Measuring metacognition: A comprehensive assessment of current methods

Dobromir Rahnev

School of Psychology, Georgia Institute of Technology, Atlanta, GA

Keywords: metacognition, confidence, signal detection theory, perceptual decision making, reliability, precision, bias

Acknowledgments

This work was supported by the National Institute of Health (award R01MH119189) and the Office of Naval Research (award N00014-20-1-2622). The author thanks Zoe Boundy-Singer for sharing scripts for estimating meta-uncertainty.

Correspondence

Dobromir Rahnev
Georgia Institute of Technology
654 Cherry Str NW
Atlanta, GA 30332

Abstract

One of the most important aspects of research on metacognition is the measurement of metacognitive ability. However, the properties of existing measures of metacognition have been mostly assumed rather than empirically established. Here I perform a comprehensive empirical assessment of many common measures (meta-d', M-Ratio, M-Diff, AUC2, Gamma, Phi, and ΔConf), as well as two recent model-based measures (meta-noise and meta-uncertainty). I also develop novel Ratio and Diff variants for the measures AUC2, Gamma, Phi, and ΔConf, resulting in a total of 17 measures of metacognition. To assess the measures, I develop a new method of determining the validity and precision of a measure of metacognition. In addition, I examine each measure's dependence on task performance, response bias, and metacognitive bias, as well as each measure's split-half and test-retest reliabilities. Finally, I examine the influence of trial number. Reassuringly, I find that all measures of metacognition investigated here are valid and most show similar levels of precision. Another reassuring finding is that all measures have very high split-half reliabilities for trial numbers over 100. However, the test-retest reliabilities are often very low with important implications for individual differences research. Finally, most measures show only weak dependence on response and metacognitive bias but many measures are strongly dependent on task performance. This comprehensive assessment paints a complex picture: no measure of metacognition is perfect and depending on the details of the experiment, different measures may be preferable. Based on these results, I make specific recommendations about the use of different measures.

Introduction

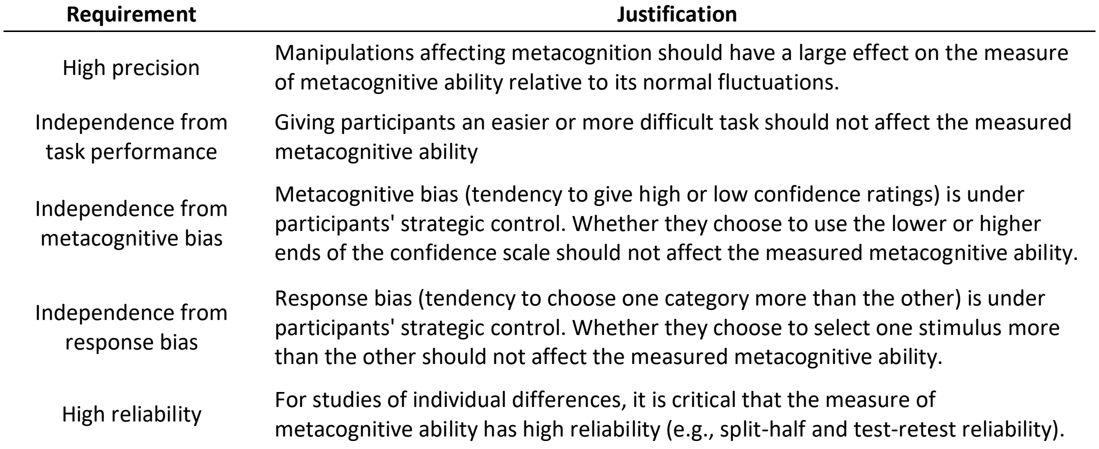
Metacognitive ability refers to the capacity to evaluate one's decisions (Metcalfe & Shimamura, 1994). High metacognitive ability allows us to have high confidence when we are correct but low confidence when we are wrong. Conversely, low metacognitive ability impairs the capacity of confidence ratings to distinguish between instances when we are correct or wrong. Metacognition is thus a critical faculty in human beings linked to our ability to learn (Guggenmos et al., 2016), make good decisions (Desender et al., 2018), interact with others (Pescetelli & Yeung, 2021), and know ourselves (Fleming, 2021). As such, it is critical that we have the tools to precisely measure this faculty in human participants.

Below, I first examine the properties that one may desire in a measure of metacognition and then review the known properties of existing measures of metacognitive ability. This brief overview demonstrates that there is little we firmly know about the properties of existing measures of metacognition. The rest of the paper aims to fill this gap by providing a comprehensive test of all critical properties of many common measures of metacognition.

The requirements for a measure of metacognition

Before one can evaluate a given measure of metacognition, it is first necessary to determine what properties are important or desirable. Since there is no existing list of desirable properties, I start by creating one here (Table 1). I believe that none of the properties listed below are controversial.

Table 1. Requirements for measures of metacognitive ability.



Requirement	Justification
High precision	Manipulations affecting metacognition should have a large effect on the measure of metacognitive ability relative to its normal fluctuations.
Independence from task performance	Giving participants an easier or more difficult task should not affect the measured metacognitive ability
Independence from metacognitive bias	Metacognitive bias (tendency to give high or low confidence ratings) is under participants' strategic control. Whether they choose to use the lower or higher ends of the confidence scale should not affect the measured metacognitive ability.
Independence from response bias	Response bias (tendency to choose one category more than the other) is under participants' strategic control. Whether they choose to select one stimulus more than the other should not affect the measured metacognitive ability.
High reliability	For studies of individual differences, it is critical that the measure of metacognitive ability has high reliability (e.g., split-half and test-retest reliability).

Validity and precision

The most important property of any measure is that it is valid: namely, it should measure what it purports to measure (Clark & Watson, 2019). A related property is precision (Luck et al., 2021): the measure should be as sensitive as possible to changes in the variable of interest relative to its inevitable fluctuations in repeated tests. However, despite the importance of validity and precision, these two properties have been largely ignored in the context of measures of metacognition. Here I develop a simple and intuitive method for assessing both validity and precision of metacognition measures. The method demonstrates that all existing measures of metacognition are valid but show some variations in precision.

Independence of nuisance variables

Perhaps the most widely appreciated desirable feature in a measure of metacognition is that it should be independent of various nuisance variables. Here a "nuisance variable" is any property of people's behavior that is not directly related to their metacognitive ability. The nuisance variable that has (rightfully) received the most attention is task performance (Fleming & Lau, 2014): a good measure of metacognition should not be affected by whether people happened to be performing an easy or a difficulty task. Task performance can be computed as *d'*, which is a measure of sensitivity derived from signal detection theory (SDT).

However, two more related features are also important. The first is independence of metacognitive bias, that is the tendency of people to be biased towards the lower or upper ranges of the confidence scale (Shekhar & Rahnev, 2021b; Xue et al., 2021). This variable can be quantified simply as the average confidence across all trials. The second is independence of response bias, that is, the tendency to select one response category more than another (Fleming & Lau, 2014). For two-choice tasks, this variable can be quantified as the decision criterion, *c*, derived from SDT. Both metacognitive and response bias are under the strategic control in that participants can freely choose to change them (e.g., they can do that in response of experimental manipulations such as expectation or reward, Rahnev & Denison, 2018). As such, measures of metacognitive ability should ideally remain independent of both types of bias.

Task performance, metacognitive bias, and response bias are arguably the primary nuisance variables that a measure of metacognitive ability should be independent of (Table 2). They are also variables that can be measured in any design that also allows the measurement of metacognitive ability. It is possible to add more variables to this list (e.g., reaction time, Desender et al., 2022) but the current paper only examines these three variables.

Table 2. Nuisance variables that can confound measures of metacognitive ability.

Measure	**Interpretation**
sensitivity (*d'*)	Task performance computed using SDT assumptions
confidence	Average confidence (measure of metacognitive bias)
criterion (*c*)	Response bias computed using SDT assumptions

Reliability

Validity, precision, and independence of nuisance variables are qualities that are important in any study. However, in studies of individual differences, it is also crucial that measures have high reliability. This paper examines both split-half and test-retest reliability.

Current measures of metacognitive ability

Several measures of metacognitive ability are relatively widely used. One popular measure is the area under the Type 2 ROC function (Clarke et al., 1959), also known as *AUC2*. Other popular measures are the Goodman–Kruskall Gamma coefficient (or just *Gamma*), which is essentially a rank correlation between trial-by-trial confidence and accuracy (Nelson, 1984) and the Pearson correlation between trial-by-trial confidence and accuracy (known as *Phi*; Kornell et al., 2007). Another simple but less frequently used measure is the difference between average confidence on correct trials and the average confidence on error trials (which I call *ΔConf*).

While all four of these traditional measures are intuitively appealing, they are all thought to be influenced by the primary task performance (Fleming & Lau, 2014). To address this issue, Maniscalco & Lau (2012) developed a new approach to measuring metacognitive ability where one can estimate the sensitivity, *meta-d'*, exhibited by the confidence ratings. Because *meta-d'* is expressed in the units of *d'*, Maniscalco and Lau then reasoned that *meta-d'* can be normalized by the observed d' to obtain either a ratio measure (*M-Ratio*, equal to *meta-d'*/*d'*) or a difference measure (*M-Diff*, equal to *meta-d'* – *d'*). These measures are often assumed to be independent of task performance (Fleming & Lau, 2014).

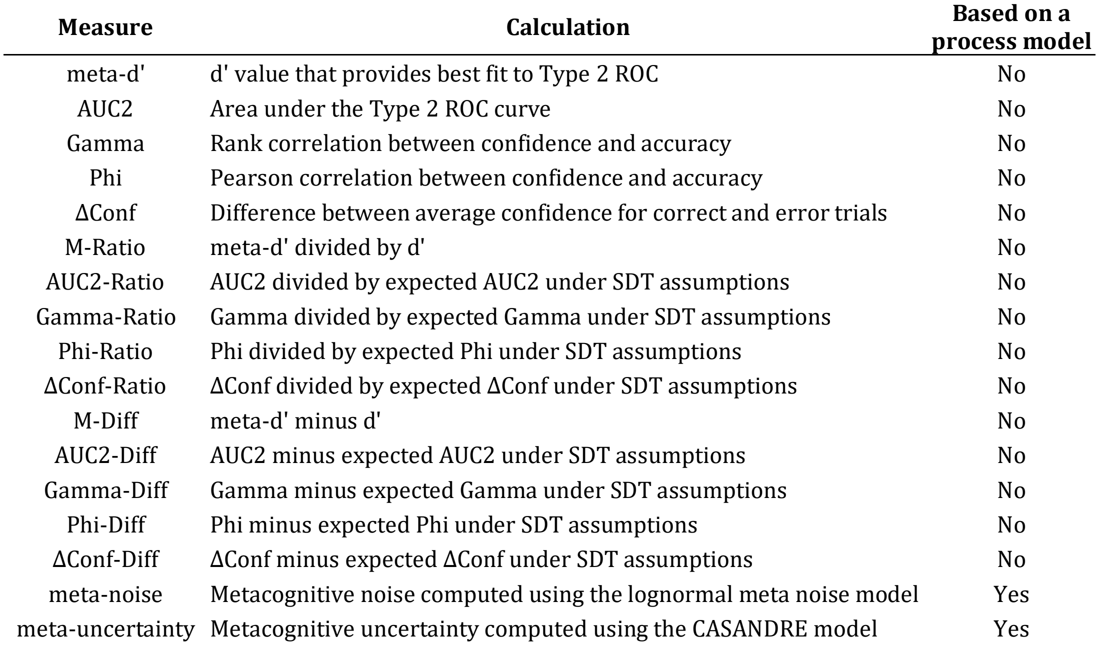
The normalization introduced by Maniscalco & Lau (2012) has only been applied to the measure *meta-d'* (resulting in the measures *M-Ratio* and *M-Diff*), but there is no theoretical reason why a conceptually similar correction cannot be applied to the traditional measures above. Consequently, here I develop eight new measures where one of the traditional measures of metacognitive ability is turned into either a ratio (*AUC2-Ratio*, *Gamma-Ratio*, *Phi-Ratio*, and *ΔConf-Ratio*) or a difference (*AUC2-Diff*, *Gamma-Diff*, *Phi-Diff*, and *ΔConf-Diff*) measure. The logic is that a given measure (e.g., *AUC2*) is computed once using the observed data (obtaining, e.g., $AUC2_{observed}$)

and a second time using the predictions of SDT given the observed sensitivity and decision criterion (obtaining, e.g., $AUC2_{expected}$). One can then take either the ratio or the difference between the observed and the SDT-predicted quantities.

Finally, one important limitation of all measures above is that they are not derived from a process model of metacognition. In other words, none of these measures are based on an explicit model of how confidence judgments may be corrupted. Recently, Shekhar & Rahnev (2021b) developed a process model of metacognition – the lognormal meta noise model – that is based on SDT assumptions but with the addition of lognormally distributed metacognitive noise. This metacognitive noise corrupts the confidence ratings but not the initial decision and, in the model, takes the form of confidence criteria that are sampled from a lognormal distribution rather than having constant values. The metacognitive noise parameter (σ_{meta}, referred here as *meta-noise*) can then be used as a measure of metacognitive ability. A similar approach was taken by Boundy-Singer et al. (2023) who developed another process model of metacognition, CASANDRE, based on the notion that people are uncertain about the uncertainty in their internal representations. The second-order uncertainty parameter (*meta-uncertainty*) thus represents another possible measure of metacognitive ability.

This paper examines the properties of all 17 measures of metacognition introduced above (for a summary, see Table 3). First, however, I briefly review the previous literature on the properties of these measures.

Table 3. Measures of metacognition examined in the current paper.



Measure	Calculation	Based on a process model
meta-d'	d' value that provides best fit to Type 2 ROC	No
AUC2	Area under the Type 2 ROC curve	No
Gamma	Rank correlation between confidence and accuracy	No
Phi	Pearson correlation between confidence and accuracy	No
ΔConf	Difference between average confidence for correct and error trials	No
M-Ratio	meta-d' divided by d'	No
AUC2-Ratio	AUC2 divided by expected AUC2 under SDT assumptions	No
Gamma-Ratio	Gamma divided by expected Gamma under SDT assumptions	No
Phi-Ratio	Phi divided by expected Phi under SDT assumptions	No
ΔConf-Ratio	ΔConf divided by expected ΔConf under SDT assumptions	No
M-Diff	meta-d' minus d'	No
AUC2-Diff	AUC2 minus expected AUC2 under SDT assumptions	No
Gamma-Diff	Gamma minus expected Gamma under SDT assumptions	No
Phi-Diff	Phi minus expected Phi under SDT assumptions	No
ΔConf-Diff	ΔConf minus expected ΔConf under SDT assumptions	No
meta-noise	Metacognitive noise computed using the lognormal meta noise model	Yes
meta-uncertainty	Metacognitive uncertainty computed using the CASANDRE model	Yes

Previous investigations into the properties of measures of metacognition

Given the importance of using measures with good psychometric properties, it is perhaps surprising that the published literature contains very little empirical investigation into the properties of the different measures of metacognition. For example, no paper to date has examined the precision of any existing measure. Several papers have used simulations to investigate some of the properties of measures of metacognition (Barrett et al., 2013; Guggenmos, 2021), but this approach is potentially problematic because it is a priori unknown how well the process models used to simulate data reflect empirical reality. One paper (Azzopardi & Evans, 2007) examined the properties of a measure, *Type-2 d'*, which was subsequently shown to be based on faulty assumptions (Galvin et al., 2003) and is therefore not investigated here. Finally, several older papers investigated the theoretical properties of several measures independent of any simulations or empirical data (Nelson, 1984) but this approach cannot be used to establish the *empirical* properties of the measures under consideration.

Only recently, Shekhar & Rahnev (2021b) examined the dependence on both task performance and metacognitive bias for five measures: *meta-d'*, *M-Ratio*, *AUC2*, *Phi*, and *meta-noise*. They found that *meta-d'*, *AUC2*, and *Phi* were strongly dependent on task performance, but *M-Ratio* and *meta-noise* were not. On the other hand, *meta-d'*, *M-Ratio*, *AUC2*, and *Phi* had a complex dependence on metacognitive bias, while only *meta-noise* appeared independent of it. Guggenmos (2021) examined both the split-half reliability and the across-subject correlation between *d'* and several measures of metacognition (*meta-d'*, *M-Ratio*, *M-Diff*, and *AUC2*) finding surprisingly low reliability and significant correlations with *d'* for all measures. Another paper developed a new technique to examine dependence on metacognitive bias and found that *meta-d'* and *M-Ratio* are not independent of metacognitive bias (Xue et al., 2021). Finally, Boundy-Singer et al. (2023) showed that *meta-uncertainty* appears to have high test-retest reliability, and only a weak dependence on task performance and metacognitive bias.

Current approach

As the brief overview above demonstrates, most previous investigations only focused on a few measures of metacognition, only examined a few of the critical properties of interest, and often did not make use of empirical data. Here I empirically examine each of the critical properties for all 17 measures of metacognition introduced above. To do so, I make use of seven large datasets (Adler & Ma, 2018; Haddara & Rahnev, 2022; Locke et al., 2020; Maniscalco et al., 2017; Rouault et al., 2018; Shekhar & Rahnev, 2021b) all made available on the Confidence Database (Rahnev et al., 2020). Details about the seven datasets are included in Table 4 and in the Methods. In addition, Table 4 indicates the dataset(s) used for each type of analysis.

Table 4. Datasets used in the current paper. The table lists details of each dataset, and indicates which analyses each dataset was used for.

Dataset	**Adler**	**Haddara**	**Locke**	**Maniscalco**	**Rouault1**	**Rouault2**	**Shekhar**
# subjects analyzed	19	70	10	22	466	484	20
# excluded subjects	0	5	0	8	32	13	0
% excluded subjects	0%	7%	0%	27%	6%	3%	0%
# trials/subject	1,916	3,000	4,900	1,000	210	210	2,800
# total trials in experiment	36,396	210,000	49,000	22,000	97,860	101,640	56,000
# difficulty levels	6	1	1	1	70	staircase	3
Criterion manipulated	—	—	✓	—	—	—	—
Original confidence scale	4-point	4-point	2-point	4-point	11-point	6-point	Continuous
Analyses on each dataset							
Precision	—	✓	—	✓	—	—	—
Dependance on task performance	✓	—	—	—	✓	✓	✓
Dependance on metacognitive bias	—	✓	—	✓	—	—	✓
Dependance on response bias	—	—	✓	—	—	—	—
Split-half reliability	—	✓	—	✓	—	—	✓
Test-retest reliability	—	✓	—	—	—	—	—

Overall, I find that no current measure of metacognitive ability is “perfect” in the sense of possessing all desirable properties. Based on these results, I make recommendations for the use of different measures of metacognition based on the specific analysis goals.

Methods

Datasets

To investigate the empirical properties of measures of metacognition, I used the datasets from the Confidence Database (Rahnev et al., 2020) that are most appropriate for each individual analysis. This process resulted in the selection of seven different datasets (Table 4), briefly discussed below in alphabetical order. In each case, subjects completed a 2-choice perceptual task and provided confidence ratings. For each dataset, I only considered trials from the main experiment and removed any staircase or practice trials that may have been included. In addition, I excluded subjects who had lower than 60% or higher than 95% accuracy, or who gave the same object-level or confidence response on more than 85% of trials. These exclusions were made because such subjects can have unstable metacognitive scores. Overall, these criteria led to excluding 58 out of 1,091 subjects (5.32% exclusion rate). Data were collected in a lab setting unless otherwise indicated.

Adler dataset

The first dataset is a combination of two very similar datasets: Task "A" from Adler_2018_Expt1 and Task "A" from Adler_2018_Expt2. The two experiments were identical except that in Expt1 subjects provided their decision and confidence with a single button press and did not receive trial-by-trial feedback, whereas in Expt2 subjects provided their decision and confidence with separate button presses and received trial-by-trial feedback. All experimental details can be found in the original publication (Adler & Ma, 2018). Briefly, the task was to determine the orientation (left vs. right) of drifting Gabor patches or ellipses. Adler_2018_Expt1 included 11 subjects and Adler_2018_Expt2 included eight subjects. Critically, each experiment featured six difficulty levels created by manipulating contrast for Gabor patches and elongation (eccentricity) for the ellipses. The combined dataset (named simply “Adler” here) consisted of 19 subjects, a total of 36,396 trials. No subjects were excluded from this dataset.

Haddara dataset

The second dataset is named “Haddara_2022_Expt2” in the Confidence Database (simplified to “Haddara” here) and consists of 75 subjects each completing 3,350 trials over seven days. Because Day 1 consisted of a smaller number of trials (350) compared to Days 2-7 (500 trials each), I only analyzed the data from Days 2-7 (3,000 trials total). All experimental details can be found in the original publication (Haddara & Rahnev, 2022). Briefly, the task was to determine the more frequent letter in a 7x7 display of X’es and O’s. Confidence was provided on a 4-point scale using a separate button press. The data collection was conducted online and half the subjects received trial-by-trial feedback (all subjects are considered jointly here). Five subjects were excluded from this dataset (6.67% exclusion rate).

Locke dataset

The third dataset is named “Locke_2020” in the Confidence Database (simplified to “Locke” here) and consists of 10 subjects each completing 4,900 trials. All experimental details can be found in the original publication (Locke et al., 2020).

Briefly, the task was to determine if a Gabor patch was tilted to the left or right of vertical. Confidence was provided on a 2-point scale using a separate button press. There were seven conditions with manipulations of both prior and reward. Rewards were manipulated by changing the payoff for correctly choosing category 1 vs. category 2 (e.g., R = 4:2 means that 4 vs. 2 points were given for correctly identifying categories 1 and 2, respectively), whereas priors were manipulated by informing subjects about the probability of category 2 (e.g., P = .75 means that there was 75% probability of presenting category 2 and 25% probability of presenting category 1). The seven categories were as follows (1) P = .5, R = 3:3, (2) P = .75, R = 3:3, (3) P = .25, R = 3:3, (4) P = .5, R = 4:2, (5) P = .5, R = 2:4, (6) P = .75, R = 2:4, and (7) P = .25, R = 4:2. There were equal number of trials (700) per condition. No subjects were excluded from this dataset.

Maniscalco dataset

The fourth dataset is named "Maniscalco_2017_expt1" in the Confidence Database (simplified to "Maniscalco" here) and consists of 30 subjects each completing 1,000 trials. All experimental details can be found in the original publication (Maniscalco et al., 2017). Briefly, the task was to determine which of two patches presented to the left and right of fixation contained a grating. A single difficulty condition was used throughout. Confidence was provided on a 4-point scale using a separate button press. Eight subjects were excluded from this dataset (26.67% exclusion rate).

Rouault1 and Rouault2 datasets

The fifth and sixth datasets are named “Rouault_2018_Expt1” and “Rouault_2018_Expt2” in the Confidence Database (simplified to “Rouault1” and “Rouault2” here). They consist of 498 and 497 subjects, respectively, each completing 210 trials. All experimental details can be found in the original publication that describes both datasets (Rouault et al., 2018). Briefly, the task was to determine which of two squares presented to the left and right of fixation contained a more dots and then rate confidence using a separate button press. The Rouault1 dataset had 70 difficulty conditions (where the difference in dot number between the two squares varied from 1 to 70) with 3 trials each. It collected confidence on a 11-point scale that goes from 1 (certainly wrong) to 11 (certainly correct). However, because the first six confidence ratings were used very infrequently, I combined them into a single rating, thus transforming the 11-point scale into a 6-point scale. On the other hand, Rouault2 used a continuously running staircase that adaptively modulated the difference in dots. It collected confidence on a 6-point scale that goes from 1 (guessing) to 6 (certainly correct), which is equivalent to the modified scale from Rouault1 and thus did not require additional modification. Data collection for both studies was conducted online. Thirty-two subjects were excluded from Rouault1 and 13 subjects were excluded from Rouault2 (6.43% and 2.62% exclusion rates, respectively).

Shekhar dataset

The final dataset is named “Shekhar_2021” in the Confidence Database (simplified to “Shekhar” here) and consists of 20 subjects each completing 2,800 trials. All experimental details can be found in the original publication (Shekhar & Rahnev, 2021b). Briefly, the task was to determine the orientation (left vs. right) of a Gabor patch presented at fixation. Subjects indicated their confidence simultaneously with the perceptual decision using a single mouse click. Confidence was provided on a continuous scale (from 50 to 100) but was binned into six levels as in the original publication. The dataset featured three different difficulty levels (manipulated by changing the contrast of the Gabor patch), which were analyzed separately. No subjects were excluded from this dataset.

Computation of each measure of metacognition

Previously proposed measures of metacognition

I computed a total of 17 measures of metacognition and provide Matlab code for their estimation. I first discuss nine of these measures that have been previously proposed: *AUC2*, *Gamma*, *Phi*, *ΔConf*, *meta-d’*, *M-Ratio*, *M-Diff*, *meta-noise*, and *meta-uncertainty*.

The first four of these measures have the longest history. *AUC2* was first proposed in the 1950’s (Clarke et al., 1959) and measures the area under the Type 2 ROC function that plots Type 2 hit rate vs. Type 2 false alarm rate. *Gamma* is perhaps the most popular measure in the memory literature and measures are the Goodman–Kruskall Gamma coefficient, which is essentially a rank correlation between trial-by-trial confidence and accuracy (Nelson, 1984). *Phi* is conceptually similar to *Gamma* but measures the Pearson correlation between trial-by-trial confidence and accuracy (Kornell et al., 2007). Finally, *ΔConf* (my terminology) measures the difference between average confidence on correct trials and the average confidence on error trials. *ΔConf* is perhaps the simplest and most intuitive measure of metacognition, but is used very infrequently in the literature.

The next three measures were developed by Maniscalco & Lau (2012). The researchers devised a new approach to measuring metacognitive ability where one can estimate the sensitivity, *meta-d'*, exhibited by the confidence ratings. Because *meta-d'* is expressed in the units of *d'*, Maniscalco and Lau then reasoned that *meta-d'* can be normalized by the observed d' to obtain either a ratio measure (*M-Ratio*, equal to *meta-d'*/*d'*) or a difference measure (*M-Diff*, equal to *meta-d'* – *d'*). These measures are often assumed to be independent of task performance (Fleming & Lau, 2014) but empirical work on this issue is scarce (though see Guggenmos, 2021).

Finally, recent years have seen a concerted effort to build models of metacognition derived from explicit process models of metacognition. Two such measures examined here were developed by Shekhar & Rahnev (2021b) and Boundy-Singer et al. (2023). Shekhar and Rahnev proposed the lognormal meta noise model, which is an SDT model with the additional assumption of lognormally distributed metacognitive noise that affects the confidence criteria. The metacognitive noise parameter (σ_{meta}, referred here as *meta-noise*) can then be used as a measure of metacognitive ability. The fitting of model to data is rather expensive because it requires the computation of many double-integrals that do not have numerical solutions. Consequently, the fitting method from Shekhar & Rahnev (2021b) takes substantially longer than other measures examined here, making the measure less practical. To address this issue, I make substantial modifications to the original code including many improvements in the efficiency of the algorithm and creating a lookup table so that values of the double integral do not need to be computed anew but can be simply loaded. These improvements reduce the computation of *meta-noise* from minutes to a few seconds, thus making the measure easy to use in practical applications. The measure developed by Boundy-Singer et al. (2023) - *meta-uncertainty* – is based on a different process model of metacognition, CASANDRE, that implements the notion that people are uncertain about the uncertainty in their internal representations. The second-order uncertainty parameter, *meta-uncertainty*, represents another possible measure of metacognition. The code for estimating *meta-uncertainty* was provided by Zoe Boundy-Singer.

New measures of metacognition

In addition to the already established measures mentioned above, I develop several new measures that conceptually follow the normalization procedure introduced by Maniscalco & Lau (2012). That normalization procedure has previously only been applied to the measure *meta-d'* (to create *M-Ratio* and *M-Diff*), but there is no theoretical reason why a conceptually similar correction cannot be applied to other traditional measures of metacognition. Consequently, here I develop eight new measures where one of the traditional measures of metacognitive ability is turned into either a ratio (*AUC2-Ratio*, *Gamma-Ratio*, *Phi-Ratio*, and *ΔConf-Ratio*) or a difference (*AUC2-Diff*, *Gamma-Diff*, *Phi-Diff*, and *ΔConf-Diff*) measure. The logic is to compute an observed and an expected value for any given measure (e.g., *AUC2*), and then use the expected value to normalize the observed value. First, a measure is computed using the observed data, thus producing what may be called, e.g., $AUC2_{observed}$. Critically, the measure is then computed again using the predictions of SDT given the observed sensitivity (*d'*) and criteria, thus obtaining what may be called, e.g., $AUC2_{expected}$. One can then take either the ratio (e.g., $AUC2_{observed}/AUC2_{expected}$) or the difference (e.g., $AUC2_{observed} – AUC2_{expected}$) between the observed and the SDT-predicted quantities to create the new measures of metacognition.

I computed the SDT expectations in the following way. First, I estimated d' using the formula:

$$d' = z(HR) - z(FAR)$$

where HR is the observed hit rate and FAR is the observed false alarm rate. Then, I estimated the location of all confidence and decision criteria using the formula:

$$c_i = -\frac{z(HR_i) + z(FAR_i)}{2}$$

In the formula above, i goes from $-(k-1)$ to $k-1$, for confidence ratings collected on an k-point scale. Intuitively, one can think of the confidence ratings $1,2,\dots k$ for category 1 being recoded to $-1,-2,\dots -k$, such that confidence goes from $-k$ to k and simultaneously indicates the decision (negative confidence values indicating a decision for category 1; positive confidence values indicating a decision for category 2). HR_i and FAR_i are then simply the proportion of times this rescaled confidence is higher or equal to i when category 2 and category 1 are presented, respectively.

Once the values of d' and c_i are computed, they can be used to generate predicted HR_i and FAR_i values (which would be slightly different from the empirically observed ones). The measures *AUC2*, *Gamma*, *Phi*, and *ΔConf* can then be straightforwardly computed based on the predicted HR_i and FAR_i values, thus enabling the computation of the new Ratio and Diff measures.

Assessing validity and precision

Any measure of metacognition should be valid and precise (Clark & Watson, 2019; Luck et al., 2021; Mueller & Knapp, 2018). However, there is no established method to assess either validity or precision of measures of metacognition. Here I develop a method to jointly assess validity and precision. The underlying idea is to artificially alter confidence to be less in line with accuracy and then assess how measures of metacognition change.

Specifically, the method corrupts confidence by decreasing confidence ratings for correct trials and increasing them for incorrect trials. For a given set of trials, the method loops over the trials starting from the first and (1) if the trial has a correct response and confidence higher than 1, then it decreases the confidence on that trial by 1 point, and (2) if the trial has an incorrect response and confidence lower than maximum (that is k on an k-point scale), then it increases the confidence on that trial by 1 point. If neither of these conditions apply, the trial is simply skipped. The method then continues to corrupt subsequent trials in the same manner until a pre-set proportion of corrupted trials is achieved. Then, all measures of metacognition are computed based on the corrupted confidence ratings. A given dataset is first split into n bins of a given trial number, and the procedure above is performed separately for each bin. Finally, to compute a measure of precision that can be compared across different measures of metacognition, I use the following formula:

$$precision = \frac{1}{n}\sum_{i=1}^{n}\frac{measureOrig_i - measureCorrupted_i}{SD}$$

where $measureOrig_i$ and $measureCorrupted_i$ are the values of a specific measure computed on the original (uncorrupted) and corrupted confidence ratings, respectively, n is the number of bins analyzed, and SD is the standard deviation of all $measureOrig_i$ for $i = 1,2,\dots,n$. Positive values of the variable $precision$ indicate valid measures of metacognition and higher values indicate more precise measures (e.g., measures more sensitive to corruption in confidence compared to background fluctuations).

I computed the precision of all 17 measures of metacognition for two datasets from the Confidence Database: Maniscalco (1 day; 1,000 trials per subject) and Haddara (6 days; 3,000 trials per subject). I separately examined the results of altering 2, 4, and 6% of all trials and computed metacognitive scores based on bins of 50, 100, 200, and 400 trials. I split the Maniscalco dataset into 20 bins of 50 trials, 10 bins of 100 trials, five bins of 200 trials, and two bins of 400 trials (by taking into consideration only the first 800 trials in this last case). I split the 500 trials from each of the six days in the Haddara dataset into 10 bins of 50 trials, five bins of 100 trials, two bins of 200 trials, and one bin of 400 trials (by taking into consideration only the first 400 trials for the 200- and 400-trial bins). Across the six days, this process resulted in 60 bins of 50 trials, 30 bins of 100 trials, 12 bins of 200 trials, and six bins of 400 trials.

Assessing dependence on task performance

To assess how task performance affects measures of metacognition, I examined whether each measure of metacognition changed across different difficulty levels in the same experiment. Specifically, I tested whether each of the 17 measures of metacognition increases or decreases for more difficult conditions. If only two difficulty conditions were present in a given dataset, I performed a paired t-test. If multiple difficulty conditions were present in a given dataset, I performed a linear regression on the values of each measure and then ran a one-sample t-test against 0 on the resulting slopes. This approach allowed me to determine which measures significantly vary with difficulty level.

This process requires datasets with (1) several difficulty conditions and (2) a large number of trials. Consequently, I selected datasets from the Confidence Database that meet these two criteria but do not include any other manipulations. This resulted in the selection of four datasets: Adler (6 difficulty levels, 19 subjects, 1,916 trials/subject, 36,404 total trials), Shekhar (3 difficulty levels, 20 subjects, 2,800 trials/subject, 56,000 total trials), Rouault1 (70 difficulty levels, 466 subjects, 210 trials/subject, 97,860 total trials), and Rouault2 (many difficulty levels, 484 subjects, 210 trials/subject, 101,640 total trials). I analyzed the Adler and Shekhar datasets using the regression approach. Because the two Rouault datasets included very few trials from each difficulty level, I instead used a median split to classify them in easy vs. difficult, and used a t-test for the analysis. In both cases, I also assessed effect sizes by computing Cohen's d. Because the most difficult conditions in some datasets produced chance-level performance for some subjects, several measures of metacognition exhibited outlier values. Consequently, for each difficulty level and each measure of metacognition, I excluded any values that deviated by more than 3*SD from the mean of that difficulty level. Finally, as a reference, I performed all of the above analyses on the measures d', c, and average confidence.

Assessing dependence on metacognitive bias

To assess how metacognitive bias affects measures of metacognition, I applied the method developed by Xue et al. (2021). In this method, confidence ratings are recoded in two different ways as to artificially induce metacognitive bias towards lower or higher confidence ratings. Specifically, an n-point scale is transformed into an (n-1)-point scale in two ways. In the first recoding, the ratings from 2 to n are all decreased by one, resulting in a bias towards low confidence. In the second recoding, only the rating of n is decreased by one, resulting in a bias towards high confidence. A measure of metacognition can then be computed for the newly obtained confidence ratings. Comparing the obtained values for the two recodings allows the assessment of whether each measure of metacognition is independent of metacognitive bias.

This process would ideally be applied to datasets with (1) a single experimental condition and (2) a large number of trials. Consequently, I selected the same two datasets used to quantify precision: Haddara (3,000 trials per subject) and Maniscalco (1,000 trials per subject). In addition, I also used the Shekhar dataset (3 difficulty levels, 2,800 trials per subject) but analyzed each difficulty level in isolation and then averaged the results across the three difficulty levels. The values of each measure of metacognition for the two recodings were compared using a paired t-test.

Assessing dependence on response bias

To assess how response bias affects measures of metacognition, I compared the values of each measure of metacognition in conditions that differed in their decision criterion. To do so, I analyzed the Locke dataset – the only dataset in the Confidence Database where response criterion is experimentally manipulated. I computed each measure of metacognition for each of the seven conditions in that dataset and conducted repeated measures ANOVAs to examine whether each measure of metacognition varied with condition. In addition, I conducted a more sensitive analysis designed to check whether stronger response bias affects each measure of metacognition. For that analysis, I averaged the obtained values for conditions 2-7 (where the criterion was experimentally biased) and compared it to the obtained value in condition 1 (where the criterion was not experimentally biased) using paired t-tests.

Assessing split-half reliability

To assess split-half reliability, I examined the correlation between the values obtained for different measures of metacognition on odd vs. even trials (Guggenmos, 2021). As with assessing precision, I estimated split-half correlations for different sample sizes, so researchers can make informed decisions about the sample sizes needed in future studies. Specifically, I used bin sizes of 50, 100, 200, and 400 trials. Note that a bin size of k here means that $2k$ trials were examined with both the odd and even trials having a sample size of k. These computations are best performed using datasets with (1) a single condition, and (2) a large number of trials per subject. Consequently, I selected the same three datasets used to examine the dependence of measures of metacognition on metacognitive bias: Haddara (3,000 trials per subject), Maniscalco (1,000 trials per subject), and Shekhar (3 difficulty levels, 2,800 trials per subject). As before, I analyzed each difficulty level in the Shekhar dataset in isolation and then averaged the results across the three difficulty levels. For a bin size of k, the computations were performed on as many as possible non-overlapping bins of $2k$ trials. The obtained r-values were then z-transformed, averaged, and the resulting average z value was transformed back to an r-value for reporting and plotting purposes.

Assessing test-retest reliability

To assess test-retest reliability, I examined the correlation between the values obtained for different measures of metacognition on different days. As with split-half reliability, I estimated test-retest correlations for sample sizes of 50, 100, 200, and 400 trials. Because test-retest computations require data from multiple days and a large number of trials per subject per day, I selected the Haddara dataset as it is the only dataset in the Confidence Database to meet these criteria. I computed test-retest correlations between all pairs of days for as many as possible non-overlapping bins. Note that, unlike for split-half analyses, analyses of bin size of k involved the selection of k trials from each day. As with the split-half analyses, the obtained r-values were then z-transformed, averaged, and the resulting average z value was transformed back to an r-value for reporting and plotting purposes.

Data and code

Code for computing all 17 measures of metacognition, as well as data and analysis code for reproducing all statistical results and plotting all figures are available at https://osf.io/y5w2d/. This study was not preregistered.

Results

Here I assess the properties of 17 measures of metacognition. Specifically, I focus on each measure's (1) validity and precision, (2) dependence on nuisance variables, and (3) reliability. To examine each of these properties, I use seven existing datasets (Table 4) from the Confidence Database. For each property, I analyze the data from between one and four of these seven datasets. In addition, I compute precision and reliabilities using 50, 100, 200, or 400 trials at a time to clarify how these measures behave for different amounts of underlying data.

Validity and precision

Perhaps the most important requirement for any measure is that it is both valid and precise (Clark & Watson, 2019; Luck et al., 2021; Mueller & Knapp, 2018). In other words, a measure should reflect the quantity it purports to measure and it should do so with a high level of sensitivity. However, despite the importance of both of these criteria, there has been no method to assess the validity and precision of measures of metacognition.

Here I develop a simple method for assessing both of these properties. The method selects a small proportion of trials and decreases confidence by 1 point for each correct trial and increases confidence by 1 point for each incorrect trial. This manipulation artificially decreases the informativeness of confidence ratings. A valid measure of metacognition should therefore show a drop when applied to these altered data. The size of the drop relative to the normal fluctuations of the measure quantifies the precision of the measure (i.e., if the drop is large relative to background fluctuations, this indicates that the measure has a high level of precision).

To quantify the precision of existing measures of metacognition, one would ideally use a dataset with very large number of trials coming from a single experimental condition because mixing conditions can strongly impact metacognitive scores (Rahnev & Fleming, 2019). Consequently, I selected the two datasets from the Confidence Database with the largest number of trials per subject that also had a single experimental condition: Haddara (3,000 trials per subject) and Maniscalco (1,000 trials per subject). In each case, I examined the results of altering 2, 4, and 6% of all trials and computed metacognitive scores using bins of 50, 100, 200, and 400 trials.

The results showed that all 17 measures are valid in that metacognitive scores decreased when confidence ratings were artificially corrupted (Figure 1). The decrease in each measure is roughly a linear function of the percent of trials corrupted. For example, in the Haddara dataset, the values of *meta-d'* decreased from an average of 1.14 without any corruption to averages of .98, .84, and .72 when 2%, 4%, and 6% of trials were corrupted, respectively (for an average drop of about .14 for every 2% of trials corrupted). However, this drop is difficult to compare between measures because different measures are on different scales (e.g., *meta-d'* takes values between 0 and ∞, whereas *AUC2* takes values between .5 and 1).

Therefore, to obtain values that are easy to interpret and compare, one can normalize the average drop after corruption by the standard deviation (SD) of the observed values across different subsets of trials in the absence of any corruption. Because the SD value is larger for smaller bin sizes – reflecting the larger noisiness of each measure when few trials are used – the results show that larger bin sizes lead to greater precision of the measures (Figure 1A). Indeed, across the 17 measures, corrupting 2% of the trials led to an average decrease of .35, .50, .70, and 1.04 SDs in the measured metacognitive ability value for bins of 50, 100, 200, and 400 trials, respectively.

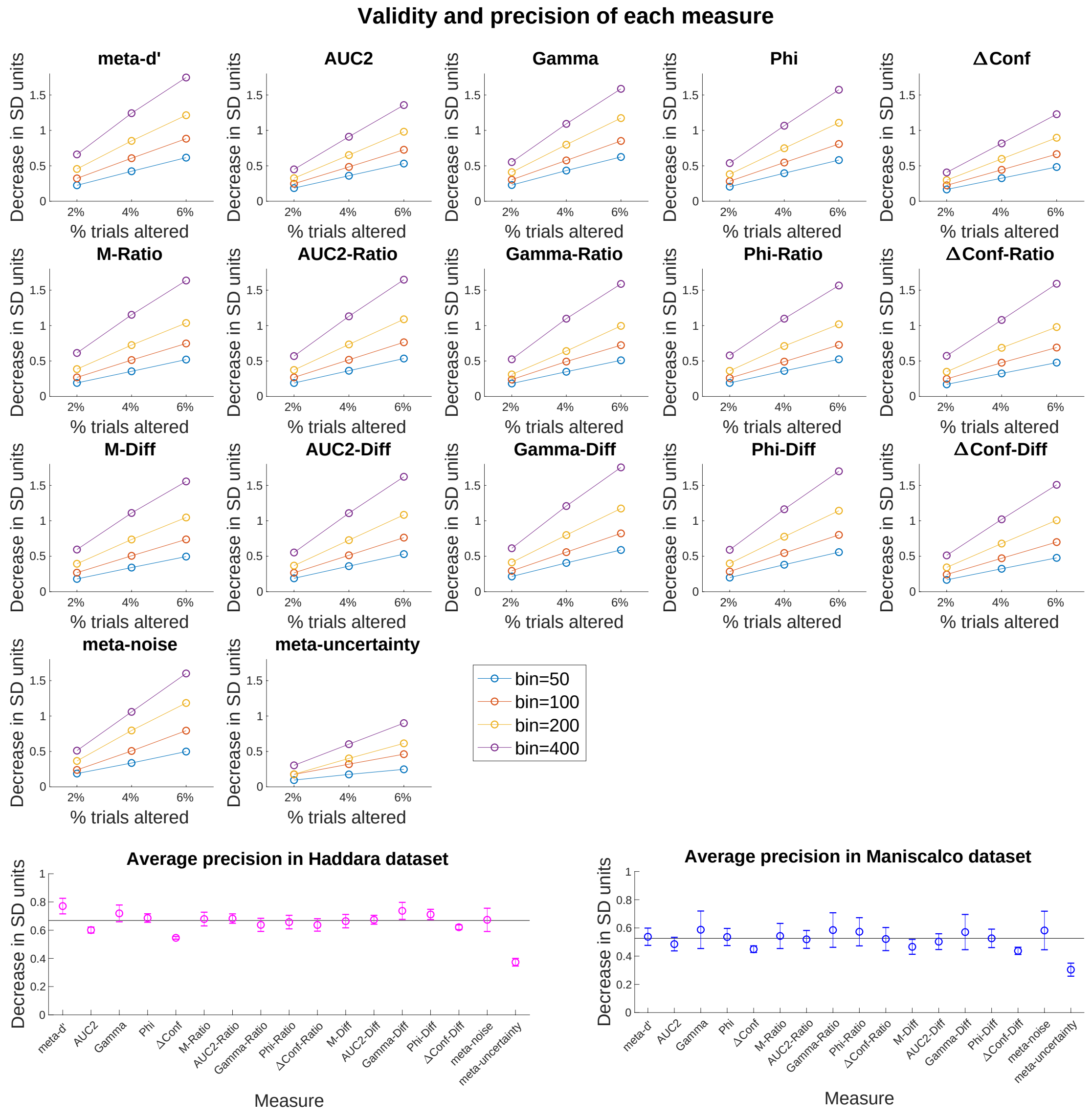


Figure 1. Validity and precision of each measure. Results of an artificial corruption of the confidence ratings where confidence for correct trials was decreased by 1, and confidence for incorrect trials was increased by 1. (A) Detailed results for the Haddara dataset. Each one of the 17 measures of metacognition showed a decrease with this manipulation. The plot shows the decrease in units of the standard deviation of the measure's fluctuations across different bins. The decrease was computed for bin sizes of 50, 100, 200, and 400 trials, as well as for 2, 4, and 6% of trials being corrupted. (B,C) Average precision in SD units for each measure in the Haddara and Maniscalco datasets averaged across the four bin sizes and the three levels of corruption. Error bars show SEM.

This technique allows us to compare the precision of different measures. To simplify the comparison, I averaged the decreases across the four different bin sizes and the three levels of corruption (2, 4, and 6%). The results revealed that most measures of metacognition had comparable levels of precision (Figure 1B,C). The one exception was *meta-uncertainty*, which had substantially lower average precision score in both the Haddara (meta-uncertainty: 0.37; average of other measures: 0.67) and the Maniscalco datasets (meta-uncertainty: 0.30; average of other measures: 0.53). Moreover, pairwise comparisons showed that the precision for *meta-uncertainty* was lower than every one of the other 16 measures in both datasets ($p < .05$ for all 32 comparisons). The differences between the remaining measures were much smaller and sometimes inconsistent across the two datasets. It should be noted that the precision scores were overall higher in the Haddara compared to the Maniscalco datasets. This difference is likely due to differences in variables such as sensitivity and metacognitive bias that are likely to vary across datasets. Therefore, the technique introduced here is useful for comparing between different measures but is unlikely to be useful if one wants to compare values across different datasets. Overall, these analyses suggest that all measures of metacognition investigated here are valid, and that most have comparable level of precision with the exception of *meta-uncertainty*, which appears to be noisier than the remaining measures.

Dependence on nuisance variables

Beyond validity and precision, another important feature for good measures of metacognition is that they should not be influenced by nuisance variables. Here I examine three nuisance variables – task performance, metacognitive bias, and response bias – and test how much each of these variable affects each of the 17 measures of metacognition.

Dependence on task performance

The most widely recognized nuisance variable for measures of metacognition is task performance (Fleming & Lau, 2014). The reason that task performance is a nuisance variable is that an ideal measure of metacognition should not be affected by whether a subject happened to be given an easier or a more difficult task. That is, the subject's estimated ability to provide informative confidence ratings should not change based on the difficulty of the object-level task that they are asked to perform.

To quantify how task performance affects measures of metacognition, one needs datasets with multiple difficulty conditions and a large number of trials (either because of including many subjects or many trials per subject). I selected the four datasets from the Confidence Database that best meet these characteristics: Adler (6 difficulty levels, 19 subjects, 1,916 trials/sub, 36,404 total trials), Shekhar (3 difficulty levels, 20 subjects, 2,800 trials/sub, 56,000 total trials), Rouault1 (70 difficulty levels, 466 subjects, 210 trials/sub, 97,860 total trials), and Rouault2 (many difficulty levels, 484 subjects, 210 trials/sub, 101,640 total trials). Both Rouault datasets have a large range of difficulty levels which I split into low/high by taking a median split. I then computed each measure separately for each difficulty level and compared them using regression slopes or t-tests.

The results showed that all traditional measures that are not normalized in any way (i.e., *meta-d'*, *AUC2*, *Gamma*, *Phi*, and *ΔConf*) are strongly dependent on task performance: they all substantially increase as the task becomes easier (p < .001 for all five measures and four datasets except the Adler dataset for *AUC2* and *Phi*; Figure 2). Further, the increase across the five measures from the most difficult to the easiest had a very large effect size (Cohen's d = 2.41, 2.24, 2.64, 1.53, and 1.80 for each of the five measures after averaging across the four datasets).

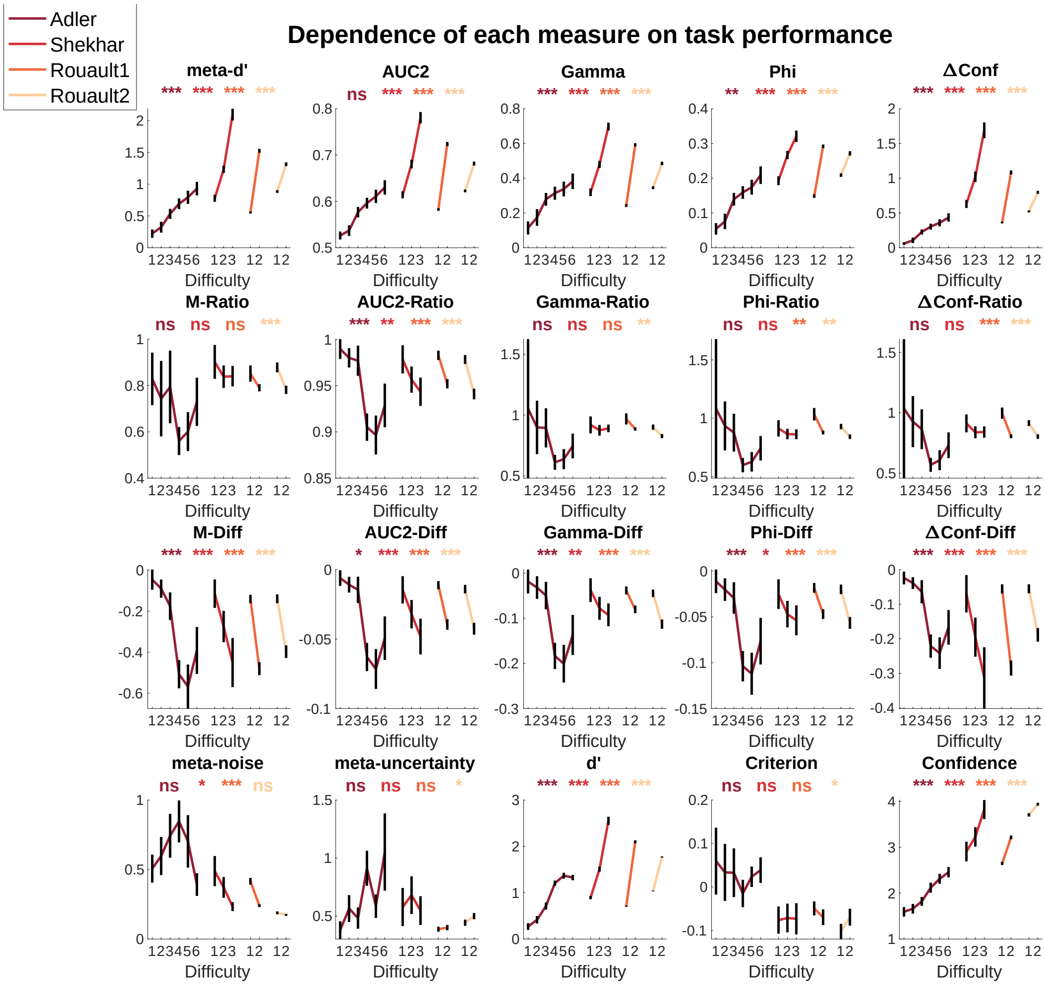


Figure 2. Dependence on task performance. Estimated metacognitive ability for all 17 measures, as well as d', criterion, and confidence for different difficulty levels in four datasets (Adler, Shekhar, Rouault1, and Rouault2). Traditional measures of metacognition (top row) all showed a strong positive relationship with task performance, whereas all Diff measures (third row) show a strong negative relationship. Ratio measures (second row) and the two model-based measures (meta-noise and meta-uncertainty) performed much better but still showed weak relationships with task performance. Note that higher numbers on the x axis indicate easier conditions. ***, p < .001; **, p < .01; *, p < .05; ns, not significant.

Having established that these five measures strongly depend on task performance, I then examined whether normalizing them removes this dependence. The more popular method of normalization – the ratio method – indeed performed well. The one exception was the measure *AUC2-Ratio*, which decreased significantly for easier conditions (*p* < .01 for all four datasets). The reason why *AUC2-Ratio* does not perform well is that *AUC2*, unlike the other four measures here, has a lower boundary of 0.5 rather than 0. None of the remaining ratio measures (*M-Ratio*, *Gamma-Ratio*, *Phi-Ratio*, and *ΔConf-Ratio*) varied significantly with difficulty level in the Adler and Shekhar datasets (all *p*'s > .05), though each of them showed a decrease for the Rouault2 dataset (all *p*'s < .01). *Phi-Ratio* and *ΔConf-Ratio* additionally showed a decrease for the Rouault1 dataset (both *p*'s < .01). Overall, all ratio measures showed a trend towards decreasing from the most difficult to the easiest condition after averaging across the four datasets. However, these effects were associated with very small Cohen's d effect sizes, except for *AUC2-Ratio* (*M-Ratio*: -0.08; *AUC2-Ratio*: -0.46; *Gamma-Ratio*: -0.11; *Phi-Ratio*: -0.16; *ΔConf-Ratio*: -0.20).

The five difference measures (*M-Diff*, *AUC2-Diff*, *Gamma-Diff*, *Phi-Diff*, and *ΔConf-Diff*) were all ineffective in removing the dependence on task performance. They all exhibited an over-correction where easier conditions led to lower scores across all five measures and four datasets (*p* < .001 for 17/20 tests; *p* < .05 for the remaining 3 tests). The Cohen's d effect sizes were medium-to-large (*M-Diff*: -0.61; *AUC2-Diff*: -0.54; *Gamma-Diff*: -0.45; *Phi-Diff*: -0.38; *ΔConf-Diff*: -0.58). These results demonstrate that the difference measures uniformly fail at their main purpose, which is to remove the dependence of metacognitive measures on task performance.

Finally, the two model-based measures (*meta-noise* and *meta-uncertainty*) showed relatively weak but still systematic relationships with task difficulty. Specifically, *meta-noise* tended to decrease for easier conditions (with the effect reaching significance for the Shekhar and Rouault1 datasets but not for Adler and Rouault2), whereas *meta-uncertainty* tended to increase for easier conditions (with the effect reaching significance for Rouault2 but not for the other three datasets). However, both of these effects were associated with small Cohen's d effect sizes that were comparable to what was observed for the ratio measures (*meta-noise*: -0.27; *meta-uncertainty*: 0.13). As such, both of the model-based measures perform as well as the ratio measures in controlling for task performance. Given that *meta-uncertainty* corrected in the opposite direction of the other viable measures (the four ratio measures and *meta-noise*), studies that feature task performance confounds may benefit from performing analyses using both *meta-uncertainty* and at least one more measure.

Dependence on metacognitive bias

A less appreciated nuisance variable is metacognitive bias: the tendency to give low or high confidence ratings for a given level of performance. Metacognitive bias can be measured simply as the average confidence in a condition. Recently, Shekhar & Rahnev (2021b) developed a method that involves recoding the original confidence ratings to examine how measures of metacognition depend on metacognitive bias. The method was further improved by Xue et al. (2021). The Xue et al. method consists of recoding confidence ratings as to artificially induce metacognitive bias towards lower or higher confidence ratings. Comparing the obtained values for a given measure of metacognition applied to the recoded confidence ratings allows us to evaluate whether the measure is independent of metacognitive bias.

Similar to quantifying precision, to quantify how metacognitive bias affects measures of metacognition, one would ideally use datasets with very large number of trials coming from a single experimental condition. Consequently, I selected the same two datasets used to quantify precision since they have the largest number of trials per subject while also featuring a single experimental condition: Haddara (3,000 trials per subject) and Maniscalco (1,000 trials per subject). In addition, I also used the Shekhar dataset (3 difficulty levels, 2,800 trials per subject) but analyzed each difficulty level in isolation and then averaged the results across the three difficulty levels. For that dataset, the continuous confidence scale was first binned into six levels as in the original publication (Shekhar & Rahnev, 2021b).

The results demonstrate that *meta-d'*, *AUC2*, *Phi*, and *ΔConf* tend to increase with higher average confidence, whereas *Gamma* tends to decrease (trend in the same direction for all three datasets and $p < .05$ for at least one of them). In other words, all five non-normalized measures of metacognition depend on metacognitive bias. Further, the average (across the three datasets) Cohen's d effect size was in the medium-to-large range for all five measures (*M-Ratio*: 0.49; *AUC2-Ratio*: 0.62; *Gamma-Ratio*: -0.64; *Phi-Ratio*: 0.95; *ΔConf-Ratio*: 0.74). All five ratio measures had a positive relationship with metacognitive bias but with smaller Cohen's d effect sizes (*M-Ratio*: 0.39; *AUC2-Ratio*: 0.08; *Gamma-Ratio*: 0.06; *Phi-Ratio*: 0.34; *ΔConf-Ratio*: 0.62). Difference measures performed similarly to ratio measures (*M-Diff*: 0.48; *AUC2-Diff*: 0.07; *Gamma-Diff*: 0.25; *Phi-Diff*: 0.10; *ΔConf-Diff*: 0.33). Finally, the two model-based measures also performed well with no statistically significant effects for any of the three datasets (all six *p*'s > .05) and with low-to-medium effect sizes that again went in opposite directions of each other (*meta-noise*: -0.30; *meta-uncertainty*: 0.34). Overall, researchers who want to control for metacognitive bias would appear to do best if they used *M-Ratio*, *AUC2-Ratio*, *Gamma-Ratio*, *Phi-Ratio*, *meta-noise*, or *meta-uncertainty*, with *AUC2-Ratio* and *Gamma-Ratio* performing the best. Given that *meta-noise* corrected in the opposite direction of the other five measures, it may be advisable for any result to be reproduced both *meta-noise* and at least one more measure from the list above.

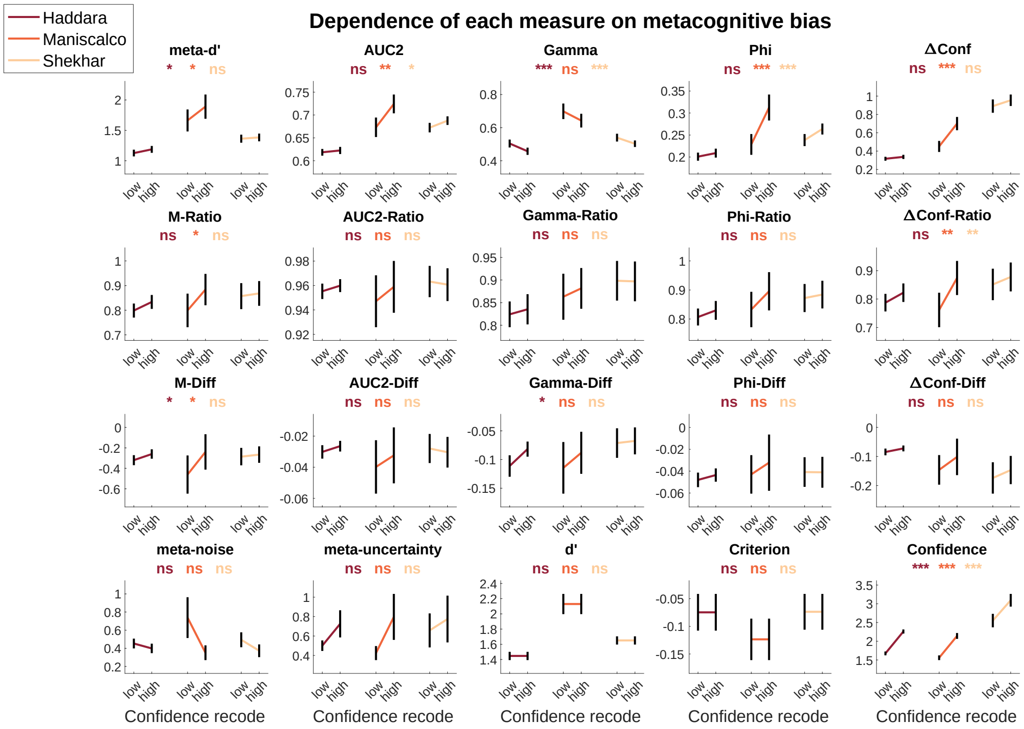


Figure 3. Dependence on metacognitive bias. Estimated metacognitive ability for all 17 measures, as well as d', criterion, and confidence for data recoded to have lower or higher confidence in three datasets (Haddara, Maniscalco, and Shekhar). Traditional measures of metacognition (top row) showed a medium-to-large positive relationship with metacognitive bias (except for *Gamma*, which showed a negative relationship). Ratio measures (second row) and the two model-based measures (meta-noise and meta-uncertainty) performed the best. ***, p < .001; **, p < .01; *, p < .05; ns, not significant.

<u>Dependence on response bias</u>

The final nuisance variable examined here is response bias. Response bias can be measured simply as the decision criterion *c* in signal detection theory. To understand how response bias affects measures of metacognition, one needs datasets where the response criterion is experimentally manipulated and confidence ratings are simultaneously collected. Very few such datasets exist and only a single such dataset is featured in the Confidence Database. The dataset –

named here Locke (Locke et al., 2020) – features seven conditions with manipulations of both prior and reward. Rewards were manipulated by changing the payoff for correctly choosing category 1 vs. category 2 (e.g., R = 4:2 means that 4 vs. 2 points were given for correctly identifying categories 1 and 2, respectively), whereas priors were manipulated by informing subjects about the probability of category 2 (e.g., P = .75 means that there was 75% probability of presenting category 2 and 25% probability of presenting category 1). The seven categories were as follows (1) P = .5, R = 3:3, (2) P = .75, R = 3:3, (3) P = .25, R = 3:3, (4) P = .5, R = 4:2, (5) P = .5, R = 2:4, (6) P = .75, R = 2:4, and (7) P = .25, R = 4:2. The Locke dataset included many trials per condition (700) but relatively few subjects (N = 10) and collected confidence on a 2-point scale.

The results suggested that none of the 17 measures of metacognition are strongly influenced by response bias (Figure 4). Indeed, while a repeated measures ANOVA revealed a very strong effect of condition on response criterion ($p = 1.3 \times 10^{-8}$), it showed no significant effect of condition on any of the measures of metacognition (all *p*'s > .13). Even a more sensitive analysis that compared the average value across all biased conditions (conditions 2-7) to the unbiased condition (condition 1) found no significant difference for any of the measures (all *p*'s > .087). While these results should be interpreted with caution given the small sample size and the fact that a 2-point confidence scale may be noisier for estimating metacognitive scores, they nonetheless suggest that response bias may not have a large biasing effect on measures of metacognition.

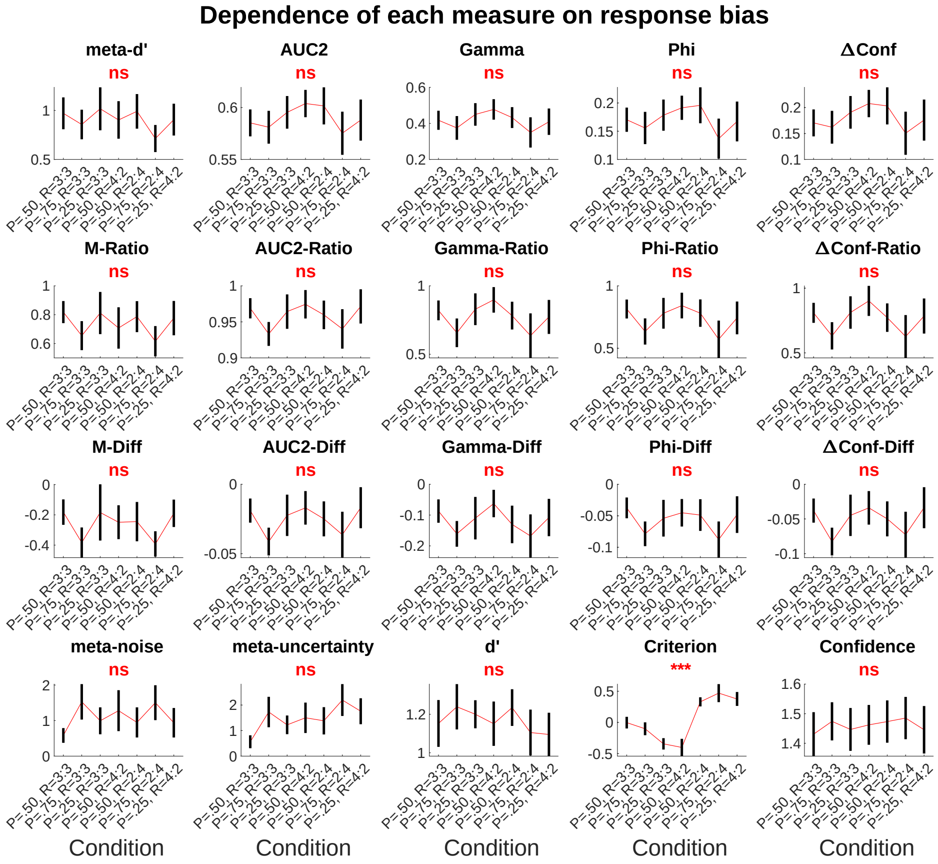


Figure 4. Dependence on response bias. Estimated metacognitive ability for all 17 measures, as well as d', criterion, and confidence for the seven conditions in the Locke dataset. While condition strongly modulated response bias, it did not significantly modulate any of the 17 measures of metacognition. ***, p < .001; ns, not significant.

Reliability

Measures of metacognition are often used in studies of individual differences to examine across-subject correlations between metacognitive ability and many different factors such as brain activity and structure (Allen et al., 2017; Fleming et al., 2010; Zheng et al., 2021), metacognitive ability in other domains (Faivre et al., 2018; Mazancieux et al., 2020), psychiatric symptom dimensions (Rouault et al., 2018), cognitive processes such as confidence leak (Rahnev et al., 2015), etc. These types of studies require that measures of metacognition have high reliability. (Note that within-subject studies of metacognition do not require high reliability – a measure that inherently depends on a large spread of scores across subjects – and instead requires high precision.)

Perhaps surprisingly, relatively little has been done to quantify the reliability of measures of metacognition (but see Guggenmos, 2021). Here I examine split-half reliability (correlation between estimates obtained from odd vs. even trials) and test-retest reliability (correlation between estimates obtained on different days).

Split-half reliability

To examine split-half reliability for different sample sizes, one needs datasets with a large number of trials per subject and a single condition (or large number of trials per condition if multiple conditions are present). Consequently, I selected the same three datasets used to examine the dependence of measures of metacognition on metacognitive bias: Haddara (3,000 trials per subject), Maniscalco (1,000 trials per subject), and Shekhar (3 difficulty levels, 2,800 trials per subject). As before, I analyzed each difficulty level in the Shekhar dataset in isolation and then averaged the results across the three difficulty levels. For each dataset, I computed each measure of metacognition based on odd and even trials separately and correlated the two. To examine how split-half reliability depends on sample size, I performed the procedure above for bins of 50, 100, 200, and 400 trials separately. Because the datasets contained multiple bins of each size, I averaged the results across all bins of a given size.

The results showed that measures of metacognition had good split-half reliability as long as the measures are computed using at least 100 trials (Figure 5). Indeed, bin sizes of 100 trials produced split-half correlations of r > .837 for all 17 measures when averaged across the three datasets with an average split-half correlation of r = .861. These numbers increased further for bin sizes of 200 (all r's > .938, average r = .946) and 400 trials (all r's > .961, average r = .965). Further, these numbers were only a little lower than the split-half correlations for d' (100 trials: r = .913; 200 trials: r = .958; 400 trials: r = .970). However, the split-half correlations strongly diminished when the measures of metacognition were computed based on 50 trials with an average r = .424 and no measure exceeding r = .6. It should be noted that while performing better, d' also had a relatively low split-half reliability of r = .685 when computed based on 50 trials. These results suggest that individual difference studies should employ 100 trials per subject at a minimum and that there is little benefit in terms of split-half reliability for using more than 200 trials.

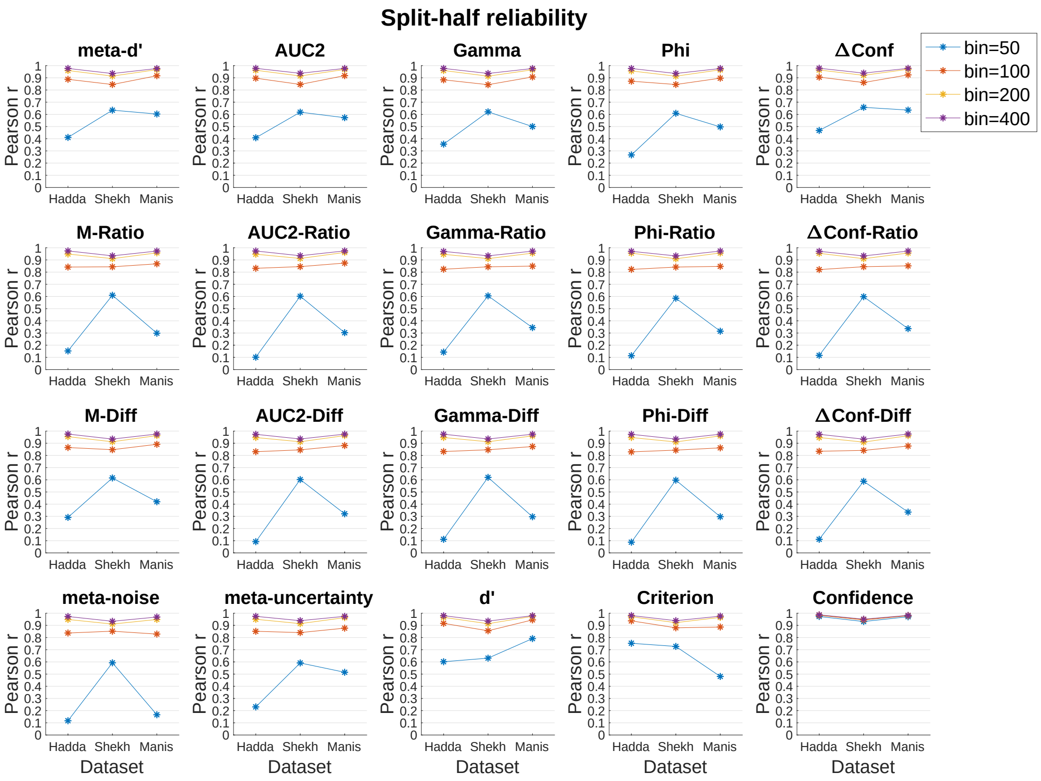


Figure 5. Split-half reliability. Correlations between each measure computed based on odd vs. even trials for sample sizes of 50, 100, 200, and 400 trials. The figure shows that split-half correlations are high when at least 100 trials are used for computations but become unacceptably low when only 50 trials are used. The x axis shows the results for three different datasets: Hadda (Haddara), Shekh (Shekhar), and Manis (Maniscalco).

Test-retest reliability

Split-half reliability is a useful measure of the intrinsic noise present in the across-subject correlations that can be expected in studies of individual differences. However, they do not account for fluctuations that could occur from day to day. These fluctuations can be examined by computing measures of metacognition obtained from different days, thus estimating what is known as test-retest reliability. Such estimation requires datasets with multiple days of testing and a large number of trials per subject per day. Only one dataset in the Confidence Database meets these criteria: Haddara (6 days; 3,000 total trials per subject; 70 subjects). I computed test-retest correlations between all pairs of days and then averaged across the different pairs.

The results showed very low test-retest reliability values (Figure 6). Even with 400 trials used for estimation, no measure of metacognition exceeded an average correlation of r = .8 and none of the measures outside of the five non-normalized and non-model-based measures (i.e., *meta-d'*, *AUC2*, *Gamma*, *Phi*, and *ΔConf*) reached correlation of r = .5. For example, the widely used measure M-Ratio had average test reliability of r = .168 (for 50 trials), .235 (for 100 trials), .293 (for 200 trials), and .422 (for 400 trials). The measure with highest test-retest correlation was *ΔConf* with r = .389 (for 50 trials), .538 (for 100 trials), .650 (for 200 trials), and .753 (for 400 trials). Notably test-retest correlations were not much higher for d' or the criterion c compared to *ΔConf* (average difference of about .1) and was only robustly high for confidence (above r = .86 regardless of sample size). These results suggest that correlations between measures of metacognition and measures that do not substantially fluctuate on a day-by-day basis (e.g., structural brain measures) are likely to be particularly noisy such that very large sample sizes may be needed to find reliable results.

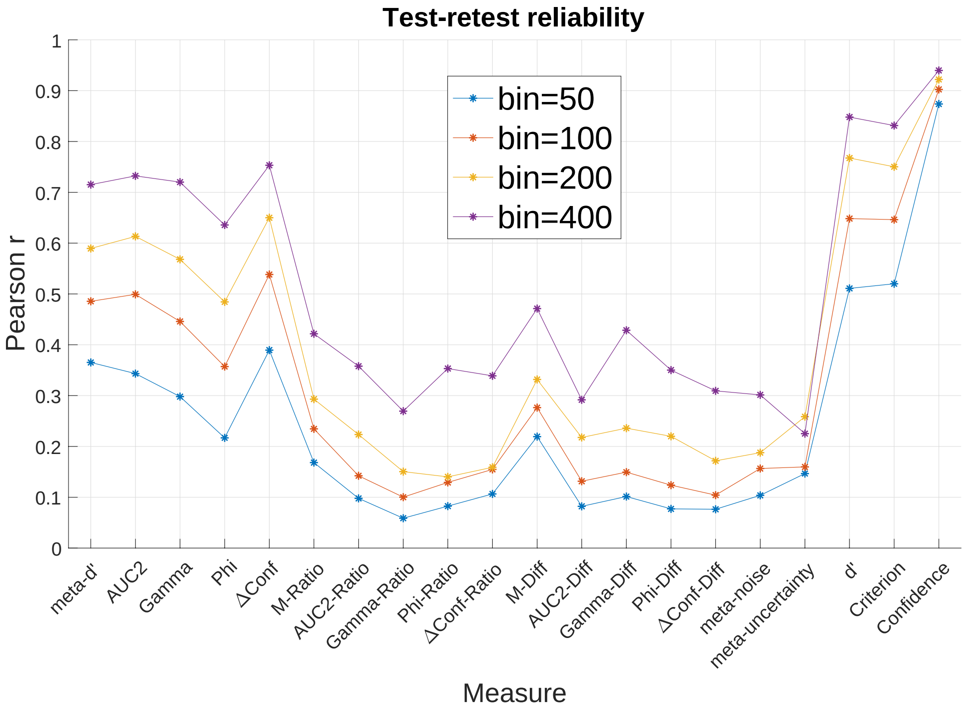
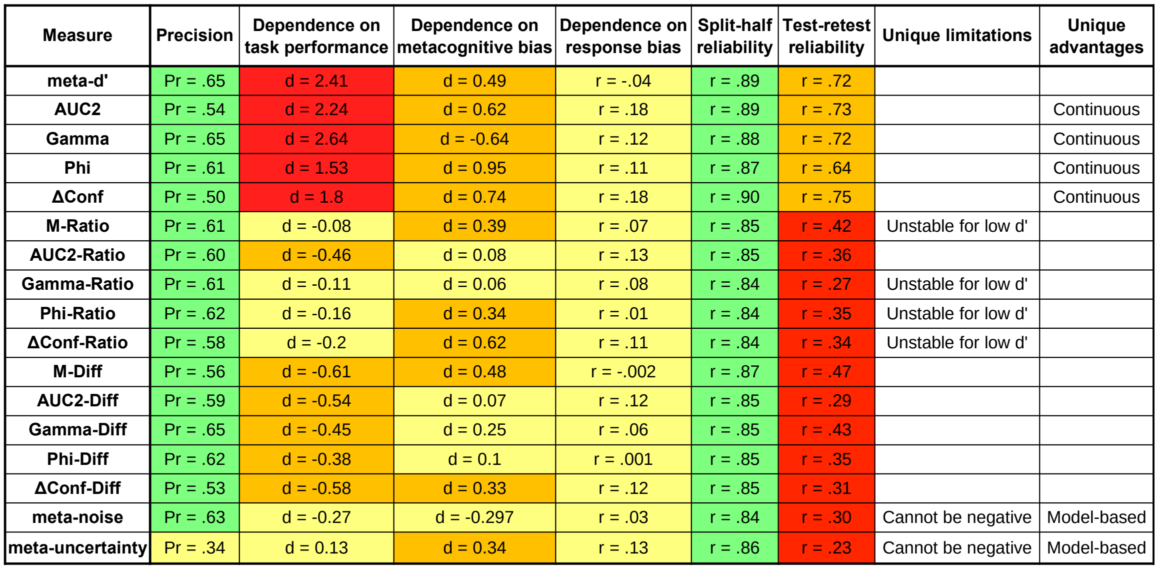


Figure 6. Test-retest reliability. Test-retest correlations in the Haddara dataset (6 days, 500 trials per day, 70 subjects) show generally low test-retest reliability. The correlations were low-to-moderate for the measures *meta-d'*, *AUC2*, *Gamma*, *Phi*, and *ΔConf* and very low for the remaining measures.

Discussion

Despite substantial interest in developing good measures of metacognition, there has been surprisingly little empirical work into the psychometric properties of current measures. Here I investigate the properties of 17 measures of metacognition, including eight new variants. I develop a method of determining the validity and precision of a measure of metacognition and also examine each measure's dependence on nuisance variables and its split-half and test-retest reliability. The results paint a complex picture. No measure of metacognition is "perfect" in the sense of having the best psychometric properties across all criteria. Researchers need to make informed decisions about which measures to use based on the empirical properties of the different measures. The results are summarized in Figure 7.



Measure	Precision	Dependence on task performance	Dependence on metacognitive bias	Dependence on response bias	Split-half reliability	Test-retest reliability	Unique limitations	Unique advantages
meta-d'	Pr = .65	d = 2.41	d = 0.49	r = -.04	r = .89	r = .72		
AUC2	Pr = .54	d = 2.24	d = 0.62	r = .18	r = .89	r = .73		Continuous
Gamma	Pr = .65	d = 2.64	d = -0.64	r = .12	r = .88	r = .72		Continuous
Phi	Pr = .61	d = 1.53	d = 0.95	r = .11	r = .87	r = .64		Continuous
ΔConf	Pr = .50	d = 1.8	d = 0.74	r = .18	r = .90	r = .75		Continuous
M-Ratio	Pr = .61	d = -0.08	d = 0.39	r = .07	r = .85	r = .42	Unstable for low d'	
AUC2-Ratio	Pr = .60	d = -0.46	d = 0.08	r = .13	r = .85	r = .36		
Gamma-Ratio	Pr = .61	d = -0.11	d = 0.06	r = .08	r = .84	r = .27	Unstable for low d'	
Phi-Ratio	Pr = .62	d = -0.16	d = 0.34	r = .01	r = .84	r = .35	Unstable for low d'	
ΔConf-Ratio	Pr = .58	d = -0.2	d = 0.62	r = .11	r = .84	r = .34	Unstable for low d'	
M-Diff	Pr = .56	d = -0.61	d = 0.48	r = -.002	r = .87	r = .47		
AUC2-Diff	Pr = .59	d = -0.54	d = 0.07	r = .12	r = .85	r = .29		
Gamma-Diff	Pr = .65	d = -0.45	d = 0.25	r = .06	r = .85	r = .43		
Phi-Diff	Pr = .62	d = -0.38	d = 0.1	r = .001	r = .85	r = .35		
ΔConf-Diff	Pr = .53	d = -0.58	d = 0.33	r = .12	r = .85	r = .31		
meta-noise	Pr = .63	d = -0.27	d = -0.297	r = .03	r = .84	r = .30	Cannot be negative	Model-based
meta-uncertainty	Pr = .34	d = 0.13	d = 0.34	r = .13	r = .86	r = .23	Cannot be negative	Model-based

Figure 7. Summary of results. The figure lists, the values obtained for each measure of metacognition for various criteria. Precision is the measure developed in this paper and the values listed are the average of the values in Figures 1B and 1C.

Higher precision values are better. For dependence of task performance and metacognitive bias, the figure lists the average Cohen's d values reported in the paper. For dependence on response bias, the figure lists the average correlation between each measure of metacognition and the absolute value of response bias ($|c|$). Lower absolute value of these dependencies is better. The reported split-half reliability is the average value across datasets obtained for a bin size of 100, whereas the reported test-retest reliability is the average value obtained for a bin size of 400. Higher reliability values are better. Color coding is meant as a general indicator but should be interpreted with caution. Green indicates very good properties, yellow indicates good properties, orange indicates problematic properties, and red indicates unacceptable properties. Colors were assigned based on the following thresholds: .5 for precision, .3 and 1 for Cohen's d, .5 for test-retest reliability. Green was not used in any of the columns regarding dependence on nuisance variables as to not give the impression that any measure is certainly independent of any of the nuisance variables. The figure also lists several unique advantages and disadvantages of each measure discussed in the main text.

Validity and precision

I found that all 17 measures of metacognition examined here are valid. With the exception of *meta-uncertainty*, all measures seem to have comparable level of precision. This result is rather surprising and suggests that precision may be limited by measurement error such that it is unlikely that any new measure of metacognition can substantially exceed the precision level found for the first 16 measures here. Nevertheless, new measures can be noisier and therefore it is critical to demonstrate their level of precision. Note that less precise measures can also appear to depend less on nuisance factors not because of their better psychometric properties but due to their noisiness.

Dependence on task performance

Task performance is arguably the most important and best appreciated nuisance variable for measures of metacognition. As has been previously suspected (Fleming

& Lau, 2014), the results here show that all traditional measures of metacognition are strongly dependent on task performance. However, the ratio method does a very good job of correcting for this dependence with *M-Ratio*, *Gamma-Ratio*, *Phi-Ratio*, and *ΔConf-Ratio* showing only very weak dependence on task performance. On the other hand, the difference method performed poorly in removing the dependence of task performance. The model-based measures *meta-noise* and *meta-uncertainty* also performed well.

Dependence on metacognitive bias

Previous research has shown that *meta-d'* and *M-Ratio* are positively correlated with metacognitive bias such that a bias towards higher confidence also leads to high values for these measures (Shekhar & Rahnev, 2021b; Xue et al., 2021). The current investigation replicated these previous results and showed that similar effects are observed for many other measures. Nevertheless, the dependence was of low to medium effect size for *M-Ratio* and comparable to newer measures such as *meta-noise* and *meta-uncertainty*.

Dependence on response bias

The results for response bias should be considered as preliminary because they are based on a single dataset that consists of 10 subjects. As such, the results should not be taken as strong evidence for an absence of dependence on response bias (hence, all measures are colored in yellow rather than green in Figure 7). Yet, it does appear that any dependencies are unlikely to be particularly strong, at least for a reasonable range of response bias strengths.

Split-half reliability

A recent paper examined many datasets in the Confidence Database and concluded that split-half reliability for *M-Ratio* is relatively poor (r ~ .7 for bin sizes between 400 and 600) (Guggenmos, 2021). (Note that the paper computes split-half reliability but it calls it test-retest reliability.) One issue with the approach by Guggenmos is that many of the analyzed datasets in the Confidence Database feature a variety of conditions, manipulations, and sample sizes. These factors may reduce the observed split-half reliability. Indeed, focusing on a select number of large datasets with a single condition at a time, the current paper finds much higher split-half reliabilities (between .84 and .9 for a bin size of 100). These results suggest that for sample sizes of 100 or more, one can expect reliable estimates of metacognition for every measure when using a single experimental condition. It is likely that studies that mix different conditions and estimate metacognitive scores across all of them would produce lower split-half reliability in line with the results of Guggenmos. Note that sample sizes of 50 produced unacceptably low reliabilities, so 100 should be considered as a rough lower boundary for the necessary number of trials when estimating metacognition in studies of individual differences.

Test-retest reliability

One of the most striking results here is the very low test-retest reliabilities observed. Besides the first five measures (*meta-d'*, *AUC2*, *Gamma*, *Phi*, and *ΔConf*), no other measure showed test-retest reliability exceeding r = .5 even for sample sizes of 400 trials. However, these five measures are strongly dependent on task performance, and thus their higher reliability may be partly (or wholly) due to the higher reliability of task performance itself (test-retest reliability of d' was .85 for a sample size of 400). Therefore, studies that match d' for all subjects may result in test-retest reliability values for these five measures of metacognition that are as low as the remaining measures. Nevertheless, these results are based on a single dataset and should therefore be replicated between very strong recommendations are made based on them. In the meantime, however, researchers who study individual differences in metacognition should be aware of the potential low test-retest reliability of measures of metacognition, which may explain previous failures to find significant correlations between metacognitive abilities across domains.

Unique advantages and disadvantages of different measures

Several measures feature unique advantages and disadvantages (Figure 7). For example, four of the Ratio measures (*M-Ratio*, *Gamma-Ratio*, *Phi-Ratio*, and *ΔConf-Ratio*) become unstable for difficult conditions because they include division by variables (d', expected Gamma, expected Phi, and expected ΔConf, respectively) that are very close to 0 in such conditions. These measures should therefore be used preferentially when performance levels are relatively high (e.g., one should aim for d' values above 1, which roughly corresponds to accuracy values above 69%).

An advantage of *AUC2*, *Gamma*, *Phi*, and *ΔConf* is that they all work well with continuous confidence scales. All other measures rely on SDT-based computations that necessitate that continuous scales are binned before analyses. Such binning may lead to loss of information but it is currently unclear how much signal-to-noise ratio may be lost by different binning methods.

The two model-based measures – *meta-noise* and *meta-uncertainty* – have unique advantages and disadvantages. Their main advantage is that all of their underlying assumptions are explicitly known. Conversely, other measures must necessarily include hidden assumptions that are difficult to reveal without linking them to a process model of metacognition (Rahnev, 2021). Another unique advantage of these measures is that they can in principle be applied much more flexibly. For example, when an experiment contains several conditions, other measures do not allow the estimation of a single measure of metacognition and simply ignoring the different conditions can lead to inflated scores (Rahnev & Fleming, 2019). Conversely, both *meta-noise* and *meta-uncertainty* allow different conditions to be modeled as part of their underlying process models and thus a single metacognitive score can be computed in a principled way across many conditions. That said, a possible disadvantage of both measures is that they can only take positive values and therefore cannot be used for situations where metacognition may contain more information than the decision itself.

Is *M-Ratio* still the gold standard for measuring metacognition?

In the last decade, *M-Ratio* has become the dominant measure of metacognition due to its assumed better psychometric properties (Fleming & Lau, 2014; Maniscalco & Lau, 2012, 2014). This status has naturally attracted greater scrutiny and many recent papers have criticized some of the properties of *M-Ratio* (Bang et al., 2019; Guggenmos, 2021; Rausch et al., 2023; Shekhar & Rahnev, 2021b; Xue et al., 2021). However, such criticisms are only meaningful in the context of how alternative measures perform on the same tests. The results here demonstrate that across all examined dimensions, there are no measures that clearly outperform *M-Ratio*. Three measures – *meta-noise*, *Gamma-Ratio*, and *Phi-Ratio* – showed very similar performance to *M-Ratio*, while all other measures appear inferior to *M-Ratio* in at least one critical dimension: they strongly depend on task performance (11 measures), have low precision (*meta-uncertainty*), or strong dependence on metacognitive bias (*ΔConf-Ratio*). The present author sees no strong argument in the present data to choose either *Gamma-Ratio* or *Phi-Ratio* over *M-Ratio*, especially given how established *M-Ratio* is contrary to *Gamma-Ratio* and *Phi-Ratio*. There are good arguments for using *meta-noise* in addition to *M-Ratio* as a way of controlling for metacognitive bias given that the two measures depend on metacognitive bias in opposite directions. Similarly, *meta-uncertainty* can also be used in addition to *M-Ratio* or *meta-noise* to control for task performance given that it depends on task performance in the opposite direction than the other two measures.

There are strong reasons for the field to eventually transition to model-based measures of metacognition (Rahnev, 2021) since model-based measures are uniquely positioned to properly capture the influence of metacognitive inefficiencies (Shekhar & Rahnev, 2021a). The measure *meta-noise* is especially promising given its good performance on the current tests and the fact that its associated model is currently the best fitting model of metacognition (Shekhar & Rahnev, 2022). That said, *meta-noise* is currently only implemented in Matlab (see codes associated with the current paper) and is more computationally intensive. Thus, although *meta-noise* or other model-based measures of metacognition should eventually supplant *M-Ratio*, for the time being it is hard to justify abandoning *M-Ratio* as the gold standard for the field.

Limitations

The present work has several limitations. First, despite the attempt to be comprehensive, several measures of metacognition have been omitted including recent model-based measures (Desender et al., 2022; Mamassian & de Gardelle, 2022), different variants of *M-Ratio* (Guggenmos, 2021), and legacy measures such as *Type-2 d'* (Azzopardi & Evans, 2007). Nevertheless, the current work should make it much easier for researchers to establish the properties of other measures of metacognition and compare them to the ones examined here. Second, while I have attempted to use multiple large datasets for each analysis, two of the analyses only included a single dataset (dependence on response bias and test-retest reliability) and should be interpreted with caution. Even in cases where multiple datasets were used, it is clear that adding more datasets would alter the values in Figure 7. As such, the values there should be understood as rough estimates that are bound to be improved upon by future work that analyzes additional large datasets.

Conclusion

The current work represents a critical step towards establishing the empirical properties of measures of metacognition. The results can help researchers make informed decisions when choosing how to measure metacognition. Overall, there are good arguments to continue to use *M-Ratio* as the gold standard in the field, but several confounds can be addressed by confirming the results using newer, model-based measures. The future of assessing metacognitive ability may lie with such model-based measures.

References

Adler, W. T., & Ma, W. J. (2018). Comparing Bayesian and non-Bayesian accounts of human confidence reports. *PLOS Computational Biology*, *14*(11), e1006572. https://doi.org/10.1371/journal.pcbi.1006572

Allen, M., Glen, J. C., Müllensiefen, D., Schwarzkopf, D. S., Fardo, F., Frank, D., Callaghan, M. F., & Rees, G. (2017). Metacognitive ability correlates with hippocampal and prefrontal microstructure. *NeuroImage*, *149*, 415–423. https://doi.org/10.1016/j.neuroimage.2017.02.008

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

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

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

Boundy-Singer, Z. M., Ziemba, C. M., & Goris, R. L. T. (2023). Confidence reflects a noisy decision reliability estimate. *Nature Human Behaviour*, *7*(1), 142–154. https://doi.org/10.1038/s41562-022-01464-x

Clark, L. A., & Watson, D. (2019). Constructing validity: New developments in creating objective measuring instruments. *Psychological Assessment*, *31*(12), 1412–1427. https://doi.org/10.1037/pas0000626

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

Desender, K., Boldt, A., & Yeung, N. (2018). Subjective Confidence Predicts Information Seeking in Decision Making. *Psychological Science*, *29*(5), 761–778. https://doi.org/10.1177/0956797617744771

Desender, K., Vermeylen, L., & Verguts, T. (2022). Dynamic influences on static measures of metacognition. *Nature Communications*, *13*(1), 4208. https://doi.org/10.1038/s41467-022-31727-0

Faivre, N., Filevich, E., Solovey, G., Kühn, S., & Blanke, O. (2018). Behavioral, Modeling, and Electrophysiological Evidence for Supramodality in Human Metacognition. *The Journal of Neuroscience*, *38*(2), 263–277. https://doi.org/10.1523/JNEUROSCI.0322-17.2017

Fleming, S. M. (2021). *Know Thyself: The Science of Self-Awareness*. Basic Books.

Fleming, S. M., & Lau, H. C. (2014). How to measure metacognition. *Frontiers in Human Neuroscience*, *8*. 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. https://doi.org/10.3758/BF03196546

Guggenmos, M. (2021). Measuring metacognitive performance: type 1 performance dependence and test-retest reliability. *Neuroscience of Consciousness*, *2021*(1). https://doi.org/10.1093/nc/niab040

Guggenmos, M., Wilbertz, G., Hebart, M. N., & Sterzer, P. (2016). Mesolimbic confidence signals guide perceptual learning in the absence of external feedback. *ELife*, *5*, e13388. https://doi.org/10.7554/eLife.13388

Haddara, N., & Rahnev, D. (2022). The Impact of Feedback on Perceptual Decision-Making and Metacognition: Reduction in Bias but No Change in Sensitivity. *Psychological Science*, *33*(2), 259–275. https://doi.org/10.1177/09567976211032887

Kornell, N., Son, L. K., & Terrace, H. S. (2007). Transfer of metacognitive skills and hint seeking in monkeys. *Psychological Science*, *18*(1), 64–71. https://doi.org/10.1111/j.1467-9280.2007.01850.x

Locke, S. M., Gaffin-Cahn, E., Hosseinizaveh, N., Mamassian, P., & Landy, M. S. (2020). Priors and payoffs in confidence judgments. *Attention, Perception, & Psychophysics*, *82*(6), 3158–3175. https://doi.org/10.3758/s13414-020-02018-x

Luck, S. J., Stewart, A. X., Simmons, A. M., & Rhemtulla, M. (2021). Standardized measurement error: A universal metric of data quality for averaged event-related potentials. *Psychophysiology*, *58*(6), e13793. https://doi.org/10.1111/psyp.13793

Mamassian, P., & de Gardelle, V. (2022). Modeling perceptual confidence and the confidence forced-choice paradigm. *Psychological Review*, *129*(5), 976–998.

https://doi.org/10.1037/rev0000312

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

Maniscalco, B., & Lau, H. (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 (Eds.), *The Cognitive Neuroscience of Metacognition* (pp. 25–66). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-45190-4_3

Maniscalco, B., McCurdy, L. Y., Odegaard, B., & Lau, H. (2017). Limited cognitive resources explain a tradeoff between perceptual and metacognitive vigilance. *The Journal of Neuroscience*, *37*(5), 2271–2213. https://doi.org/10.1523/JNEUROSCI.2271-13.2016

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

Metcalfe, J., & Shimamura, A. P. (1994). *Metacognition: Knowing about Knowing*. MIT Press.

Mueller, R. O., & Knapp, T. R. (2018). Reliability and Validity. In G. R. Hancock, L. M. Stapleton, & R. O. Mueller (Eds.), *The Reviewer's Guide to Quantitative Methods in the Social Sciences* (2nd editio, p. 5). Routledge.

Nelson, T. O. (1984). A comparison of current measures of the accuracy of feeling-of-knowing predictions. *Psychological Bulletin*, *95*(1), 109–133. http://www.ncbi.nlm.nih.gov/pubmed/6544431

Pescetelli, N., & Yeung, N. (2021). The role of decision confidence in advice-taking and trust formation. *Journal of Experimental Psychology: General*, *150*(3), 507–526. https://doi.org/10.1037/xge0000960

Rahnev, D. (2021). Visual metacognition: Measures, models, and neural correlates. *American Psychologist*, *76*(9), 1445–1453. https://doi.org/10.1037/amp0000937

Rahnev, D., & Denison, R. N. (2018). Suboptimality in Perceptual Decision Making. *Behavioral and Brain Sciences*, *41*(e223), 1–66. https://doi.org/10.1017/S0140525X18000936

Rahnev, D., Desender, K., Lee, A. L. F., Adler, W. T., Aguilar-Lleyda, D., Akdoğan, B., Arbuzova, P., Atlas, L. Y., Balcı, F., Bang, J. W., Bègue, 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*(3), 317–325. https://doi.org/10.1038/s41562-019-0813-1

Rahnev, D., & Fleming, S. M. (2019). How experimental procedures influence estimates of metacognitive ability. *Neuroscience of Consciousness*, *2019*(1), niz009. https://doi.org/10.1093/nc/niz009

Rahnev, D., Koizumi, A., McCurdy, L. Y., D'Esposito, M., & Lau, H. (2015). Confidence Leak in Perceptual Decision Making. *Psychological Science*, *26*(11), 1664–1680. https://doi.org/10.1177/0956797615595037

Rausch, M., Hellmann, S., & Zehetleitner, M. (2023). Measures of metacognitive efficiency across cognitive models of decision confidence. *PsyArXiv*. https://doi.org/10.31234/osf.io/kdz34

Rouault, M., Seow, T., Gillan, C. M., & Fleming, S. M. (2018). Psychiatric Symptom Dimensions Are Associated With Dissociable Shifts in Metacognition but Not Task Performance. *Biological Psychiatry*, *84*(6), 443–451. https://doi.org/10.1016/j.biopsych.2017.12.017

Shekhar, M., & Rahnev, D. (2021a). Sources of Metacognitive Inefficiency. *Trends in Cognitive Sciences*, *25*(1), 12–23. https://doi.org/10.1016/j.tics.2020.10.007

Shekhar, M., & Rahnev, D. (2021b). 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

Xue, K., Shekhar, M., & Rahnev, D. (2021). Examining the robustness of the relationship between metacognitive efficiency and metacognitive bias. *Consciousness and Cognition*, *95*, 103196. https://doi.org/10.1016/j.concog.2021.103196

Zheng, Y., Wang, D., Ye, Q., Zou, F., Li, Y., & Kwok, S. C. (2021). Diffusion property and functional connectivity of superior longitudinal fasciculus underpin human metacognition. *Neuropsychologia*, *156*, 107847. https://doi.org/10.1016/j.neuropsychologia.2021.107847