



Evaluating false positive rates of standard and hierarchical measures of metacognitive accuracy

Manuel Rausch^{1,2} · Michael Zehetleitner¹

Received: 12 April 2023 / Accepted: 28 July 2023 / Published online: 9 September 2023
© The Author(s) 2023

Abstract

A key aspect of metacognition is metacognitive accuracy, i.e., the degree to which confidence judgments differentiate between correct and incorrect trials. To quantify metacognitive accuracy, researchers are faced with an increasing number of different methods. The present study investigated false positive rates associated with various measures of metacognitive accuracy by hierarchical resampling from the confidence database to accurately represent the statistical properties of confidence judgements. We found that most measures based on the computation of summary-statistics separately for each participant and subsequent group-level analysis performed adequately in terms of false positive rate, including gamma correlations, meta- d' , and the area under type 2 ROC curves. Meta- d'/d' is associated with a false positive rate even below 5%, but log-transformed meta- d'/d' performs adequately. The false positive rate of HMeta- d depends on the study design and on prior specification: For group designs, the false positive rate is above 5% when independent priors are placed on both groups, but the false positive rate is adequate when a prior was placed on the difference between groups. For continuous predictor variables, default priors resulted in a false positive rate below 5%, but the false positive rate was not distinguishable from 5% when close-to-flat priors were used. Logistic mixed model regression analysis is associated with dramatically inflated false positive rates when random slopes are omitted from model specification. In general, we argue that no measure of metacognitive accuracy should be used unless the false positive rate has been demonstrated to be adequate.

Keywords Metacognition · Metacognitive sensitivity · Metacognitive accuracy · Meta- d' · HMeta- d · False positive rate

✉ Manuel Rausch
manuel.rausch@hochschule-rhein-waal.de

¹ Catholic University of Eichstätt-Ingolstadt, Eichstätt, Germany

² Faculty of Society and Economics, Rhine-Waal University of Applied Sciences, Marie-Curie-Straße 1, 47533 Kleve, Germany

Introduction

Metacognition is traditionally defined as knowledge or cognition about cognitive phenomena (Flavell, 1979). A key aspect of metacognition is metacognitive accuracy, defined as the degree to which confidence judgments separate between correct and incorrect trials (Fleming & Lau, 2014). Sometimes metacognitive accuracy is also referred to as metacognitive sensitivity (Maniscalco & Lau, 2012), resolution of confidence (Baranski & Petrusic, 1994), or type 2 sensitivity (Galvin et al., 2003). To measure metacognitive accuracy, a plethora of different methodological options have been proposed, including but not limited to

- gamma correlation coefficients (Nelson, 1984),
- confidence slopes (Yates, 1990),
- logistic mixed model regression (Sandberg et al., 2010),
- area under type 2 receiver operating characteristics (Fleming et al., 2010),
- meta- d' and meta- d'/d' (Maniscalco & Lau, 2014),
- HMeta- d (Fleming, 2017),
- Meta-SDT-regression (Kristensen et al., 2020),
- σ_{meta} (Shekhar & Rahnev, 2021),
- confidence efficiency (Mamassian & de Gardelle, 2021),
- metacognitive noise (Guggenmos, 2022), and
- meta-uncertainty (Boundy-Singer et al., 2022).

Unfortunately, such a vast number of methodological options may be harmful to progress in the field: If different measures do not converge to the same results, it is unclear which measure of metacognitive accuracy researchers should trust. Moreover, if there is a large number of analysis options, some researchers may be tempted to run multiple analyses and report only those analyses that 'worked', thus publishing effects that in fact do not exist (Gelman & Loken, 2014; Simmons et al., 2011; Steegen et al., 2016). Many researchers have recently accepted meta- d'/d' as the gold standard to measure metacognitive accuracy (e.g. Alkan et al., 2020; Barrientos et al., 2022; Davies et al., 2018). Meta- d'/d' is widely believed to allow for a more straightforward interpretation of the results than other methods because meta- d'/d' was designed to control for task performance, choice bias, and confidence criteria (Maniscalco & Lau, 2012, 2014). Unfortunately, it has recently been demonstrated that the control meta- d'/d' provides is not necessarily effective (Boundy-Singer et al., 2022; Guggenmos, 2021; Rahnev, 2023; Rausch et al., 2023; Shekhar & Rahnev, 2021; Zhu et al., 2023). While controlling for theoretically irrelevant variables is an important consideration when choosing an adequate measure of metacognitive accuracy, the statistical properties of measures of metacognitive accuracy have not yet received the same amount of attention from the field. The present study examined one criterial statistical property of previously proposed measures of metacognitive accuracy: the false positive rate, i.e., the probability that a measure of metacognitive accuracy will lead to the detection of an effect that does in fact not exist.

Summary-statistics vs. hierarchical measures of metacognitive accuracy

Measures of metacognitive accuracy fall into two categories depending on how the method deals with the clustered statistical structure of data typical of studies of metacognition. Participants in studies measuring metacognitive accuracy typically perform multiple trials of a perceptual, memory, or cognitive task and report their confidence in being correct in each single trial. As a consequence, datasets in metacognition research consist of many observations which are clustered within subjects and conditions. We refer to the two approaches to deal with the clustered nature of the data as *summary-statistics and hierarchical analysis*. Summary-statistics are by far the most common analysis approach in Psychology (Judd et al., 2017) and can be used with gamma correlation, confidence slopes, type 2 ROCs, meta- d' , and meta- d'/d' . For the summary-statistics approach, two levels of analysis are required (McNabb & Murayama, 2021): For the first level of analysis, a coefficient quantifying metacognitive accuracy is computed separately for each participant in each condition. For the second level of analysis, the coefficients of metacognitive accuracy obtained during the first analysis step are subjected to a standard statistical test, such as a t-test or an ANOVA. However, when the data deviate substantially from normal distributions, it is possible for standard statistical tests to produce false positive rates other than the nominal alpha level. In particular for meta- d'/d' and meta- d_a/d_a , two ratio-based measures, the distribution is expected to be non-normal, hence log transformation is recommended (Fleming & Lau, 2014). To our knowledge, it has not been investigated whether the various summary statistic-based measures of metacognitive accuracy tend to produce normal distributions and, if not, whether the deviations are such that the false positive rates are no longer at the nominal alpha level.

For hierarchical analysis, the statistical test is performed on the level of single observations, not on the summary statistics. The clusters in the data caused by different participants are accounted for by specifying fixed and random effects in the regression model: Fixed effects are factors whose levels are experimentally determined or whose interest lies in the specific effects of each level. Fixed effects are represented by parameters that are assumed to be constant within one condition. In contrast, random effects are factors whose levels are assumed to be sampled from a larger population, or whose interest lies in the variation among them rather than the specific effects of each level. Random effects are represented by parameters that assess the variability associated with the random effect (Bolker et al., 2008).

In principle, hierarchical analysis seems to be well-suited to account for the statistical properties of common data sets in metacognition studies (Fleming, 2017; Kristensen et al., 2020; Murayama et al., 2014; Paulewicz & Blaut, 2020). In line with this intuition, simulations showed that when there was a random effect of item, the false positive rate of mixed-model logistic regression is robust, but the false positive rate of gamma correlation coefficients is inflated (Murayama et al., 2014). In general, hierarchical analyses is more powerful to detect true effects because they separate between trial-to-trial variance and between-subject-variance, whereas analyses based on summary-statistics misinterpret trial-to-trial variance as between-subject variance and as a consequence are biased against the alternative hypothesis (Boehm et al., 2018). However, hierarchical models are only able to maintain an acceptable false positive rate if the random effect structure is correctly specified (Barr et al., 2013; Hesselmann, 2018). Specifically, there is a risk of increased false positive rates if random slopes are omitted from model specification (Oberauer, 2022). Unfortunately, there is no consensus on how complex the random effect structure needs

to be (Barr et al., 2013; Matuschek et al., 2017; McNabb & Murayama, 2021). However, while model misspecification in hierarchical models is associated with a risk of false positives in general, to our knowledge, it is not known if this risk applies to data sets with statistical properties characteristic for the study of metacognition. In addition, for many previously proposed hierarchical and summary-statistic based measures of metacognitive accuracy, it has never been empirically investigated if these measures maintain an acceptable false positive rate or not.

Rationale of the present study

The aim of the present study is to investigate empirically the false positive rates associated with summary-statistics and hierarchical measures of metacognitive accuracy. To assess the false positive rate associated with each measure of metacognitive accuracy, we performed simulations where a grouping variable and a continuous predictor variable were randomly sampled independently from confidence and responses. A previous study used a mathematical model to randomly to generate data to investigate false positive rates of gamma correlations and logistic mixed model regression (Murayama et al., 2014), but there are many different proposals how to model confidence judgments (Aitchison et al., 2015; Boundy-Singer et al., 2022; Guggenmos, 2022; Hellmann et al., 2023; Mamassian & de Gardelle, 2021; Moran et al., 2015; Pereira et al., 2021; Peters et al., 2017; Pleskac & Busemeyer, 2010; Ratcliff & Starns, 2009, 2013; Rausch et al., 2018, 2020; Reynolds et al., 2020; Shekhar & Rahnev, 2021, 2022) and a clear consensus is still pending (Rahnev et al., 2022). In addition, the statistical properties of confidence judgments may vary heavily across different datasets (Rahnev et al., 2020). Therefore, the present study used hierarchical bootstrapping (Saravanan et al., 2020) based on the confidence database, an online collection of datasets (Rahnev et al., 2020), to simulate datasets with realistic statistical properties. Thus, each simulated dataset represents a justified guess what kind of data may be expected in new experiments without subscribing to a specific mathematical model of confidence.

We estimated the false positive rate for those measures that can be computed in an experimental design with a binary stimulus, a binary response, and a confidence judgment and for which computation is sufficiently robust and efficient to perform a large number of simulations: gamma correlation coefficients (Nelson, 1984), confidence slopes (Yates, 1990), type 2 receiver operating characteristics (Fleming et al., 2010), meta- d' and meta- d'/d' (Maniscalco & Lau, 2014), meta- d_a and meta- d_a/d_a (Maniscalco & Lau, 2012), HMeta- d (Fleming, 2017) and logistic mixed model regression (Sandberg et al., 2010). It should be noted that the authors of the meta- d' and meta- d_a method argued later on that meta- d' should be preferred over meta- d_a (Maniscalco & Lau, 2014); nevertheless, meta- d_a was included into our analysis as well because the false positive rate is a crucial piece of information for the interpretation of already published studies using meta- d_a . For logistic mixed-model regression, we tested two different model specifications, one with a random effect of participant on the intercept (Sandberg et al., 2010) and one with random slopes (Wierzbuch et al., 2019).

Methods

The confidence database

The confidence database is a collection of openly available datasets with a broad range of experimental paradigms, participant populations, and fields of study (Rahnev et al., 2020). All data sets from the confidence database include a measurement of participants' confidence in their response to a task. To compare the false positive rate of different measures of metacognitive accuracy based on identical datasets, we restricted our simulations to a subset of datasets because not all datasets of the confidence database allow the computation of all measures of metacognitive accuracy. Specifically, we used only those datasets where stimulus and task response can be classified into two categories, because binary stimulus and response categories are required to compute meta- d' , meta- d_a , and HMeta- d . Finally, we included only those datasets where there were at least 20 participants with at least 20 trials each, because we considered 20 observations to be the bare minimum to represent the variance between subjects as well as within-subjects. Overall, 46 datasets of the database met the inclusion criteria and were used for our simulations. The majority of these 46 studies involved a perceptual task (e.g. a masked orientation discrimination task; Rausch et al., 2018), but there were also cognitive (e.g. a general knowledge task; Mazancieux et al., 2020), and memory tasks (e.g. old vs. new word recognition task; Kantner & Lindsay, 2012).

Hierarchical resampling

The 46 datasets from the confidence database were used to simulate 5000 experiments. In each simulation, we first randomly selected one of the datasets. Second, if the study included multiple experimental manipulations or difficulty, we randomly selected only one of these conditions. Then, we created two groups of simulated participants: For each group, we sampled from the participants of the original study as many subjects as in the original study with replacement. For each simulated participant, we sampled as many trials as there were in the original study from the data of the corresponding participant in the original study. To keep computation time manageable, we capped the number of subjects per simulated study at 200 and the number of trials per subject at 500. In addition, we simulated a continuous between-subject predictor variable by randomly sampling one value for each participant from an identical Gaussian distribution. Then, we computed measures of metacognitive accuracy (see below) and performed two analyses per simulated data set. First, we tested if there was a difference between the two simulated groups with respect to each measure of metacognitive accuracy. Second, we tested if the continuous predictor variable was associated with metacognitive accuracy.

Measures of metacognitive accuracy

Gamma correlation

The gamma correlation coefficient, which was proposed by Nelson (1984) as a measure of metacognitive accuracy, is a nonparametric measurement of association between two

binary or ordered variables. To compute the gamma coefficient, it is necessary to categorize all possible pairs of trials as concordant pair, discordant pair, or tied pair. A pair is concordant if the ordering between the two trials with respect to accuracy of the primary task response is consistent with the ordering of the same two trials with respect to confidence. A pair is discordant if the ordering between the two trials with respect to accuracy of the primary task response is inconsistent with the ordering of the same two trials with respect to confidence. Finally, a pair is tied if the two trials have the same value either in terms of accuracy or in terms of confidence. Then, the frequency N_s of concordant pairs and the frequency N_d of discordant pairs are counted, and Gamma is computed as

$$G = \frac{N_s - N_d}{N_s + N_d}. \quad (1)$$

G , like a standard correlation coefficient, takes on values between -1 and 1, with values larger than zero indicating a positive association between confidence and accuracy.

Confidence slopes

The confidence slope proposed by Yates (1990) assumes a linear regression to predict confidence rating with dichotomous variable of correct/incorrect as a predictor. It is calculated according to the expression

$$\text{slope} = \mu_{S=R} - \mu_{S \neq R} \quad (2)$$

with $\mu_{S=R}$ the mean of all confidence ratings when the task response was correct, and $\mu_{S \neq R}$ the mean of all confidence ratings when the task response was incorrect.

Area under the type 2 ROC curve

The area under type II ROC curve is a measure derived from type II signal detection theory (Clarke et al., 1959; Galvin et al., 2003; Pollack, 1959), an extension of signal detection theory (Green & Swets, 1966; Peterson et al., 1954; Tanner & Swets, 1954). The aim of type II ROC curves is to quantify participants' ability to differentiate between correct and incorrect trials irrespective of rating criteria (Fleming et al., 2010; Fleming & Lau, 2014; but see Shekhar & Rahnev, 2021). To construct a type II ROC curve, it is necessary to determine type II hit rates and type II false alarm rates associated with multiple criteria. Type II hit rate is defined the proportion of high confidence trials when the participants is correct, and type II false alarm rate is the proportion of high confidence trials when the participants is incorrect. Type II hit rates and false alarm rates based on multiple criteria are obtained by means of rating scales with multiple confidence levels: It is assumed that there is a criterion that separates each confidence level from the adjacent confidence levels on the rating scale. For example, for a four-level confidence scale, there is a liberal criterion that assigns low confidence only to the first confidence level and high confidence to the other three confidence levels, then a higher criterion that assigns low confidence to the first and the second confidence level, and so on. For each criterion, type II hit rate and type II false alarm rate are plotted as individual points with type II hit rate plotted on the y-axis and type II false alarm rate on the x-axis. The curve that passes through these different points is referred to as type II ROC curve. The area under the type 2 ROC curve is a measure of the ability to differentiate between correct and incorrect trials; it is 1 if confidence

ratings differentiate perfectly between correct and incorrect trials and 0.5 if confidence ratings do not differentiate between correct and incorrect trials at all. The area under the type II ROC curve A_{ROC} can be calculated based on the formula

$$A_{ROC} = 0.5 + \frac{1}{4} \times \sum_{i=1}^n \left[(h_{i+1} - f_i)^2 - (h_i - f_{i+1})^2 \right] \quad (3)$$

with h_i the type II hit rate associated with confidence level i , f_i the type II false alarm rate associated with confidence level i , and n the number of confidence levels measured by the confidence rating scale (Song et al., 2011).

Meta- d' , meta- d'/d' , meta- d_a and meta- d_a/d_a

The conceptual idea of meta- d' and meta- d_a is to quantify metacognitive accuracy in terms of discrimination sensitivity in a hypothetical signal detection model inferred from confidence judgments assuming participants had perfect access to the sensory evidence underlying the discrimination choice and were perfectly consistent in placing their confidence criteria (Maniscalco & Lau, 2012, 2014). Meta- d' and meta- d_a can therefore be directly compared to d' and d_a respectively, the corresponding measures of task performance: If meta- d' equals d' or meta- d_a equals d_a , it means that metacognitive accuracy is exactly as good as expected from task performance. If meta- d' is lower than d' , it means that metacognitive accuracy is worse than expected from task performance. The computation meta- d' and meta- d_a is based on a hypothetical signal detection model of confidence judgments (Maniscalco & Lau, 2014). The underlying model assumes that observers select a binary response $R \in \{0, 1\}$ about a stimulus characterized by two classes $S \in \{0, 1\}$ and as well as a confidence rating out of an ordered set of confidence categories $C \in \{1, 2, \dots, C_{max}\}$. To estimate meta- d' and meta- d_a , we used an R implementation of matlab code provided by Brian Maniscalco (<http://www.columbia.edu/~bsm2105/type2sdt>, last accessed 2021-09-20).

Meta- d' The algorithm to compute meta- d' involved the following computational steps: First, frequency of each confidence category was determined depending on the stimulus class and the accuracy of the response. To correct for extreme proportions, $1/(2C_{max})$ was added to each cell of the frequency table. Second, discrimination sensitivity d' and discrimination criterion c were calculated using standard formulae

$$d' = \Phi^{-1}\left(\frac{n_{S1R1}}{n_{S1}}\right) - \Phi^{-1}\left(\frac{n_{S0R1}}{n_{S0}}\right) \quad (4)$$

$$c = -\frac{1}{2} \times \left(\Phi^{-1}\left(\frac{n_{S1R1}}{n_{S1}}\right) + \Phi^{-1}\left(\frac{n_{S0R1}}{n_{S0}}\right) \right) \quad (5)$$

with n_{S1} the number of trials when $S = 1$, n_{S0} the number of trials when $S = 0$, n_{S1R1} the number of trials when $S = 1$ and $R = 1$, n_{S0R1} the number of trials when $S = 0$ and $R = 1$, and Φ^{-1} the quantile function of the standard normal distribution. For this purpose, a maximum likelihood optimization procedure was used with respect to the confidence data given stimulus, and response as well as the parameters determined at previous steps, i.e. d' and c . The model included a free parameter for meta- d'

d_{meta} as well as the rating criteria $\theta_1, \theta_2, \dots, \theta_{C_{max}-1}, \theta_{C_{max}+1}, \theta_{C_{max}+2}, \dots, \theta_{2C_{max}-1}, \theta_{C_{max}}$ was fixed at $d_{meta} \times c \div d'$. To enforce that the criteria were ordered, all free criteria were parametrized as the log of the distance to the adjacent criterion. The probability for a specific confidence rating given stimulus and response can be computed as

$$p(C|S, R = 0) = \frac{\Phi\left(\theta_{(C_{max}-C+1)}|\mu = (S - \frac{1}{2}) \times d_{meta}\right) - \Phi\left(\theta_{(C_{max}-C)}|\mu = (S - \frac{1}{2}) \times d_{meta}\right)}{\Phi\left(\theta_{C_{max}}|\mu = (S - \frac{1}{2}) \times d_{meta}\right)} \tag{6a}$$

$$p(C|S, R = 1) = \frac{\Phi\left(\theta_{(C+C_{max})}|\mu = (S - \frac{1}{2}) \times d_{meta}\right) - \Phi\left(\theta_{(C+C_{max}-1)}|\mu = (S - \frac{1}{2}) \times d_{meta}\right)}{1 - \Phi\left(\theta_{C_{max}}|\mu = (S - \frac{1}{2}) \times d_{meta}\right)} \tag{6b}$$

where Φ indicates the cumulative gaussian density function with mean μ variance of 1, θ_0 is $-\infty$ and $\theta_{2C_{max}}$ is ∞ . Finally, meta-d' is equal to the parameter d_{meta} .

Meta-d_a The computation of meta-d_a was similar to the computation of meta-d' but with the following differences: First, to determine the parameter a , which quantifies the ratio of the standard deviations of the signal associated with the two classes of the stimulus, an auxiliary signal detection rating model was fitted to the binary response and confidence data using a maximum likelihood maximation procedure. The signal detection rating model included the parameters d , a , and a set of criteria $\theta_1, \theta_2, \dots, \theta_{2C_{max}-1}$ as free parameters. The probability of the data according to the auxiliary signal detection rating model was calculated as

$$p(C, R = 0|S) = \Phi\left(\theta_{(C_{max}-C+1)}|\mu = (S - \frac{1}{2}) \times d, \sigma = a^{-S}\right) - \Phi\left(\theta_{(C_{max}-C)}|\mu = (S - \frac{1}{2}) \times d, \sigma = a^{-S}\right) \tag{7a}$$

$$p(C, R = 1|S) = \Phi\left(\theta_{(C+C_{max})}|\mu = (S - \frac{1}{2}) \times d, \sigma = a^{-S}\right) - \Phi\left(\theta_{(C+C_{max}-1)}|\mu = (S - \frac{1}{2}) \times d, \sigma = a^{-S}\right) \tag{7b}$$

where Φ indicates the cumulative gaussian density function with mean μ and the standard deviation σ , θ_0 is $-\infty$ and $\theta_{2C_{max}}$ is ∞ . To enforce that the criteria were ordered, all criteria except for θ_1 were parametrized as the log distance to the adjacent more negative criterion during the fitting procedure. Having obtained a , we computed discrimination sensitivity d_a ,

$$d_1 = \frac{1}{a} \times \Phi^{-1}\left(\frac{n_{S1R1}}{n_{S1}}\right) - \Phi^{-1}\left(\frac{n_{S0R1}}{n_{S0}}\right) \tag{8}$$

$$d_a = d_1 \times a \times \sqrt{2/(1 + a^2)} \tag{9}$$

as well as the discrimination criterion c_1 ,

$$c_1 = \frac{-1}{1+a} \times \left(\Phi^{-1} \left(\frac{n_{S1R1}}{n_{S1}} \right) + \Phi^{-1} \left(\frac{n_{S0R1}}{n_{S0}} \right) \right). \tag{10}$$

The next step included fitting the meta- d_a model, which included the same parameters as the meta- d' model, but $\theta_{C_{max}}$ was fixed at $d_{meta} \times c_1 \div d_1$. According to the meta- d_a -model, the probability for a specific confidence rating given stimulus and response is calculated as

$$p(C|S, R = 0) = \frac{\Phi \left(\theta_{(C_{max}-C+1)} \mid \mu = (S - \frac{1}{2}) \times d_{meta}, \sigma = a^{-S} \right) - \Phi \left(\theta_{(C_{max}-C)} \mid \mu = (S - \frac{1}{2}) \times d_{meta}, \sigma = a^{-S} \right)}{\Phi \left(\theta_{C_{max}} \mid \mu = (S - \frac{1}{2}) \times d_{meta}, \sigma = a^{-S} \right)} \tag{11a}$$

$$p(C|S, R = 1) = \frac{\Phi \left(\theta_{(C+C_{max})} \mid \mu = (S - \frac{1}{2}) \times d_{meta}, \sigma = a^{-S} \right) - \Phi \left(\theta_{(C+C_{max}-1)} \mid \mu = (S - \frac{1}{2}) \times d_{meta}, \sigma = a^{-S} \right)}{1 - \Phi \left(\theta_{C_{max}} \mid \mu = (S - \frac{1}{2}) \times d_{meta}, \sigma = a^{-S} \right)} \tag{11b}$$

Finally, meta- d_a can be calculated from d_{meta} and a as follows:

$$meta - d_a = d_{meta} \times a \times \sqrt{2/(1+a^2)} \tag{12}$$

HMeta-d

HMeta-d provides an estimate of the meta- d'/d' ratio based on a hierarchical Bayesian model (Fleming, 2017) and thus its computation is closely related to the computation of meta- d' . To estimate HMeta-d, we used R code provided by Steve Fleming (<https://github.com/metacoglab/HMeta-d>, last accessed 2020-08-04), which relies on the free software jags to sample from the posterior distribution (Plummer, 2003). Sampling was performed in three separate Markov Chains to allow computation of Gelman and Rubin's convergence diagnostic \hat{R} (Gelman & Rubin, 1992). When $\hat{R} \geq 1.1$, the corresponding data set was discarded from the analysis.

According to the HMeta-d method, just as for standard meta- d' , discrimination performance d' and discrimination criterion c were computed first using formulae (4) and (5) separately for each participant and then submitted to Jags as constants. The hierarchical estimation procedure was used only for the meta- d'/d' and confidence criteria. For this purpose, the absolute frequency of each confidence rating of participant j given stimulus and response $f(C|S, R)$ was modelled as a multinomial distribution \mathcal{M} ,

$$f(C_i|S, R) \sim \mathcal{M}(n = n_{SR}, p = p(C|S, R)) \tag{13}$$

where n_{SR} is the number of trials with stimulus S and response R , and $p(C|S, R)$ calculated using formula (6). However, for HMeta-d, unlike meta- d' , $\theta_{C_{max}}$ is fixed at c . $p(C_i|S, R)$ depends on the free parameters d_{meta} and a set of criteria θ . The priors placed on the parameters depended on whether we examined the false positive rare associated with a grouping variable or with a continuous predictor variable.

HMeta-d with a grouping variable According to the computer code published by Fleming, the effect of group on HMeta-d can be assessed by using the HMeta-d algorithm to sample from the posterior for the average meta-d'/d' separately for each group. Then, each sample of the posterior of group 2 is subtracted from the corresponding sample of the posterior of group 1 to obtain a posterior of the group difference in terms of mean meta-d'/d'. For this purpose, on the level of a single participant j , the priors of the parameters were specified as follows:

$$\begin{aligned}\theta_{jR0} &\sim \text{tr}\mathcal{N}(\mu = -\mu_\theta, \sigma = \sigma_\theta, a = -\infty, b = c_j) \\ \theta_{jR1} &\sim \text{tr}\mathcal{N}(\mu = \mu_\theta, \sigma = \sigma_\theta, a = c_j, b = \infty) \\ \log(d_{\text{meta}}/d')_j &= \mu_M + \xi_M \times \delta_j \\ \delta_j &\sim \mathcal{N}(0, \sigma_\delta)\end{aligned}\quad (14)$$

where θ_{jR0} indicates the confidence criteria of participant j when the response was 0, θ_{jR1} indicates the confidence criteria of participant j when the response was 1, $\text{tr}\mathcal{N}$ indicates a truncated Gaussian distribution with the location parameter μ and the scale parameter σ , lower bound a , and upper bound b , μ_θ , and σ_θ are the parameters of the prior distribution of criteria on the group level, μ_M is the mean of the prior distribution of $\log(d_{\text{meta}}/d')$ on the group level, σ_δ controls the variability of $\log(d_{\text{meta}}/d')$ on the group level, and ξ_M and δ_j are redundant multiplicative parameters to facilitate sampling from the posterior via parameter expansion. On the group level, priors were specified as follows:

$$\begin{aligned}\mu_\theta &\sim \mathcal{N}(\mu = 0, \sigma = 10) \\ \sigma_\theta &\sim \text{tr}\mathcal{N}(\mu = 0, \sigma = 10, a = 0, b = \infty) \\ \mu_M &\sim \mathcal{N}(\mu = 0, \sigma = 1) \\ \xi_M &\sim \text{Beta}(\alpha = 1, \beta = 1) \\ \sigma_\delta &\sim \text{tr}\mathcal{N}(\mu = 0, \sigma = 1, a = 0, b = \infty)\end{aligned}\quad (15)$$

Placing independent prior distributions on both groups may not be considered a valid approach because the prior biases estimates of single participants towards the group average and thus the difference between the two groups is increased artificially. Therefore, we repeated this analysis using a prior placed on the difference between groups. For this purpose, we recoded the two groups as -0.5 and 0.5 and used the computer code (including default priors) Fleming proposed for continuous predictor variables (see below): The regression coefficients obtained in this way can therefore be interpreted as the difference between the groups in terms of the logarithm of meta-d'/d'.

HMeta-d with a continuous predictor variable According to the default implementation of HMeta-d for continuous predictor variables, on the level of a single participant j , the priors of the parameters were specified as follows:

$$\begin{aligned}\theta_{jR0} &\sim \text{tr}\mathcal{N}(\mu = -\mu_\theta, \sigma = \sigma_\theta, a = -\infty, b = c_j) \\ \theta_{jR1} &\sim \text{tr}\mathcal{N}(\mu = \mu_\theta, \sigma = \sigma_\theta, a = c_j, b = \infty) \\ \log(d_{\text{meta}}/d')_j &= x_j \times \beta_j + M_j, \\ M_j &\sim \mathcal{N}(\mu = \mu_M, \sigma = \sigma_M) \\ \beta_j &\sim \mathcal{N}(\mu = \mu_\beta, \sigma = \sigma_\beta)\end{aligned}\quad (16)$$

where again θ_{jR0} indicates the confidence criteria of participant j when the response was 0, θ_{jR1} indicates the confidence criteria of participant j when the response was 1, $\text{tr}\mathcal{N}$

indicates the truncated gaussian distribution with location parameter μ , scale parameter σ , lower bound a , and upper bound b ; x_j is the continuous predictor variable, β_j is the subject-specific effect of the continuous predictor variable on $\log(d_{meta}/d')$, M_j is the subject-specific intercept, μ_M and σ_M are the parameters of the distribution of M_j on the group level, μ_β quantifies the overall effect of the continuous predictor variable on $\log(d_{meta}/d')$, and σ_β quantifies the variability of the effect across subjects.

On the group level, we used three different sets of hyperpriors. First, we used the standard priors for HMeta-d as specified in the computer code provided by Fleming, which uses informative priors for μ_M and μ_β and relatively flat priors for all other parameters.

$$\begin{aligned}\mu_\theta &\sim \mathcal{N}(\mu = 0, \sigma = 10) \\ \sigma_\theta &\sim \text{tr}\mathcal{N}(\mu = 0, \sigma = 10, a = 0, b = \infty) \\ \mu_M &\sim \mathcal{N}(\mu = 0, \sigma = 1) \\ \sigma_M &\sim \text{tr}\mathcal{N}(\mu = 0, \sigma = 10, a = 0, b = \infty) \\ \mu_\beta &\sim \mathcal{N}(\mu = 0, \sigma = 1,) \\ \sigma_\beta &\sim \text{tr}\mathcal{N}(\mu = 0, \sigma = 10, a = 0, b = \infty)\end{aligned}\quad (17)$$

Second, we used close-to-flat prior for all parameters.

$$\begin{aligned}\mu_\theta &\sim \mathcal{N}(\mu = 0, \sigma = \sqrt{1000}) \\ \sigma_\theta^{-2} &\sim \Gamma(\alpha = 0.001, \beta = 0.001) \\ \mu_M &\sim \mathcal{N}(\mu = 0, \sigma = \sqrt{1000}) \\ \sigma_M^{-2} &\sim \Gamma(\alpha = 0.001, \beta = 0.001) \\ \mu_\beta &\sim \mathcal{N}(\mu = 0, \sigma = \sqrt{1000}) \\ \sigma_\beta^{-2} &\sim \Gamma(\alpha = 0.001, \beta = 0.001)\end{aligned}\quad (18)$$

Finally, we examined the false positive rate with a set of more narrow prior distributions:

$$\begin{aligned}\mu_\theta &\sim \mathcal{N}(\mu = 0, \sigma = 1) \\ \sigma_\theta^{-2} &\sim \Gamma(\alpha = 1, \beta = 1) \\ \mu_M &\sim \mathcal{N}(\mu = 0, \sigma = 1) \\ \sigma_M^{-2} &\sim \Gamma(\alpha = 1, \beta = 1) \\ \mu_\beta &\sim \mathcal{N}(\mu = 0, \sigma = 1) \\ \sigma_\beta^{-2} &\sim \Gamma(\alpha = 1, \beta = 1)\end{aligned}\quad (19)$$

Logistic mixed-model regression

Logistic regression is a specific case of a generalized linear regression model (Bolker et al., 2008). In general, it is a method to quantify the relationship between a binary outcome variable and one or several dichotomous or continuous predictors. There is quite a variety of different logistic regression models to measure metacognitive accuracy (Barthelmé & Mamassian, 2009; Murayama et al., 2014; Rahnev et al., 2020; Sandberg et al., 2010; Wierzchoń et al., 2019), which is why we selected two logistic regression models for the present study with wider applicability (Sandberg et al., 2010; Wierzchoń et al., 2019). In both models, the probability of being correct in the primary task $p(T)$ is modelled as

a linear function of a confidence rating C . A linear relationship between confidence and accuracy is obtained by transforming the probability of being correct into the logarithm of the odds of the primary response being correct to being incorrect. The two logistic regression models used in the present study differed in terms of the random effect structure: According to the regression model proposed by Sandberg et al. (2010), the clustered nature of the data is modelled by a random effect of participant on the intercept:

$$\log\left(\frac{p(T)}{1-p(T)}\right) = \gamma_0 + \mu_j + \beta_1 \times C \quad (20)$$

where γ_0 is the overall intercept, μ_j is the random intercept, and β_1 is the slope of the effect of confidence and considered to be fixed across participants. In contrast, according to the regression model by Wierzchoń et al. (2019), there is not only a random effect on the intercept, but also on the slope:

$$\log\left(\frac{p(T)}{1-p(T)}\right) = \gamma_0 + \mu_{0j} + (\gamma_1 + \mu_{1j}) \times C \quad (21)$$

where γ_0 is the overall intercept, μ_{0j} is the random intercept, γ_1 is the overall slope of the effect of confidence, and μ_{1j} is the random slope of participant. All logistic regression models were fit using the lme4 library in R (Bates et al., 2015).

Statistical analysis

All analyses were conducted using the free software R (R Core Team, 2020).

For each simulated data set, we first excluded simulated subjects whose performance was below chance level. Summary-statistics to measure metacognitive accuracy were computed separately for each participant. We tested whether the summary-statistics were normally distributed using a series of Kolmogorov–Smirnov tests. Then we compared the two simulated groups using two-sample t-tests. For meta-d'/d' and meta-d_a/d_a, we performed two additional tests because ratio distributions are often non-gaussian, a non-parametric Mann–Whitney-U-test as well as t-tests on log-transformed meta-d'/d' and meta-d_a/d_a. For the logistic regression models, we used Wald z-tests to test the interaction effect between the fixed effect of group and the fixed effect of confidence. For all measures except for HMeta-d, the false positive rate was estimated by dividing the number of simulations with a significant effect of group by the total number of simulations. For HMeta-d, a false positive was defined as a simulated experiment where the 95% CI interval of the difference between the two simulated groups in terms of mean meta-d'/d' excluded zero.

For each simulation, we also tested if the Pearson correlation between the measure of metacognitive accuracy and the continuous predictor variable was significant. For meta-d'/d' and meta-d_a/d_a, we also tested Spearman's ρ as well as Pearson's r with log-transformed meta-d'/d' and meta-d_a/d_a. For the logistic regression models, we again used Wald z-tests to test the interaction effect between the fixed effect of predictor and the fixed effect of confidence. For all measures except for HMeta-d, the estimated false positive rate was obtained by dividing the number of simulations with a significant effect of predictor by the number of simulations after exclusion of convergence errors. For HMeta-d, a false positive was defined as a simulated experiment where the 95% CI interval of the regression coefficient excluded zero.

Statistical evidence if the false positive rate was different from the nominal alpha level of 0.05 was quantified using Bayes factors using default priors and the BayesFactor package (Morey & Rouder, 2015). Bayes factors were interpreted according to standard guidelines (Lee & Wagenmakers, 2013). We assumed a logistic prior distribution of logit-transformed false positive error rate with a location parameter corresponding to a false positive rate of 0.05 and a scale parameter of 0.5. The prior distribution implied a 95% prior probability that the empirical false positive rate would fall between 0.8% and 25.0%. This prior distribution represents the belief that false positive error rates close to the nominal alpha level of 0.05 are more likely a priori than more extreme false positive error rates. The same prior distribution was used to construct posterior distributions of the false positive rate for each measure of metacognitive accuracy. As explanatory analysis, we assessed the relationship between false positives and trial number as well as between false positives and subject number using logistic regression, which we converted into Bayes factors using the BFpack package (Mulder et al., 2021).

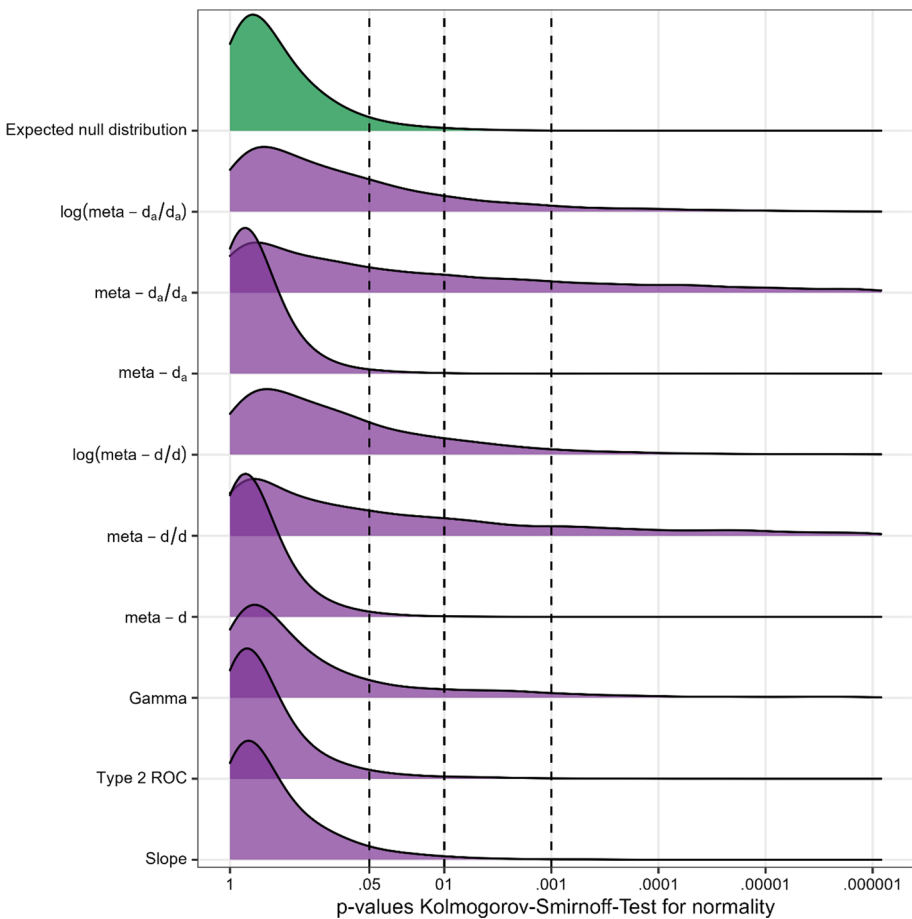


Fig. 1 P-values of Kolmogorov–Smirnov tests for normality of summary-statistic based measures of metacognitive accuracy. *Note.* The expected null distribution was simulated by sampling from a Gaussian distributions using the same sample sizes as in the other simulations

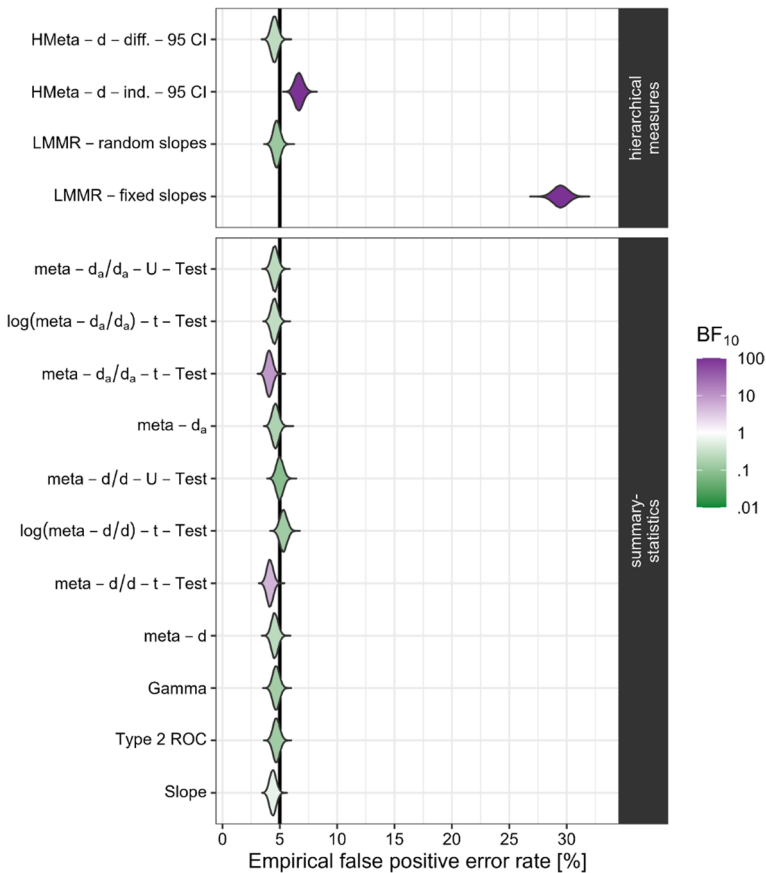


Fig. 2 Posterior distributions of false positive rate in comparisons between with two groups associated with different summary-statistics and hierarchical measures of metacognitive accuracy. *Note.* Colours indicate Bayes factors quantifying the strength of evidence that the false positive rate is different from 5%

Results

Testing summary-statistic based measures for normality

The distributions of meta- d'/d' and meta- d_a/d_a were often heavily leptokurtic (Kurtosis for meta- d'/d' $Mdn=8.53$ and for meta- d_a/d_a $Mdn=8.89$) but only slightly skewed (skewness for meta- d'/d' $Mdn=0.68$ and for meta- d_a/d_a $Mdn=0.58$, see Supplementary Fig. 1 for example distributions). We detected a significant deviation from the normal distributions in 47.54% of simulations for meta- d'/d' and 51.46% for meta- d_a/d_a (see Fig. 1). Log-transforming meta- d'/d' and meta- d_a/d_a reduced the kurtosis to $Mdn=5.63$ for meta- d'/d' and $Mdn=5.87$ for meta- d_a/d_a , without strongly affecting skewness (meta- d'/d' $Mdn=-1.01$; meta- d_a/d_a $Mdn=-0.92$), resulting in a reduced number of significant deviations from normality (28.24% for meta- d'/d' and 30.06% for meta- d_a/d_a ; see Supplementary Fig. 2). Gamma correlations also frequently deviated from normality (18.84%). The deviations appeared to be related to the fact that Gamma is bounded between -1 and 1 (see

Supplementary Fig. 3). For the other measures, kurtosis ranged between $Mdn=2.67$ and $Mdn=3.12$ and skewness between $Mdn=0.00$ and $Mdn=0.31$, resulting in between 1.26% (meta- d_a) and 5.66% (slopes) significant deviations from normality.

Grouping variable

Figure 2 shows the posterior distribution of the false positive rate associated with different summary-statistics and hierarchical measures of metacognitive accuracy, assuming comparisons between two independent groups.

Table 1 provides Bayes factors quantifying the evidence if the observed false positive rate of different measures of metacognitive accuracy is consistent with an alpha frequency of 5%. There was moderate evidence that the empirical false positive rate is identical to 5% for gamma correlations, type 2 ROC curves, meta- d' , and meta- d_a . For confidence slopes the evidence was not conclusive.

Concerning meta- d'/d' , there was moderate evidence that the empirical false positive rate is below 5% when significance testing was performed using t-tests. There was moderate evidence that the false positive rate is 5% when meta- d'/d' was log-transformed before

Table 1 Estimated false positive rates for different measures of metacognitive accuracy in comparisons between two separate groups

Method	Estimated false positive rate [%]	95% CI		BF_{10}	$p(H_0 Data)$
		Lower	Upper		
Area under the Type-II ROC curve	4.70	4.14	5.30	0.13	0.885
Confidence slope	4.40	3.86	4.99	0.61	0.621
Gamma	4.68	4.12	5.28	0.14	0.877
HMeta-d					
- independent priors, 95% CI	6.68	5.94	7.42	8.29×10^3	<0.001
- priors on the group difference, 95% CI	4.54	3.96	5.16	0.25	0.80
Logistic mixed model regression					
- fixed slopes	29.52	28.23	30.75	4.74×10^{678}	<0.001
- random slopes	4.74	4.17	5.35	0.12	0.893
meta- d'	4.56	4.01	5.16	0.24	0.807
meta- d'/d'					
- t-Test	4.14	3.62	4.72	5.27	0.160
- log-transformed, t-Test	5.32	4.72	5.96	0.13	0.885
- U-Test	5.00	4.42	5.62	0.08	0.926
meta- d_a	4.62	4.06	5.22	0.18	0.848
meta- d_a/d_a					
- t-Test	4.08	3.56	4.65	9.76	0.091
- log-transformed, t-Test	4.54	3.99	5.14	0.26	0.794
- U-Test	4.56	4.01	5.16	0.24	0.807

BF_{10} is a Bayes factor quantifying the evidence that the empirical false positive rate is different from 5%, assuming a logistic prior distribution of the logit-transformed false positive error rate with a location parameter that corresponded to a false positive rate of 5%, as well as a scale parameter of 0.5. 95% credible intervals were based on the same prior. $p(H_0|Data)$ is the posterior probability that the false positive rate is 5% given the data, assuming that the prior probability that the false positive rate is 5% is equal to 0.5

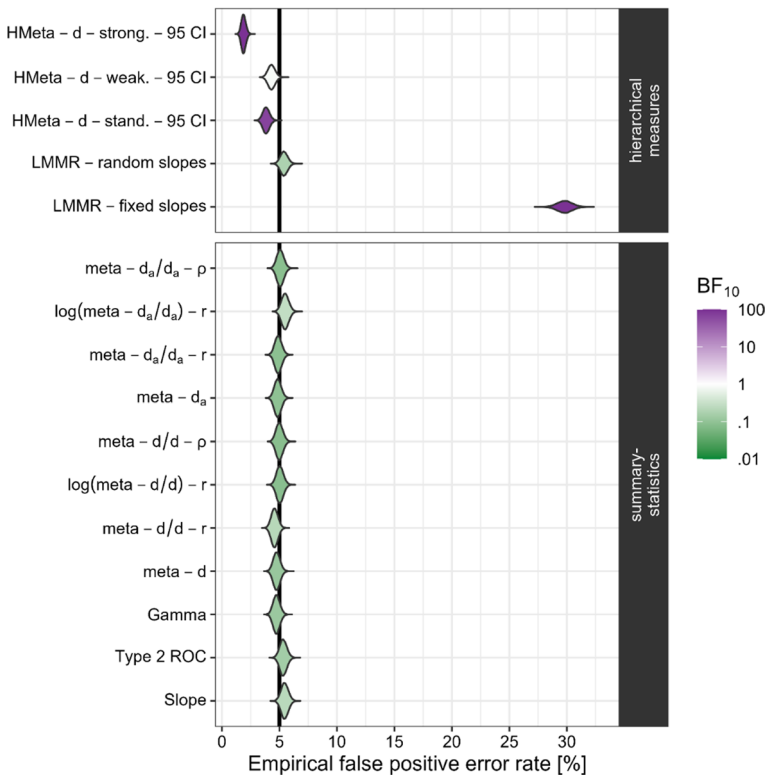


Fig. 3 Posterior distributions of false positive rate associated with different summary-statistics and hierarchical measures of metacognitive accuracy in simulations with a continuous predictor variable. *Note.* Colours indicate Bayes factors quantifying the strength of evidence that the false positive rate is different from 5%

using t-tests and strong evidence that the false positive rate is 5% when significance testing was based on the Mann–Whitney-U-test. Likewise, for meta-d_a/d_a, there was moderate evidence that the empirical false positive rate is lower than 5% when t-tests were used. However, there was moderate evidence that the empirical false positive rate is 5% when meta-d_a/d_a was log-transformed before using t-tests and when Mann–Whitney-U-tests were used.

For hierarchical measures, there was extremely strong evidence that the false positive rate of logistic mixed model regression with a random effect on intercepts is larger than 5%, but there was also moderate evidence that the false positive rate is 5% when logistic mixed model regression included random slopes. There was extremely strong evidence that the false positive rate is larger than 5% for HMeta-d when independent priors were placed on both groups. However, when a prior was placed on the difference between groups in terms of meta-d/d', there was a moderate amount of evidence that the false positive rate is 5%.

An exploratory analysis investigated if false positives are associated with mean accuracy in a simulated experiment, the number of trials and the number of participants. Supplementary Table S1 shows that the correlation with mean accuracy, the number of trials and the number of subjects was negligibly small for all measures, but with one exception:

For logistic mixed model regression with fixed slopes, there was very strong evidence that the probability for a false positive increased with average accuracy and with trial number.

Continuous predictor variables

Figure 3 shows the posterior distribution of the false positive rate associated with different summary-statistics and hierarchical measures of metacognitive accuracy when assessing the relationship to a continuous predictor variable.

Table 2 provides Bayes factors quantifying the evidence that the observed false positive rate of different measures of metacognitive accuracy is consistent with 5%. Concerning summary-statistic based measures, there was moderate evidence that the empirical false positive rate is 5% for type 2 ROC curves, confidence slopes, gamma correlations, meta-d', meta-d'/d' (tested using Pearson's r), and log-transformed meta-d_a/d_a. For meta-d_a, log-transformed meta-d'/d', meta-d'/d' (tested using Spearman's ρ), and meta-d_a/d_a (both when tested using Pearson's r and Spearman's ρ), there was even strong evidence that the observed false positive rate is 5%.

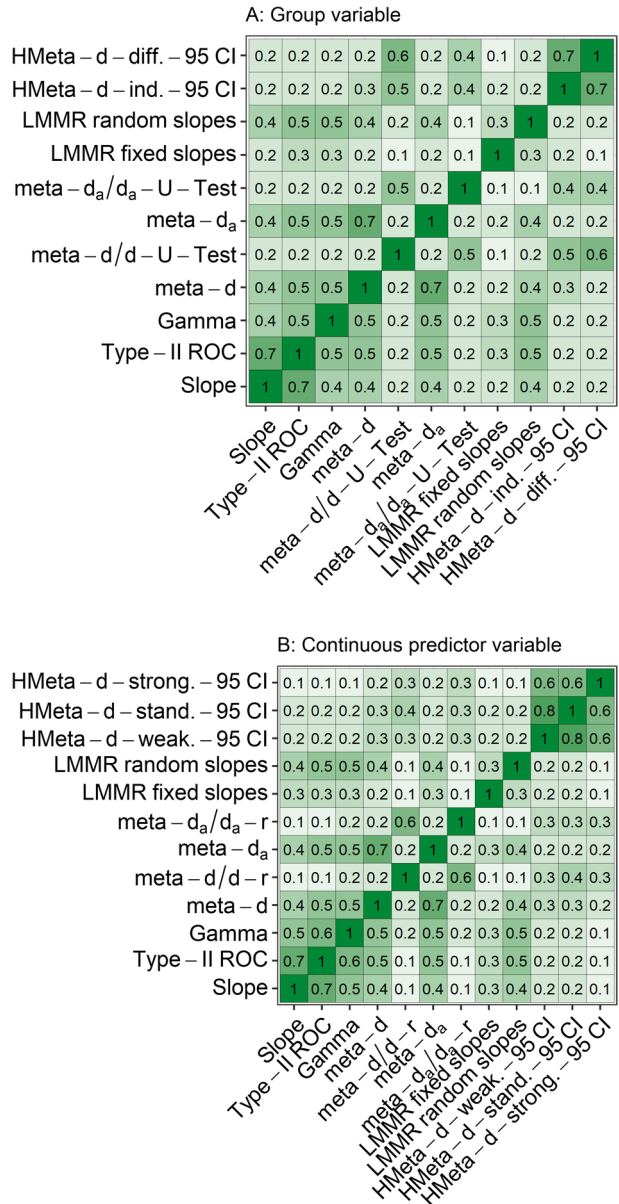
Table 2 Estimated false positive rates for different measures of metacognitive accuracy in in simulations with a continuous predictor

Method	Estimated false positive rate [%]	95% CI		BF_{10}	$p(H_0 Data)$
		Lower	Upper		
Area under the Type-II ROC curve	5.34	4.73	5.97	0.14	0.877
Confidence slope	5.46	4.85	6.10	0.23	0.813
Gamma	4.74	4.17	5.35	0.12	0.893
HMeta-d					
- Original standard prior	3.83	3.29	4.44	68.62	0.010
- Weak priors	4.31	3.73	4.94	0.85	0.541
- Strong priors	1.85	1.51	2.29	3.92×10^{26}	<0.001
Logistic mixed model regression					
- fixed slopes	29.86	28.56	31.09	1.08×10^{694}	<0.001
- random slopes	5.40	4.79	6.04	0.18	0.848
meta-d'	4.76	4.19	5.37	0.11	0.901
meta-d'/d'					
- Pearson's r	4.58	4.02	5.18	0.22	0.820
- Log-transformed, Pearson's r	5.04	4.45	5.66	0.08	0.926
- Spearman's ρ	5.00	4.42	5.62	0.08	0.926
meta-d _a	4.86	4.28	5.47	0.09	0.917
meta-d _a /d _a					
- Pearson's r	4.90	4.32	5.51	0.09	0.917
- Log-transformed, Pearson's r	5.50	4.88	6.14	0.27	0.787
- Spearman's ρ	5.08	4.49	5.7	0.08	0.926

BF_{10} is a Bayes factor quantifying the evidence that the empirical false positive rate is different from 5%, assuming a logistic prior distribution of the logit-transformed false positive error rate with a location parameter that corresponded to a false positive rate of 5%, as well as a scale parameter of 0.5. 95% credible intervals were based on the same prior. $p(H_0|Data)$ is the posterior probability that the false positive rate is 5% given the data, assuming that the prior probability that the false positive rate is 5% is equal to 0.5

Concerning hierarchical measures, there was extremely strong evidence that the false positive rate of logistic mixed model regression with a random effect on intercepts is larger than 5%, but there was also moderate evidence that the false positive rate of logistic mixed model regression with random slopes is identical to 5%. We also found strong evidence that HMeta-d using default priors is associated with a false positive rate below 5%. HMeta-d with stronger priors on the parameters of HMeta-d resulted an even smaller false positive

Fig. 4 Phi-correlations between the outcome of the statistical test between each pair possible pair of metacognitive accuracy. *Note.* The outcome of the statistical test was coded as 1 if the 95% CI excluded zero (HMeta-d) or if the test was significant (all other measures) and zero otherwise. (A) Correlation between the outcome of statistical tests with respect to the grouping variable. (B) Correlation between the outcome of statistical tests with respect to the continuous predicting variable



rate. In contrast, the evidence was not conclusive if the false positive rate was different from 5% when we tested a variant of HMeta-d with close-to-flat priors.

Again, we performed an exploratory analysis to see if false positives are associated with mean accuracy, the number of trials or the number of participants. Supplementary Table S2 shows that only for logistic mixed model regression with fixed slopes, the probability of a false positive increased with mean accuracy and with trial number. For all other measures, we detected no association with the probability of a false positive.

P-hacking by selecting measures of metacognitive accuracy

Next, we examined if it is possible to p-hack results by computing several measures of metacognitive accuracy and select the measure of metacognitive accuracy based on the outcome of the statistical test without correction of the alpha level for multiple comparisons. P-hacking by post-hoc selection of a measure of metacognitive accuracy is possible if the results of the corresponding statistical tests are not perfectly correlated. Figure 4 shows that for most pairs of measures of metacognitive accuracy, the correlations between the outcomes of statistical tests is only moderate.

Finally, we examined the false positive rate if researchers compute n different measures of metacognitive accuracy and report only those that yielded a significant result without correction for multiple comparisons. Because the purpose of the analysis was specifically the impact of switching between different measures of metacognitive accuracy and not switching between statistical methods, we did not include all variants of testing meta- d'/d' and meta- d_a/d_a for this specific analysis; instead, we assumed for this specific analysis that meta- d'/d' and meta- d_a/d_a were tested using Mann-Whitney-U-tests or Spearman's rho, respectively. We also excluded measures for which we had evidence that the false positive rate is not 5%. Figure 5 shows that the false positive rate is close to 5% when only one measure is randomly selected. However, randomly selecting two different measures

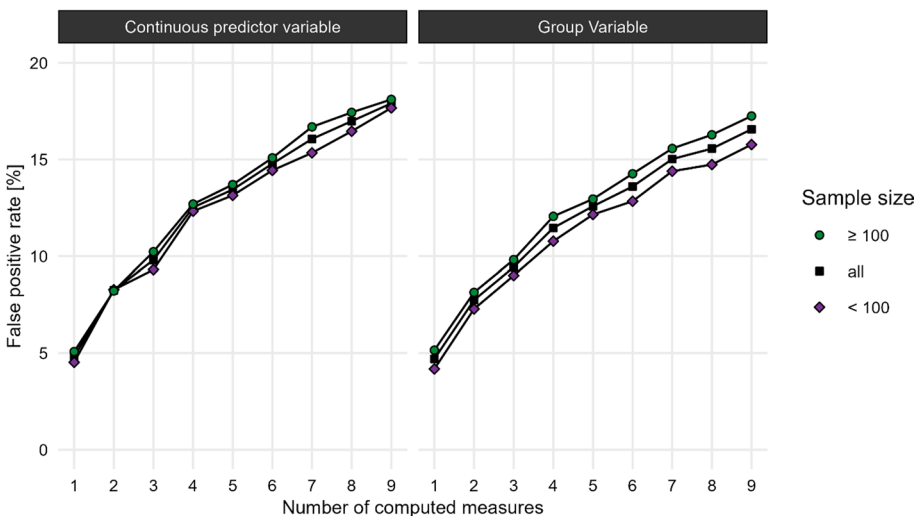


Fig. 5 Frequency of false positive results if researchers compute n measures of metacognitive accuracy and report only favorable outcomes

of metacognitive accuracy without correction is sufficient to increase the false positive to approximately 8%. When all nine different measures are computed, the false positive rate increases to 16.6% and 17.9%, respectively. Figure 5 also shows that unreported changes of the measure of metacognition do not increase the false positive rate to any lesser extent when the sample size is at least 100 subjects.

Discussion

We found that the false positive rates associated with most summary-statistic based measures of metacognitive accuracy were not distinguishable from 5%, including gamma correlations, type 2 ROC curves, meta- d' , and meta- d_a . For group comparisons using meta- d'/d' or meta- d_a/d_a , the false positive rate is too low when meta- d'/d' or meta- d_a/d_a is directly subjected to a t-test. However, the problem can be solved by either log transformation or by using the Mann–Whitney-U-test. In contrast, hierarchical measures of metacognitive accuracy did not perform consistently well: For logistic mixed model regression, the false positive rates are dramatically inflated when random slopes are omitted from the model specification. For HMeta- d , the false positive rate is slightly increased for group comparisons with independent priors on both groups, but the false positive rate is adequate when the prior is placed on the difference between groups. When the standard priors of HMeta- d are used with a continuous predictor variable, the false positive rate is below 5%; the false positive rate cannot be distinguished from 5% if close-to-flat priors are used. We also found that false positive rates increase dramatically when researchers choose a measure of metacognitive accuracy based on the results of statistical tests without correction of the alpha level for multiple corrections. In general, the present study demonstrates that it should not be assumed a priori that the false positive rate associated with a measure of metacognitive accuracy is acceptable; we strongly recommend that only measures for which it has been demonstrated that the false positive rate is appropriate should be used.

Robustness of measures of metacognitive accuracy

In the present study, none of the summary-statistic based measures of metacognitive accuracy were associated with an increased false positive rate. Thus, we did not replicate the increased false positive rate of gamma correlations reported in a previous simulation study (Murayama et al., 2014). In the present study, only hierarchical measures were found to have an increased false positive rate. First, the error rates associated with logistic mixed model regression with fixed slopes are extremely high. Previous studies have already shown that omitting random slopes from the model specification carries the risk of an increased false positive rate (Oberauer, 2022). Our study shows that such a high false positive rate not only occurs in ad-hoc simulations, but is also found in simulations with very similar characteristics to real datasets.

The error rate associated with HMeta- d in simulations with two groups and independent priors is slightly above 5%, which is within the range of empirical false positive rates that are considered to be robust (Bradley, 1978). Nevertheless, from the perspective of the ongoing reproducibility crisis in psychology (Nelson et al., 2018; Nosek et al., 2022; Pashler & Harris, 2012), it seems to us that even such a small excess in empirical false positive rate is too large: If the true false positive rate is 6.7% instead of 5%, the risk of published studies reporting statistical evidence for a non-existent effect increases by about 25%. For

this reason, we recommend placing a prior on the between-group difference, which in our stimulation is sufficient to avoid an increased false positive rate.

Considerations beyond false positive rates

The false-positive rate is only one factor that researchers need to consider when choosing a measure of metacognitive accuracy. A more commonly discussed factor is the validity of different measures of metacognitive accuracy: Measures of metacognitive accuracy should not be confounded by other theoretically important variables. Specifically, previous studies have examined whether measures of metacognitive accuracy are contaminated by task performance, task criteria, and confidence criteria (Barrett et al., 2013; Guggenmos, 2021; Masson & Rotello, 2009; Rahnev, 2023; Rausch & Zehetleitner, 2017; Rausch et al., 2023; Shekhar & Rahnev, 2021). Measures based on type II signal detection theory have become very popular because it was promised that type 2 signal detection theory isolates measures of metacognitive accuracy from some of these confounds (Fleming & Lau, 2014). For example, type 2 ROC curves have been designed to control for the criteria that participants apply when reporting their level of confidence (Fleming et al., 2010). Unfortunately, type 2 ROC curves do not control for task performance and task criteria (Fleming & Lau, 2014), and even the control for confidence criteria is not necessarily robust (Shekhar & Rahnev, 2021). A reanalysis of seven experiments showed a medium-sized correlation between type 2 ROC curves and task performance and only a very small correlation between type 2 ROC curves with task criteria and with confidence criteria (Rahnev, 2023). Meta- d'/d' was designed to explicitly control task performance, task criterion, and confidence criteria simultaneously without assuming a specific generative model underlying confidence judgments (Maniscalco & Lau, 2014). Unfortunately, it has been repeatedly shown that the control provided by meta- d'/d' is not necessarily effective (Guggenmos, 2021; Rahnev, 2023; Shekhar & Rahnev, 2021) and depends on the generative model underlying confidence judgments (Boundy-Singer et al., 2022; Rausch et al., 2023; Zhu et al., 2023). In addition, meta- d'/d' is influenced by the dynamics of the decision process (Desender et al., 2022). However, a recent study showed that the correlation between meta- d'/d' and task performance and between meta- d'/d' and task criteria is very small, although there is a medium-sized correlation between meta- d'/d' and confidence criteria (Rahnev, 2023). Others have proposed measures of metacognitive accuracy that depend on a specific generative model of confidence (Boundy-Singer et al., 2022; Desender et al., 2022; Guggenmos, 2022; Mamassian & de Gardelle, 2021; Shekhar & Rahnev, 2021, 2022); however, there is currently no consensus on which model best describes human confidence for the maximum number of data sets (Rahnev et al., 2022). Overall, the appropriate choice for a measure of metacognitive accuracy depends on the theoretical variables necessary to control for a specific research question, as well as the statistical properties of confidence judgments in a specific experiment.

A second factor that researchers need to consider is statistical power. Although the present study did not explicitly address statistical power, the present results are informative about the statistical power of meta- d'/d' and meta- d_a/d_a in group designs with t-tests, and of HMeta-d with continuous predictors and standard priors, because the false positive rate was even lower than 5%. An empirical false positive rate below the nominal alpha level indicates that the statistical power is too low (Bradley, 1978). This is because the alpha level is a criterion that researchers use to make statistical inferences, controlling both power and false positive rate. If researchers are willing to accept a higher false positive rate

by increasing the alpha level, they are more likely to detect effects if they exist. Likewise, if researchers keep the false positive rate excessively small, there will also be no large power to detect effects even though they exist. The low power of meta- d'/d' and meta- d_a/d_a can be explained by a high number of stimulations in which meta- d'/d' and meta- d_a/d_a were not normally distributed. Importantly, low statistical power also paradoxically implies that the probability that a published significant research finding is true is low (Ioannidis, 2005). Thus, previously reported results based on meta- d'/d' and meta- d_a/d_a in group designs (assuming parametric statistics and without log-transformation) or results based on HMeta- d with regression analysis and standard priors should be interpreted with caution.

For group designs, the false positive rate of meta- d'/d' and meta- d_a/d_a is just as expected when the non-parametric Mann–Whitney-U-test is used or when meta- d'/d' and meta- d_a/d_a are log-transformed, so we recommend using one of these procedures in future studies. For HMeta- d , the false positive rate was closer to 5% when a close-to-flat prior was used, but a close-to-flat prior may not be considered as reasonable choice as it gives too much weight to extremely large meta- d'/d' ratios that are not plausible from a theoretical point of view. Instead, it is preferable to ensure sufficient statistical power of HMeta- d through sample size planning using Bayesian power analysis (Kruschke, 2014).

Preregistration is recommended

Regardless of the measure researchers choose to use, we recommend that researchers pre-register their choice of measure of metacognitive accuracy prior to data collection, especially in cases where more than one choice for a measure of metacognitive accuracy is defensible to a sceptical audience. The reason is that the false positive rate is unacceptably high if the measure of metacognitive accuracy is chosen based on the outcome of the statistical test. It is well known that undisclosed flexibility in data analysis causes an unacceptable high number of false positive results (Simmons et al., 2011). The present study demonstrates that the measure of metacognitive accuracy is a choice that gives researchers the opportunity to engage in so-called p-hacking (Nelson et al., 2018). This means that it is crucial that researchers select measures of metacognitive accuracy independently of the outcome of their statistical tests. The best way to demonstrate that analysis procedures have been selected independently from the results is by preregistration before data collection (Chambers, 2013; Nosek & Lakens, 2014; Wagenmakers et al., 2012). Unfortunately, if measures disagree if an effect is present or not, researchers may be tempted to convince oneself that the significant measure is the more accurate or powerful measure. Thus, preregistration is an important tool for researchers to try and be objective about their own results (Oberauer & Lewandowsky, 2019).

Conclusion

The present analysis showed that gamma correlations, confidence slopes, type 2 ROC curves, meta- d' , and meta- d_a are all adequate in terms of false positive rate. The false positive rate associated with meta- d'/d' is adequate when meta- d'/d' is log-transformed or when significance is assessed by a non-parametric statistical test. For logistic regression, the false positive rate is too high if the random effect of slope is omitted from the model. For HMeta- d , the false positive rate is slightly too high if independent priors are placed on two different groups. Selecting measures of metacognitive accuracy based on the outcome of

statistical tests inflates false positive rates, which is why preregistration is recommended for future studies investigating metacognitive accuracy. Finally, because it cannot be assumed a priori that newly proposed measures of metacognitive accuracy are adequate in terms of false positive rates, we recommend that any new method to assess metacognitive accuracy should be carefully validated in terms of false positive rate.

Open practices statement

The data sets underlying the present simulations are available from the confidence database (<https://osf.io/s46pr>; Rahnev et al., 2020). All simulated data and analysis code are available at <https://osf.io/4pfrs>. This study's design and its analysis were not pre-registered.

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1007/s11409-023-09353-y>.

Funding Open Access funding enabled and organized by Projekt DEAL. This research was in parts supported by the Deutsche Forschungsgemeinschaft (DFG, grant numbers ZE 887/8–1, RA2988/3–1). The funders had no role in study design, data collection, analysis, decision to publish, or preparation of the manuscript.

Declarations

Conflicts of interest None.

Ethical approval None. The study is based on simulations and data freely available on the web.

Informed consent Not applicable.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Aitchison, L., Bang, D., Bahrami, B., & Latham, P. E. (2015). Doubly Bayesian analysis of confidence in perceptual decision-making. *PLoS Computational Biology*, *11*(10), e1004519. <https://doi.org/10.1371/journal.pcbi.1004519>
- Alkan, E., Davies, G., Greenwood, K., & Evans, S. L. H. (2020). Brain structural correlates of metacognition in first-episode psychosis. *Schizophrenia Bulletin*, *46*(3), 552–561. <https://doi.org/10.1093/schbul/sbz116>
- Baranski, J. V., & Petrusic, W. M. (1994). The calibration and resolution of confidence in perceptual judgments. *Perception & Psychophysics*, *55*(4), 412–428. <https://doi.org/10.3758/BF03205299>
- Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278. <https://doi.org/10.1016/j.jml.2012.11.001>
- 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>

- Barrientos, M. S., Valenzuela, P., Hojman, V., & Reyes, G. (2022). Students with high metacognition are favourable towards individualism when anxious. *Frontiers in Psychology, 13*, 910132. <https://doi.org/10.3389/fpsyg.2022.910132>
- Barthelmé, S., & Mamassian, P. (2009). Evaluation of objective uncertainty in the visual system. *PLoS Computational Biology, 5*(9), e1000504. <https://doi.org/10.1371/journal.pcbi.1000504>
- Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1–48. <https://doi.org/10.18637/jss.v067.i01>
- Boehm, U., Annis, J., Frank, M. J., Hawkins, G. E., Heathcote, A., Kellen, D., Kryptos, A.-M., Lerche, V., Logan, G. D., Palmeri, T. J., van Ravenzwaaij, D., Servant, M., Singmann, H., Starns, J. J., Voss, A., Wiecki, T. V., Matzke, D., & Wagenmakers, E.-J. (2018). Estimating across-trial variability parameters of the Diffusion Decision Model: Expert advice and recommendations. *Journal of Mathematical Psychology, 87*, 46–75. <https://doi.org/10.1016/j.jmp.2018.09.004>
- Bolker, B. M., Brooks, M. E., Clark, C. J., Geange, S. W., Poulsen, J. R., Stevens, M. H. H., & White, J.-S.S. (2008). Generalized linear mixed models: A practical guide for ecology and evolution. *Trends in Ecology and Evolution, 24*(3), 127–135. <https://doi.org/10.1016/j.tree.2008.10.008>
- Boundy-Singer, Z. M., Ziemia, C. M., & Goris, R. L. T. (2022). Confidence reflects a noisy decision reliability estimate. *Nature Human Behaviour, 7*(1), 142–154. <https://doi.org/10.1038/s41562-022-01464-x>
- Bradley, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology, 31*(2), 144–152. <https://doi.org/10.1111/j.2044-8317.1978.tb00581.x>
- Chambers, C. D. (2013). Registered reports: A new publishing initiative at Cortex. *Cortex, 49*, 609–610.
- Clarke, F. R., Birdsall, T. G., & Tanner, W. P. (1959). Two types of ROC curves and definitions of parameters. *The Journal of the Acoustical Society of America, 31*(5), 629–630. <https://doi.org/10.1121/1.1907764>
- Davies, G., Rae, C. L., Garfinkel, S. N., Seth, A. K., Medford, N., Critchley, H. D., & Greenwood, K. (2018). Impairment of perceptual metacognitive accuracy and reduced prefrontal grey matter volume in first-episode psychosis. *Cognitive Neuropsychiatry, 23*(3), 165–179. <https://doi.org/10.1080/13546805.2018.1444597>
- Desender, K., Vermeylen, L., & Verguts, T. (2022). Dynamic influences on static measures of metacognition. *Nature Communications, 13*(1), 1–30. <https://doi.org/10.1038/s41467-022-31727-0>
- Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive—Developmental inquiry. *American Psychologist, 34*(10), 906–911. <https://doi.org/10.1037/0003-066X.34.10.906>
- Fleming, S. M. (2017). HMeta-d: Hierarchical Bayesian estimation of metacognitive efficiency from confidence ratings. *Neuroscience of Consciousness, 1*, 1–14. <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>
- Fleming, S. M., Weil, R. S., Nagy, Z., Dolan, R. J., & Rees, G. (2010). Relating introspective accuracy to individual differences in brain structure. *Science, 329*(5998), 1541–1543. <https://doi.org/10.1126/science.1191883>
- 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.
- Gelman, A., & Loken, E. (2014). The statistical crisis in science. *American Scientist, 102*(6), 460–465. <https://doi.org/10.1511/2014.111.460>
- Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science, 7*(4), 457–511.
- Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. Wiley.
- Guggenmos, M. (2021). Measuring metacognitive performance: Type 1 performance dependence and test-retest reliability. *Neuroscience of Consciousness, 7*(1), 1–14. <https://doi.org/10.1093/nc/niab040>
- Guggenmos, M. (2022). Reverse engineering of metacognition. *Elife, 11*, 1–29. <https://doi.org/10.7554/eLife.75420>
- Hellmann, S., Zehetleitner, M., & Rausch, M. (2023). Simultaneous modeling of choice, confidence, and response time in visual perception. *Psychological Review, 130*(1), 1–14. <https://doi.org/10.1037/rev0000411>
- Hesselmann, G. (2018). Applying linear mixed effects models (LMMs) in within-participant designs with subjective trial-based assessments of awareness—A caveat. *Frontiers in Psychology, 9*, 788. <https://doi.org/10.3389/fpsyg.2018.00788>
- Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine, 2*(8), 696–701. <https://doi.org/10.1371/journal.pmed.0020124>
- Judd, C. M., Westfall, J., & Kenny, D. A. (2017). Experiments with more than one random factor: Designs, analytic models, and statistical power. *Annual Review of Psychology, 68*, 601–625. <https://doi.org/10.1146/annurev-psych-122414-033702>

- Kantner, J., & Lindsay, D. S. (2012). Response bias in recognition memory as a cognitive trait. *Memory & Cognition*, 40(8), 1163–1177. <https://doi.org/10.3758/s13421-012-0226-0>
- Kristensen, S. B., Sandberg, K., & Bibby, B. M. (2020). Regression methods for metacognitive sensitivity. *Journal of Mathematical Psychology*, 94(102297), 1–17. <https://doi.org/10.1016/j.jmp.2019.102297>
- Kruschke, J. (2014). *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan*. Academic Press.
- Lee, M. D., & Wagenmakers, E.-J. (2013). *Bayesian cognitive modeling: A practical course*. Cambridge University Press.
- Mamassian, P., & de Gardelle, V. (2021). Modeling perceptual confidence and the confidence forced-choice paradigm. *Psychological Review*, 1–23. <https://doi.org/10.1037/rev0000312>
- Maniscalco, B., & Lau, H. (2012). A signal detection theoretic method for estimating metacognitive sensitivity from confidence ratings. *Consciousness and Cognition*, 21(1), 422–430.
- Maniscalco, B., & Lau, H. C. (2014). Signal detection theory analysis of type 1 and type 2 data: Meta-d', response-specific meta-d', and the unequal variance SDT model. In S. M. Fleming & C. D. Frith (Hrsg.), *The cognitive neuroscience of metacognition* (S. 25–66). Springer. https://doi.org/10.1007/978-3-642-45190-4_3
- 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>
- Matuschek, H., Kliegl, R., Vasishth, S., Baayen, H., & Bates, D. (2017). Balancing Type I error and power in linear mixed models. *Journal of Memory and Language*, 94, 305–315. <https://doi.org/10.1016/j.jml.2017.01.001>
- Mazancieux, A., Dinze, C., Souchay, C., & Moulin, C. J. A. (2020). Metacognitive domain specificity in feeling-of-knowing but not retrospective confidence. *Neuroscience of Consciousness*, 2020(1), niaa001. <https://doi.org/10.1093/nc/niaa001>
- McNabb, C. B., & Murayama, K. (2021). Unnecessary reliance on multilevel modelling to analyse nested data in neuroscience: When a traditional summary-statistics approach suffices. *Current Research in Neurobiology*, 2, 1–7. <https://doi.org/10.1016/j.crneur.2021.100024>
- Moran, R., Teodorescu, A. R., & Usher, M. (2015). Post choice information integration as a causal determinant of confidence: Novel data and a computational account. *Cognitive Psychology*, 78, 99–147. <https://doi.org/10.1016/j.cogpsych.2015.01.002>
- Morey, R. D., & Rouder, J. N. (2015). *BayesFactor: Computation of Bayes factors for common designs*. R package version 0.9.10–1. [Software]. <https://cran.r-project.org/package=BayesFactor>
- Mulder, J., Williams, D. R., Gu, X., Tomarken, A., Böing-Messing, F., Olsson-Collentine, A., Meijerink, M., Menke, J., van Aert, R., Fox, J.-P., Hoijsink, H., Rosseel, Y., Wagenmakers, E.-J., & van Lissa, C. (2021). BFpack: Flexible Bayes factor testing of scientific theories in R. *Journal of Statistical Software*, 100(18). <https://doi.org/10.18637/jss.v100.i18>
- Murayama, K., Sakaki, M., Yan, V. X., & Smith, G. (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>
- Nelson, L. D., Simmons, J., & Simonsohn, U. (2018). Psychology's renaissance. *Annual Review of Psychology*, 69, 511–534. <https://doi.org/10.1146/annurev-psych-122216-011836>
- Nelson, T. O. (1984). A comparison of current measures of the accuracy of feeling-of-knowing predictions. *Psychological Bulletin*, 95(1), 109–133.
- Nosek, B. A., Hardwicke, T. E., Moshontz, H., Allard, A., Corker, K. S., Dreber, A., Fidler, F., Hilgard, J., Struhl, M. K., Nuijten, M. B., Rohrer, J. M., Romero, F., Scheel, A. M., Scherer, L. D., Schönbrodt, F. D., & Vazire, S. (2022). Replicability, robustness, and reproducibility in psychological science. *Annual Review of Psychology*, 73, 719–748.
- Nosek, B. A., & Lakens, D. (2014). Registered reports: A method to increase the credibility of published results. *Social Psychology*, 45(3), 137–141. <https://doi.org/10.1027/1864-9335/a000192>
- Oberauer, K. (2022). The importance of random slopes in mixed models for Bayesian hypothesis testing. *Psychological Science*, 33(4), 648–665. <https://doi.org/10.1177/09567976211046884>
- Oberauer, K., & Lewandowsky, S. (2019). Addressing the theory crisis in Psychology. *Psychonomic Bulletin & Review*, 26, 1596–1618. <https://doi.org/10.3758/s13423-019-01645-2>
- Pashler, H., & Harris, C. R. (2012). Is the replicability crisis overblown? Three arguments examined. *Perspectives on Psychological Science*, 7(6), 531–536. <https://doi.org/10.1177/1745691612463401>
- Paulewicz, B., & Blaut, A. (2020). The bhsdr package: A general-purpose method of Bayesian inference for signal detection theory models. *Behavior Research Methods*, 52(5), 2122–2141. <https://doi.org/10.3758/s13428-020-01370-y>

- Pereira, M., Megevand, P., Tan, M. X., Chang, W., Wang, S., Rezai, A., Seeck, M., Corniola, M., Momjian, S., Bernasconi, F., Blanke, O., & Faivre, N. (2021). Evidence accumulation relates to perceptual consciousness and monitoring. *Nature Communications*, *12*(3261). <https://doi.org/10.17605/OSF.IO/YHXDB>
- Peters, M. A. K., Thesen, T., Ko, Y. D., Maniscalco, B., Carlson, C., Davidson, M., Doyle, W., Kuzniecky, R., Devinsky, O., Halgren, E., & Lau, H. (2017). Perceptual confidence neglects decision-incongruent evidence in the brain. *Nature Human Behaviour*, *1*(0139), 1–21. <https://doi.org/10.1038/s41562-017-0139>
- Peterson, W. W., Birdsall, T. G., & Fox, W. C. (1954). The theory of signal detectability. *Transactions of the IRE Professional Group on Information Theory*, *4*(4), 171–212. <https://doi.org/10.1109/TIT.1954.1057460>
- Pleskac, T. J., & Busemeyer, J. R. (2010). Two-stage dynamic signal detection: A theory of choice, decision time, and confidence. *Psychological Review*, *117*(3), 864–901. <https://doi.org/10.1037/a0019737>
- Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In K. Hornik, F. Leisch & A. Zeileis (Eds.) *Proceedings of the 3rd international workshop on distributed statistical computing (DSC 2003)*. Vienna, Austria. <https://www.ci.tuwien.ac.at/Conferences/DSC-2003/>
- Pollack, I. (1959). On indices of signal and response discriminability. *Journal of the Acoustical Society of America*, *31*, 1031.
- R Core Team. (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. <https://www.r-project.org/>
- Rahnev, D. (2023). *Measuring metacognition: A comprehensive assessment of current methods* [Preprint]. PsyArXiv. <https://doi.org/10.31234/osf.io/waz9h>
- Rahnev, D., Balsdon, T., Charles, L., de Gardelle, V., Denison, R., Desender, K., Faivre, N., Filevich, E., Fleming, S. M., Jehee, J., Lau, H., Lee, A. L. F., Locke, S. M., Mamassian, P., Odegaard, B., Peters, M. A. K., Reyes, G., Rouault, M., Sackur, J., ... Zylberberg, A. (2022). Consensus goals in the field of visual metacognition. *Perspectives on Psychological Science*, *17*(6), 1746–1765. <https://doi.org/10.1177/1745691622110756>
- Rahnev, D., Desender, K., Lee, A. L. F., Adler, W. T., Aguilar-Lleyda, D., Akdoğan, B., Arbuzaova, P., Atlas, L. Y., Balci, F., Bang, J. W., Bégue, I., Birney, D. P., Brady, T. F., Calder-Travis, J., Chetverikov, A., Clark, T. K., Davranche, K., Denison, R. N., Dildine, T. C., ... Zylberberg, A. (2020). The confidence database. *Nature Human Behaviour*, *4*, 317–325. <https://doi.org/10.1038/s41562-019-0813-1>
- Ratcliff, R., & Starns, J. J. (2009). Modeling confidence and response time in recognition memory. *Psychological Review*, *116*(1), 59–83. <https://doi.org/10.1037/a0014086>
- Ratcliff, R., & Starns, J. J. (2013). Modeling confidence judgments, response times, and multiple choices in decision making: Recognition memory and motion discrimination. *Psychological Review*, *120*(3), 697–719. <https://doi.org/10.1037/a0033152>
- Rausch, M., Hellmann, S., & Zehetleitner, M. (2018). Confidence in masked orientation judgments is informed by both evidence and visibility. *Attention, Perception, and Psychophysics*, *80*(1), 134–154. <https://doi.org/10.3758/s13414-017-1431-5>
- Rausch, M., Hellmann, S., & Zehetleitner, M. (2023). *Measures of metacognitive efficiency across cognitive models of decision confidence* [Preprint]. PsyArXiv. <https://doi.org/10.31234/osf.io/kdz34>
- Rausch, M., & Zehetleitner, M. (2017). Should metacognition be measured by logistic regression? *Consciousness and Cognition*, *49*, 291–312. <https://doi.org/10.1016/j.concog.2017.02.007>
- Rausch, M., Zehetleitner, M., Steinhäuser, M., & Maier, M. E. (2020). Cognitive modelling reveals distinct electrophysiological markers of decision confidence and error monitoring. *NeuroImage*, *218*(116963), 1–14. <https://doi.org/10.1016/j.neuroimage.2020.116963>
- Reynolds, A., Kvam, P. D., Osth, A. F., & Heathcote, A. (2020). Correlated racing evidence accumulator models. *Journal of Mathematical Psychology*, *96*, 102331. <https://doi.org/10.1016/j.jmp.2020.102331>
- Sandberg, K., Timmermans, B., Overgaard, M., & Cleeremans, A. (2010). Measuring consciousness: Is one measure better than the other? *Consciousness and Cognition*, *19*(4), 1069–1078. <https://doi.org/10.1016/j.concog.2009.12.013>
- Saravanan, V., Berman, G. J., & Sober, S. J. (2020). Application of the hierarchical bootstrap to multi-level data in neuroscience. *Neurons, Behavior, Data Analysis and Theory*, *3*(5), 1–25. <https://doi.org/10.1101/819334>
- Shekhar, M., & Rahnev, D. (2021). The nature of metacognitive inefficiency in perceptual decision making. *Psychological Review*, *128*(1), 45–70. <https://doi.org/10.1037/rev0000249>
- Shekhar, M., & Rahnev, D. (2022). How do humans give confidence? A comprehensive comparison of process models of metacognition. *PsyArXiv*. <https://doi.org/10.31234/osf.io/cwrnt>

- Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, 22(11), 1359–1366. <https://doi.org/10.1177/0956797611417632>
- Song, C., Kanai, R., Fleming, S. M., Weil, R. S., Schwarzkopf, D. S., & Rees, G. (2011). Relating inter-individual differences in metacognitive performance on different perceptual tasks. *Consciousness and Cognition*, 20(4), 1787–1792. <https://doi.org/10.1016/j.concog.2010.12.011>
- Steege, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing transparency through a multi-verse analysis. *Perspectives on Psychological Science*, 11(5), 702–712. <https://doi.org/10.1177/1745691616658637>
- Tanner, W. P., Jr., & Swets, J. A. (1954). A decision-making theory of visual detection. *Psychological Review*, 61(6), 401–409. <https://doi.org/10.1037/h0058700>
- Wagenmakers, E.-J., Wetzels, R., Borsboom, D., van der Maas, H. L. J., & Kievit, R. A. (2012). An Agenda for Purely Confirmatory Research. *Perspectives on Psychological Science*, 7(6), 632–638. <https://doi.org/10.1177/1745691612463078>
- Wierchoń, M., Anzulewicz, A., Hobot, J., Paulewicz, B., & Sackur, J. (2019). In search of the optimal measure of awareness: Discrete or continuous? *Consciousness and Cognition*, 75, 102798. <https://doi.org/10.1016/j.concog.2019.102798>
- Yates, J. F. (1990). *Judgment and decision making*. Prentice Hall.
- Zhu, J.-Q., Sundh, J., Spicer, J., Chater, N., & Sanborn, A. N. (2023). The autocorrelated Bayesian sampler: A rational process for probability judgments, estimates, confidence intervals, choices, confidence judgments, and response times. *Psychological Review*. <https://doi.org/10.1037/rev0000427>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.