**Analyses with different inclusion criteria**

We have analyzed 5160 individual cases by using inclusion criteria different from those in the main manuscript. For the gaussian SDTs, we have included the cases for which both gaussian d’ and meta-d’ were estimated to be greater than 0.5. For the logistic SDTs, we have included the cases of logistic d’ and meta-d’ > 0.8. This is because gaussian d’ of 0.5 and logistic d’ of 0.8 approximately correspond to the type-1 accuracy of 0.6 under the optimal type-1 criterion setting.

Table S1 summarizes the results, where paired t tests showed that mean gaussian meta-d’ was significantly smaller than mean gaussian d’ \(t\) (3142) = -9.58, \(p < .001\), while mean logistic meta-d’ was not significantly different from mean logistic d’ \(t\) (3538) = 1.29, \(p < .198\).

|                      | \(d'\) | \(\text{meta-d}'\) | \(m\)-ratio | case converged | case analyzed |
|----------------------|--------|---------------------|-------------|----------------|---------------|
| Gaussian meta-SDT    | 1.500  | 1.398               | 0.977       | 4056/5160      | 3143/4056     |
| Logistic meta-SDT    | 2.510  | 2.537               | 1.061       | 4355/5160      | 3539/4355     |
| Gaussian type-1 SDT  | 1.436  | 1.436               | 1           | 4208/5160      | 3899/4208     |
| Logistic type-1 SDT  | 2.421  | 2.421               | 1           | 4442/5160      | 4131/4442     |

Performance measures were averaged across the cases where each model converged and satisfied the inclusion criteria. Logistic estimates are shown at approximately 1.81 times the scale than gaussian estimates.
Dependency of metacognitive accuracy on raw confidence criteria

As in the main manuscript, we have conducted analyses based on the 26927 binary reformatted data, for which the gaussian and logistic meta-SDTs converged and exhibited above-chance type-1 and type-2 performances. Here, instead of considering the normalized confidence criteria, we have evaluated the dependence of meta-d’ on raw values of estimated criteria. For each binary reformatted data, an estimate of confidence criterion was obtained respectively for S1 and S2 responses. As there is no unique way to integrate these two estimates, we have evaluated the criterion dependency respectively for each response class.

We conducted linear regression to explain meta-d’ values from raw criteria estimates and tested if the slope is significantly different from 0 (26927 data points are aggregated). The gaussian meta-d’ was estimated to be smaller for higher confidence criteria for the response class of S1 ($t = 28.98, p < .001$) and S2 ($t = 28.99, p < .001$). The logistic meta-d’ tended to be larger for higher criteria for S1 ($t = 19.92, p < .001$) and S2 ($t = 19.69, p < .001$). Namely, the analyses replicated the reversed criterion dependency between the models.
The present study is grounded in the analyses of the existing datasets, which are characterized by a number of different features. Motivated by previous literature (Rahnev & Fleming, 2019), we have explored the impact of the cognitive domain (perception, memory, other) and the presence of trial-by-trial varying difficulty (yes, no, na) on metacognitive performances (these features are summarized in Supplementary Material 1).

Table S2 reports model estimates sorted by these two factors, where “yes” means that an explicit trial-by-trial manipulation was made on stimulus difficulty (e.g., jittered contrast in visual discrimination), “no” means that stimulus difficulty was constant across trials, and “na” means that variability was rather unintentionally invited by the use of naturalistic stimuli (e.g., variable memorability for naturalistic face or scene stimuli).

It seems safe to say that there was little evidence for the effect of the difficulty manipulation at least in the present datasets (comparison between “perception no” and “perception yes”). Also, metacognition seems rather accurate in those cases that employed naturalistic stimuli, although it is not certain if this enhancement comes from varying trial difficulty or other properties equipped in naturalistic stimuli. Note that the results need to be interpreted with caution because the datasets differ in other features than these two dimensions.

### Table S2. Model estimates averaged across individual participants.

| perception | gaussian | | | | logistic | | | |
|------------|----------|------|-----|-------|----------|------|-----|-------|
|            | d'       | meta-d' | m-ratio | frequency | d'       | meta-d' | m-ratio | frequency |
| yes        | 1.282    | 1.084  | 0.872  | 1751     | 2.161    | 1.995  | 0.960  | 1947     |
| no         | 1.461    | 1.111  | 0.876  | 821      | 2.482    | 2.190  | 1.029  | 860      |
| na         | 1.031    | 0.942  | 1.016  | 163      | 1.654    | 1.782  | 1.182  | 182      |
| memory     | | | | | | | | |
| na         | 1.469    | 1.593  | 1.278  | 530      | 2.512    | 2.965  | 1.394  | 538      |
| other      | | | | | | | | |
| yes        | 1.217    | 0.973  | 0.858  | 348      | 2.023    | 1.730  | 0.922  | 384      |
| na         | 2.411    | 2.469  | 1.137  | 205      | 4.222    | 4.684  | 1.246  | 216      |

Performance measures were averaged across the cases where each model converged and showed above-chance type-1 and type-2 performances. Logistic estimates are shown at approximately 1.81 times the scale than gaussian estimates.
Analyses on “perception no” datasets

We have explored the datasets belonging to the “perception no” condition, which is free from the potential confounding effect of varying difficulty. We have identified 989 individual cases of the condition, to which we have fitted the four different SDT models (Table S3). For model comparison, we have examined the number of the cases for which each model declared the best AIC/BIC fit (Table S4). As was found in the main manuscript, the logistic models were generally favored over the gaussian counterparts, and the type-1 logistic SDT was revealed to be a clear victor.

Table S3. Summary of the model fits.

| Model               | d’     | meta-d’ | m-ratio | case converged | case above-chance |
|---------------------|--------|---------|---------|----------------|-------------------|
| Gaussian meta-SDT   | 1.461  | 1.111   | 0.876   | 914/989        | 821/914           |
| Logistic meta-SDT   | 2.482  | 2.190   | 1.029   | 947/989        | 860/947           |
| Gaussian type-1 SDT | 1.406  | 1.406   | 1       | 922/989        | 891/922           |
| Logistic type-1 SDT | 2.410  | 2.410   | 1       | 961/989        | 929/961           |

Performance measures were averaged across the cases where each model converged and showed above-chance type-1 and type-2 performances. Logistic estimates are shown at approximately 1.81 times the scale than gaussian estimates.

Table S4. Number of the cases for which each model showed the AIC/BIC fit.

| Model               | best AIC fits | best BIC fits |
|---------------------|---------------|---------------|
| Gaussian meta-SDT   | 140           | 61            |
| Logistic meta-SDT   | 214           | 105           |
| Gaussian type-1 SDT | 221           | 279           |
| Logistic type-1 SDT | 399           | 529           |
Additional subset analysis

We have further narrowed down the analysis and identified three datasets of the “perception no” condition, for which all the four models converged for every included participant (“Maniscalco, 2017, experiment 1”, “Maniscalco, 2017, experiment 2, contrast 3”, and “Massoni, unpub, study 1, difficulty 2”). These datasets include 90 individual cases in total, and summed AIC/BIC measures generally favored the logistic models over the gaussian counterparts (Table S5). All the four models showed above-chance type-1 and type-2 performances in 85/90 cases, for which a paired t test showed that mean gaussian d’ (95% CI = [1.491, 1.918]) was significantly larger than mean gaussian meta-d’ (95% CI = [1.185, 1.564]) (t (84) = 4.51, p < .001). However, we found no significant difference between mean logistic d’ (95% CI = [2.545, 3.332]) and mean logistic meta-d’ (95% CI = [2.304, 3.088]) (t (84) = 1.68, p = .097).

Table S5. Summed AIC and BIC difference relative to the best model.

| Model Type            | AIC Difference | BIC Difference |
|-----------------------|----------------|----------------|
| Gaussian meta-SDT     | 51             | 129            |
| Logistic meta-SDT     | 0              | 78             |
| Gaussian type-1 SDT   | 621            | 384            |
| Logistic type-1 SDT   | 237            | 0              |
Model fits to aggregated data

For each of the 105 targeted datasets, we have fitted the SDT models by aggregating data from individual participants (Table S6). Although not free from aggregation artifacts, fittings to aggregated data can supplement the individual analysis, allowing for better model convergence and fuller use of available data (e.g., Cohen, Sanborn, & Shiffrin, 2008; Kellen, Klauer, & Bröder, 2013).

The models converged and showed above-chance performances for all the 105 aggregated datasets, which enabled us to use summed information criteria for quantitative model comparisons. Summed AIC and BIC best favored the logistic meta-SDT model, which was closely followed by the gaussian meta-SDT model (Figure S1). Unlike the analyses on the individual data, the meta-SDT models were much favored over the type-1 SDT models. This should be partly because aggregated data offer greater statistical power to justify extra model complexity. It would also be possible that data aggregation across individuals, each of whose placement of confidence criteria should be somewhat idiosyncratic, artifactually impairs the diagnosticity for confidence rating to discriminate correct and incorrect type-1 decisions; this makes meta-d’ much smaller than d’ and the type-1 models suffered difficulty in explaining data under the constraint of meta-d’ = d’. In the comparison between the type-1 models, the logistic model is far ahead of the gaussian counterpart, presumably because it can naturally capture the zROC nonlinearity.

Table S6. Model fits to aggregated data.

|                      | d’      | meta-d’  | m-ratio | case converged | case above-chance |
|----------------------|---------|----------|---------|----------------|------------------|
| Gaussian meta-SDT    | 1.256   | 0.954    | 0.771   | 105/105        | 105/105          |
| Logistic meta-SDT    | 2.062   | 1.863    | 0.962   | 105/105        | 105/105          |
| Gaussian type-1 SDT  | 1.256   | 1.256    | 1       | 105/105        | 105/105          |
| Logistic type-1 SDT  | 2.062   | 2.062    | 1       | 105/105        | 105/105          |

Performance measures were averaged across all the 105 cases. Logistic estimates are shown at approximately 1.81 times the scale than gaussian estimates.

Figure S1. Summed AIC and BIC difference relative to the best model. The measures were aggregated across the 105 datasets.
Generalized gaussian SDT fits to aggregated data

The comparisons of the gaussian and logistic SDT frameworks indicate that the kurtosis of the underlying distribution has major impact on the estimation of metacognitive accuracy (kurtosis is 0 for the gaussian distribution while the logistic distribution has the kurtosis of 1.2). To verify this insight, we have employed generalized gaussian distributions, which allowed us to systematically modulate the kurtosis parameter (calculation was implemented by gnorm package on R).

We have fitted generalized gaussian SDT models to datasets aggregated across participants. We set the $\beta$ parameter of the distribution at 2, 1.576, and 1.34, which gave the kurtosis approximately of 0, 0.6, and 1.2; these $\beta$ values also gave the standard deviation approximately of 0.71, 0.83, and 0.96, which is not relevant in the present analysis. Table S7 summarizes the estimation results, representing the substantial impact of distribution kurtosis in the measurement of metacognitive efficiency.

Figures S2 and S3 display summed AIC and BIC measures relative to the best model. The best fits were achieved with the intermediate kurtosis for the meta-SDT models. The impact of kurtosis was more pronounced for the type-1 models, as seen in the considerable misfits under the kurtosis of 0.

As expected, the criterion dependency of metacognitive performances was modulated by distribution kurtosis (Figure S4). Under the kurtosis of 0, negative criterion dependency was found for meta-d' ($t = -3.83, p < .001$) and m-ratio ($t = -4.63, p < .001$). The kurtosis of 0.6 gave no significant criterion-dependency for meta-d' ($t = 1.04, p = .298$) and m-ratio ($t = 1.38, p = .168$). With the kurtosis of 1.2, positive criterion dependency was observed for meta-d' ($t = 5.19, p < .001$) and m-ratio ($t = 6.45, p < .001$).

Table S7. Model estimates averaged across 105 individual participants.

|           | d'  | meta-d' | m-ratio | case converged | case above-chance |
|-----------|-----|---------|---------|----------------|------------------|
| **meta-SDT** |     |         |         |                |                  |
| kurtosis = 0 | 0.888 | 0.674  | 0.771   | 105/105        | 105/105          |
| kurtosis = 0.6 | 0.959 | 0.871  | 0.940   | 105/105        | 105/105          |
| kurtosis = 1.2 | 1.042 | 1.051  | 1.077   | 105/105        | 105/105          |
| **type-1 SDT** |     |         |         |                |                  |
| kurtosis = 0 | 0.888 | 0.888  | 1       | 105/105        | 105/105          |
| kurtosis = 0.6 | 0.959 | 0.959  | 1       | 105/105        | 105/105          |
| kurtosis = 1.2 | 1.042 | 1.042  | 1       | 105/105        | 105/105          |

Performance measures were averaged across the 105 cases.
Figure S2. Summed AIC difference relative to the best model.

Figure S3. Summed BIC difference relative to the best model.
Figure S4. Metacognitive performances evaluated at different confidence criteria.
**Analysis on non-2AFC datasets**

We have analyzed left/right dot motion discrimination data (Gherman, 2018) and left/right gabor orientation discrimination data (Mazor, 2020), which are not incorporated with trial-by-trial varying difficulty. The results are consistent with our main findings made on the 2AFC data in that the logistic meta-SDT gave a greater averaged m-ratio with slightly better model convergence than the gaussian meta-SDT.

**Table S8.** Estimates averaged across individual participants.

|                | gaussian d’ | gaussian meta-d’ | logistic d’ | logistic meta-d’ | case converged | case above-chance |
|----------------|-------------|------------------|-------------|------------------|----------------|------------------|
| Gherman, 2018  | 1.52        | 1.35             | 2.55        | 2.60             | 22/24 (gaussian) | 22/22 (gaussian) |
|                |             |                  |             |                  | 22/24 (logistic)| 22/22 (logistic)|
| Mazor, 2020,   | 1.47        | 1.40             | 2.36        | 2.61             | 43/46 (gaussian) | 42/43 (gaussian) |
| discrimination |             |                  |             |                  | 46/46 (logistic)| 45/46 (logistic)|

Performance measures were averaged across the cases where each model converged and showed above-chance type-1 and type-2 performances.
References

Cohen, A. L., Sanborn, A. N., & Shiffrin, R. M. (2008). Model evaluation using grouped or individual data. *Psychonomic Bulletin & Review, 15*, 692–712.

Kellen, D., Klauer, K. C., & Bröder, A. (2013). Recognition memory models and binary-response ROCs: A comparison by minimum description length. *Psychonomic Bulletin & Review, 20*, 693–719.