Do people know when they are good at spotting liars? – Metacognitive efficiency in Lie Detection

Nadia Said[1+]*, Sarah Volz[2+], Marc-André Reinhard[2], Patrick Müller[3], Markus Huff[1,4]

[1] Department of Psychology, University of Tübingen; Schleichstr. 4, 72076 Tübingen, Germany.

[2] Department of Psychology, University of Kassel; Holländische Str. 36-38, 34127 Kassel, Germany.

[3] Stuttgart University of Applied Sciences; Schellingstr. 24, 70174 Stuttgart, Germany.

[4] Leibniz-Institut für Wissensmedien; Schleichstraße 6, 72076 Tübingen, Germany.

*Corresponding author. Email: nadia.said@uni-tuebingen.de

+These authors contributed equally to this work

Abstract: We investigated whether the confidence in lie detection judgments is a signal for the accuracy of judgments. We argue that previous methods in tackling this question are inadequate as the assessment of judgment accuracy and confidence is confounded with response bias and lie detection performance. We addressed this confidence-accuracy puzzle by applying a hierarchical Bayesian approach based on Signal-Detection Theory to estimate metacognitive efficiency. Metacognitive efficiency describes individuals' insight into the accuracy of their judgments about truth and deception, but unlike previous measures, it is free of bias and independent of lie detection performance. In re-analyses of 12 studies (N=2817 participants in total), metacognitive efficiency was on average only about 23% of what would have been expected given participants' discrimination performance. Hence, individuals largely lack metacognitive insight into the quality of their judgments, which is particularly problematic because they cannot reliably discriminate between lies and truths.

The ability to detect lies is essential in professional settings and daily interactions (Ekman, 2009; Vrij, 2008). For instance, professionals in forensic contexts (e.g., police officers) or at the airport security must detect lies to prevent dangerous situations, recruiters need to identify candidates' lies to ensure candidates actually have the claimed experiences and qualifications, and individuals in romantic relationships may want to identify cases in which their partner is dishonest. A large body of research suggests that humans cannot reliably discriminate between lies and truths in such ad hoc veracity judgments; a meta-analysis revealed an average accuracy rate of 54% (Bond & DePaulo, 2006). However, it is still unclear whether humans have insight into their (lacking) ability to detect lies through their confidence in these judgments.

Unlike previously used measures of metacognition, the measure we apply in this article (metacognitive efficiency) investigates the confidence-accuracy puzzle in lie detection for the first time free of response bias and independent of lie detection performance. We use M_{ratio} as a universal measure of metacognitive efficiency, which has already been used in various areas of research such as perception, memory (Bang, J. W., Shekhar, M., & Rahnev, D., 2019; Folke, Ouzia, Bright, Martino, & Filippi, 2016; Hainguerlot, Vergnaud, & Gardelle, 2018; Mazancieux, Fleming, Souchay, & Moulin, 2020; Muthesius et al., 2022; Palmer, David, & Fleming, 2014; Reyes et al., 2020), and knowledge (Fischer, Amelung, & Said, 2019; Fischer, Huff, & Said, 2021). In re-analyses of twelve lie detection studies with N=2817 participants in total, we calculate M_{ratio} by applying a hierarchical Bayesian approach implemented by Fleming (2017).

Measuring metacognition. A veracity judgment can be more or less accurate depending on its correspondence to the veracity of the statement in question. Similarly, confidence in a veracity judgment can be more or less accurate depending on its correspondence to the accuracy of the veracity judgment. To illustrate, believing that the sender of a deceptive message (i.e., a liar) is lying constitutes an accurate veracity judgment. Making such an accurate

veracity judgment with high confidence constitutes more accurate confidence than making it with lower confidence. By contrast, believing that a liar is telling the truth constitutes an inaccurate veracity judgment. Making such an inaccurate veracity judgment with low confidence constitutes more accurate confidence than making it with higher confidence. Individuals may exhibit metacognitive biases; they are overconfident (underconfident) when they overestimate (underestimate) the accuracy of their judgments, that is, when they are more (less) confident than the accuracy of their judgments would warrant.

Determining whether people have insight into their ability to discern lies from truths is especially important given that judgmental confidence bares the potential to influence decision-making and behaviors (Hadar, Sood, & Fox, 2013; Jackson & Kleitman, 2014; Meyer, Payne, Meeks, Rao, & Singh, 2013; Podbregar et al., 2001). Accurate confidence is associated with making appropriate decisions in domains characterized by high uncertainty, whereas overconfidence is associated with diagnostic error, for instance, in medical decision-making (Berner & Graber, 2008; Meyer et al., 2013; Simon & Houghton, 2003). Excessive confidence that one knows the truth can lead to a decreased tendency to integrate feedback about the accuracy of one's judgments (Mannes & Moore, 2013). Moreover, not only the person making a judgment can use confidence as an indication of judgment accuracy and adjust behaviors accordingly, but also other individuals. For example, judges in court hearings might integrate the confidence of police investigators in their veracity judgments to determine the further course of a trial. Especially because humans cannot reliably make accurate veracity judgments (Bond & DePaulo, 2006), it is crucial to investigate whether they are aware of the limits of their ability (and could potentially use this information to make better judgments).

So far, research results on the confidence-accuracy puzzle in lie detection have been mixed; some studies reported a confidence-accuracy relation under certain conditions (Reinhard, Sporer,

& Scharmach, 2013; Smith & Leach, 2019), while other studies did not (DePaulo, Charlton, Cooper, Lindsay, & Muhlenbruck, 1997; Hartwig, Voss, Brimbal, & Wallace, 2017; Jaffé, Reinhard, Ask, & Greifeneder, 2018; Kassin & Fong, 1999; Meissner & Kassin, 2002; Sporer, Masip, & Cramer, 2014). These lie detection studies addressed the confidence-accuracy puzzle with various analysis methods (see Volz, Reinhard, & Müller, 2022, for an overview). For instance, correlations between the accuracy of a veracity judgment (correct vs. incorrect) and the confidence with which the judgment was made were calculated, either across all judgments made in a study or separately within each judge (e.g., DePaulo et al., 1997; Reinhard et al., 2013; Sporer et al., 2014). Other studies correlated per-judge averages of confidence and accuracy (e.g., Jaffé et al., 2018; Kassin & Fong, 1999; Meissner & Kassin, 2002). Calibration analyses were used to determine the alignment of confidence with accuracy, that is, whether judgments made with 100% confidence were accurate in 100% of cases and judgments made with 90% confidence were also accurate in 90% of cases and so on (e.g., Hartwig et al., 2017; Reinhard et al., 2013; Smith & Leach, 2019). Volz et al. (2022) applied the above-described measures to a set of twelve lie detection studies. Some discrepancies between the results that the different measures yielded for individual data sets were found. The authors explained these discrepancies in part by conceptual differences between the measures, as the measures capture different aspects of the relationship between confidence and accuracy.

Most measures used so far aimed to examine metacognitive sensitivity, meaning that they were intended to address the ability of individuals to distinguish their accurate from their inaccurate veracity judgments by their confidence judgments. However, there is a substantial body of research that shows that the assessment of judgment accuracy and confidence is confounded with response bias (Evans & Azzopardi, 2007; Galvin, Podd, Drga, & Whitmore, 2003; Macmillan & Creelman, 2005). That is, primary task performance (judging whether a

message is truthful or deceptive) is biased with the tendency of participants to judge a message as truth (see also Bond & DePaulo, 2006). The respective confidence measures are biased with the tendency of participants to report high confidence. Furthermore, measures of the relation between confidence and accuracy (metacognitive sensitivity) are confounded with primary task performance (Barrett, Dienes, & Seth, 2013; Maniscalco & Lau, 2012; Fleming, 2017; Fleming & Lau, 2014). To illustrate, despite having the same ability to distinguish their correct and their incorrect veracity judgments in their confidence, individuals who differ in their lie detection performance might also differ on measures of metacognitive sensitivity. Put differently, metacognitive sensitivity is influenced by participants' lie detection performance. This influence becomes particularly problematic considering that lie detection performance is determined mainly by the senders' ability to lie (see e.g., Bond & DePaulo, 2008). To determine whether individuals can rely on their confidence as a proxy for accuracy, independent of the senders' usually unknown ability to lie, the measure of the confidence-accuracy relation must be independent of lie detection performance.

Applying methods from Signal Detection Theory (SDT). To address the above issues, methods from signal detection theory have been used to factor out the influence of response bias and task performance (Barrett et al., 2013; Fleming, 2017; Maniscalco & Lau, 2012). This approach allows to calculate three different measures: (i) participants' *discrimination performance*, that is how good participants are at distinguishing between truths and lies, independent of their truth bias, (ii) participants' *metacognitive sensitivity*, that is the degree to which participants' confidence judgments reflect accurate versus false veracity judgments, independent of participants' tendency to give high confidence ratings, and (iii) participants' *metacognitive efficiency*, that is participants' metacognitive sensitivity controlled for their

performance at discriminating between truthful and deceptive messages (discrimination performance).

To compute a bias-free measurement of judgment accuracy, peoples' *discrimination performance d'* is computed as the difference between the true positives ('yes' (it is a lie) responses to messages that were lies) and the false positives ('yes' (it is a lie) responses to messages that were true).

To compute a bias-free measurement of the confidence-accuracy relationship, peoples' *metacognitive sensitivity meta-d'* is calculated "as the *d'* of the type 1 SDT model that maximizes the likelihood of the observed type 2 ROC data" (Maniscalco & Lau, 2012, p.426).[1] Thus, meta-d' provides information about the extent to which a confidence judgment contains information about the accuracy of the corresponding veracity judgment.

Because d' and meta-d' are captured in the same signal-to-noise ratio units, they can be directly compared, and M_{ratio} as the relative confidence sensitivity or *metacognitive efficiency* can be calculated (meta-d'/d'). M_{ratio} assesses participants' ability to discern correct from incorrect veracity judgments in their confidence judgments while controlling for their performance at discriminating between truthful and deceptive messages. Put differently, M_{ratio} allows quantifying the lack of confidence accuracy that a lack of discrimination performance cannot explain. Cognitively ideal values of M_{ratio} suggest that participants could use all of the information they used for the lie detection task when estimating their confidence (Fleming, 2017); hence, optimal observers make the best out of their confidence judgments, given their lie detection performance.

1 One straightforward approach to factor out response bias (tendency to report high confidence) would be to define *a'* as the difference between the high confidence-correct rate and the high confidence-incorrect rate: a' = Z(H) – Z(F) with H = correct identification with high confidence, F = incorrect identification with high confidence (Evans & Azzopardi, 2007). However, it was shown that *a'* is confounded by bias (Evans & Azzopardi, 2007, Galvin et al., 2003). To address this issue, Maniscalco & Lau (2012) introduced meta-d' as the observed type 2 sensitivity. For an extensive description of meta-d' and a comparison to previous calculations see Barrett et al. (2013).

To illustrate, Figure 1 (adapted from Fleming, 2017) displays two simulated data sets with the same task performance (d' = 2.00) but two different values of metacognitive sensitivity (meta-d'$_A$ = 2.02, meta-d'$_B$ = 1.28). Although participants showed the same task performance, they differed in their metacognitive abilities. In panel A, confidence ratings were more often lower for incorrect answers (orange) and higher for correct answers (blue), resulting in a meta-d' of 2.02 and a M_{ratio} of about 1, suggesting optimal metacognitive efficiency. In panel B, the confidence ratings reflect correct and incorrect answers to a much lower degree, resulting in a meta-d' of 1.28 and a M_{ratio} of about 0.64. Hence, metacognitive efficiency was only about 64% of what would have been expected given the discrimination performance.

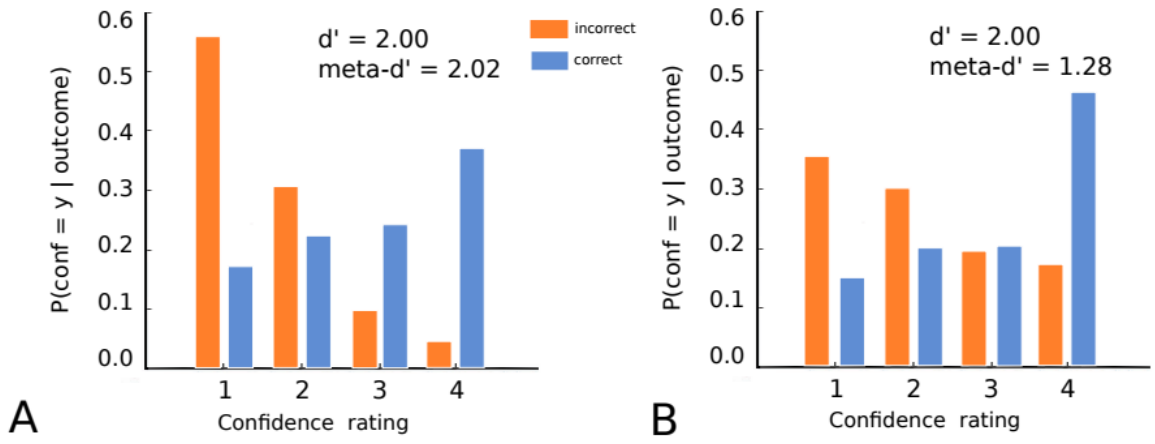

Figure 1: Example of data from a signal detection model with discrimination performance d' = 2.00. The y-axis displays the conditional probability P(conf = y | outcome) of a specific confidence rating, given that the response was correct (blue) or incorrect (orange). Panel A displays the results for metacognitive sensitivity meta-d' = 2.02. In this case metacognitive efficiency is optimal, that is M_{ratio} = 1.01. Panel B displays the results for metacognitive sensitivity meta-d' = 1.28. In this case, metacognitive efficiency is about 64% of what would have been expected given the discrimination performance, that is, M_{ratio} = 0.64. The figure was adapted from Fleming (2017).

A hierarchical Bayesian approach to estimate M_{ratio} (metacognitive efficiency).

As mentioned before, since the introduction of M_{ratio} (meta-d'/d') by Maniscalco and Lau (2012), this more general measure of metacognitive ability has been applied to a range of different domains (Fleming et al., 2015; Folke et al., 2016; McCurdy et al., 2013; Palmer et al., 2014; Ruby et al., 2017). While those assessments of M_{ratio} have been calculated by using maximum

likelihood estimations (MLE) or minimization of sum-of-squared error (SSE) approaches, Fleming (2017) introduced a *hierarchical Bayesian* approach to calculate metacognitive efficiency. This approach has the advantage of (i) enabling accurate estimation of M_{ratio} in case of limited confidence rating data and or limited participant numbers, (ii) providing group-level fits, and (iii) M_{ratio} is *directly* estimated from participants' judgments and confidence ratings which allows controlling for discrimination performance without correcting d'. Put differently, while other approaches estimate meta-d' and d' on an individual level, the hierarchical Bayesian method allows for directly estimating metacognitive efficiency at the group level, with d' being treated as a subject-level nuisance parameter.

To investigate parameter recovery of known meta-d'/d' ratios, Fleming compared this approach to the standard computations (MLE & SSE) in a series of simulation runs for N = 20 participants, with varying trial numbers ranging from 20 to 400 as well as different d' (0.5, 1.0, 2.0) and type 2 criteria. He showed that applying a hierarchical Bayesian approach was superior to the classical approaches, especially for low trial numbers. In a more recent simulation analysis, Harrison et al. (2021) demonstrated the superiority of the hierarchical Bayesian approach compared to single-subject estimations (MLE and single Bayes fits) again by adequately recovering M_{ratio} values between 0.25 and 2.0 (with 0.25 steps) for N = 60 participants with trial numbers ranging between 20 and 60.

The present research

In the present research, we focus on investigating *metacognitive efficiency* because it is bias-free and goes beyond metacognitive sensitivity by factoring out the performance at discriminating between lies and truths. This is important as individuals' metacognitive sensitivity varies with discrimination performance which is mostly determined by a sender's ability to lie in lie detection (see e.g., Bond & DePaulo, 2008). Yet, a sender's ability to lie is usually unknown in

ad hoc veracity judgments. Hence, to determine whether individuals can rely on confidence as a proxy for accuracy regardless of a sender's ability to lie, the measurement must be independent of discrimination performance. As metacognitive efficiency is independent of discrimination performance (Fleming, 2017), it constitutes a universal and bias-free measurement to investigate the confidence-accuracy puzzle in lie detection. Moreover, it may be beneficial in group comparisons by eliminating possible intergroup differences in lie detection performance.

As mentioned above, metacognitive efficiency has been assessed in a wide range of different domains, for instance, in perception tasks: M_{ratio} ~100% (Mazancieux et al., 2020; Palmer et al., 2014), in memory tasks: M_{ratio} ~73% (Palmer et al., 2014) and M_{ratio} ~50% (Mazancieux et al., 2020), as well as in knowledge tasks: M_{ratio} ~100% for science knowledge, and M_{ratio} ~50% for climate change knowledge (Fischer et al., 2019), M_{ratio} ~86 % for COVID-19 knowledge (Fischer et al., 2021). To illustrate, for COVID-19 knowledge, participants did not use 14% of the evidence they used for the knowledge task when making confidence judgments. So far, no study has assessed metacognitive efficiency in lie detection. For this project, we applied M_{ratio} to re-analyze twelve lie detection studies. Thereby we not only introduce the bias-free measure of metacognitive efficiency to lie detection research, we also report first results for this measure to examine whether confidence may be a proxy for accuracy in the many domains in which ad hoc lie detection is essential.

Method

We re-analyzed twelve studies with a total of N=2817 participants (sample sizes ranging between 138 and 625). Studies with videotaped messages and binary veracity judgments were selected from the studies that measured confidence and were available in the research group. Further, only studies with complete confidence data were chosen (i.e., no missing confidence ratings). The selected studies employed seven different stimulus materials (676 messages in

total) with typical lie detection paradigms (Bond & DePaulo, 2006). Considering the large impact senders have on lie detection accuracy (Bond & DePaulo, 2008), the variety of stimulus materials should increase the generalization of the findings across judge samples and across sender samples.

In the following, we describe the general procedure that all analyzed studies followed (see Table 1 in the supplemental material for additional characteristics of individual studies and stimulus materials). APA ethical standards were followed in the conduct of all studies reported in this article.

Stimulus Material

To generate the different stimulus materials, studies had been conducted in which participants, referred to as senders, were video-recorded while talking truthfully or deceptively about a particular topic. Senders either recorded only one message (i.e., they were randomly assigned to lie or tell the truth), or recorded several messages (i.e., they were randomly assigned to a recording order, that is, whether to lie first and then tell the truth or the reverse). Within a stimulus material, messages were assigned to sets. The sets within a stimulus material were of equal size, but set sizes varied between stimulus materials. The sets contained a 50:50 ratio of lies and truths and each sender was featured only once per set. The creation of sets allowed stimulus replication within individual studies.

Procedure

Judges were randomly assigned to one of the sets of messages. For each message within the respective set, judges made a binary veracity judgment followed by a confidence rating. Hence, judges stated whether they believed the respective sender was telling the truth or not and

indicated how confident they were about that judgment. This procedure was repeated for all messages contained in the assigned set. The confidence scales differed between the studies, with all but one study measuring confidence as a percentage value (see Table 1 in the supplemental material).

Analysis

Discrimination performance. To measure participants' performance at judging whether the sender of a message is lying or telling the truth, we determined participants' *discrimination performance d'*. As specified in a signal detection theory framework, we calculated the difference of the Z-standardized true positive rate (deceptive messages judged as lies) and the Z-standardized false positive rate (truthful messages judged as lies) with a correction for true and false positives values of 1 or 0 (Macmillan & Creelman, 2005), with Z being the inverse cumulative density function of the normal distribution.

Metacognitive efficiency. To measure participants' ability to discern correct from incorrect veracity judgments in their confidence judgments while controlling for their performance at discriminating between truthful and deceptive messages, we determined *metacognitive efficiency* M_{ratio}. To compute M_{ratio}, we applied a *hierarchical Bayesian approach* developed by Fleming (2017). The code our computations are based on is provided at https://github.com/smfleming/HMeta-d. The model was implemented in JAGS (Plummer, 2003). To estimate the joint posterior distribution of model parameters, we used Markov chain Monte Carlo (MCMC) sampling with three chains of 10000 samples and 1000 burn-in steps. Prior choices were taken from Fleming (2017). A more detailed overview of the model is given in Fleming (2017).

Confidence level. The hierarchical Bayes analysis necessary for determining M_{ratio} is computation-intensive, and computation time increases with the number of participants, messages, and confidence levels. To run the analyses within a reasonable time frame (i.e., analysis for one data set does not exceed 24 hours), confidence ratings measured on confidence scales in steps of 1% were summarized. For those scales, ten confidence ratings were combined into one level (e.g., confidence ratings 0-9% were combined as the lowest confidence level).

Results

We calculated discrimination performance d' and, more importantly, metacognitive efficiency M_{ratio} in the twelve studies. d' and M_{ratio} with 95% credible intervals, and the number of (summarized) confidence levels are listed for each study in Table A1 in the appendix. On average, discrimination performance d' was 0.09 (ranging from -0.11 to 0.27), overall accuracy was about 52% (ranging from 47% to 55%); participants performed slightly above chance level, replicating earlier work (cf. Bond & DePaulo, 2006).

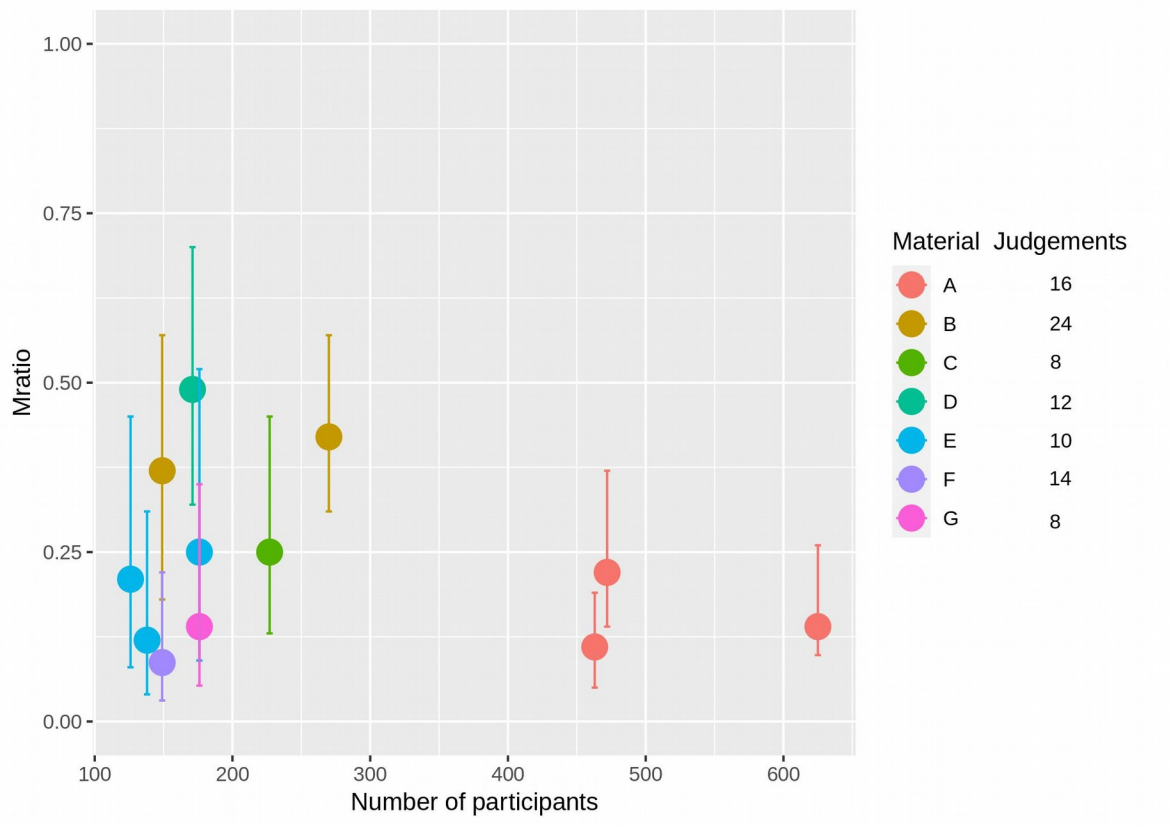

Figure 2: Metacognitive efficiency Mratio for the twelve data sets. The figure displays participants' metacognitive efficiency as a function of the number of participants with 95% credible intervals.

Figure 2 displays participants' *metacognitive efficiency* M_{ratio} as a function of the number of participants. M_{ratio} was on average 0.23; participants' metacognitive efficiency was only about 23% of what would have been expected given their lie detection performance. Results showed that metacognitive efficiency mainly was independent of the material used and estimates became in most cases more precise with an increasing number of participants.

Discussion

Detecting lies is an essential skill for various areas like forensic contexts, recruiting, or relationships (Ekman, 2009; Vrij, 2008). So far, research has shown that people's ability to

detect lies in ad hoc veracity judgments is relatively low (Bond & DePaulo, 2006). But can people at least rely on their confidence to signal the accuracy of judgments? As there is a substantial body of evidence suggesting that confidence is often used as signal to guide behavior (Balsdon, Wyart, & Mamassian, 2020; Hainguerlot et al., 2018; van den Berg, Zylberberg, Kiani, Shadlen, & Wolpert, 2016) this makes the question of whether people can rely on their confidence while making those judgments of utmost importance. Until now, it has been unclear whether there is a relationship between confidence and judgment accuracy in lie detection, with previous research finding mixed evidence. When individuals have little insight into the effectiveness of their performance, here into their performance when discriminating between truth and deception, their metacognitive sensitivity is considered low (Fleming & Lau, 2014).

In our analysis, we went one step further than calculating metacognitive sensitivity. By employing a hierarchical Bayesian approach to estimate *metacognitive efficiency*, we obtained a bias-free measurement and a measurement that allowed us to directly compare the twelve studies independent of participants' discrimination performance. Results showed that metacognitive efficiency was on average 23%; participants did not use 77% of the evidence they used for the veracity judgments when making the confidence judgments. However, two important limitations need to be considered: First, even accounting for discrimination performance, there was still variability within the M_{ratio} estimates (ranging from 0.09 to 0.49), but the credibility intervals were large and overlapped in most cases. Given the relatively strong influence of stimulus materials on research results in lie detection (see, e.g., Bond & DePaulo, 2006; Levine, Daiku, & Masip, 2022), M_{ratio} may not fully control for the influence of specific stimulus materials and discrimination performance in this specific research area, which could explain some of the variance in M_{ratio} estimates found here.

Second, in five studies, average discrimination performance d' was close to zero (in one study, d' = 0), and three had negative d' values (indicating performance near chance level). Even though this did not directly affect M_{ratio} estimates, as in the hierarchical Bayesian approach M_{ratio} is estimated at the group level and not as single subject estimates (meta-d'/d'), an overall close to zero discrimination performance is still problematic; the information available for introspection that is necessary for making an accurate confidence judgment is very low. Because of this, the results should be interpreted with caution; however, if only the studies with d'>0.20 are considered ($n_{studies}$=4, N_{total}=766), M_{ratio} remains low at 36%. In those studies, participants did not use 64% of the evidence they used for the veracity judgments when making the confidence judgments.

This finding helps answer the fundamental question of whether confidence is a proxy for judgment accuracy in lie detection. By applying a radically new approach on a large data basis, including a wide variety of stimuli, we showed that people cannot rely on their confidence when determining who is lying to them.

References

Balsdon, T., Wyart, V., & Mamassian, P. (2020). Confidence controls perceptual evidence accumulation. *Nature Communications*, *11*(1), 1753. https://doi.org/10.1038/s41467-020-15561-w

Bang, J. W., Shekhar, M., & Rahnev, D. (2019). Sensory noise increases metacognitive efficiency. *Journal of Experimental Psychology: General*, *148*(3), 437–452. https://doi.org/10.1037/xge0000511

Barrett, A. B., Dienes, Z., & Seth, A. K. (2013). Measures of metacognition on signal-detection theoretic models. *Psychological Methods*, *18*(4), 535–552. https://doi.org/10.1037/a0033268

Berner, E. S., & Graber, M. L. (2008). Overconfidence as a cause of diagnostic error in medicine. *The American Journal of Medicine*, *121*(5 Suppl), S2-23. https://doi.org/10.1016/j.amjmed.2008.01.001

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

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

Ekman, P. (2009). *Telling Lies: Clues to Deceit in the Marketplace, Politics, and Marriage (Revised Edition)*: WW Norton & Company.

Evans, S., & Azzopardi, P. (2007). Evaluation of a 'bias-free' measure of awareness. *Spatial Vision*, *20*(1-2), 61–77. https://doi.org/10.1163/156856807779369742

Fischer, H., Amelung, D., & Said, N. (2019). The accuracy of German citizens' confidence in their climate change knowledge. *Nature Climate Change*, *9*(10), 776–780. https://doi.org/10.1038/s41558-019-0563-0

Fischer, H., Huff, M., & Said, N. (2021). Insight into the accuracy of COVID-19 beliefs predicts behavior during the pandemic. https://doi.org/10.31234/osf.io/x2qv3

Fleming, S. M. (2017). Hmeta-d: Hierarchical Bayesian estimation of metacognitive efficiency from confidence ratings. *Neuroscience of Consciousness*, *2017*(1), nix007. https://doi.org/10.1093/nc/nix007

Fleming, S. M., & Lau, H. C. (2014). How to measure metacognition. *Frontiers in Human Neuroscience*, *8*, 443. https://doi.org/10.3389/fnhum.2014.00443

Fleming, S. M., Maniscalco, B., Ko, Y., Amendi, N., Ro, T., & Lau, H. (2015). Action-specific disruption of perceptual confidence. *Psychological Science*, *26*(1), 89–98. https://doi.org/10.1177/0956797614557697

Folke, T., Ouzia, J., Bright, P., Martino, B. de, & Filippi, R. (2016). A bilingual disadvantage in metacognitive processing. *Cognition*, 150, 119–132. https://doi.org/10.1016/j.cognition.2016.02.008

Galvin, S. J., Podd, J. V., Drga, V., & Whitmore, J. (2003). Type 2 tasks in the theory of signal detectability: Discrimination between correct and incorrect decisions. *Psychonomic Bulletin & Review*, *10*(4), 843–876. https://doi.org/10.3758/bf03196546

Hadar, L., Sood, S., & Fox, C. R. (2013). Subjective knowledge in consumer financial decisions. *Journal of Marketing Research*, *50*(3), 303–316. https://doi.org/10.1509/jmr.10.0518

Hainguerlot, M., Vergnaud, J.-C., & Gardelle, V. de (2018). Metacognitive ability predicts learning cue-stimulus associations in the absence of external feedback. *Scientific Reports*, *8*(1), 5602. https://doi.org/10.1038/s41598-018-23936-9

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

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

Jaffé, M. E., Reinhard, M.-A., Ask, K., & Greifeneder, R. (2018). Truth or tale? How construal level and judgment mode affect confidence and accuracy in deception detection. *Open Psychology*, *1*(1), 12–24. https://doi.org/10.1515/psych-2018-0002

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*(5), 499–516. https://doi.org/10.1023/A:1022330011811

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

Macmillan, N. A., & Creelman, C. D. (2005). *Detection theory: A user's guide* (2nd ed.). New York: Psychology Press.

Maniscalco, B., & Lau, H. (2012). A signal detection theoretic approach for estimating metacognitive sensitivity from confidence ratings. *Consciousness and Cognition*, *21*(1), 422–430. https://doi.org/10.1016/j.concog.2011.09.021

Mannes, A. E., & Moore, D. A. (2013). A behavioral demonstration of overconfidence in judgment. *Psychological Science*, *24*(7), 1190–1197. https://doi.org/10.1177/0956797612470700

Mazancieux, A., Fleming, S. M., Souchay, C., & Moulin, C. J. A. (2020). Is there a G factor for metacognition? Correlations in retrospective metacognitive sensitivity across tasks. *Journal of Experimental Psychology: General*, *149*(9), 1788–1799. https://doi.org/10.1037/xge0000746

McCurdy, L. Y., Maniscalco, B., Metcalfe, J., Liu, K. Y., Lange, F. P. de, & Lau, H. (2013). Anatomical coupling between distinct metacognitive systems for memory and visual perception. *Journal of Neuroscience*, *33*(5), 1897–1906. https://doi.org/10.1523/JNEUROSCI.1890-12.2013

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

Meyer, A. N. D., Payne, V. L., Meeks, D. W., Rao, R., & Singh, H. (2013). Physicians' diagnostic accuracy, confidence, and resource requests: A vignette study. *JAMA Internal Medicine*, *173*(21), 1952–1958. https://doi.org/10.1001/jamainternmed.2013.10081

Muthesius, A., Grothey, F., Cunningham, C., Hölzer, S., Vogeley, K., & Schultz, J. (2022). Preserved metacognition despite impaired perception of intentionality cues in schizophrenia. *Schizophrenia Research. Cognition*, *27*, 100215. https://doi.org/10.1016/j.scog.2021.100215

Palmer, E. C., David, A. S., & Fleming, S. M. (2014). Effects of age on metacognitive efficiency. *Consciousness and Cognition*, *28*, 151–160. https://doi.org/10.1016/j.concog.2014.06.007

Plummer, M. (2003, March). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In *Proceedings of the 3rd international workshop on distributed statistical computing* (Vol. 124, No. 125.10, pp. 1-10).

Podbregar, M., Voga, G., Krivec, B., Skale, R., Pareznik, R., & Gabrscek, L. (2001). Should we confirm our clinical diagnostic certainty by autopsies? *Intensive Care Medicine*, *27*(11), 1750–1755. https://doi.org/10.1007/s00134-001-1129-x

Reinhard, M.-A., Sporer, S. L., & Scharmach, M. (2013). Perceived familiarity with a judgmental situation improves lie detection ability. *Swiss Journal of Psychology*, *72*(1), 43–52. https://doi.org/10.1024/1421-0185/a000098

Reyes, G., Vivanco-Carlevari, A., Medina, F., Manosalva, C., Gardelle, V. de, Sackur, J., & Silva, J. R. (2020). Hydrocortisone decreases metacognitive efficiency independent of perceived stress. *Scientific Reports*, *10*(1), 14100. https://doi.org/10.1038/s41598-020-71061-3

Ruby, E., Giles, N., & Lau, H. (2017). Finding domain-general metacognitive mechanisms requires using appropriate tasks. BioRxiv, 211805. https://doi.org/10.1101/211805

Simon, M., & Houghton, S. M. (2003). The relationship between overconfidence and the introduction of risky products: Evidence from a field study. *Academy of Management Journal*, *46*(2), 139–149. https://doi.org/10.5465/30040610

Smith, A. M., & Leach, A.-M. (2019). Confidence can discriminate between accurate and inaccurate lie decisions. *Perspectives on Psychological Science*, 1-10. https://doi.org/10.1177/1745691619863431

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? *The American Journal of Psychology*, *127*(1), 43–61. https://doi.org/10.5406/amerjpsyc.127.1.0043

van den Berg, R., Zylberberg, A., Kiani, R., Shadlen, M. N., & Wolpert, D. M. (2016). Confidence is the bridge between multi-stage decisions. *Current Biology*, *26*(23), 3157–3168. https://doi.org/10.1016/j.cub.2016.10.021

Volz, S., Reinhard, M.-A., & Müller, P. (2022). The Confidence-Accuracy Relation– A Comparison of Metacognition Measures in Lie Detection. *Manuscript Under Revision*.

Vrij, A. (2008). *Detecting Lies and Deceit: Pitfalls and Opportunities* (2nd ed.). *Wiley series in the psychology of crime, policing and law*. Chichester, England: John Wiley & Sons Ltd.

Acknowledgments: We thank Daniel Benz, Nina Reinhardt, Simon Schindler, and Kristin Wenzel 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 stimulus material A.

Author contributions:

Conceptualization: NS, SV, MR, PM, MH
Data curation: NS, SV, MH
Formal analysis: NS
Investigation: SV, MR, PM
Methodology: NS, SV, MR, PM, MH
Project administration: NS, SV
Resources: MR, PM
Visualization: NS
Writing – original draft: NS, SV
Writing – review & editing: NS, SV, MR, PM, MH

Competing interests: Authors declare that they have no competing interests.

Data and materials availability: The analysis code (R) that produces all results and figures of this article will be made available at 10.6084/m9.figshare.14376683 latest by the time of the final publication. The data sets on which the calculations and plots are based will be made available to reviewers upon request and they will be openly accessible at the latest by the time of the final publication. Stimulus material A is available at http://hdl.handle.net/2374.MIA/6067. The other stimulus materials cannot be made publicly available due to insufficient consent from the individuals included in these materials.

Appendix

Table A1. Results for discrimination performance d' and metacognitive efficiency M_{ratio} with 95% credible intervals for the twelve studies.

Stimulus Material Type	Participants	Judgments per Judge	Aggregated Confidence Level	d'/ 95% CI		M_{ratio}/ 95% CI	
A_1	625	16	10	0.08	[0.03, 0.12]	0.14	[0.10, 0.26]
A_2	472	16	10	0.08	[0.03, 0.13]	0.22	[0.14, 0.37]
A_3	463	16	10	0.04	[-0.01, 0.09]	0.11	[0.05, 0.19]
B_1	270	24	10	0.23	[-0.16, 0.29]	0.42	[0.31, 0.57]
B_2	149	24	11	0.27	[-0.15, 0.40]	0.37	[0.18, 0.57]
C	227	8	7	0.00	[-0.10, 0.11]	0.25	[0.13, 0.45]
D	171	12	10	0.23	[0.15, 0.30]	0.49	[0.32, 0.70]
E_1	138	10	5	-0.03	[-0.14, 0.08]	0.12	[0.04, 0.31]
E_2	126	10	5	-0.11	[-0.23, 0.07]	0.21	[0.08, 0.45]
E_3	176	10	5	-0.03	[-0.04, 0.19]	0.25	[0.09, 0.52]
F	149	14	11	0.08	[-0.04, 0.19]	0.09	[0.03, 0.22]
G	176	8	5	0.22	[-0.13, 0.31]	0.14	[0.05, 0.35]