**Supplemental Material**

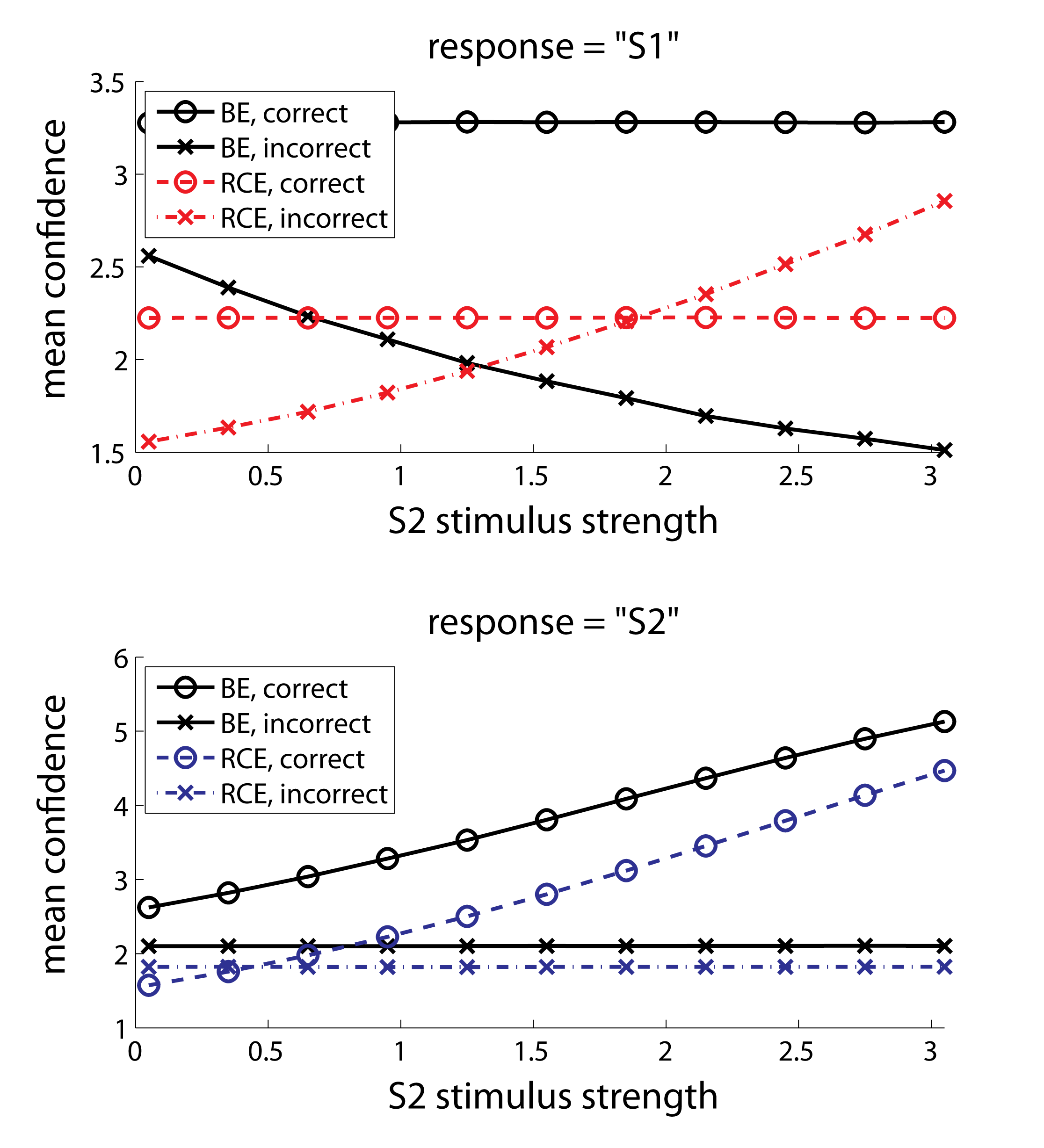
*Model Simulation Details*

We performed simulations of task performance and metacognition in the context of two-dimensional SDT models of the Balance of Evidence and Response-Congruent Evidence rules (Fig. 3). Following the logic of Fig. 4, we varied the mean of eS2 values while holding the S1 distribution constant.  For all simulations, for the S2 distribution, mean eS2 ranged from 0.05 to 3.05 in steps of 0.3 while mean eS1 was set to 0. For the S1 distribution, mean eS1 was set to 0.95 and mean eS2 was set to 0.  The covariance matrix for S1 and S2 distributions was set to the 2 x 2 identity matrix (entailing a uniform standard deviation of 1 with no correlation between eS1 and eS2).  For each level of S2 strength, 100,000 samples of (eS1, eS2) were drawn from each of the S1 and S2 distributions.  Five evenly-spaced confidence criteria were used to rate confidence for each evidence sample.  Note that dividing the evidence space into N confidence regions requires setting 2N + 1 decision criteria: one for the stimulus judgment, and N-1 to divide the confidence space for each of the “S1” or “S2” responses into N possible ratings.  See main text for simulation results.

The above simulations were also conducted using values for confidence criteria that were adjusted for S2 stimulus strength, rather than remaining constant across all conditions. Likewise, simulations were repeated while adjusting the covariance matrix to create a correlation between eS1 and eS2 values of 0.3 (moderate) and 0.7 (strong). In all cases, results similar to the original simulation were obtained. In particular, every permutation of the simulations exhibited the distinct linear cross-over effect of the response-specific meta-d’ curves shown in Figure 4A. Thus, the predicted dissociation between task performance and response-specific metacognition under the Response-Congruent Evidence rule appears to be quite robust against changes in the parameters of the model.

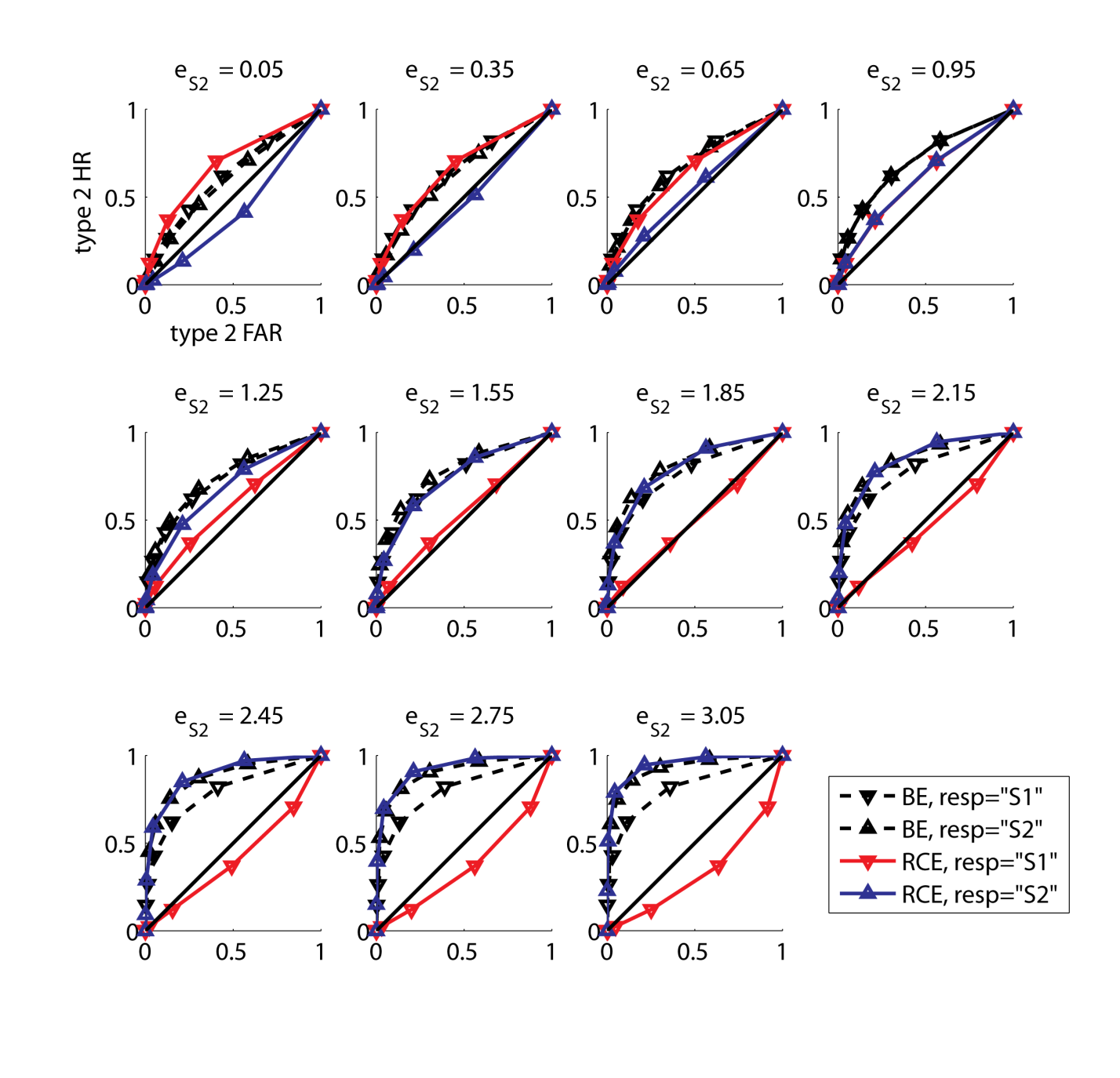
*Discussion of Model Visualizations*

We first examined predictions for average confidence for correct and incorrect responses (Figure S1). For the Response-Congruent Evidence decision rule, average confidence for incorrect “S1” responses increases even as average confidence for correct “S1” responses remains constant, for the reasons explained in Fig. 4. This pattern is not evident for “S2” responses, or for the Balance of Evidence rule.



**Figure S1.** Simulated values for response-specific mean confidence as a function of S2 stimulus strength. For trials on which the observer responds “S1” (top panel), the Balance of Evidence rule (BE) and Response-Congruent Evidence rule (RCE) make opposite predictions about the relationship between mean confidence for incorrect responses and S2 stimulus strength. The RCE decision rule uniquely predicts that as S2 becomes stronger, and therefore as task performance increases, the mean confidence for incorrect “S1” responses should increase even as confidence for correct “S1” responses remains constant. By contrast, the BE decision rule predicts that confidence for incorrect “S1” responses should decrease with increasing task performance.

However, one concern in the simpler approach first adopted above is that average confidence depends on the values used for the confidence criteria, and whether confidence criteria are fixed across all conditions.  So, in addition to Figure S1, which assumes fixed criteria across conditions, in Figure S2 we plot type 2 ROC curves (type 2 hit rate – p(high confidence | correct discrimination) – vs. type 2 false alarm rate – p(high confidence | incorrect discrimination)), which do not depend on the exact placement of confidence criteria. The area under the type 2 ROC curve (AUC) provides a metric of metacognitive performance, varying from 0.5 (chance performance) to 1 (perfect performance).  Chance-level metacognitive sensitivity is denoted by the diagonal line type 2 HR = type 2 FAR, i.e. the region of the ROC plot where high confidence responses are equally likely regardless of correct versus incorrect stimulus discrimination.  This corresponds to an AUC = .5.  In contrast, if an observer rates high confidence only for correct responses, and low confidence for incorrect ones, she will show 100% type 2 HR at 0% type 2 FAR, meaning that AUC = 1.  As S2 stimulus strength increases, AUC for “S1” responses progressively decreases under the Response-Congruent Evidence decision rule, whereas AUC for “S2” responses increases.



**Figure S2.** Simulated values for response-specific type 2 ROC curves as a function of S2 stimulus strength. As S2 stimulus strength increases, the Balance of Evidence rule (BE) predicts that area under the type 2 ROC curve (a measure of metacognitive performance) should increase for both “S1” and “S2” responses. By contrast, the Response-Congruent Evidence rule (RCE) predicts that area under the type 2 ROC curve for “S2” responses should increase whereas the area for “S1” responses should decrease. This demonstrates that the idiosyncratic patterns of confidence rating predicted by the RCE model are not attributable to response bias, but rather manifest as differences in metacognitive sensitivity.

A limitation of this approach is that AUC for the type 2 ROC depends on d’ (Galvin, Podd, Drga, & Whitmore, 2003; Maniscalco & Lau, 2012), and so the exact relationship between task performance and metacognition remains somewhat obscure.  So, in Figs. 4 – 7 (main text) we rely on plotting meta-d’ as a function of d’.  Meta-d’ measures metacognitive sensitivity such that, if confidence ratings follow their expected patterns under SDT, meta-d’ = d’ (Maniscalco & Lau, 2012). This visualization confirms the above analyses: meta-d’ for “S1” and “S2” responses exhibit a distinct linear cross-over pattern under the Response-Congruent Evidence rule. For most values of d’, meta-d’ for “S1” and “S2” responses also underperforms SDT expectation under the Response-Congruent Evidence rule. Conversely, under the Balance of Evidence rule, meta-d’ tracks d’ faithfully, in accordance with SDT expectation.  These predicted behavioral patterns held despite variations on simulation implementation.

*Diffusion model simulation*

Is it possible that the dissociation between task performance and metacognition observed in Experiment 1 (Fig. 4B) could be accounted for by a simple drift diffusion model (Ratcliff, 1978) that posits that confidence ratings are inversely related to reaction time on the discrimination task? Here we consider this possibility. Note that we are not interested here in evaluating the ability of the diffusion model to capture other aspects of the data besides the performance/metacognition dissociation, and also note that we are only aiming to evaluate a very basic kind of diffusion model that posits that confidence ratings and reaction times are inversely related. Thus, we make no claims about how well the diffusion model captures more general features of this data set aside from the dissociation, and we also make no claims about diffusion models (or other models of the dynamics of perceptual decision making) that allow for confidence to vary independently from reaction time.

The Response-Congruent Evidence rule predicts a dissociation between task performance and metacognitive sensitivity because, as S2 stimulus strength increases, task performance (d’) increases even as confidence for errors on S2 trials increases (Figs. 2 – 4, S1). Yet, as S2 stimulus strength increases and S2 stimuli become easier to distinguish from S1 stimuli, we might expect that reaction time for S2 stimuli also decreases—a commonly observed inverse relationship between task performance and reaction time that is captured by drift diffusion models (Pleskac & Busemeyer, 2010). In turn, if subjects base confidence on reaction time, then we might expect that a simple drift diffusion model would be sufficient to account for the empirically observed dissociation between task performance and metacognition (Fig. 4B).

In Fig. S3, we plot mean confidence as a function of response type, accuracy, and S2 stimulus strength for Experiments 1 and 2. In the data for Experiment 1 (Fig. S3A), note that mean confidence for incorrect “S1” responses increases for the two highest levels of S2 stimulus strength. This pattern is congruent with the pattern of mean confidence predicted by the Response-Congruent Evidence model (Fig. S1) and accounts for why meta-d’ for “S1” responses decreases at the two highest levels of S2 stimulus strength (Fig. 4B)—average confidence for incorrect “S1” responses increases even as average confidence for correct “S1” responses remains the same, leading to a reduction in area under the type 2 ROC curve for “S1” responses (Fig. S2) and thus a decrease in meta-d’ (Fig. 4). By contrast, an increase in average confidence for incorrect “S1” responses with increasing S2 stimulus strength is not evident in Experiment 2 (Fig. S3B), corresponding to the absence of the meta-d’ dissociation in the average taken across all four subjects (Fig. 7).

Thus, if the drift diffusion model is to explain the observed meta-d’ dissociation, it must do so by positing that

1. reaction time decreases with increasing S2 stimulus strength for incorrect “S1” responses, and
2. confidence is inversely related to reaction time.

To test the drift diffusion model account, we conducted a simple simulation. On each trial of the simulation, two accumulators of evidence in favor of responding “S1” and “S2” were defined as

where xS1(t) and xS2(t) are the evidence in favor of responding “S1” or “S2” at time t, dxS1 and dxS2 are the drift rates for S1 and S2, and ηS1(t) and ηS2(t) are independent noise terms drawn randomly at each time step t from a normal distribution with mean zero. We set xS1(0) = xS2(0) = 0 and dxS1 > 0 if the stimulus on the simulated trial was S1 and dxS1 = 0 otherwise (and similarly for dxS2). At each time step we evaluated the following decision rule:

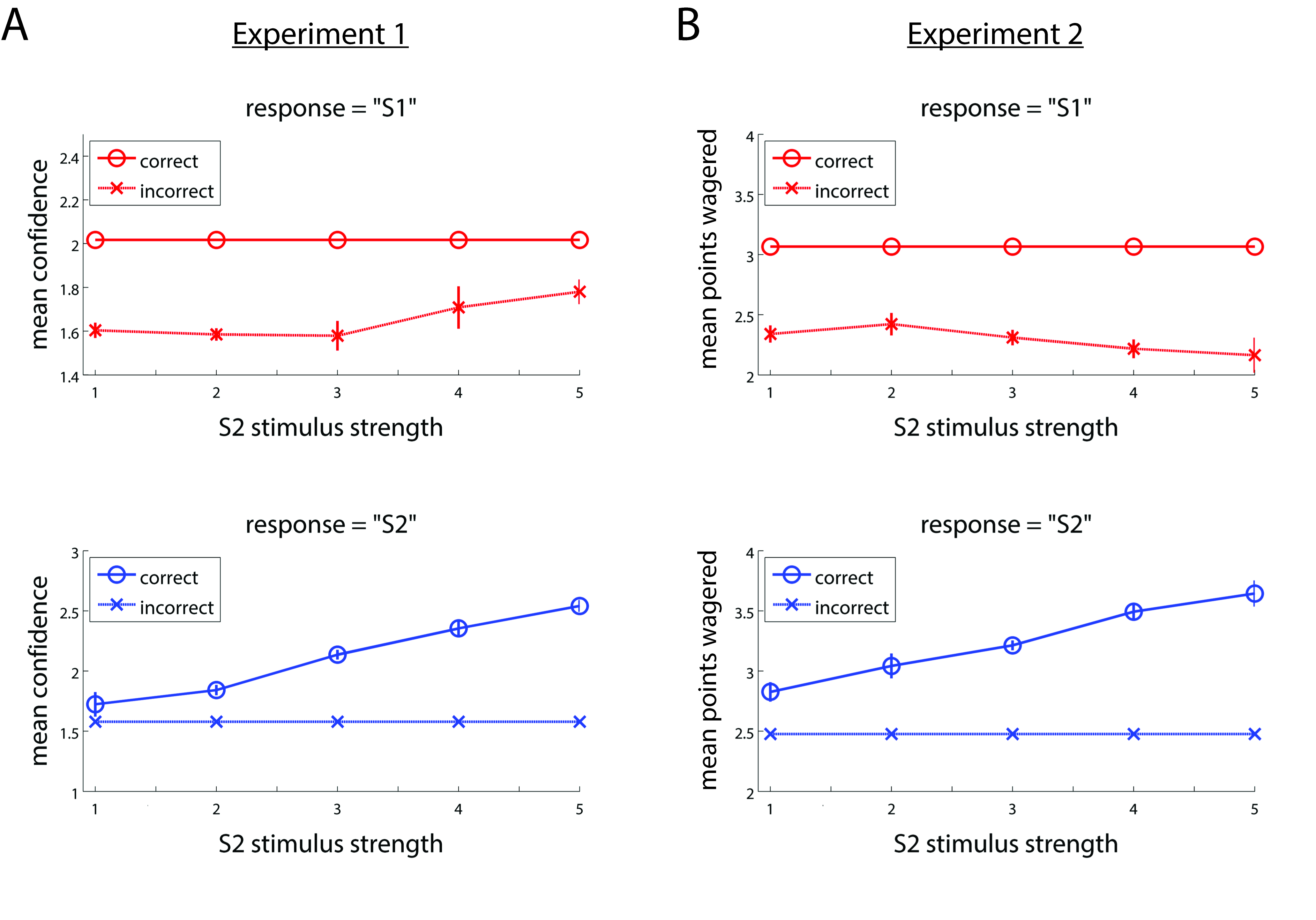
where ±b represents the values of the decision boundary, such that a response is elicited at time t if the difference in accumulated evidence for S1 and S2 exceeds one of the decision bounds.

In our simulation we used one level of S1 stimulus strength and seven levels of S2 stimulus strength, and simulated 5000 trials for each level of S1 and S2 stimulus strength. We selected parameter values for the model such that when dxS1 = dxS2 then d’ ≈ 1.5, and the minimum and maximum values of dxS2 yielded d’ values of approximately 1 and 2.5, respectively. Thus the range of d’ values produced by the diffusion model across different levels of S2 stimulus strength closely matched those probed in the Response-Congruent Evidence rule simulation (Fig. 4A) and in the empirical data (Fig. 4B). We measured reaction time on each trial as the number of time steps required to reach a decision bound, and summarized reaction time across trials by taking the median value.

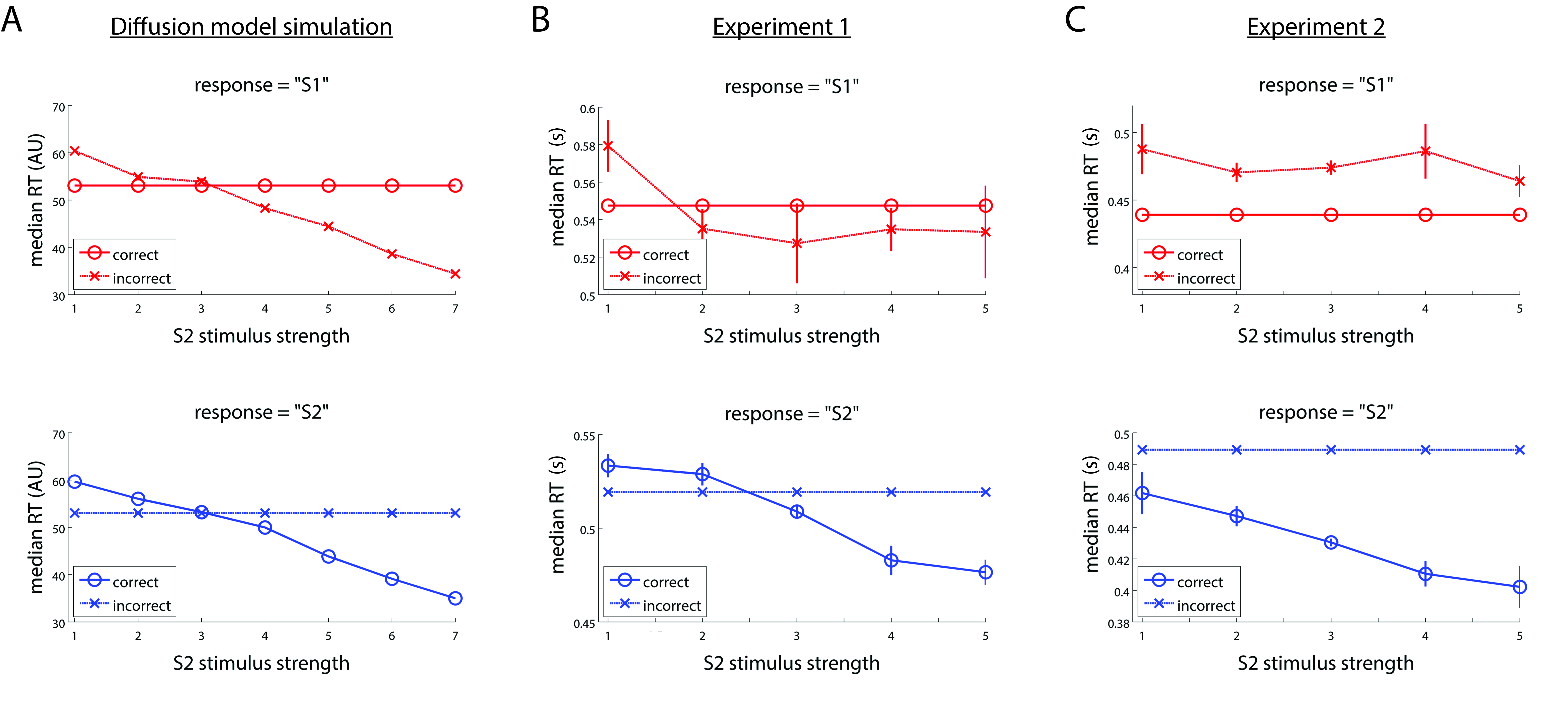
In Fig. S4A we plot the median reaction time as a function of response type, accuracy, and S2 stimulus strength for the diffusion simulation, and in Fig. S4B-C we plot the empirical median reaction time data from Experiments 1 and 2. Fig. S4A demonstrates that the diffusion model simulation yielded decreasing reaction time for incorrect “S1” responses with increasing S2 stimulus strength. Thus, the diffusion model satisfies requirement (1) listed above. (We verified that the patterns exhibited in Fig. S4A were preserved even when changing the evidence accumulation starting point by setting xS1(0) = 0.25, 0.5, or 0.75 while keeping xS2(0) = 0, and setting xS2(0) = 0.25, 0.5, or 0.75 while keeping xS1(0) = 0. We also observed similar patterns when adding time steps due to non-decision related processes drawn randomly from a normal distribution on each trial.)

However, in the empirical reaction time data for Experiment 1, it can be seen that reaction time for incorrect “S1” responses is constant, rather than decreasing, across the four highest levels of S2 stimulus strength (Fig. S4B). In particular, although mean confidence for incorrect “S1” responses increases for the highest levels of S2 stimulus strength (Fig. S3A), which drives the reduction in meta-d’ for “S1” responses at these same levels of S2 stimulus strength (Fig. 4B), a concomitant decrease in reaction time for incorrect “S1” responses is not observed in the empirical data (Fig. S4B). Thus, the diffusion model account of the metacognitive dissociation fails criterion (2) listed above, since the empirical data do not support the claim that confidence for incorrect “S1” responses is driven by reaction time for incorrect “S1” responses.

In the preceding, we have only intended to assess a very simple sort of diffusion model that posits that confidence and reaction time must be inversely related. More sophisticated models that allow for confidence to vary independently of reaction time could potentially account for the present findings. However, an extended consideration of more complex families of diffusion models is beyond the scope of the current work, and so we leave such investigations to future work.



**Figure S3.** Average confidence in Experiment 1, and average points wagered in Experiment 2, as a function of response type, accuracy, and S2 stimulus strength. Compare panel A with Fig. S1. Confidence for incorrect “S1” responses increases with increasing S2 stimulus strength even as confidence for correct “S1” responses remains constant. This has the effect of driving down area under the type 2 ROC curve (Fig. S2) and meta-d’ (Fig. 4) for “S1” responses even as d’ increases due to the increase in stimulus strength.



**Figure S4.** Median reaction time as a function of response type, accuracy, and S2 stimulus strength for (A) drift diffusion model simulation, (B) Experiment 1, (C) Experiment 2. The drift diffusion model predicts that RT for incorrect “S1” responses should decrease with increasing S2 stimulus strength. Thus, if confidence and RT are inversely related, then the diffusion model would predict a pattern like that observed in Fig. S1 and could therefore provide a plausible alternative account for the dissociation observed in Fig. 4. However, empirical RT in Experiment 1 does not decrease at the highest levels of S2 stimulus strength. Thus, the empirical RT does not exhibit the relationship with confidence data required by the drift diffusion model account.

**Supplemental References**

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. doi:10.3758/BF03196546

Maniscalco, B., & Lau, H. (2012). A signal detection theoretic approach for estimating metacognitive sensitivity from confidence ratings. *Consciousness and Cognition*, *21*(1), 422–430.

Pleskac, T. J., & Busemeyer, J. R. (2010). Two-stage dynamic signal detection: a theory of choice, decision time, and confidence. *Psychological Review*, *117*(3), 864–901. doi:10.1037/a0019737

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological review*, *85*(2), 59.