceptive messages as a result of receiving more feedback and adjusting their strategies accordingly (see, e.g., Serota et al., 2022).

As also noted by Levine et al. (2011), sender demeanor is likely not "completely trait-like" (p. 380) and subjected to situational influences—for instance, even individuals who appear consistently credible cannot lie convincingly about particular topics (e.g., when claiming to have walked on Mars as in the above example). Hence, the situated message may also be a source of variation, which would imply that senders appear more credible in some situations than in others. When situations are standardized, as oftentimes is the case in lie-detection studies (e.g., given topic, standardized preparation and interview questions, time limits), the situation should play a minor role compared with the person effects of judges and senders. Hence, when disentangling judge, sender, and message as sources of variation in lie-versus-truth classifications, we assumed the person effects of judges and senders to be the main sources of variation rather than the message (individual-differences hypothesis).

Research Question 2: What Determines Whether a Judgment Is Correct or Incorrect?

Previous studies have consistently found that judges are a less important source of variation in the accuracy of judgments than senders (e.g., Bond & DePaulo, 2008; Levine et al., 2010, 2011, 2022). Likewise, judge characteristics that predict accuracy are rare (for meta-analyses, see Aamodt & Custer, 2006; Bond & DePaulo, 2006, 2008), which underlines that judges play a minor role regarding judgment accuracy.

Again, because of the sender–message entanglement, it is unclear whether the variation previously attributed to senders is due to a global person effect of the sender, idiosyncratic message features leading to detectability differences, or, at least in part, variation stemming from the veracity effect. Theoretically, an effect of the individual messages resulting from idiosyncratic features seems plausible; factors not inherent in the sender (e.g., situational factors) or characteristics of the sender that are not stable across messages might make one message of a sender easier to detect than other messages of the sender. For instance, senders' motivation, cognitive load, or fatigue when delivering a specific message, senders' expertise with the topic of the message, preparation, base-rate effects (as in the Mars and soccer examples above), or specific question strategies could influence the individual message's detectability (see, e.g., Bond & DePaulo, 2006; Hartwig & Bond, 2011; Levine et al., 2010; Vrij et al., 2008, 2017). In addition to such unique message effects, a message's veracity might systematically be a predictor of accuracy, as indicated by research on truth bias (e.g., Bond & DePaulo, 2006) and related work on the veracity effect (e.g., Levine et al., 1999). Because judges overall tend to rate messages more often as truths than as lies, truths should be judged correctly more often.

Suggesting a global person effect of senders, Levine (2010) argued that there are "a few transparent liars" (p. 43) who are easier to detect than other senders (see also Levine, 2016). For instance, senders who lie very little (see Serota et al., 2022) may be easier to detect than more frequent liars, either because they do not have practice and feedback to adjust their strategies or because they refrain from lying as a result of getting caught often. A person effect of senders would imply that the demeanor (or credibility) of some senders should be linked to veracity. In other words, these senders appear either less credible when they lie or more credible when they tell the truth, or both; this would make them easier to be detected across their messages. Because such transparent senders are assumed to be rare (see Levine, 2010, 2016), the demeanor of the majority of senders should not be linked to veracity. When demeanor is independent of veracity and consistent across a sender's messages, the (mis)match of demeanor and actual veracity should determine accuracy (see also Levine, 2016; Levine et al., 2011). Thus, the detectability of individual senders should differ between their lies and truths. Put differently, an honest-appearing sender should be easy to detect when telling the truth but difficult to detect when lying (higher truth than lie accuracy) and vice versa for dishonest-appearing senders.

In summary, when disentangling judge, sender, and message as sources of variation in judgment accuracy, little variation, if any, should be attributable to judges. Because we assumed a general truth bias in the sample, we predicted higher truth than lie accuracy (veracity-effect hypothesis). If a sender's demeanor is relatively stable across situations and if Levine et al.'s (2011) argument applies that a (mis)match of demeanor and veracity determines accuracy, there should be a person effect of the sender that depends on veracity. Further, if there are "a few transparent liars" (see Levine, 2010, 2016) whose demeanor is (partly) linked to veracity, we should find a global person effect of senders across their messages. If characteristics or behaviors of senders that are not stable or situational factors affect the detectability of only single messages, there should be a message effect rather than a global sender effect (idiosyncratic-message-accuracy hypothesis).

Research Question 3: What Determines the Confidence With Which a Judgment Is Made?

To our knowledge, sources of variation in the confidence in veracity judgments have not been systematically examined. Smith and Leach (2019) argued that messages with stronger evidence for a lie should elicit higher levels of confidence; that is, the more a message looks like a lie, the higher the confidence in the veracity judgment made about it. If this theoretical argument applied, messages should be a crucial factor for the variation in confidence. A large body of literature from outside lie-detection research provides less support for this argument and rather suggests high variability among judges. Individuals' confidence ratings were found to be consistent within and across domains, suggesting a general self-confidence factor (e.g., Ais et al., 2016; Blais et al., 2005; Jackson & Kleitman, 2014; Jonsson & Allwood, 2003; Kantner & Dobbins, 2019; Kleitman & Stankov, 2007; Navajas et al., 2017; Pallier et al., 2002). Hence, confidence may reflect judges' assessments of their own performance rather than characteristics of the task itself or situational factors. Accordingly, research has identified multiple individual (judge) differences such as personality traits (e.g., Pallier et al., 2002; Pulford & Sohal, 2006; Wolfe & Grosch, 1990), need for cognition (e.g., Jonsson & Allwood, 2003), or gender (e.g., Vajapey et al., 2020) to predict confidence in domains other than lie detection. For lie detection in particular, judges' sex was identified as a predictor of confidence (e.g., DePaulo et al., 1997), as well as judges' level of shyness, social anxiety (both negative correlations), and extraversion (positive correlation; Vrij & Baxter, 1999). Moreover, Curci et al. (2018) found that judges' average confidence ratings vary more than average confidence ratings made about individual messages. Thus, we assumed that most variation in confidence should be due to judges (judge–confidence hypothesis).

The Current Research

In a reanalysis of data from four lie-detection studies, we partition the variation in lie detection attributable to judges, senders, and messages by using multiple truthful and deceptive messages from each sender and having them judged by multiple judges. In most previous studies (Bond & DePaulo, 2008; Levine, 2016; Levine et al., 2011; Masip et al., 2020), variability in lie-detection variables has been computed separately for judge dimensions (ability and credulity) and sender dimensions (detectability and credibility). To illustrate, the standard deviation of judges' ability was calculated across all senders, and, in a separate analysis, the standard deviation of senders' detectability was calculated across all judges. Thus, these analyses could not fully separate variability resulting from judges and senders because the calculation for the judge dimension was not independent of the variability on the sender dimension and vice versa. In addition, the sender–message entanglement in some previous studies did not allow differentiation between person effects of the sender and effects of the situated message.

Here, we used mixed-effects models to simultaneously estimate variability attributable to judges, senders, and individual messages, thus cutting through the knot of the sender-message entanglement. Thereby we could differentiate between person effects of judges and of senders and effects of the messages. Mixed-effects models use the single judgment as the unit of analysis, avoiding the aggregation of data that has been done in the past (see also Watkins & Martire, 2015). Instead, these models allow the contribution of judges, senders, and messages to be determined in terms of whether a lie or truth judgment is made and whether that judgment is correct or incorrect (i.e., consistent with actual veracity). For the first time, we also systematically examine the variability of confidence in veracity judgments.

Method

We reanalyzed the data of four lie-detection studies that used a stimulus material in which each sender delivered multiple messages (Lloyd et al., 2019). Here we present an overview of the study procedure that all four studies followed; further individual characteristics of the studies are outlined in Appendix A. All studies were conducted in accordance with the American Psychological Association's ethical standards.

Stimulus material

All studies used the messages from the Miami University Deception Detection Database (for detailed information on these materials, see Lloyd et al., 2019). To create the messages, Black and White female and male students and staff members from Miami University were invited to the laboratory; they were told that videos of lies and truths would be recorded there. The study used a fully factorial mixed design: 2 (Race: Black vs. White) × 2 (Gender: male vs. female) × 2 (Valence: positive vs. negative) × 2 (Veracity: honest vs. dishonest). Race and gender were between-subjects factors, and valence and veracity were within-subjects factors. Hence, each sender recorded four messages: one positive truth (talking positively about a person they liked), one negative lie (talking negatively about the person they liked), one negative truth (talking negatively about a person they disliked), and one positive lie (talking positively about the person they disliked). Participants had to describe why they liked or disliked the person and outlined their positive or negative qualities for a maximum of 45 s per message. Senders could neither view the videos nor redo the recording.

Lloyd et al. (2019) selected 20 senders of each of the four sender demographic categories (i.e., Black female, Black male, White female, White male) according to a priori inclusion criteria, resulting in a collection of 80 senders.[3] Characteristics of the selected senders can be found in Appendix B. In the four reanalyzed studies, the 320 messages were randomly assigned to one of 20 sets of 16 messages each. Each sender was featured

Table 1. Descriptive Statistics for Judges' Ability, Credulity, and Confidence of the Reanalyzed Studies

Study	No. of judges analyzed	No. of judges excluded	Per-judge means (*SD*) in %		
			Ability[a]	Credulity[b]	Confidence[c]
1	619	6	51.53 (12.02)	60.18 (15.17)	69.27 (11.64)
2	467	5	51.62 (11.21)	66.05 (17.01)	66.71 (11.85)
3	462	1	50.60 (10.83)	72.59 (18.37)	68.90 (12.63)
4	1,355	—	50.67 (11.05)	73.09 (18.91)	69.98 (12.23)
Overall	2,903	12	50.99 (11.26)	69.12 (18.54)	69.13 (12.16)

[a]Percentage of messages judged correctly. [b]Truth-judgment rates across all messages judged. [c]Measured on a percentage scale ranging from 0% to 100% in steps of 1 and averaged across all messages judged.

only once per set, and each set contained one video of each message condition. That is, each set included a positive truth, a negative truth, a positive lie, and a negative lie from one sender per demographic category, resulting in 16 messages per set.

Procedure of judgment studies

Informed consent was obtained from judges in all studies (for sociodemographic characteristics of judges, see Appendix C). In three of the reanalyzed studies, judges were randomly assigned to an experimental condition and worked on the respective task for the manipulation. Next, they engaged in the lie-detection task. In the fourth study, judges worked on the lie-detection task immediately because it was a correlative study with the respective variables measured after the lie-detection task (for more details on the individual studies, see Appendix A).

In each study, judges were randomly assigned to one of the 20 sets and judged the 16 messages therein. After having watched a video, judges stated whether they thought the sender was lying or telling the truth (binary judgment). In addition, they indicated how confident they were in that judgment on a percentage scale ranging from 0% to 100% in steps of 1. This procedure was repeated until judges had judged all messages of the assigned set.

Results

Data from all four studies were merged to one long-format data set. To control for possible influences of the different studies' experimental treatments, we included the seven-level factor "manipulation" with one level for each treatment (two levels for each of Studies 1–3 and one level for Study 4, which had no experimental treatment). For each research question, we estimated a set of mixed-effects models in R using the lme4 package (Bates et al., 2015) and compared the models using chi-squared likelihood-ratio tests from the anova() function. Because these model comparisons do not allow for missing data, data from 12 judges who did not report their gender were excluded. This was done to keep the number of observations analyzed constant across models, even when the judges' gender was included. The exclusion of these individuals did not affect the conclusions regarding the partitioning of the variation between judge, sender, and message. The number of excluded judges per study is displayed in Table 1. Further, Table 1 shows per-judge means and standard deviations of the three dependent variables for each of the reanalyzed studies. Whereas judges' overall truth-classification rate of almost 70% was above the truth-classification rate of approximately 56% found in the Bond and DePaulo (2006) meta-analysis, the overall accuracy of 51% was below the meta-analysis average of 54%. Plots of the distribution of judge, message, and sender scores are included in the Supplemental Material available online. Interested readers can find the data set on which the following analyses are based on OSF at https://osf.io/f7wbd/.

Research question 1: What determines whether a message is classified as lie or truth?

To determine the variance components of whether a message is classified as lie or as truth, we estimated several logistic mixed-effects models with the judgment-type variable (0 = lie, 1 = truth) as the dependent variable. Model results are displayed in Table 2, including model statistics and the results of the likelihood-ratio tests of model comparisons.

In Model 1, we added a random intercept for the manipulation to account for the potential treatment effects of the four studies. The experimental manipulations

Table 2. Logistic Mixed-Effects Models and Model Comparison Statistics for the Judgment-Type Variable (0 = Lie, 1 = Truth)

	Model 1				Model 2				Model 3				Model 4				Model 5				Model 6			
		No. of groups	*SD*	*ICC*		No. of groups	*SD*	*ICC*		No. of groups	*SD*	*ICC*		No. of groups	*SD*	*ICC*		No. of groups	*SD*	*ICC*		No. of groups	*SD*	*ICC*
Random effects																								
Manipulations		7	0.26	0.02		7	0.27	0.02		7	0.27	0.02		7	0.31	0.02		7	0.31	0.02		7	0.31	0.02
Messages						320	0.39	0.04		320	0.29	0.02		320	0.32	0.02		320	0.30	0.02		320	0.30	0.02
Senders										80	0.26	0.02		80	0.29	0.02		80	0.29	0.02		80	0.29	0.02
Judges														2,903	0.82	0.16		2,903	0.82	0.16		2,903	0.82	0.16
	Estimate	*SE*	*z*	*p*	Estimate	*SE*	*z*	*p*	Estimate	*SE*	*z*	*p*	Estimate	*SE*	*z*	*p*	Estimate	*SE*	*z*	*p*	Estimate	*SE*	*z*	*p*
Fixed effects[a]																								
(*Intercept*)	0.77	0.10	7.99	< .001	0.80	0.10	7.70	< .001	0.80	0.11	7.45	< .001	0.91	0.13	7.25	< .001	0.91	0.13	7.27	< .001	0.91	0.12	7.30	< .001
Veracity																	0.05	0.02	2.59	.010	0.05	0.02	2.59	.010
Valence																	0.07	0.02	3.73	< .001	0.07	0.02	3.73	< .001
Senders' gender																	0.07	0.04	1.75	.080	0.07	0.04	1.75	.080
Senders' race																	−0.01	0.04	−0.25	.799	−0.01	0.04	−0.25	.799
Judges' gender																					0.02	0.02	1.02	.308
Model fit																								
BIC				56,863				56,006				55,967				54,081				54,101				54,111
Deviance				56,842				55,974				55,924				54,027				54,005				54,004
χ^{2b}								868.19				49.97				1,896.16				22.70				1.03
df								1				1				1				4				1
p								< .001				< .001				< .001				< .001				.310

Note: BIC = Bayesian information criterion.

[a]Effect coded: message veracity (−1 = lie, 1 = truth), valence (−1 = negative, 1 = positive), senders' gender (−1 = female, 1 = male), senders' race (−1 = Black, 1 = White), and judges' gender (−1 = female, 1 = male). [b]Statistics of model comparisons from the respective model to the previous model.

accounted for 2% of the variance in the judgment-type variable (see Table 2). In Model 2, we added a random intercept for messages that explained another 4% of the variance. In Model 3, we added a random intercept for senders that explained 2% of the variance and reduced the variance explained by messages to 2%. In Model 4, a random intercept for judges was entered that explained 16% of the variance. In Model 5, we added fixed effects for the stimulus-specific variables (effect-coded): message veracity (−1 = lie, 1 = truth), valence (−1 = negative, 1 = positive), senders' gender (−1 = female, 1 = male), and senders' race (−1 = Black, 1 = White). The fixed effects of veracity, $b = 0.05$, $SE = 0.02$, $z = 2.59$, $p = .010$, and valence, $b = 0.07$, $SE = 0.02$, $z = 3.73$, $p < .001$, were significant. Truthful messages, $M = 72.40\%$, 95% confidence interval (CI) = [67.15%, 77.09%], were overall more likely to be judged as truthful compared with deceptive messages, $M = 70.27\%$, 95% CI = [64.82%, 75.20%], $OR = 1.11$, 95% CI = [1.03, 1.20]. Positive messages, $M = 72.85\%$, 95% CI = [67.65%, 77.49%], were overall more likely to be judged as truthful compared with negative messages, $M = 69.79\%$, 95% CI = [64.30%, 74.77%], $OR = 1.16$, 95% CI [1.07, 1.26]. In Model 6, we added judges' gender as a fixed effect (−1 = female, 1 = male), which was not significant.

Model comparisons revealed that Models 2 to 4 significantly improved the model fit compared with the previous model. Entering the stimulus-specific variables in Model 5 also significantly improved the model fit; however, the lower Bayesian information criterion (BIC) of Model 4 as opposed to Model 5 indicates that this might have been due to the higher number of predictors in Model 5. When adding the judges' gender in Model 6, the model fit was not improved significantly.

Most of the variance in judgment type was indeed attributable to judges (16%); senders and messages were less relevant, each accounting for 2% of the variance. As predicted in the individual-differences hypothesis, whether a message was classified as a lie or as true depended primarily on the person making the judgment and to a lesser extent on the particular sender; however, messages explained a similar share of the variance as senders. The model fit was significantly improved when we added a random intercept for senders to the model that already contained a random intercept for messages. This suggests that not only the message itself contributed to whether it was believed to be true but also, in line with the idea of sender demeanor (see, e.g., Levine et al., 2011), the respective person delivering the message. Neither stimulus-specific fixed effects nor judges' gender accounted for much of the variance of judgment type.

Research question 2: What determines whether a judgment is correct or incorrect?

To determine the variance components of the accuracy of judgments, we estimated several logistic mixed-effects models with the accuracy variable (0 = judgment incorrect, 1 = judgment correct) as the dependent variable. Model results are displayed in Table 3, including model statistics and the results of the likelihood-ratio tests of model comparisons.

In Model 1, the experimental manipulations accounted for less than 0.01% of the variance in the accuracy variable. In Model 2, we added a random intercept for messages that explained 21% of the variance. In Model 3, we added a random intercept for senders to see how much variance senders would explain on top of the variance explained by the specific messages (i.e., whether some senders are generally easier to detect than others). The random intercept for senders explained less than 0.01% of the variance in this model. In Model 4, a random intercept for judges was entered that accounted for less than 0.01% of the variance. In Model 5, we added the fixed effects for the stimulus-specific variables. Veracity was the only significant predictor in the model, $b = 0.83$, $SE = 0.02$, $z = 35.58$, $p < .001$. As predicted by the veracity-effect hypothesis, truths, $M = 70.63\%$, 95% CI = [69.26%, 71.97%], were overall more likely to be judged correctly compared with lies, $M = 31.29\%$, 95% CI = [29.91%, 32.71%], $OR = 5.28$, 95% CI = [4.82, 5.79]. In Model 6, we added judges' gender as a fixed effect, which was not significant.

Model comparisons indicated that the random intercept for messages entered in Model 2 and the stimulus-specific variables entered in Model 5 improved the model fit. Because the variance explained by messages was reduced by about 17 percentage points when veracity was included in Model 5, it appears that the veracity of a message is a good predictor of whether it will be judged correctly (see also Levine et al., 1999). To examine the variance resulting from veracity and the variance explained by other idiosyncratic message features more thoroughly, we estimated Model 6 without the random intercept for messages and compared it to Model 6. The model with the random intercept for messages explained significantly more variance than the model without the random intercept, $\chi^2(1) = 781.03$, $p = < .001$. The pseudo R^2 for the fixed effects of the models suggested that message veracity explained about 16.8% of the variance in judgment accuracy, whereas other idiosyncratic characteristics of messages accounted for 4% of the variance. This suggests that messages differed in their likelihood of being judged

Table 3. Logistic Mixed-Effects Models and Model Comparison Statistics for the Accuracy Variable (0 = Judgment Incorrect, 1 = Judgment Correct)

	Model 1				Model 2				Model 3				Model 4				Model 5				Model 6				Model 7 (random-slope model)			
		No. of groups	*SD*	*ICC*		No. of groups	*SD*	*ICC*		No. of groups	*SD*	*ICC*		No. of groups	*SD*	*ICC*		No. of groups	*SD*	*ICC*		No. of groups	*SD*	*ICC*		No. of groups	*SD*	*ICC*
Random effects																												
Manipulations		7	0.00	0.00		7	0.01	0.00		7	0.01	0.00		7	0.01	0.00		7	0.01	0.00		7	0.00	0.00		7	0.00	0.00
Messages						320	0.92	0.21		320	0.92	0.21		320	0.92	0.21		320	0.38	0.04		320	0.38	0.04		320	0.27	0.02
Senders										80	0.00	0.00		80	0.00	0.00		80	0.00	0.00		80	0.00	0.00		80	0.00	0.00
Judges														2,903	0.00	0.00		2,903	0.00	0.00		2,903	0.00	0.00		2,903	0.00	0.00
Veracity | senders																											0.26	
	Estimate	*SE*	*z*	*p*	Estimate	*SE*	*z*	*p*	Estimate	*SE*	*z*	*p*	Estimate	*SE*	*z*	*p*	Estimate	*SE*	*z*	*p*	Estimate	*SE*	*z*	*p*	Estimate	*SE*	*z*	*p*
Fixed effects[a]																												
(*Intercept*)	0.04	0.01	4.29	< .001	0.04	0.05	0.84	.400	0.04	0.05	0.84	.400	0.04	0.05	0.84	.400	0.05	0.02	1.93	.054	0.05	0.02	2.00	.045	0.05	0.02	2.56	.011
Veracity																	0.83	0.02	35.58	< .001	0.83	0.02	35.58	< .001	0.83	0.03	24.47	< .001
Valence																	0.00	0.02	0.11	.912	0.00	0.02	0.11	.912	0.00	0.02	0.15	.885
Senders' gender																	−0.04	0.02	−1.50	.133	−0.04	0.02	−1.50	.133	−0.03	0.02	−1.86	.063
Senders' race																	0.02	0.02	0.75	.451	0.02	0.02	0.75	.451	0.02	0.02	0.96	.335
Judges' gender																					−0.02	0.01	−1.50	.133	−0.02	0.01	−1.52	.129
Model fit																												
BIC				64,394				57,114				57,125				57,136				56,662				56,671				56,640
Deviance				64,372				57,082				57,082				57,082				56,566				56,563				56,511
χ^{2b}								7,290.14				0.00				0.00				516.36				2.24				52.80
df								1				1				1				4				1				2
p								< .001				1.00				1.00				< .001				.134				< .001

Note: BIC = Bayesian information criterion.

[a] Effect coded: message veracity (−1 = lie, 1 = truth), valence (−1 = negative, 1 = positive), senders' gender (−1 = female, 1 = male), senders' race (−1 = Black, 1 = White), and judges' gender (−1 = female, 1 = male). [b] Statistics of model comparisons from the respective model to the previous model.

correctly not only because some of them were true and others were lies but also because of other idiosyncratic characteristics of the messages.

To test the idea of a (mis)match of demeanor and veracity determining the accuracy of a judgment, we estimated Model 7, in which we entered a random slope of veracity for senders.[4] In doing so, we could test whether lie and truth detectability differed at the level of the individual sender. Model comparisons showed that the inclusion of this random slope improved the model fit compared with the previous model; that is, lie and truth detectability differed at the level of the individual sender, supporting the idea of a (mis)match of demeanor and veracity determining the accuracy of a judgment (see also Levine et al., 2011).

In line with the veracity-effect hypothesis, lie and truth accuracy differed at a global level; truth accuracy was higher than lie accuracy, and veracity was the best predictor of judgment accuracy. Supporting the idiosyncratic-message-accuracy hypothesis, messages explained a significant part of the variance of judgment accuracy, even when controlling for veracity. There was not a global person effect of senders, and the distribution of the senders' detectability scores did not suggest the existence of transparent liars (see Levine, 2010) in the analyzed data (for details on distributions, see Supplemental Material).[5] Judges' gender and stimulus-specific fixed effects other than veracity did not explain much variance.

Research question 3: What determines the confidence with which a judgment is made?

To determine the variance components of confidence, we estimated several linear mixed-effects models with confidence as the dependent variable. Model results are displayed in Table 4, including model statistics and the results of the likelihood-ratio tests of model comparisons.

The random intercepts for manipulation, messages, and senders entered in Models 1 to 3 each explained less than 1% of the variance in confidence. In Model 4, a random intercept for judges was entered that accounted for 46% of the variance. In Model 5, we added the fixed effects for the stimulus-specific variables. Valence was the only significant predictor in the model, b = 0.15, SE = 0.08, t = 2.01, p =.046. Positive messages, M = 69.16%, 95% CI = [68.24%, 70.07%], were given higher confidence ratings than negative messages, M = 68.85%, 95% CI = [67.94%, 69.77%]. In Model 6, we added judges' gender as a fixed effect, which was not significant.

Model comparisons revealed that with each added random intercept, the model fit was significantly improved. The stimulus-specific variables entered in Model 5 also significantly improved the model fit; however, the lower BIC of Model 4 as opposed to Model 5 indicates that this might have been due to the higher number of predictors in Model 5. Adding judges' gender in Model 6 did not significantly improve the model fit.

In line with the judge-confidence hypothesis, confidence was mainly determined by the person who was making the veracity judgment and, thus, on the person who was also giving the confidence rating. Although the random intercepts for senders and messages improved the model fit, the part of the variance explained by them was smaller than 1%. Hence, compared with the 46% variance attributable to judges, senders and messages played a subordinate role for confidence. Neither stimulus-specific fixed effects nor judges' gender accounted for much of the variance in confidence.

Discussion

In a reanalysis of four studies with more than 45,000 veracity judgments, we partitioned the variance in lie detection attributable to judges, senders, and messages. Contrary to previous studies, we separated the variance being due to a global person factor of the sender from the variance being due to the individual messages these senders conveyed.

Research question 1: What determines whether a message is classified as lie or truth?

In line with previous studies that found higher judge variation than stimulus variation (e.g., Levine et al., 2022; Masip et al., 2020) but in contrast to those that found the opposite pattern (e.g., Bond & DePaulo, 2008; Levine, 2016), variance in whether a message was classified as a lie or as truth was mainly due to judges (16%). Senders and messages each explained 2% of the variance. Hence, the results were in line with the individual-differences hypothesis in that person effects played the largest role in whether a message was judged as a lie or as truth. Our results support both the idea that lie-versus-truth classifications are determined by judges' individual tendencies to rate messages as true and, although to a lesser extent, the idea that lie-versus-truth classifications are determined by sender demeanor (see, e.g., Levine et al., 2011). The 2% variance explained by senders indicated that some senders were more likely to be believed to be truthful than others across their messages, indicating a more sincere demeanor of these

Table 4. Mixed-Effects Models and Model Comparison Statistics for the Confidence Variable (0% to 100%)

	Model 1				Model 2				Model 3				Model 4				Model 5				Model 6			
		No. of groups	*SD*	*ICC*		No. of groups	*SD*	*ICC*		No. of groups	*SD*	*ICC*		No. of groups	*SD*	*ICC*		No. of groups	*SD*	*ICC*		No. of groups	*SD*	*ICC*
Random effects																								
Manipulations		7	1.09	0.00		7	1.09	0.00		7	1.09	0.00		7	0.99	0.00		7	0.99	0.00		7	0.99	0.00
Messages						320	1.17	0.00		320	0.82	0.00		320	0.90	0.00		320	0.86	0.00		320	0.86	0.00
Senders										80	0.83	0.00		80	0.94	0.00		80	0.92	0.00		80	0.92	0.00
Judges														2,903	11.70	0.46		2,903	11.70	0.46		2,903	11.70	0.46
	Estimate	*SE*	*t*	*p*	Estimate	*SE*	*t*	*p*	Estimate	*SE*	*t*	*p*	Estimate	*SE*	*t*	*p*	Estimate	*SE*	*t*	*p*	Estimate	*SE*	*t*	*p*
Fixed effects[a]																								
(*Intercept*)	68.97	0.42	164.26	< .001	68.97	0.42	162.30	< .001	68.97	0.43	158.75	< .001	69.01	0.46	150.00	< .001	69.01	0.46	150.22	< .001	68.99	0.46	149.71	< .001
Veracity																	0.13	0.08	1.76	.080	0.13	0.08	1.76	.080
Valence																	0.15	0.08	2.01	.046	0.15	0.08	2.01	.046
Senders' gender																	−0.18	0.13	−1.40	.164	−0.18	0.13	−1.40	.164
Senders' race																	0.11	0.13	0.85	.399	0.11	0.13	0.85	.399
Judges' gender																					0.17	0.23	0.73	.465
Model fit																								
BIC				396,043				396,005				395,999				374,837				374,870				374,880
Deviance				396,011				395,962				395,945				374,772				374,763				374,762
χ^{2b}								49.14				16.51				21,172.84				9.66				0.53
df								1				1				1				4				1
p								< .001				< .001				< .001				.047				.465

Note: BIC = Bayesian information criterion.

[a]Effect coded: message veracity (−1 = lie, 1 = truth), valence (−1 = negative, 1 = positive), senders' gender (−1 = female, 1 = male), senders' race (−1 = Black, 1 = White), and judges' gender (−1 = female, 1 = male). [b]Statistics of model comparisons from the respective model to the previous model.

senders. As also noted by Levine et al. (2011) but so far not tested, sender demeanor does not seem to be "completely trait-like" (p. 380). In addition, situational factors captured in the message (2% explained variance) influenced lie-versus-truth classifications beyond the person effect of senders. This indicates that the previously found variation on the stimulus side was likely not only due to a person effect of senders but also determined by the specific messages used from these senders. Further research is needed to investigate what stable sender characteristics and behaviors may be responsible for a sender's global credibility and how the situation as well as specific behaviors displayed in individual messages play into this. To this end, studies in which senders record a variety of messages (e.g., high- vs. low-stakes lies, interview situations with follow-up questions vs. simple statements, short vs. long messages) might be particularly useful for examining the interaction of senders and situations.

Research question 2: What determines whether a judgment is correct or incorrect?

The variance in judgment accuracy was mainly due to messages, particularly message veracity (about 16.8% of the total variance). Other idiosyncratic characteristics of messages explained an additional 2% to 4% of the variance, whereas judges and senders explained less than 0.01% each. For judges, this finding aligns with our predictions and previous work suggesting that individuals, if at all, vary only little in their ability to discriminate between truth and deception (see, e.g., Bond & DePaulo, 2008; Levine, 2016). As hypothesized, the accuracy of a judgment depended mostly on the message, on idiosyncratic features of the messages (idiosyncratic-message-accuracy hypothesis), and especially on its veracity (veracity-effect hypothesis). In line with Levine et al. (1999), the veracity of a message was "the single best predictor of detection accuracy" (p. 139); a message was most often judged correctly when it was a true message. When controlling for message veracity and sender-specific effects of veracity, the random intercept for messages still explained 2% of the variance in judgment accuracy. Thus, apart from the veracity of the message, messages appear to have characteristics that are associated with higher or lower detectability, for example, situational factors, or the state that the senders are in when they convey the message.

Contrary to the predicted global sender effect across messages, senders did not differ much in detectability across their messages. There was not a significant amount of senders whose credibility was linked to actual veracity (i.e., who appear credible when they tell the truth and uncredible when they lie; see also Levine et al., 2010), and the descriptive data did not suggest the existence of a transparent liar in the stimulus material that was used. Instead, the detectability of senders seemed to differ on an individual level between their truthful and deceptive messages. Given that the senders also varied to some extent in their credibility (2% explained variance in lie-vs.-truth classifications), the variance explained by the sender-veracity combination supports Levine et al.'s (2011) argument that a (mis) match of sender demeanor and veracity determines accuracy. This theory has so far not been tested across multiple truthful and deceptive messages from a sender and still requires further research with systematic variation of the situation. Such research would also allow testing whether there are transparent liars or transparent lies and could help in understanding how the interplay of senders, idiosyncratic message properties, and message veracity determines judgment accuracy.

Because the global sender effect that has been implied by previous work was not found here, we would adapt the conclusion from the meta-analysis by Bond and DePaulo (2008) that "the accuracy of a deception judgment depends more on the liar than the judge" (p. 486). Having cut through the sender-message entanglement knot, the analyzed data suggest that the accuracy of a veracity judgment may depend more on the message than on the judge and the sender, and given the typical truth bias, especially on whether the message is a lie or truth. Only the combination of sender and veracity, but not senders alone, seems to contribute to whether a judgment is correct.

Research question 3: What determines the confidence with which a judgment is made?

As predicted in the judge-confidence hypothesis and in line with the idea of a general confidence factor, most variance in confidence was due to judges (46%), and only little was due to senders and messages (less than 1% each); hence, confidence mostly depends on the person making the judgment. This finding aligns with research suggesting that individuals are largely consistent in their confidence ratings (e.g., Ais et al., 2016; Jackson & Kleitman, 2014; Kantner & Dobbins, 2019; Pallier et al., 2002) and with studies that identified interindividual judge differences as predictors for confidence (see, e.g., Jonsson & Allwood, 2003; Vajapey et al., 2020; Vrij & Baxter, 1999). Our results do not support the theoretical claim based on signal-detection theory that messages with more evidence of a lie (higher detectability) also elicit higher confidence ratings (see Smith & Leach, 2019). Even though some messages were easier to detect than others, these messages likely had not generally elicited higher confidence ratings. The variation in confidence was mainly due to judges, but they did not vary in their ability to detect lies; whereas some judges were more confident than others across their judgments, this higher confidence was overall not supported by higher accuracy scores. Thus, like several previous studies (e.g., DePaulo et al., 1997; Hartwig et al., 2017; Volz et al., 2022), our research gives little reason to believe that there is a relationship between confidence and accuracy, neither from the perspective of the message nor from the perspective of the judge.

Implications and suggestions for future research

Relating the results for the three variables, an ironic picture emerges. As senders, it is most important to individuals that their truths be recognized as truths and that their lies remain undetected; that is, they strive to be seen as truthful regardless of whether they lie or tell the truth. However, in our data, senders varied little in their credibility; the variance in the classification of lies and truth was mainly due to judges. As judges, it is most important for individuals to make correct judgments; that is, they strive to correctly discern whether someone is telling the truth or lying to them. However, judges varied little, if at all, in their ability to distinguish between truth and deception; instead, messages were the most important source of variation in whether a judgment was correct. Although likely not more accurate, some judges were more confident in their judgments than others, suggesting that at least some judges were not aware of their lack of lie-detection ability.

The reanalyzed studies that led to the above conclusions are prototypical for many lie-detection studies (see, e.g., Bond & DePaulo, 2006). The stimulus material depicted an everyday-life social situation, and there was no special form of motivation for senders; the lies were relatively low stakes, from a nonforensic context, and sanctioned (i.e., induced by the experimental procedure). Judges made binary ad hoc veracity judgments about these rather short (< 1 min) videotaped messages. Judges had no direct interaction with senders, no exposure to senders' baseline behavior, and no prior information about them. The message recording was rather standardized and gave constraints on senders (e.g., specified maximum talking time, no choice whether to lie but instead assignment to veracity condition for a given topic). Senders were not interviewed, meaning there were no follow-up questions on what they had conveyed. Here, we demonstrated what sources of variance may look like in such typical lie-detection studies with a sender sample recruited at a university and an independent, more diverse judge sample. Although some of the factors listed above altered the variance explained by individual sources in Bond and DePaulo's (2008) moderator analyses (Table 1), the main results were nonetheless generalizable across the moderators analyzed; that is, judge ability showed the least variation across conditions, and sender credibility showed the highest variation. These analyses did not treat senders and messages as separate entities and should therefore be interpreted with caution, still, they show that there might be important moderators to be understood when trying to determine variation in lie detection. Regardless of their generalizability, our results show that messages play a larger role than previously acknowledged, at least in fairly prototypical lie-detection studies as those analyzed here. It remains to be determined whether this is also the case in studies with modifications of the procedure, samples, and stimulus material and whether the results can be generalized to other contexts (e.g., politics), to high-stakes lies, or to scenarios in which judges and senders know each other.

For prototypical lie-detection studies similar to the ones analyzed here and as characterized above, we believe that our results will replicate when the sender-message entanglement is accounted for. Especially for confidence and accuracy we have reason to believe that the results will replicate despite variations in the study procedure. Note that with replication, we do not mean the specific amount of variance explained by each source but the main source of variance for the particular dependent variable. For confidence judgments, we assume high replicability of the results because judges accounted for almost half of the variance. Increasing our confidence in the replicability of this result, research in other fields suggests the existence of a general self-confidence factor because individuals' confidence ratings have been shown to be consistent within and across domains (e.g., Ais et al., 2016; Blais et al., 2005; Jackson & Kleitman, 2014; Jonsson & Allwood, 2003; Kantner & Dobbins, 2019; Kleitman & Stankov, 2007; Navajas et al., 2017; Pallier et al., 2002).

For accuracy, we assume that the lack of judge variation will replicate because it is in line with past findings (e.g., Bond & DePaulo, 2008; Levine, 2016), there are rarely any judge variables that predict lie-detection ability (see, e.g., Aamodt & Custer, 2006; Bond & DePaulo, 2006, 2008), and results were stable across the four large judge samples. Therefore, a search for manipulations on the judge side and judge variables associated with increased lie-detection ability hardly seems worthwhile.

Instead, interventions and changes in the study procedure that address the message and the situation as well as verbal-deception detection methods will likely have the most impact on accuracy (see, e.g., Vrij, 2015; Vrij et al., 2011). Given the robust finding of a truth bias, we assume that the veracity effect will replicate but might be smaller because the truth-judgment rates here were overall about 13 percentage points higher than the roughly 56% typically found in similar studies (for a meta-analysis, see Bond & DePaulo, 2006). Judge groups with a lower tendency to rate messages as true (e.g., police officers; see Masip et al., 2005) may reduce the veracity effect further or may even make it disappear. Whether there are transparent liars should be tested in studies with multiple truthful and deceptive messages of each sender from a variety of contexts and situations.

For lie-versus-truth classifications, we assume not only that the findings of judges being the main source of variance will replicate but also that senders in particular might gain some influence. Whereas the judge sample was rather heterogeneous (e.g., in age, professional experience, and education), the sender sample was not. More homogeneous judge groups, such as police officers, may reduce the judge variance resulting from similar base-rate assumptions. Manipulations affecting judges' information-processing strategies could also have an impact on the variation in lie-versus-truth classifications. Although we found only little sender variation here, we would not conclude that senders generally do not differ in credibility. The senders were young adults, probably predominantly from WEIRD populations (Western, educated, industrialized, rich, and democratic; Henrich et al., 2010); many of them were psychology students and probably of above-average intelligence. Characteristics found to affect credibility such as social and emotional skills (e.g., Riggio et al., 1987) or attractiveness (e.g., Patzer, 1983; Zebrowitz et al., 1996) may have varied little in this sender sample, restricting the variance explained by senders. Including groups of individuals who may be seen as less credible such as socially anxious or autistic individuals (e.g., Lim et al., 2022) or less proficient speakers (Da Silva & Leach, 2013; Evans & Michael, 2014) likely increases sender variation. Further, individuals who feel uncomfortable with being videorecorded or with lying (in front of a camera) may not have engaged in the videorecording study, or they may have acted in a way that led to them being excluded from the final material (e.g., talking very shortly). If these sender characteristics or metacognitions about one's ability to lie are related to perceived credibility, the absence or exclusion of these individuals may have limited the variation in sender credibility. Further, if transparent liars avoid putting themselves in situations in which they have to lie, as suggested by Levine et al. (2010), studies including the abovementioned individuals could also shed light on the existence of a few transparent liars.

We have presented and tested different theoretical ideas to discern sources of variation in lie detection. More research is needed to uncover the underpinnings of this variation and to explore potential interactions between the factors. One focus could be on the interaction of senders and situations to better understand senders' credibility across situations. Situations may vary in the difficulty for senders to appear credible, especially when lying; in more difficult situations, sender variation is probably higher because only some senders succeed in coming across as credible, whereas most senders succeed in easy situations. In highly restrictive situations in which messages vary only little in their content (e.g., when senders deny having committed a crime; see also Vrij & Baxter, 1999), senders might be a more important source of variation in lie-versus-truth classifications. When what is said varies only little between messages, judges might rely more on different sender attributes. Crucial sender attributes determining their credibility in such situations may lay in senders' general appearance, such as their attractiveness (see also Patzer, 1983; Zebrowitz et al., 1996). Although many situations in everyday life appear rather restrictive, there are also situations that grant more freedom to senders (e.g., when getting to know someone in a dating context). In such rather unrestricted situations, senders are freer to decide whether to lie or not, in choosing the topic they want to talk (or lie) about, and in how they want to present their message (e.g., how long they want to talk). Some senders may be generally more successful than others in making these decisions, resulting in higher credibility even when lying. In addition, some senders may be better than others at monitoring other people's reactions to them and in adjusting their strategy accordingly. This would make these senders appear more credible across unrestricted situations, but not necessarily across more restricted situations. Not only Sender × Message but also Sender × Judge interactions seem possible, especially if judges have a prior attitude toward or a relation with senders (e.g., being classmates or friends).

In addition to message veracity, idiosyncratic message characteristics and the combination of the particular sender and the veracity of the message determined judgment accuracy. To examine the underpinnings of these effects, research is needed with stimulus materials that include systematic within-sender manipulations of the situation and of veracity. This allows identifying characteristics and behaviors that determine (a) the credibility of a sender across situations/messages and (b) the credibility of individual messages. Equally important, such research will help in understanding how these characteristics and behaviors, in combination with veracity, contribute to judgment accuracy. To illustrate with a simplified example for sender credibility, one stereotype is that liars avoid eye contact (e.g., Global Deception Research Team, 2006). Therefore, senders who tend to avoid eye contact across situations might be perceived as liars more often. When these senders lie, they would be more likely to be judged correctly than when they tell the truth. A similar phenomenon could also occur at the message level, which might explain differences in message detectability. Even though a lot of research on cues to deception in messages has been conducted (for a meta-analysis, see DePaulo et al., 2003), there is doubt about the actual diagnostic utility of these cues (see Luke, 2019). Instead of such systematic differences between truth and deception, differences in message detectability could be a rather random product of the combination of message characteristics and veracity, similar to the proposition by Levine et al. (2011) that the combination of sender demeanor and veracity determines accuracy. To illustrate again with a simplified example, suppose the consistency of messages varies randomly and independently of actual veracity. One stereotype is that lies are inconsistent (e.g., Global Deception Research Team, 2006), so a message that is inconsistent may be more often judged correctly when it is a lie than when it is a true message. Thus, a lie would be correctly judged particularly often if it fits the stereotype of a lie, and a truth would be correctly judged particularly often if it fits the stereotype of a truth.

As touched on above, our results provide guidance for directing future research efforts, including possible steps to improve the accuracy of lie detection in practice. Because individuals (i.e., judges and senders) explained almost no variance in judgment accuracy, it may be more promising to focus on (situated) messages rather than interindividual differences. In other words, developing message-oriented approaches to increase message detectability seems more promising than attempting to identify "good lie detectors" and deploying them in relevant positions (e.g., in forensic contexts). Situations, procedures, and questions should be identified that lead to increased detectability of messages, or existing procedures should be further developed (see, e.g., Hartwig et al., 2014; Levine et al., 2010; Vrij et al., 2009). Although we recommend focusing on the part that explains most of the variance in accuracy, namely the message, this should not translate into a complete disregard for the individuals making veracity judgments. Usually, it still comes down to individuals applying message-oriented approaches and, in the end, making the veracity judgment. Accordingly, they should be enabled to make the best use of the information from message-oriented approaches. To illustrate, the cognitive-load approach (see, e.g., Vrij et al., 2006, 2008, 2011) tries to create a diagnostic situation by placing additional demands on senders' cognitive resources to amplify differences between truthful and deceptive messages. To make this approach as successful as possible, judges should then know how to best exploit the resulting differences between truthful and deceptive messages (see also Mac Giolla & Luke, 2021).

Levine et al. (2022) recommended using larger sender samples in lie-detection studies to avoid idiosyncratic results and to increase the stability and replicability of findings. However, because the study did not differentiate between sender and messages, the study leaves open whether more senders, more messages, or more of both are required. Because senders compared with messages appear to play a smaller role for the variation in judgment accuracy, our data suggest that researchers could collect multiple messages from each sender (ideally balanced for veracity) to increase stimulus variability when examining judgment accuracy. This approach requires relatively little additional effort compared with an approach in which each sender records only one message (i.e., in which the number of senders equals the number of messages; see also Levine et al., 2022). Given the limitations of the results discussed above, further research is needed to determine how to ensure stimulus variability in a resource-efficient manner. For now, it seems that collecting multiple messages per sender from a variety of situations (e.g., high-stakes and low-stakes lies, different topics) is a good approach when researching accuracy. Depending on the research question, scholars could also combine messages from different stimulus materials in a single judgment study to increase stimulus variability and to reduce the risk that findings do not generalize across situations. When exploring confidence or lie-versus-truth classifications, the number of judges and senders is also crucial to consider because they also contribute to the variation in these variables. However, because the number of messages can limit the power of a study even when the number of judges approaches infinity, the number of messages must also be taken into account in a priori power calculations (for guidance, see, e.g., Aarts et al., 2014; Westfall et al., 2014).

The reported results demonstrate the importance of separating the concepts of judges, senders, and messages, whether in developing theories, in deriving hypotheses, in study designs, or in statistical analyses. Our results suggest that there are considerable proportions of variance attributable to judges, senders, and messages. Thus, this study is also a call to consider the hierarchical data structure of lie-detection studies in statistical methods. As shown by Judd et al. (2012) for studies using stimulus material in general and by Watkins and Martire (2015) for deception-detection studies in particular, mixed-effects models can accommodate both judge-side and stimulus-side variance. Accounting for these variances can increase the replicability of findings by reducing the impact of idiosyncratic stimulus materials on the overall results of studies and can help to prevent Type I error inflation (Aarts et al., 2014; Judd et al., 2012; Westfall et al., 2015). In addition, mixed-effects models can avoid the loss of information resulting from data aggregation (Judd et al., 2012; Watkins and Martire, 2015), for instance, when calculating per-judge accuracy scores. They also allow investigating more complex research designs (e.g., determining the relation between judgments and individual message characteristics), which can be beneficial when determining why one message is easier to detect than another.

Conclusion

In this article, we highlighted the importance of thinking precisely about the deception-detection process and the components involved in it. Treating the message and the sender as two distinct entities allowed us to shed new light on previously held beliefs. For instance, the belief that the accuracy of a veracity judgment depends mainly on the sender did not stand up to closer scrutiny in the data we analyzed: Senders did not generally differ in their detectability. Instead, some messages were easier to detect than others. We do not claim, however, that the results presented can be generalized to all types of lie-detection situations. Rather, we want to raise awareness that messages have often been underestimated as a factor in typical lie-detection studies to date and need to be properly considered in the future. Being clear about the concepts of judges, senders, and messages in theoretical work and using the appropriate methods to test the derived hypotheses can help to guide research efforts efficiently. Because the variance proportions may not be as definite and stable as previously thought, we will probably have to discard some of our previous assumptions, which will advance the field of lie detection considerably.

Appendix A. Manipulation and Measures Used Before the Lie-Detection Task of the Reanalyzed Studies

Study	Manipulation	Measures
1	Uncertainty salience manipulation: two open-ended questions in which participants should write about feeling uncertain (experimental condition) or about feeling certain (control condition) about themselves	Manipulation check PANAS
2	—	—
3	Mortality salience manipulation: two open-ended questions in which participants should write about their own death (experimental condition) or about their favorite food (control condition)	PANAS
4	Manipulation of ostracism using the cyberball game	Honesty-humility

Note: PANAS = Positive and Negative Affect Schedule.

Appendix B. Sociodemographic Characteristics of Senders

Characteristics of senders	Values
Gender, *n* (%)	
Female	40 (50.0)
Male	40 (50.0)
Race, *n* (%)	
Black	40 (50.0)
White	40 (50.0)
Employment, *n* (%)	
Students majoring in psychology	33 (41.3)
Students majoring in other fields	45 (56.3)
Staff member	1 (1.3)
Not indicated	1 (1.3)
Age, *M* (*SD*)	20.20 (1.5)

Appendix C. Sociodemographic Characteristics of Judges (Across Analyzed Studies)

Characteristics of judges	Values
Gender, *n* (%)	
Female	1,317 (45.4)
Male	1,586 (54.6)
Education, *n* (%)	
Less than a high school diploma	45 (1.6)
High school degree or equivalent	45 (1.6)
Associate degree	128 (4.4)
Bachelor's degree	1,669 (57.5)
Master's degree	724 (24.9)
Professional degree	55 (1.9)
Doctorate	28 (1.0)
Some college, no degree	203 (7.0)
Not indicated	6 (0.2)
Employment, *n* (%)	
Employed full time (≥ 40 h per week)	2,393 (82.4)
Employed part time (≤ 39 h per week)	218 (7.5)
Unemployed and currently looking for work	36 (1.2)
Unemployed and not currently looking for work	7 (0.2)
Student	16 (0.6)
Retired	48 (1.7)
Homemaker	39 (1.3)
Self-employed	131 (4.5)
Unable to work	11 (0.4)
Not indicated	4 (0.1)
Ethnicity, *n* %	
American Indian/Alaska Native	26 (0.9)
Asian American	169 (5.8)
African American/Black	336 (11.6)
Native Hawaiian/ Pacific Islander	6 (0.2)
Caucasian American/White	2,206 (76.0)
Bi- or multiracial	21 (0.7)
Hispanic	103 (3.5)
Not indicated	36 (1.2)
Age, *M* (*SD*)	37.55 (11.96)
Political attitudes (measured on 11-point scales),[a] *M* (*SD*)	
Scale from 1 = *democratic* to 11 = *republican*	5.56 (3.66)
Scale from 1 = *left* to 11 = *right*	6.12 (3.52)

[a] Political attitudes were not assessed in Study 4; mean and standard deviations were calculated only across the 1,548 participants from the other three studies.

Transparency

Action Editor: Timothy J. Pleskac
Editor: Klaus Fiedler
Declaration of Conflicting Interests
The author(s) declared that there were no conflicts of interest with respect to the authorship or the publication of this article.

ORCID iD

Sarah Volz https://orcid.org/0000-0002-5958-5002

Acknowledgments

We thank Daniel Benz, Nina Reinhardt, and Simon Schindler for giving us access to their studies. We further thank E. Paige Lloyd, Jason C. Deska, Kurt Hugenberg, Allen R. McConnell, Brandon T. Humphrey, and Jonathan W. Kunstman for providing the stimulus material for these studies.

Supplemental Material

Additional supporting information can be found at http://journals.sagepub.com/doi/suppl/10.1177/17456916221149943

Notes

1. Note that in this article we refer to situations in which senders cannot choose the topic they lie about, as is oftentimes the case in everyday life (e.g., when being accused of a crime, there is only little room to freely decide what to lie about). Further, it is assumed that judges and senders have no prior knowledge about each other, and they do not interact.
2. One study (Levine et al., 2005) investigated sender and message effects, but judges rated senders' behaviors rather than making veracity judgments.
3. Lloyd et al. (2019) reported the following a priori inclusion criteria:

> Participants responded to all four prompts, spoke for at least 20 s in each of the four videos, followed directions (e.g., spoke about a friend as opposed to an actor or a political figure), did not disrupt the camera with excessive movement (e.g., banging on the table), and remained in frame throughout all four videos. When more than 20 targets met the a priori inclusion criteria, we selected the 20 targets with the best video quality, and when the video quality was not visibly different, targets were chosen on the basis of random selection. (p. 433)

4. We report the random-slope model (Model 7) only for the judgment-accuracy variable because it has no theoretical relevance to lie-versus-truth classifications and confidence. Accordingly, the random-veracity slope neither significantly improved the model fit nor changed the overall conclusions for these two dependent variables.
5. Despite the overall accuracy of 50.99%, in purely mathematical terms, there could have been at least two transparent senders if one assumes a detection rate of 80% for transparent senders and an average detection rate of 50% for nontransparent senders.

References

Aamodt, M. G., & Custer, H. (2006). Who can best catch a liar? *Forensic Examiner*, *15*(1), 6–11. https://search.proquest.com/openview/bb696fa3ce2bffbf2e657a63ff1b54c9/1?pq-origsite=gscholar&cbl=25766

Aarts, E., Verhage, M., Veenvliet, J. V., Dolan, C. V., & van der Sluis, S. (2014). A solution to dependency: Using multilevel analysis to accommodate nested data. *Nature Neuroscience*, *17*(4), 491–496. https://doi.org/10.1038/nn.3648

Ais, J., Zylberberg, A., Barttfeld, P., & Sigman, M. (2016). Individual consistency in the accuracy and distribution of confidence judgments. *Cognition*, *146*, 377–386. https://doi.org/10.1016/j.cognition.2015.10.006

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

Blais, A.-R., Thompson, M. M., & Baranski, J. V. (2005). Individual differences in decision processing and confidence judgments in comparative judgment tasks: The role of cognitive styles. *Personality and Individual Differences*, *38*(7), 1701–1713. https://doi.org/10.1016/j.paid.2004.11.004

Bond, C. F., & DePaulo, B. M. (2006). Accuracy of deception judgments. *Personality and Social Psychology Review*, *10*(3), 214–234. https://doi.org/10.1207/s15327957pspr1003_2

Bond, C. F., & DePaulo, B. M. (2008). Individual differences in judging deception: Accuracy and bias. *Psychological Bulletin*, *134*(4), 477–492. https://doi.org/10.1037/0033-2909.134.4.477

Buller, D. B., & Burgoon, J. K. (1996). Interpersonal deception theory. *Communication Theory*, *6*(3), 203–242. https://doi.org/10.1111/j.1468-2885.1996.tb00127.x

Curci, A., Lanciano, T., Battista, F., Guaragno, S., & Ribatti, R. M. (2018). Accuracy, confidence, and experiential criteria for lie detection through a videotaped interview. *Frontiers in Psychiatry*, *9*, Article 748. https://doi.org/10.3389/fpsyt.2018.00748

Da Silva, C. S., & Leach, A.-M. (2013). Detecting deception in second-language speakers. *Legal and Criminological Psychology*, *18*(1), 115–127. https://doi.org/10.1111/j.2044-8333.2011.02030.x

DePaulo, B. M., Charlton, K., Cooper, H., Lindsay, J. J., & Muhlenbruck, L. (1997). The accuracy-confidence correlation in the detection of deception. *Personality and Social Psychology Review*, *1*(4), 346–357. https://doi.org/10.1207/s15327957pspr0104_5

DePaulo, B. M., Lindsay, J. J., Malone, B. E., Muhlenbruck, L., Charlton, K., & Cooper, H. (2003). Cues to deception. *Psychological Bulletin*, *129*(1), 74–118. https://doi.org/10.1037/0033-2909.129.1.74

DePaulo, B. M., & Rosenthal, R. (1979). Telling lies. *Journal of Personality and Social Psychology*, *37*(10), 1713–1722. https://doi.org/10.1037/0022-3514.37.10.1713

Evans, J. R., & Michael, S. W. (2014). Detecting deception in non-native English speakers. *Applied Cognitive Psychology*, *28*(2), 226–237. https://doi.org/10.1002/acp.2990

Frank, M. G., & Ekman, P. (2004). Appearing truthful generalizes across different deception situations. *Journal of Personality and Social Psychology*, *86*(3), 486–495. https://doi.org/10.1037/0022-3514.86.3.486

Garrido, E., Masip, J., & Herrero, C. (2004). Police officers' credibility judgments: Accuracy and estimated ability. *International Journal of Psychology*, *39*(4), 254–275. https://doi.org/10.1080/00207590344000411

Global Deception Research Team. (2006). A world of lies. *Journal of Cross-Cultural Psychology*, *37*(1), 60–74. https://doi.org/10.1177/0022022105282295

Hartwig, M., & Bond, C. F. (2011). Why do lie-catchers fail? A lens model meta-analysis of human lie judgments. *Psychological Bulletin*, *137*(4), 643–659. https://doi.org/10.1037/a0023589

Hartwig, M., Granhag, P. A., & Luke, T. (2014). Strategic use of evidence during investigative interviews. In D. C. Raskin, C. R. Honts, & John C. Kircher (Eds.), *Credibility assessment: Scientific research and applications* (pp. 1–36). Academic Press. https://doi.org/10.1016/B978-0-12-394433-7.00001-4

Hartwig, M., Voss, J. A., Brimbal, L., & Wallace, D. B. (2017). Investment professionals' ability to detect deception: Accuracy, bias and metacognitive realism. *Journal of Behavioral Finance*, *18*(1), 1–13. https://doi.org/10.1080/15427560.2017.1276069

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *The Behavioral and Brain Sciences*, *33*(2–3), 61–83. https://doi.org/10.1017/s0140525x0999152x

Jackson, S. A., & Kleitman, S. (2014). Individual differences in decision-making and confidence: Capturing decision tendencies in a fictitious medical test. *Metacognition and Learning*, *9*(1), 25–49. https://doi.org/10.1007/s11409-013-9110-y

Jonsson, A.-C., & Allwood, C. M. (2003). Stability and variability in the realism of confidence judgments over time, content domain, and gender. *Personality and Individual Differences*, *34*(4), 559–574. https://doi.org/10.1016/S0191-8869(02)00028-4

Judd, C. M., Westfall, J., & Kenny, D. A. (2012). Treating stimuli as a random factor in social psychology: A new and comprehensive solution to a pervasive but largely ignored problem. *Journal of Personality and Social Psychology*, *103*(1), 54–69. https://doi.org/10.1037/a0028347

Kantner, J., & Dobbins, I. G. (2019). Partitioning the sources of recognition confidence: The role of individual differences. *Psychonomic Bulletin & Review*, *26*(4), 1317–1324. https://doi.org/10.3758/s13423-019-01586-w

Kleitman, S., & Stankov, L. (2007). Self-confidence and metacognitive processes. *Learning and Individual Differences*, *17*(2), 161–173. https://doi.org/10.1016/j.lindif.2007.03.004

Levine, T. R. (2010). A few transparent liars explaining 54% accuracy in deception detection experiments. *Annals of the International Communication Association*, *34*(1), 41–61. https://doi.org/10.1080/23808985.2010.11679095

Levine, T. R. (2016). Examining sender and judge variability in honesty assessments and deception detection accuracy:

Evidence for a transparent liar but no evidence of deception-general ability. *Communication Research Reports*, *33*(3), 188–194. https://doi.org/10.1080/08824096.2016.1186629

Levine, T. R., Daiku, Y., & Masip, J. (2022). The number of senders and total judgments matter more than sample size in deception-detection experiments. *Perspectives on Psychological Science*, *17*(1), 191–204. https://doi.org/10.1177/1745691621990369

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*(3), 203–217. https://doi.org/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*(2), 125–144. https://doi.org/10.1080/03637759909376468

Levine, T. R., Serota, K. B., Shulman, H., Clare, D. D., Park, H. S., Shaw, A. S., Shim, J. C., & Lee, J. H. (2011). Sender demeanor: Individual differences in sender believability have a powerful impact on deception detection judgments. *Human Communication Research*, *37*(3), 377–403. https://doi.org/10.1111/j.1468-2958.2011.01407.x

Levine, T. R., Shaw, A., & Shulman, H. C. (2010). Increasing deception detection accuracy with strategic questioning. *Human Communication Research*, *36*(2), 216–231. https://doi.org/10.1111/j.1468-2958.2010.01374.x

Lim, A., Young, R. L., & Brewer, N. (2022). Autistic adults may be erroneously perceived as deceptive and lacking credibility. *Journal of Autism and Developmental Disorders*, *52*(2), 490–507. https://doi.org/10.1007/s10803-021-04963-4

Lloyd, E. P., Deska, J. C., Hugenberg, K., McConnell, A. R., Humphrey, B. T., & Kunstman, J. W. (2019). Miami University deception detection database. *Behavior Research Methods*, *51*(1), 429–439. https://doi.org/10.3758/s13428-018-1061-4

Luke, T. J. (2019). Lessons from Pinocchio: Cues to deception may be highly exaggerated. *Perspectives on Psychological Science*, *14*(4), 646–671. https://doi.org/10.1177/1745691619838258

Lykken, D. T. (1974). Psychology and the lie detector industry. *American Psychologist*, *29*(10), 725–739. https://doi.org/10.1037/h0037441

Mac Giolla, E., & Luke, T. J. (2021). Does the cognitive approach to lie detection improve the accuracy of human observers? *Applied Cognitive Psychology*, *35*(2), 385–392. https://doi.org/10.1002/acp.3777

Markowitz, D. M. (2022). Toward a deeper understanding of prolific lying: Building a profile of situation-level and individual-level characteristics. *Communication Research*. Advance online publication. https://doi.org/10.1177/00936502221097041

Markowitz, D. M., & Hancock, J. T. (2018). Deception in mobile dating conversations. *Journal of Communication*, *68*(3), 547–569. https://doi.org/10.1093/joc/jqy019

Masip, J., Alonso, H., Garrido, E., & Anton, C. (2005). Generalized communicative suspicion (GCS) among police officers: Accounting for the investigator bias effect. *Journal of Applied Social Psychology*, *35*(5), 1046–1066. https://doi.org/10.1111/j.1559-1816.2005.tb02159.x

Masip, J., Levine, T. R., Somastre, S., & Herrero, C. (2020). Teaching students about sender and receiver variability in lie detection. *Teaching of Psychology*, *47*(1), 84–91. https://doi.org/10.1177/0098628319888116

Meissner, C. A., & Kassin, S. M. (2002). "He's guilty!": Investigator bias in judgments of truth and deception. *Law and Human Behavior*, *26*(5), 469–480. https://doi.org/10.1023/A:1020278620751

Navajas, J., Hindocha, C., Foda, H., Keramati, M., Latham, P. E., & Bahrami, B. (2017). The idiosyncratic nature of confidence. *Nature Human Behaviour*, *1*(11), 810–818. https://doi.org/10.1038/s41562-017-0215-1

Pallier, G., Wilkinson, R., Danthiir, V., Kleitman, S., Knezevic, G., Stankov, L., & Roberts, R. D. (2002). The role of individual differences in the accuracy of confidence judgments. *The Journal of General Psychology*, *129*(3), 257–299. https://doi.org/10.1080/00221300209602099

Patzer, G. L. (1983). Source credibility as a function of communicator physical attractiveness. *Journal of Business Research*, *11*(2), 229–241. https://doi.org/10.1016/0148-2963(83)90030-9

Pulford, B. D., & Sohal, H. (2006). The influence of personality on HE students' confidence in their academic abilities. *Personality and Individual Differences*, *41*(8), 1409–1419. https://doi.org/10.1016/j.paid.2006.05.010

Riggio, R. E., Tucker, J., & Throckmorton, B. (1987). Social skills and deception ability. *Personality & Social Psychology Bulletin*, *13*(4), 568–577. https://doi.org/10.1177/0146167287134013

Sagarin, B. J., Rhoads, K. v. L., & Cialdini, R. B. (1998). Deceiver's distrust: Denigration as a consequence of undiscovered deception. *Personality and Social Psychology Bulletin*, *24*(11), 1167–1176. https://doi.org/10.1177/01461672982411004

Semrad, M., Scott-Parker, B., & Nagel, M. (2019). Personality traits of a good liar: A systematic review of the literature. *Personality and Individual Differences*, *147*, 306–316. https://doi.org/10.1016/j.paid.2019.05.007

Serota, K. B., Levine, T. R., & Docan-Morgan, T. (2022). Unpacking variation in lie prevalence: Prolific liars, bad lie days, or both? *Communication Monographs*, *89*(3), 307–331. https://doi.org/10.1080/03637751.2021.1985153

Smith, A. M., & Leach, A.-M. (2019). Confidence can discriminate between accurate and inaccurate lie decisions. *Perspectives on Psychological Science*, *14*(6), 1062–1071. https://doi.org/10.1177/1745691619863431

Solbu, A., & Frank, M. G. (2019). Evolution and development of deception in modern times. In T. Docan-Morgan (Ed.), *Springer eBook collection. The Palgrave handbook of deceptive communication* (pp. 41–66). Palgrave Macmillan. https://doi.org/10.1007/978-3-319-96334-1_3

Street, C. N. H., & Richardson, D. C. (2015). Lies, damn lies, and expectations: How base rates inform lie-truth judgments. *Applied Cognitive Psychology*, *29*(1), 149–155. https://doi.org/10.1002/acp.3085

Trovillo, P. V. (1939). A history of lie detection. *Journal of Criminal Law and Criminology*, *29*(6), 848–881. https://doi.org/10.2307/1136489

Vajapey, S. P., Weber, K. L., & Samora, J. B. (2020). Confidence gap between men and women in medicine: A systematic review. *Current Orthopaedic Practice*, *31*(5), 494–502. https://doi.org/10.1097/BCO.0000000000000906

Volz, S., Reinhard, M.-A., & Müller, P. (2022). The confidence-accuracy relation—A comparison of metacognition measures in lie detection. *Applied Cognitive Psychology*, *36*(3), 673–684. https://doi.org/10.1002/acp.3953

Vrij, A. (2015). Verbal lie detection tools: Statement validity analysis, reality monitoring and scientific content analysis. In P.-A. Granhag, A. Vrij, & B. Verschuere (Eds.), *Detecting deception: Current challenges and cognitive approaches* (pp. 3–36). John Wiley & Sons.

Vrij, A., & Baxter, M. (1999). Accuracy and confidence in detecting truths andlies in elaborations and denials: Truth bias, lie bias and individual differences. *Expert Evidence*, *7*(1), 25–36. https://doi.org/10.1023/A:1008932402565

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

Vrij, A., Fisher, R. P., Mann, S. A., & Leal, S. (2006). Detecting deception by manipulating cognitive load. *Trends in Cognitive Sciences*, *10*(4), 141–142. https://doi.org/10.1016/j.tics.2006.02.003

Vrij, A., Granhag, P.-A., Mann, S. A., & Leal, S. (2011). Outsmarting the liars: Toward a cognitive lie detection approach. *Current Directions in Psychological Science*, *20*(1), 28–32. https://doi.org/10.1177/0963721410391245

Vrij, A., Leal, S., Granhag, P. A., Mann, S., Fisher, R. P., Hillman, J., & Sperry, K. (2009). Outsmarting the liars: The benefit of asking unanticipated questions. *Law and Human Behavior*, *33*(2), 159–166. https://doi.org/10.1007/s10979-008-9143-y

Vrij, A., Mann, S. A., Fisher, R. P., Leal, S., Milne, R., & Bull, R. (2008). Increasing cognitive load to facilitate lie detection: The benefit of recalling an event in reverse order. *Law and Human Behavior*, *32*(3), 253–265. https://doi.org/10.1007/s10979-007-9103-y

Watkins, I. J., & Martire, K. A. (2015). Generalized linear mixed models for deception research: Avoiding problematic data aggregation. *Psychology, Crime & Law*, *21*(9), 821–835. https://doi.org/10.1080/1068316X.2015.1054384

Westfall, J., Judd, C. M., & Kenny, D. A. (2015). Replicating studies in which samples of participants respond to samples of stimuli. *Perspectives on Psychological Science*, *10*(3), 390–399. https://doi.org/10.1177/1745691614564879

Westfall, J., Kenny, D. A., & Judd, C. M. (2014). Statistical power and optimal design in experiments in which samples of participants respond to samples of stimuli. *Journal of Experimental Psychology: General*, *143*(5), 2020–2045. https://doi.org/10.1037/xge0000014

Wolfe, R. N., & Grosch, J. W. (1990). Personality correlates of confidence in one's decisions. *Journal of Personality*, *58*(3), 515–534. https://doi.org/10.1111/j.1467-6494.1990.tb00241.x

Zebrowitz, L. A., Voinescu, L., & Collins, M. A. (1996). "Wide-eyed" and "crooked-faced": Determinants of perceived and real honesty across the life span. *Personality & Social Psychology Bulletin*, *22*(12), 1258–1269. https://doi.org/10.1177/01461672962212006