Received: 24 September 2021 | Revised: 28 January 2022 | Accepted: 20 April 2022

DOI: 10.1002/acp.3953

RESEARCH ARTICLE

WILEY

The confidence-accuracy relation – A comparison of metacognition measures in lie detection

Sarah Volz[1] | Marc-André Reinhard[1] | Patrick Müller[2]

[1]Department of Psychology, University of Kassel, Kassel, Germany

[2]Faculty of Civil Engineering, Building Physics, and Business, University of Applied Sciences Stuttgart, Stuttgart, Germany

Correspondence
Sarah Volz, Department of Psychology, University of Kassel, Holländische Straße 36-38, 34127 Kassel, Germany.
Email: sarah.volz@uni-kassel.de

Abstract

Previous research has produced mixed results on the question of whether confidence in ad hoc veracity judgments can be used as an indicator of judgment accuracy. These studies have used a variety of measures to analyze the confidence-accuracy relationship; however, they have rarely explicitly addressed why a particular measure was chosen and what its properties are. We theoretically and empirically examined previously used measures of metacognition in lie detection and report the results these measures yielded in re-analyses of 12 lie detection studies (total $N = 2817$ participants). Regardless of the measure, none of the studies found a confidence-accuracy relationship. Discrepancies between the measures are likely due to conceptual differences between them, emphasizing the importance of carefully selecting appropriate measures for the research question at hand. More work on the underlying processes of confidence judgments in lie detection is needed to improve the assessment of confidence and the selection of appropriate measures.

KEYWORDS
accuracy, calibration analysis, confidence, lie detection, metacognition

1 | INTRODUCTION

Research investigated whether individuals can assess their own performance through their confidence in different areas of psychology, such as perception (e.g., Balsdon et al., 2020; Hainguerlot et al., 2018), knowledge (e.g., Fischer et al., 2019), memory (e.g., Mazancieux et al., 2020; Palmer et al., 2014), and decision-making (e.g., Berner & Graber, 2008; Meyer et al., 2013; Simon & Houghton, 2003). Similarly, lie detection research examined whether the confidence in ad hoc veracity judgments reflects their accuracy (e.g., DePaulo et al., 1997; Smith & Leach, 2019). Do individuals have metacognitive insight into the quality of their judgments through their confidence? Research on the confidence-accuracy relation in lie detection has yielded mixed results (e.g., DePaulo et al., 1997; Masip et al., 2006; Smith & Leach, 2019) and did so with a wide variety of measures such as correlation or calibration analyses.

While research on the relationship between confidence and accuracy in other research areas has also focused on identifying which measures are most appropriate for examining this relationship (e.g., Fleming & Lau, 2014; Juslin et al., 1996; Vuorre & Metcalfe, 2021), the few discussions that exist on this topic in lie detection research have only been undertaken recently (e.g., Said et al., 2022; Smith & Leach, 2019). From an empirical perspective, such discussions are important because applying different measures to the same data set can lead to different conclusions about the existence of a relationship between confidence and accuracy (see, e.g., Smith & Leach, 2019). From a theoretical perspective, such discussions are important because even if two measures yield a relationship between confidence and accuracy for a data set, that relationship may mean something different depending on the measure because some of them are conceptually dissimilar. With this in mind, purposefully selecting a measure for the research question at hand is

important; yet, research reports only rarely provide an explicit rationale for the chosen measure (see, e.g., Smith & Leach, 2019, for an exception). With this article, we aim to advance the discussion of metacognition measures in lie detection at both the theoretical and empirical levels to assist future research in the purposeful selection of appropriate measures. First, we review previously used measures of the confidence-accuracy relation in lie detection and highlight their theoretical implications and specific properties. Second, we provide an empirical investigation of these measures by applying them to 12 lie detection studies and comparing the results of the different measures for the individual studies. Third, we deepen the empirical investigation of measures by presenting indicators of internal agreement of individual measures and indicators of external agreement between the measures. Indicators of internal agreement provide information about the measures' stability and thus about the reproducibility of the results obtained with the respective measure. Indicators of external agreement provide information about the degree of correspondence between the measures and shed light on how the results obtained from the different measures might differ.

Note, that the focus of this article is on data analysis rather than study design or data collection; therefore, the latter are addressed only when relevant to the choice of analysis method.

1.1 | Previously used metacognition measures in lie detection

1.1.1 | Correlational measures

Most of the previously employed measures were correlative in nature. In a meta-analysis, DePaulo et al. (1997) introduced the terms within-judge and between-judge correlation for two of these correlational approaches. *Within-judge correlations* are correlations between the binary variable of whether a judgment was correct or incorrect and the confidence in that judgment, calculated individually for each judge and subsequently averaged across judges. Thus, within-judge correlations inform whether judges on average make higher confidence ratings when making accurate judgments and lower confidence ratings when making inaccurate judgments. *Between-judge correlations* are correlations between judges' overall accuracy rates (percentage of correct judgments across all messages judged) and overall confidence ratings (mean confidence rating across all messages judged). Thus, between-judge correlations inform whether overall more confident judges also make more accurate judgments. In addition to between- and within-judge correlations, overall-correlations have been calculated between the binary variable of whether a judgment was correct or incorrect and the confidence in that judgment across all judgments made in a study (e.g., Sporer et al., 2014).

1.1.2 | Calibration analyses

In calibration analyses, confidence judgments are framed as probabilities of judgments being correct (see e.g., Fleming & Lau, 2014). Hence, calibration analyses examine the correspondence between confidence levels and the proportion of correct judgments made at the respective confidence level (e.g., are judgments made with 90% confidence also correct with a probability of 90%?). To our knowledge, three lie detection studies employed calibration analyses so far (Hartwig et al., 2017; Reinhard et al., 2013; Smith & Leach, 2019), each of them in a slightly different manner.

For calibration analyses, confidence judgments are usually grouped into confidence categories. This can be done by employing a confidence scale with rather broad categories to ensure enough ratings per category when collecting the data or by aggregating confidence ratings after data collection to broader categories. Hartwig et al. (2017) seemingly chose the first approach but collapsed the high confidence judgments (90%, 95%, and 100%) after data collection due to only few judgments falling into these categories. Smith and Leach (2019) and Reinhard et al. (2013) aggregated confidence ratings after data collection; Smith and Leach chose three fixed categories (low: <70%; medium: 70%–89%; high: >90%), whereas Reinhard et al. divided the confidence judgments into four categories of roughly equal size based on the collected data. For this article, we focus on calibration analyses using *data-based confidence categories* as carried out by Reinhard et al. (2013) as well as on calibration analyses using *fixed confidence categories* as carried out by Smith and Leach (2019). Unlike the approach used by Hartwig et al. (2017), who adjusted the confidence categories after data collection, these approaches can be applied to different types of confidence scales. In all three studies, which used calibration analyses, accuracy scores for the defined confidence categories were visually inspected using calibration plots, but only Reinhard et al. additionally reported formal calibration indices (see the analysis section for more details and formulas). Here, we calculate the formal calibration indices not only for the approach using data-based confidence categories (as in Reinhard et al., 2013) but also for the approach using fixed confidence categories as carried out by Smith and Leach (2019).

1.1.3 | M_{ratio} (metacognitive efficiency)

Most measures described above addressed metacognitive sensitivity, that is, judges' ability to discern correct from incorrect veracity judgments in their confidence judgments. Said et al. (2022) introduced M_{ratio} as a measure of *metacognitive efficiency* to lie detection research. Metacognitive efficiency goes beyond metacognitive sensitivity by factoring out judges' task performance from metacognitive sensitivity. In other words, metacognitive efficiency is judges' ability to discern correct from incorrect veracity judgments in their confidence judgments independent of judges' ability to discriminate between lies and truths. This independence can be useful because measures of the confidence-accuracy relation can be confounded by an individuals' task performance (here, lie detection performance) as well as by their metacognitive biases, that is, by their tendency to give generally high or low confidence ratings (e.g., Fleming, 2017; Fleming & Lau, 2014; Galvin et al., 2003; Maniscalco & Lau, 2012). In

other words, measures of metacognitive sensitivity for two individuals might differ even though they have the same ability to tell apart their correct and their incorrect veracity judgments in their confidence judgments. This can occur, for example, when two individuals perform equally well on the lie detection task but have different metacognitive biases or when they have the same metacognitive bias, but one individual performs better on the lie detection task. M_{ratio} as a measure of metacognitive efficiency is independent of metacognitive biases and task performance; thus, it takes a slightly different stance on the confidence-accuracy relation compared to the other measures reviewed above (see Said et al., 2022 for a more detailed explanation).

1.1.4 | Mixed effects models

Recently, researchers advocated mixed effects models to analyze the data of psychological studies, which employ stimulus material, to account for the variability of stimuli (see e.g., Judd et al., 2012; Westfall et al., 2015; Wolsiefer et al., 2017). Following these suggestions, some lie detection studies were analyzed using mixed effects models (e.g., Hudson et al., 2020; Volz et al., 2020; Watkins & Martire, 2015). Although we are not aware of studies in which mixed effects models were used specifically to analyze the confidence-accuracy relation, we have included mixed effects models in our comparison to assist researchers who might consider using them also for the confidence-accuracy relation (see also Murayama et al., 2014, for the use of mixed effects models in research on metamemory accuracy).

2 | EMPIRICAL PART – COMPARING MEASURES FOR 12 LIE DETECTION STUDIES

For the empirical comparison of the measures described above, we re-analyzed the data from 12 lie detection studies from our lab. In addition to comparing the results for the individual studies obtained from the different measures, we computed indicators of internal agreement of individual measures and indicators of external agreement between the measures. For internal agreement, we examined the extent to which each measure yields similar results when the composition of the sample varies slightly. If a measure yields different results when the sample varies slightly, the results of that measure might have a lower chance of being replicated. For external agreement, we examine the extent to which different measures yield similar results when applied to the same data set. If measures yielded different results for the same data set, this would reinforce the importance of carefully selecting a measure in light of the research question. It would also imply that measures should not be randomly substituted.

The 12 studies used in this article were also re-analyzed by Said et al. (2022) who introduced M_{ratio} as a measure for the confidence-accuracy relation in lie detection. The studies used seven different stimulus materials (a total of 676 messages) with paradigms commonly used in lie detection studies (see e.g., Bond & DePaulo, 2006). Here we give a short overview of the procedure all 12 studies followed (more information on individual characteristics of the studies and stimulus materials can be found in Appendix A). All studies complied with the APA ethical standards.

2.1 | Stimulus material

For each of the seven employed stimulus materials, a study was conducted in which participants (senders) lied or told the truth about a given topic while being video recorded. In some studies, senders recorded only one message and were randomly assigned to either the lie condition or the truth condition. In other studies, senders recorded multiple messages and were randomly assigned to lie or to tell the truth first. Following data collection for the stimulus materials, the messages of each stimulus material were assigned to sets of the same size with a 50:50 ratio of truthful and deceptive messages. Each sender was featured only once per set.

2.2 | Procedure of the judgment studies

Judges were randomly assigned to one set of messages from the stimulus material employed in the respective study. They watched all messages from the assigned set and indicated for each message whether they thought the sender was lying or telling the truth (binary judgment) and how confident they were in that judgment. Confidence was measured as a percentage value in all but one study. The percentage scales varied regarding the scale points (e.g., confidence from 0% to 100% in steps of 1% vs. steps of 10%, see Appendix A for details of the confidence scales from each study).

2.3 | Analysis

We chose a bootstrapping approach to estimate the above outlined measures including confidence intervals for each of the 12 studies. We considered bootstrapping a useful approach here to increase confidence in the comparisons. Bootstrapping allows to examine the measures for a large number of data sets similar to the original one, that is, for data sets that could have occurred instead of the original data set. For each study, we generated 5000 bootstrap samples, which we used to estimate all measures except for M_{ratio}. Estimates of a measure were averaged across the 5000 samples to obtain the bootstrapped value for the respective study; for correlational measures, estimates were Fisher's Z-transformed before being averaged and subsequently transformed back into correlations.

Because M_{ratio} was estimated using a hierarchical Bayesian approach, the 5000 bootstrap samples could not be used for M_{ratio}; running this procedure 5000 times for each of the 12 studies would have been too computationally expensive. Therefore, the final estimates from the hierarchical Bayesian models of each original data set and the respective 95% credible interval were used for M_{ratio} (further details below).

TABLE 1 Estimates of the measures for the 12 studies. Estimates are displayed in bold when the 95% confidence interval does not include 0 for correlations or 1 for odds ratios

No.	Mean accuracy (in %)	Overall-correlation	Within-correlation	Between-correlation	Calibration (data-based confidence categories)[a]	Calibration (fixed confidence categories)[a]	M_{ratio}	Odds ratio of mixed effects model
1	51.64	0.00	0.01	−0.04	0.05	0.04	0.22	1.04
		[−0.02, 0.02]	[−0.01, 0.03]	[−0.10, 0.01]	[0.04, 0.06]	[0.04, 0.05]	[0.14, 0.37]	[0.77, 1.38]
2	51.55	0.02	**0.03**	−0.01	0.06	0.06	0.14	**1.32**
		[0.00, 0.04]	[0.02, 0.05]	[−0.05, 0.04]	[0.06, 0.07]	[0.05, 0.06]	[0.10, 0.26]	[1.03, 1.66]
3	50.61	0.00	−0.01	0.03	0.06	0.05	0.11	0.99
		[−0.02, 0.03]	[−0.03, 0.01]	[−0.04, 0.10]	[0.05, 0.06]	[0.05, 0.06]	[0.05, 0.19]	[0.70, 1.38]
4	55.31	**0.05**	**0.08**	0.00	0.04	0.03	0.42	**1.67**
		[0.03, 0.08]	[0.05, 0.10]	[−0.07, 0.07]	[0.03, 0.05]	[0.03, 0.04]	[0.31, 0.57]	[1.24, 2.22]
5	55.12	**0.05**	**0.07**	0.00	0.07	0.07	0.37	1.78
		[0.01, 0.08]	[0.03, 0.10]	[−0.10, 0.08]	[0.06, 0.08]	[0.06, 0.08]	[0.18, 0.57]	[0.92, 3.09]
6	50.13	−0.02	−0.05	0.05	0.07	0.05	0.25	0.65
		[−0.07, 0.02]	[−0.09, 0.00]	[−0.04, 0.13]	[0.06, 0.08]	[0.04, 0.06]	[0.13, 0.45]	[0.41, 1.00]
7	55.36	0.00	0.00	0.00	0.06	0.05	0.49	0.88
		[−0.04, 0.03]	[−0.04, 0.03]	[−0.08, 0.08]	[0.05, 0.07]	[0.04, 0.06]	[0.32, 0.70]	[0.51, 1.39]
8	48.70	**−0.06**	−0.01	−0.10	0.09	0.09	0.12	**0.46**
		[−0.11, −0.01]	[−0.06, 0.04]	[−0.21, 0.01]	[0.08, 0.11]	[0.07, 0.11]	[0.04, 0.31]	[0.18, 0.98]
9	47.14	−0.03	**−0.06**	0.02	0.10	0.10	0.21	0.70
		[−0.09, 0.02]	[−0.11, −0.01]	[−0.07, 0.11]	[0.08, 0.13]	[0.08, 0.12]	[0.08, 0.45]	[0.25, 1.59]
10	49.09	−0.04	**−0.07**	0.00	0.08	0.08	0.25	0.55
		[−0.09, 0.01]	[−0.11, −0.03]	[−0.09, 0.10]	[0.06, 0.09]	[0.06, 0.09]	[0.09, 0.52]	[0.23, 1.10]
11	51.21	0.00	0.03	−0.09	0.10	0.10	0.09	1.82
		[−0.04, 0.04]	[−0.01, 0.06]	[−0.19, 0.00]	[0.09, 0.11]	[0.09, 0.11]	[0.03, 0.22]	[0.71, 3.88]
12	55.47	**0.06**	**0.14**	−0.08	0.05	0.05	0.14	1.97
		[0.01, 0.11]	[0.10, 0.19]	[−0.17, 0.02]	[0.04, 0.06]	[0.03, 0.06]	[0.05, 0.35]	[0.77, 4.17]

[a]Low calibration scores show good calibration.

2.3.1 | Calculation of measures

Overall-correlations

Within each of the 5000 bootstrap samples, correlations were calculated between whether a judgment was correct or incorrect (coded 1 vs. 0) and the confidence with which the judgment was made across all judgments.

Within-correlations

Unlike overall-correlations, which were calculated across all judgments regardless of judges, within-correlations were calculated separately *within each judge*. Hence, the correlation between whether a judgment was correct or incorrect (coded 1 vs. 0) and the confidence with which the judgment was made was calculated for each judge. Fisher's *Z* transformed values of these per-judge correlations were averaged across judges to obtain the value of a bootstrap sample for the respective study.

Between-correlations

Average confidence rates and average accuracy rates of judges were correlated.

Calibration analysis with data-based confidence categories

Judgments were categorized into four confidence categories of roughly equal size as in Reinhard et al. (2013). This categorization was made within each bootstrap sample separately so that limits of confidence categories could differ between bootstrap samples. For visual inspection, the proportion of correct judgments was plotted for each confidence category in calibration plots. Panel A of Figure 1 depicts such a calibration plot across the 5000 bootstraps samples from Study 4 as it yielded the best formal calibration score (calibration plots of the other studies can be found in the online Supplementary material). The calibration score (*C*) as well as over−/underconfidence scores (*O*/*U*), and normalized resolution indexes (NRI) were calculated using the following formulas from Reinhard et al. (2013):

$$C = \frac{1}{n}\sum_{j=1}^{J} n_j (c_j - a_j)^2$$

$$O/U = \frac{1}{n}\sum_{j=1}^{J} n_j (c_j - a_j)$$

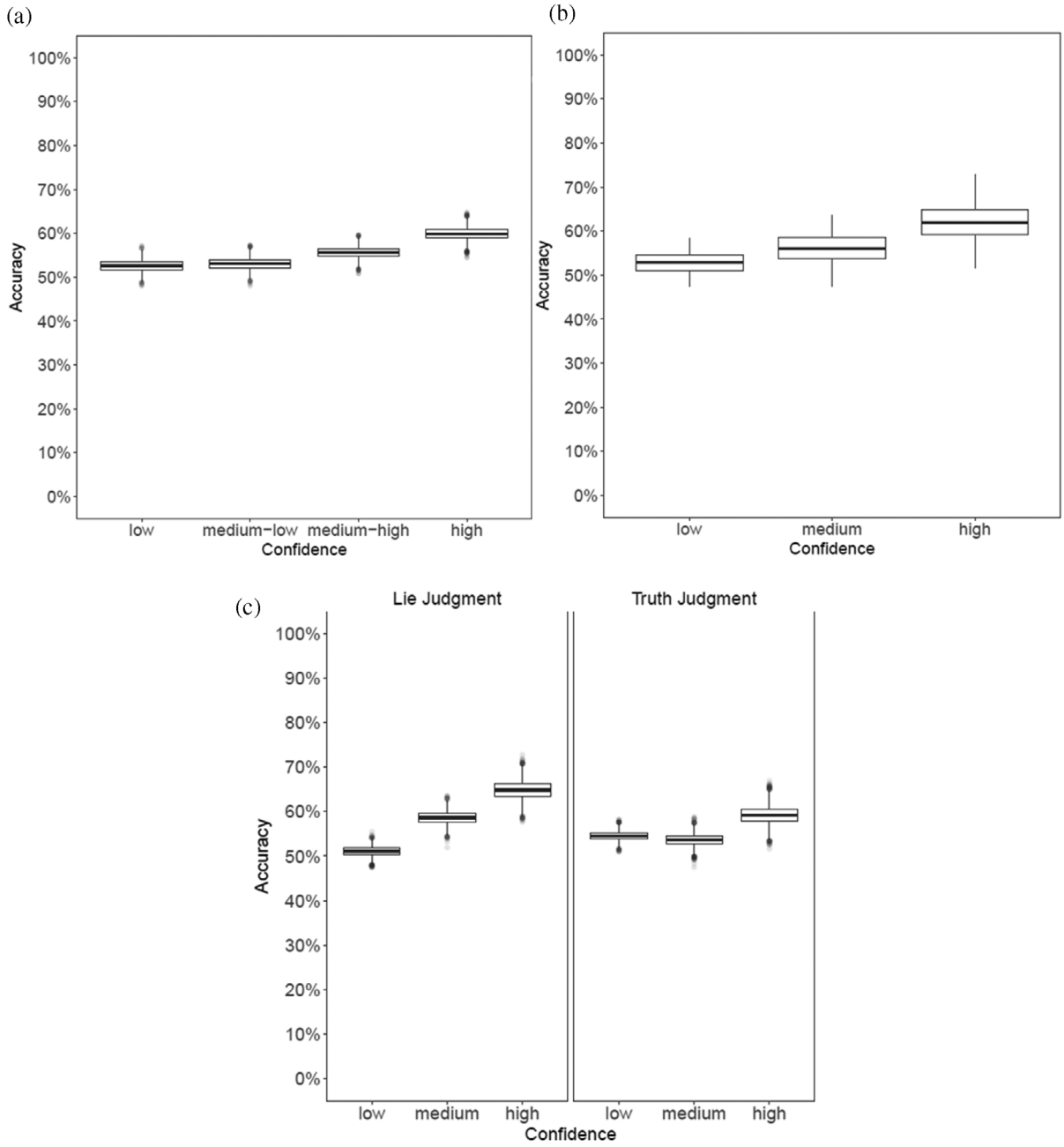

FIGURE 1 Plots for calibration analyses with the data-based confidence categories (panel a) and fixed confidence categories (panel b and panel c) across the 5000 bootstrap samples of Study 4

$$\mathrm{NRI} = \left[\frac{1}{n}\sum_{j=1}^{J} n_j (a_j - a)^2\right] / a(1-a)$$

Within the *j*th category, n_j is the frequency of confidence scores, c_j is the mean confidence, and a_j is the proportion of correct responses. a is the overall accuracy. The calibration score captures the discrepancy between the mean confidence and the mean accuracy of judgments made in each confidence category and can range from 0 (perfect calibration) to 1. Over−/underconfidence scores capture the direction of a potential miscalibration and can range from −1 (underconfidence) to +1 (overconfidence). The NRI describes the extent to which confidence can discern correct from incorrect judgments. Here we focus on calibration scores; results for the over−/underconfidence scores and the NRI can be found in the online Supplementary material.

Calibration analysis with fixed confidence categories

Following the categorization by Smith and Leach (2019), judgments were categorized into three fixed confidence categories (low: <70%;

[Correction added on 12 May 2022, after first online publication: the formula on the fifth page has been corrected in this version.]

medium: 70%–89%; high: >89%). As for the approach with data-based confidence categories, we calculated calibration scores, over−/underconfidence scores, and NRI based on the formulas in Reinhard et al. (2013) to facilitate comparison of the methods. Results for the over−/underconfidence scores and the NRI can be found in the online Supplementary material.

As with the data-based confidence categories, Study 4 achieved the best formal calibration score and is therefore plotted in Figure 1 (see online Supplementary material for the plots of all studies). Following Smith and Leach (2019), the calibration plot in Panel C contains separate lines for judgments in which the judge thought the sender is lying (lie judgments) and for judgments in which the judge thought the sender is telling the truth (truth judgments). Panel B shows the calibration plot without the distinction between lie and truth judgments.

M_{ratio} *(metacognitive efficiency)*

A hierarchical Bayesian approach established by Fleming (2017) was used to estimate M_{ratio}. Said et al. (2022) used the measure of M_{ratio} in lie detection for the first time. In the present article, we re-analyzed the same 12 studies as Said et al. (2022) did in their article. Therefore, we adopted their M_{ratio} estimates and the corresponding 95% credible intervals. Unlike other estimation approaches of M_{ratio}, the hierarchical Bayesian approach used by Said et al. provides reliable estimates even when the number of judges or the number of judgments per judge is low or when d' is low (see Fleming, 2017), which is typically the case in lie detection. Moreover, this approach directly estimates M_{ratio} without prior calculation of an individual judge's d', so that corrections of d' are not necessary. More details on how M_{ratio} was estimated can be found in Said et al. (2022).

Mixed effects models

A logistic mixed effects model with random intercepts for judges and senders was estimated for each of the bootstrapped samples. The binary variable of whether a judgment was correct or incorrect (coded 1 vs. 0) was entered as outcome variable and confidence was entered as fixed effect. The fixed effect estimate of confidence was exponentiated, resulting in the odds ratios displayed Table 1, which describe the change in accuracy associated with a one-unit change in confidence.

2.3.2 | Comparison of results across the 12 studies

In most cases, the confidence intervals included 0 for correlations or 1 in the case of odds ratios. Accordingly, most measurements showed no significant link between confidence and accuracy. If 0 (for correlations) or 1 (for odds ratios) were not included in the confidence intervals, there were still only small effects; the largest correlation was 0.14, and the largest odds ratio that did not include 1 in the confidence interval was 1.67.

Here we outline some interesting results where the estimates of the different measures seemed to be contradictory. For example, Study 7 had the highest (i.e., best) M_{ratio} value, while the between-, within-, and overall correlations for this study were zero. Study 4 had the best calibration values, the highest odds ratio (excluding 1 in the confidence

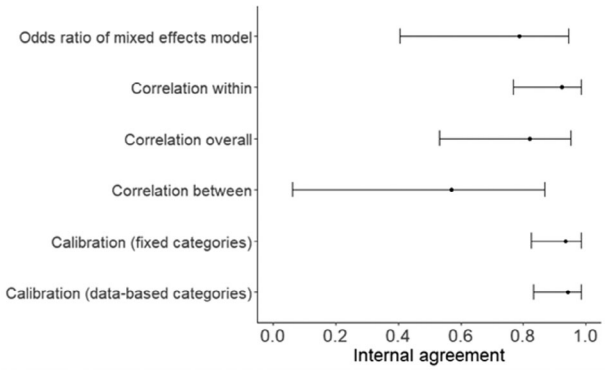

FIGURE 3 Internal agreement (including 95% confidence interval) based on the mean spearman correlation between the order of the 12 studies across the 5000 bootstrap samples for the measures

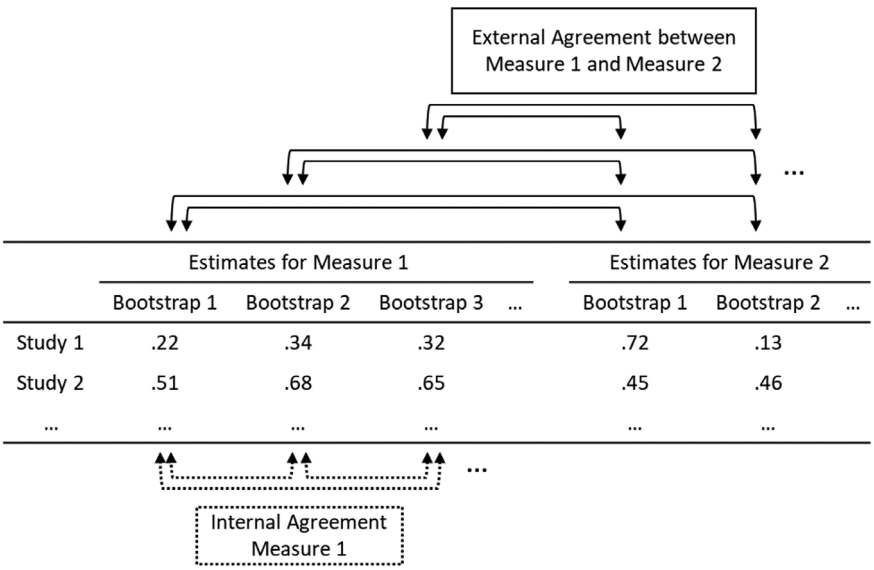

FIGURE 2 Illustration of the calculation of spearman correlations for external and internal agreement of the measures (fictitious values)

TABLE 2 External agreement based on the mean rank correlations of the order of the 12 studies between the measures averaged over bootstrap samples

	1	2	3	4	5	6
1. M_{ratio}[a]	–					
2. Calibration (data-based categories)[b]	−0.30	–				
3. Calibration (fixed categories)[b]	−0.37	0.91	–			
4. Between-correlation	0.35	−0.05	−0.12	–		
5. Overall-correlation	0.17	−0.59	−0.49	−0.04	–	
6. Within-correlation	0.00	−0.51	−0.39	−0.33	0.80	–
7. Odds ratio of mixed effects model	−0.01	−0.43	−0.32	−0.17	0.75	0.79

[a]Due to the different estimation approach for M_{ratio}, the external agreements with M_{ratio} are the average rank correlations of all bootstraps of the respective measure with the final M_{ratio} estimates.
[b]Low calibration scores indicate good calibration.

interval), and the second best M_{ratio} value, while the between-correlation was zero. Study 12 showed a low positive within-correlation and a, yet lower, overall correlation. Study 12 also had rather good calibration values, but at the same time a comparatively poor M_{ratio} value and a between-correlation not suggesting a positive confidence-accuracy relation. In summary, it is possible that the measures imply different conclusions about the general existence of a confidence-accuracy relationship. However, the analyses also suggest that when a measure does find an association, it is only a small effect; none of the studies showed a relationship between confidence and accuracy that exceeded a small effect size, regardless of the measure.

2.3.3 | Coherence of measures

To further enhance and systematize the empirical comparison of the measures, we calculated the internal agreement and external agreement of the measures. The calculations of these indicators had to meet the following criteria: They had to be combinable across studies as well as across bootstraps, so that for each measure (or in the case of external agreement for each pair of measures) only one value resulted. Due to the different (distributional) properties of the measures, the calculation should be nonparametric. Hence, we based calculations of internal and external agreement on rank correlations.

Internal agreement of measures

We estimated the internal agreement as proxy for the stability of a measure. The internal agreement should give an impression of how sensitive the respective measure is to small variations in the data. For measures with a lower level of internal agreement, small variations in the data likely have a larger impact on the overall result compared to measures with a higher level of internal agreement. Hence, measures with lower levels of internal agreement are less likely to replicate compared to measures with higher levels of internal agreement. For each measure, we investigated how stable the rank order of the 12 studies would be across the 5000 bootstraps samples for a measure.

We calculated the average Spearman correlation between the order of the 12 studies across all possible combinations of the 5000 bootstraps. A schematic representation of the calculation is shown in Figure 2. The averages of these 12,497,500 correlations for the measures are displayed in Figure 3. For M_{ratio}, this value could not be calculated due to the different estimation approach. Most measures were internally stable; for between-correlations, the order of the studies changed more frequently (lower internal agreement), indicating a lower stability for them. To a lesser extent, this was also true for odds ratios of mixed models and overall correlations.

External agreement between measures

Complementing the analysis of theoretical and conceptual differences between the measures, the external agreement should provide an empirical indicator of the degree of correspondence between the measures. Pairs of measures with lower external agreement likely yield more diverse results than pairs of measures with higher external agreement. To calculate external agreement, we compared the order of the 12 studies obtained from the measures. That is, we calculated Spearman correlations between the order of the 12 studies obtained from one measure with the order of the 12 studies obtained from another measure across all possible combinations of the 5000 bootstraps. The averages of these 25,000,000 correlations are displayed in Table 2 for all pairings of measures. External agreement with M_{ratio} was calculated as the average rank correlation between the orders obtained from the 5000 bootstraps of a measure with the order obtained from the final M_{ratio} estimates, hence, the average of 5000 correlations. External agreement between measures varied from no agreement (e.g., overall- and between-correlation) to high agreement (e.g., calibration with data-based and fixed confidence categories).

3 | DISCUSSION

To investigate measures of the confidence-accuracy relation in lie detection research, we have applied previously used measures to the data of 12 lie detection studies. In none of the studies did we find anything more than a small relationship between confidence and accuracy, regardless of the measure. Nevertheless, there were cases where one of the measures indicated a small relationship between confidence and accuracy, while other measures did not find this relationship for the same study. For instance, this was the case for within- and between-correlations in Study 12. Similar discrepancies between within- and between-correlations have been previously reported by DePaulo

et al. (1997). Because within- and between-correlations capture different aspects of the confidence-accuracy relation, discrepancies between measures might reflect conceptual differences between them.

In conjunction with the conceptual differences of the measures, the external agreement between the measures across studies showed that some pairings of measures yielded more similar results than other pairings of measures (see Table 2). Especially, measures that are conceptually and computationally more similar (e.g., mixed effects models, within-correlations, and overall-correlations as well as the two calibration approaches) yielded higher levels of external agreement than measures that are conceptually and computationally more dissimilar (e.g., between- and within-correlations). However, note that low external agreement between measures does not necessarily constitute a problem if measures are not treated as interchangeably. The cases of low external agreement rather underline the importance of considering specific properties of measures and conceptual differences between measures when selecting a suitable measure for a research question (see Table 3 for a summary of these properties). Ideally, the choice of measure(s) is not only well grounded in theory, but also preregistered along with the associated hypothesis tests.

3.1 | Implications and recommendations for the usage of measures

Between-correlations appeared problematic in both the theoretical and empirical analyses. Between-correlations showed a low level of internal agreement. Therefore, they might be more difficult to replicate compared to other measures. Even more problematic at the theoretical level, between-correlations can only indicate whether more confident judges also make more accurate judgments. Typically, however, one is interested in the correspondence between the accuracy of individual judgments and confidence in that judgment, that is, whether confidence in a veracity judgment can be a proxy for its accuracy (see also

TABLE 3 Comparison of measures

Measure	Interpretation	Consider for application
Overall-correlation	Is the confidence in a judgment related to its accuracy?	• *p*-values should be treated with special caution due to the higher degrees of freedom • Rather internally unstable • Can be biased with task performance and response bias
Within-correlation	Do judges on average make higher confidence ratings when making correct judgments?	• Can be biased with task performance and response bias • Adjustments/exclusion of judges might be necessary when one of the variables (confidence or judgment correct vs. incorrect) lacks variation within a judge; especially problematic when number of judgments per judge is low
Between-correlation	Do overall more confident judges also make more accurate judgments?	• Internally unstable • Does not provide information on correspondence between trial-by-trial confidence and accuracy which is most often of interest
Calibration	Are judgments made at a certain confidence level (e.g., 90%) also correct with a probability similar to the respective confidence level (i.e., 90%)?	• Confidence judgments are framed as probabilities of judgments being correct • Provides information on whether people over−/underestimate the accuracy of their judgments • Preregistration (e.g., for data aggregation to confidence categories) is recommended to limit researcher degree of freedom and to increase replicability • Formal calibration scores increase cross-study comparability and allow formal hypothesis tests (e.g., between conditions)
M_{ratio}	To what extent are judges able to discern correct from incorrect veracity judgments in their confidence judgments independent of their lie detection ability?	• Confidence scales with fewer scale points are more efficient in terms of computation time than scales with many points • Measure of metacognitive efficiency that can disentangle whether low metacognitive performance is due to low lie detection performance or due to low metacognitive insight into the quality of one's judgments (see also Fleming & Lau, 2014) • Can be beneficial for group comparisons, especially when differences in lie detection performance between groups are expected (see also Fleming & Lau, 2014) • Computation can be more complex than for other measurements
Odds ratio of mixed effects model	Is the confidence in a judgment related to its accuracy (regardless of judge- and sender-specific variance)?	• Odds ratios depend on the confidence scale; this should be considered when comparing odds ratios from different studies • Rather internally unstable • Can be biased with task performance and response bias • Allows to model judge-specific and message-specific variance • Further (control) variables can be included; can be useful for group comparisons (see also Murayama et al., 2014)

Smith & Leach, 2019). Because between-correlations, unlike the other measures, do not refer to the judgment level, low levels of external agreement with the other measures seem to be a logical consequence. Hence, we see no benefit in using between-correlations unless a research question explicitly requires the use of them.

Overall and within-correlations are more suitable than between-correlations for most research questions from a theoretical viewpoint. Both address the confidence-accuracy relationship at the judgment level; within-correlations examine this relationship within judges, whereas overall correlations do not consider individual judges and examine the confidence-accuracy relationship across all judgments made in a study. Yet, correlations have shown to be problematic on an empirical level because both high and low confidence-accuracy correlations are compatible with perfect calibration (see e.g., Juslin et al., 1996). Moreover, correlations cannot be computed when the variance of either the confidence or the accuracy variable is zero. This can lead to missing values especially when the number of judgments per judge is low, and this is oftentimes the case in lie detection (see e.g., Bond & DePaulo, 2006).

In our view, calibration analyses are more useful than correlations; they provide information not only on whether confidence levels are consistent with accuracy levels, but also on whether judges are overconfident or underconfident, thus, whether their confidence level exceeds or falls short of their accuracy level. As can be seen in the different approaches for creating confidence categories, calibration analyses can be conducted in a variety of ways; specifying one of these ways in a preregistration could increase the replicability of the research. In addition, to strengthen the objectivity of calibration analyses, formal calibration scores can facilitate cross-study comparisons and allow for more formal hypothesis testing as opposed to purely visual inspection of calibration plots.

Investigating the confidence-accuracy relation with odds ratios of mixed effects models can be beneficial when additional variables are of interest or need to be controlled for (see also Murayama et al., 2014). Mixed effects models also account for the variance that is due to judges and senders, which increases the generalizability of the results across judge and sender samples (see e.g., Judd et al., 2012). Yet, in comparison to most other measures, odds ratios showed a lower level of internal agreement. Because odds ratios describe the change in accuracy for a one-unit change on the confidence scale and a one-unit change can mean different magnitudes of change (e.g., a 1% change on a percentage scale vs. a one-point change on a verbalized seven-point scale), caution is required when interpreting or comparing odds ratios of different studies.

When a research question requires insight into metacognitive performance independent of task performance, we recommend using M_{ratio}. As a measure of metacognitive efficiency, it is free of metacognitive bias and factors out task performance. Even though the estimation of M_{ratio} is computationally expensive, it can provide valuable insights, for instance, when comparing metacognitive performance of different groups, leaving aside potential differences in their lie detection performance. It is recommended to think about the choice of the confidence scale when employing M_{ratio}, because confidence scales with fewer scale points (e.g., 10 vs. 100) are more efficient in terms of computation time.

Summarizing the above recommendations, it is important to select a theoretically meaningful measure for the respective research question because measures might sometimes yield different results for the same data. Moreover, we strongly encourage researchers to record the commitment to a theoretically meaningful measure in a preregistration. If several measures seem theoretically suitable, it may also be useful to compare the results obtained with different preregistered measures. Such comparisons can increase the confidence in the results because the reviewed measures have different advantages and disadvantages. Also, in the case of comparing the results of different measures, preregistrations are important. They not only reduce the degree of freedom of the researcher in the selection of a measure, but also in the application of the measure, which is particularly relevant in calibration analyses.

3.2 | Limitations and suggestions for future research

Note that the set of measures included in this article is not exhaustive for measures potentially suitable for analyzing the confidence-accuracy relation in lie detection. We focused only on the measures used to date, identified potential problems, and made suggestions to address those problems. In other areas of psychological research, methodological development concerning the relationship between confidence and accuracy is more advanced than in lie detection research (e.g., Fleming & Lau, 2014; Juslin et al., 1996; Luna & Martín-Luengo, 2012; Masson & Rotello, 2009; Tekin et al., 2015); findings from these areas could also benefit lie detection research. For example, lie detection research has largely disregarded the problem that veracity judgments typically have a guessing probability of 50%. Even when individuals are not at all confident in their veracity judgments, just by chance, they will still be correct in 50% of the cases. This high guessing probability might distort measures of metacognition, making it even more important to incorporate metacognition measures that factor out task performance (see also Maniscalco & Lau, 2012; Vuorre & Metcalfe, 2021). Because the cognitive processes underlying confidence judgments in lie detection differ from those in memory, learning, or perception, results from other areas, of course, cannot be transferred to lie detection directly. Nevertheless, valuable insights can be obtained from other domains (e.g., insights on mathematical problems of measures) that can be used as a basis for further research in lie detection.

No confidence-accuracy relationship with an effect size larger than a small effect was found for any of the studies, regardless of the measure used. Due to these low scores, even small differences between scores could have led to shifts in the order of the studies when calculating internal and external agreement. Such shifts reduce the internal and external agreement of the measurements (since they are based on rank correlations), although none of the scores would have suggested a relationship between confidence and accuracy. Despite this issue, high levels of internal and external agreement were found for some of the measures, suggesting that some measures were nonetheless more stable than others and that some pairs of measures had higher levels of agreement than others.

The studies reported here had a mean accuracy rate of about 50%, which is typically the case in lie detection (see Bond & DePaulo, 2006, for a meta-analysis). One could argue that it is therefore not surprising that only small relationships, if at all, were found between confidence and accuracy. However, high task performance is in principle no prerequisite for metacognitive performance. Consider a task in which individuals answer knowledge questions and indicate their confidence in each of the given answers. An individual who answers most questions correctly and indicates high confidence for the answers should receive good metacognitive performance scores. The same applies to an individual who answers most questions incorrectly but does so with low confidence in the answers. Hence, if individuals have insight into the limits of their performance, good metacognitive performance scores can still result.

Whether individuals can at all have insight into their lie detection performance is a question that needs to be addressed at the theoretical level. Recently, a new theoretical idea has been put forward by Smith and Leach (2019) who suggested that confidence may be a proxy for accuracy when lie-tellers are easy to detect (i.e., display many signs of deception). From a practitioner's perspective, however, confidence would still be of limited use because knowledge of the sender's detectability is required to determine whether confidence in a veracity judgment about that sender may reflect its accuracy. In other words, only when one knows that a sender is easy to detect, could one use confidence as a proxy for accuracy. Because this knowledge about a senders' detectability is usually not available, measures of metacognitive efficiency such as M_{ratio} that factor out task performance (i.e., lie detection performance) could further deepen the understanding of metacognition in lie detection (see also Said et al., 2022).

The re-analyzed studies employed different kinds of confidence scales, which is a crucial point especially for the odds ratios of the mixed effects models as outlined above. When deciding on a confidence scale, theoretical considerations as well as computation-related considerations (e.g., many scale points heavily increase the computation time for M_{ratio}) may play a role. In lie detection studies, confidence is often assessed using fine-grained percentage scales; yet, it is unclear how individuals form confidence judgments and how confidence in veracity judgments is captured best. Not only the type of scale (e.g., verbalized or numeric percentage scale; see also Tekin et al., 2015) but also the number of scale points may be crucial in this regard. Individuals might, for instance, form binary confidence judgments (confident vs. not confident) rather than elaborate percentage scores (e.g., 79% confident). If the commonly used confidence scales were inadequate and distorted the assessment of confidence, this would be problematic for all measures of the confidence-accuracy relation. Hence, further work is needed to deepen the understanding of the processes involved in the formation of confidence which would enable the application of theoretically suitable scales and measures for theory testing.

4 | CONCLUSION

Although discrepancies occurred between the different metacognition measures, no relationship between confidence and accuracy could be assumed for any of the studies, regardless of the measure. Discrepancies were likely due to conceptual differences, meaning that researchers should carefully select the appropriate measure to test their hypotheses. Because the variety of available measures implies high researcher degrees of freedom, we suggest that the choice of measure be specified in advance in preregistrations, along with a description of how each measure will be conducted.

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. The data that support the findings of this study are openly available in OSF at http://doi.org/10.17605/OSF.IO/KE7V8.

CONFLICT OF INTEREST

The research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

DATA AVAILABILITY STATEMENT

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.

ORCID

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

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

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–S23. 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

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

Dickhäuser, O., Reinhard, M.-A., & Marksteiner, T. (2012). Accurately detecting students' lies regarding relational aggression by correctional instructions. *Educational Psychology*, *32*(2), 257–271. https://doi.org/10.1080/01443410.2011.645271

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

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), 1–9. https://doi.org/10.3389/fnhum.2014.00443

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

Hainguerlot, M., Vergnaud, J.-C., & Gardelle, V. D. (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

Hudson, C. A., Vrij, A., Akehurst, L., Hope, L., & Satchell, L. P. (2020). Veracity is in the eye of the beholder: A lens model examination of consistency and deception. *Applied Cognitive Psychology*, *34*(5), 996–1004. https://doi.org/10.1002/acp.3678

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

Juslin, P., Olsson, N., & Winman, A. (1996). Calibration and diagnosticity of confidence in eyewitness identification: Comments on what can be inferred from the low confidence–accuracy correlation. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *22*(5), 1304–1316. https://doi.org/10.1037/0278-7393.22.5.1304

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

Luna, K., & Martín-Luengo, B. (2012). Confidence-accuracy calibration with general knowledge and eyewitness memory cued recall questions. *Applied Cognitive Psychology*, *26*(2), 289–295. https://doi.org/10.1002/acp.1822

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

Masip, J., Garrido, E., & Herrero, C. (2006). Observers' decision moment in deception detection experiments: Its impact on judgment, accuracy, and confidence. *International Journal of Psychology*, *41*(4), 304–319. https://doi.org/10.1080/00207590500343612

Masson, M. E. J., & Rotello, C. M. (2009). Sources of bias in the Goodman-Kruskal gamma coefficient measure of association: Implications for studies of metacognitive processes. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *35*(2), 509–527. https://doi.org/10.1037/a0014876

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

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

Murayama, K., Sakaki, M., Yan, V. X., & Smith, G. M. (2014). Type I error inflation in the traditional by-participant analysis to metamemory accuracy: A generalized mixed-effects model perspective. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*(5), 1287–1306. https://doi.org/10.1037/a0036914

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

Reinhard, M.-A. (2010). Need for cognition and the process of lie detection. *Journal of Experimental Social Psychology*, *46*(6), 961–971. https://doi.org/10.1016/j.jesp.2010.06.002

Reinhard, M.-A., & Schwarz, N. (2012). The influence of affective states on the process of lie detection. *Journal of Experimental Psychology. Applied*, *18*(4), 377–389. https://doi.org/10.1037/a0030466

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

Reinhard, M.-A., Sporer, S. L., Scharmach, M., & Marksteiner, T. (2011). Listening, not watching: Situational familiarity and the ability to detect deception. *Journal of Personality and Social Psychology*, *101*(3), 467–484. https://doi.org/10.1037/a0023726

Said, N., Volz, S., Reinhard, M.-A., Müller, P., & Huff, M. (2022). *Do people know when they are good at spotting liars? – Metacognitive efficiency in lie detection*. PsyArXiv. https://doi.org/10.31234/osf.io/v6nbd

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*, 1062–1071. 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

Tekin, S., Granhag, P. A., Strömwall, L., Giolla, E. M., Vrij, A., & Hartwig, M. (2015). Interviewing strategically to elicit admissions from guilty suspects. *Law and Human Behavior*, *39*(3), 244–252. https://doi.org/10.1037/lhb0000131

Volz, S., Reinhard, M.-A., & Müller, P. (2020). Why don't you believe me? Detecting deception in messages written by nonnative and native speakers. *Applied Cognitive Psychology*, *34*(1), 256–269. https://doi.org/10.1002/acp.3615

Vuorre, M., & Metcalfe, J. (2021). Measures of relative metacognitive accuracy are confounded with task performance in tasks that permit guessing. *Metacognition and Learning*. https://doi.org/10.1007/s11409-020-09257-1

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: A Journal of the Association for Psychological Science*, *10*(3), 390–399. https://doi.org/10.1177/1745691614564879

Wolsiefer, K., Westfall, J., & Judd, C. M. (2017). Modeling stimulus variation in three common implicit attitude tasks. *Behavior Research Methods*, *49*(4), 1193–1209. https://doi.org/10.3758/s13428-016-0779-0

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website. [Correction added on 12 May 2022, after first online publication: the Supporting information file has been corrected in this version.]

How to cite this article: 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

APPENDIX A: OVERVIEW OF RE-ANALYZED STUDIES AND THEREIN EMPLOYED STIMULUS MATERIALS

Study	Stimulus material	Number of			Description stimulus material	Confidence scale
		Judges	Judgments per judge	Messages		
1	A	472	16	320	Each of the 80 Black and White individuals (female and male) recorded four messages: talking honestly (vs. dishonestly) about a person they liked (vs. disliked). See Lloyd et al. (2019) for more details on the material	0%–100% steps of 1%
2	A	625	16	320	See above	0%–100% steps of 1%
3	A	463	16	320	See above	0%–100% steps of 1%
4	B	270	24	72	Senders were randomly assigned to lie or tell the truth about liking or disliking a movie/TV series, yielding four different kinds of messages, see Reinhard (2010, Experiment 3) for more details on the material	0%–100% steps of 1%
5	B	149	24	72	See above	50%–100% steps of 5%
6	C	227	8	16	Male senders recorded two messages in random order, one talking about an internship they had in fact (truth condition), and one randomly assigned internship they did not have (lie condition). The procedure for generating the material paralleled that of Reinhard and Schwarz (2012, Experiment 1)	Scale from 1 (not confident at all) to 7 (very confident)
7	D	171	12	204	Senders denied having written a mobbing email which they actually wrote (lie condition) or did not write (truth condition) after doing an exercise either on (a) mindfulness, (b) mind wandering, or (c) positive thinking. Each judge rated two truthful and two deceptive messages from each exercise condition	0%–100% steps of 1%
8	E	138	10	20	Messages in which senders denied they had taken money from a purse which they actually took (lie condition) or did not take (truth condition), see Reinhard et al. (2011, Experiment 1) for more details on the material	50%–100% steps of 1%
9	E	126	10	20	See above	50%–100% steps of 1%
10	E	176	10	20	See above	50%–100% steps of 1%
11	F	149	14	28	Senders recorded two messages in random order, one talking about a student job they had in fact had (truth condition), and one randomly assigned student job they did not have (lie condition). See Reinhard and Schwarz (2012, Experiment 1) for more information on the stimulus material	50%–100% steps of 5%
12	G	176	8	16	Eight male and eight female students denied in an interview that they wrote an offending letter to another student. Half of them actually wrote the letter (lie condition), the other half did not (truth condition). See Dickhäuser et al. (2012) for more information on the stimulus material	50%–100% steps of 1%