Further details and results on meta-d' on alternative SDT models

(Supplemental material for "Measures of metacognition on signal-detection theoretic models")

Adam B. Barrett, Zoltan Dienes and Anil K. Seth
University of Sussex

Mathematical description of degrading signal model

The general mathematical description of the type II evidence on this model is as follows. The degraded type II evidence when the stimulus is absent is

$$X_0^{(\mathrm{II})} \sim \mathcal{N}(a_0 x_0, s_0^2)\,, \tag{1}$$

where x_0 is the outcome of the type I evidence, and $0 < a_0 < 1$ and s_0 are free parameters. Similarly, when the stimulus is present the degraded type II evidence is

$$X_1^{(\mathrm{II})} \sim \mathcal{N}(a_1 x_1, s_1^2)\,. \tag{2}$$

We denote the type I threshold by θ and the type II thresholds by $\tau_\pm$ as above, but note that due to the degradation of the signal, the constraint $\tau_- < \theta < \tau_+$ is not needed. The type II hit rate for positive responses is then given by

$$\begin{aligned} H_+ &= P(X_1^{(\mathrm{II})} > \tau_+ | X_1 > \theta) && (3)\\ &= \frac{1}{h}\int_\theta^\infty P(X_1^{(\mathrm{II})} > \tau_+ | X_1 = x)\cdot P_{X_1}(x)\mathrm{d}x && (4)\\ &= 1 - \frac{1}{h}\int_\theta^\infty \phi_{d',\sigma}(x)\Phi_{a_1 x, s_1}(\tau_+)\mathrm{d}x\,. && (5)\end{aligned}$$

Similarly,

$$\begin{aligned} F_+ &= 1 - \frac{1}{f}\int_\theta^\infty \phi_0(x)\Phi_{a_0 x, s_0}(\tau_+)\mathrm{d}x\,, && (6)\\ H_- &= \frac{1}{1-f}\int_{-\infty}^\theta \phi_0(x)\Phi_{a_0 x, s_0}(\tau_-)\mathrm{d}x\,, && (7)\\ F_- &= \frac{1}{1-h}\int_{-\infty}^\theta \phi_{d',\sigma}(x)\Phi_{a_1 x, s_1}(\tau_-)\mathrm{d}x\,. && (8)\end{aligned}$$

Mathematical description of enhancing signal model

The general mathematical description of the type II evidence on this model is as follows. When the stimulus is absent, the type II evidence is given by

$$X_0^{(\mathrm{II})} \sim \mathcal{N}(x_0, b_0^2)\,, \tag{9}$$

where x_0 is the outcome of the type I evidence, and b_0 is a free parameter. Thus, some additional variance is added, reflecting an increase in noise, but the evidence remains the same on average. When the stimulus is present, the enhanced type II evidence is given by

$$X_1^{(\mathrm{II})} \sim \mathcal{N}(x_1 + b_1 d', b_1^2\sigma^2)\,, \tag{10}$$

where x_1 is the outcome of the type I evidence, and b_1 is a free parameter. The type II hit rates and false alarm rates are computed similarly to on the degrading signal model, such that

$$\begin{aligned}
H_+ &= 1 - \frac{1}{h}\int_\theta^\infty \phi_{d',\sigma}(x)\Phi_{x+b_1 d',b_1\sigma}(\tau_+)\mathrm{d}x\,, && (11)\\
F_+ &= 1 - \frac{1}{f}\int_\theta^\infty \phi_0(x)\Phi_{x,b_0}(\tau_+)\mathrm{d}x\,, && (12)\\
H_- &= \frac{1}{1-f}\int_{-\infty}^\theta \phi_0(x)\Phi_{x,b_0}(\tau_-)\mathrm{d}x\,, && (13)\\
F_- &= \frac{1}{1-h}\int_{-\infty}^\theta \phi_{d',\sigma}(x)\Phi_{x+b_1 d',b_1\sigma}(\tau_-)\mathrm{d}x\,. && (14)
\end{aligned}$$

Examples with unequal variances

Here we illustrate the behaviour of meta-d' measures on degrading and enhancing signal models with type I evidence distributions of unequal variance ($\sigma = 2$). Figure S1 shows behaviour on the degrading signal model and Figure S2 illustrates the enhancing signal model. These figures correspond to Figures 6 and 7 for the equal variance case.

Model with type I criterion jitter

SDT models often assume that the decision threshold remains stable over time; however it has been argued that trial-to-trial jitter in the decision threshold may exist (Ashby & Maddox, 1993; Mueller & Weidemann, 2008; Benjamin, Diaz, & Wee, 2009). Here we examine a model with type I criterion jitter to test whether this affects the independence of meta-d' from types I and II response bias. On this model, the type I and type II evidence are generated following the standard SDT model. However, while the type II thresholds $\tau_\pm$ remain constant across trials, the type I threshold is jittered according to an independent Gaussian random variable on each trial.

The mathematical description of this model is as follows. We denote the jittered type I threshold by $\Theta \sim \mathcal{N}(\theta, \eta^2)$. We denote the distance between 'present' and 'absent' distributions by d, dropping the prime since the actual d' [as measured by performance according to (1)] is affected by the jitter

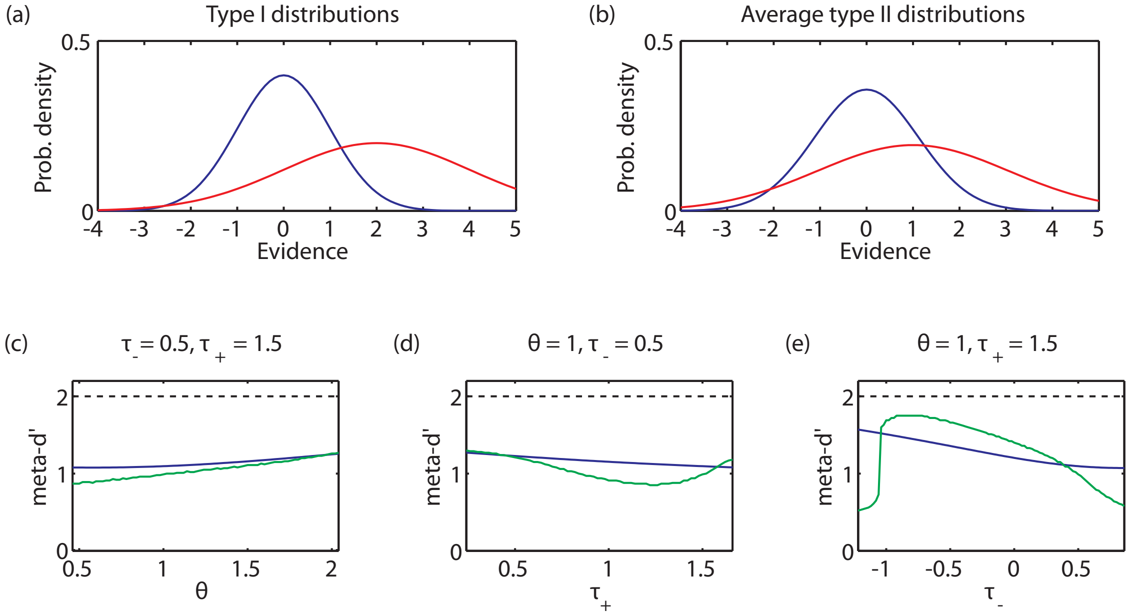

Figure S1. Meta-d' on a degrading signal model with unequal variances ($a_0 = a_1 = s_0 = s_1 = 1/2$, $d' = 2$, $\sigma = 2$). Top row: evidence distributions for (a) the type I response and (b) the type II response; stimulus absent in blue, stimulus present in red. Bottom row: behaviour of meta-d' for varying (c) θ, (d) τ_+, and (e) τ_-. Blue curves show $\tilde{d}'_{\mathrm{b}}$ and green curves show $\tilde{d}'_{\mathrm{SSE}}$. Dashed lines show the constant value of $d' = 2$. In each panel, the threshold being varied is taken across the full range that satisfies the inclusion criterion (29).

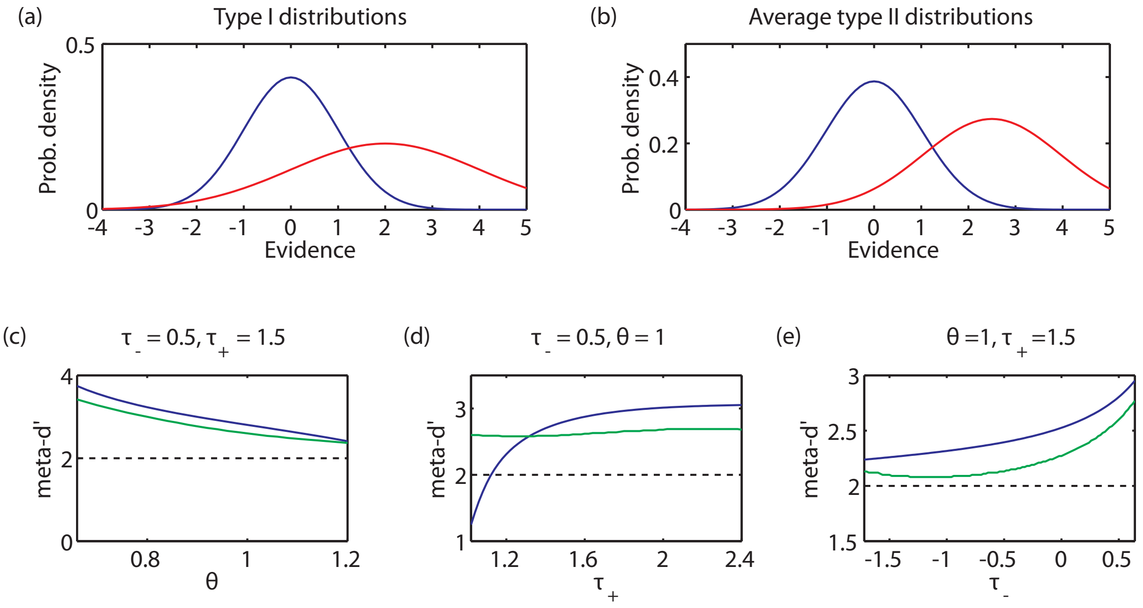

Figure S2. Meta-d' on an enhancing signal model with unequal variances ($b_0 = b_1 = 1/4$, $d' = 2$, $\sigma = 2$). Top row: evidence distributions for (a) the type I response and (b) the type II response; stimulus absent in blue, stimulus present in red. Bottom row: behaviour of meta-d' for varying (c) θ, (d) τ_+, and (e) τ_-. Blue curves show $\tilde{d}'_{\mathrm{b}}$ and green curves show $\tilde{d}'_{\mathrm{SSE}}$. Dashed lines show the constant value of $d' = 2$. In each panel, the threshold being varied is taken across the full range that satisfies the inclusion criterion (29).

and is less than d. It can be shown that the type I hit rate and false alarm rate are given by

$$\begin{aligned} h &= 1 - \Phi_0\left(-\frac{d-\theta}{\sqrt{\sigma^2+\eta^2}}\right), && (15)\\ f &= 1 - \Phi_0\left(\frac{\theta}{\sqrt{1+\eta^2}}\right), && (16) \end{aligned}$$

and hence

$$d' = \sigma\frac{d-\theta}{\sqrt{\sigma^2+\eta^2}} + \frac{\theta}{\sqrt{1+\eta^2}}\,. \tag{17}$$

The type II quantities are derived as follows:

$$\begin{aligned} H_+ = &= P(X_1 > \tau_+ | X_1 > \Theta) && (18)\\ &= \int_{-\infty}^{\infty} \mathrm{d}\theta' P(X_1 > \tau_+ | X_1 > \theta) P_\Theta(\theta') && (19)\\ &= \int_{-\infty}^{\tau_+} \mathrm{d}\theta' \Phi_{\theta,\eta}(\theta') \frac{1-\Phi_{d,\sigma}(\tau_+)}{1-\Phi_{d,\sigma}(\theta')} + \int_{\tau_+}^{\infty} \mathrm{d}\theta' \Phi_{\theta,\eta}(\theta')\,, && (20) \end{aligned}$$

and similarly

$$\begin{aligned} F_+ &= \int_{-\infty}^{\tau_+} \mathrm{d}\theta' \Phi_{\theta,\eta}(\theta') \frac{1-\Phi_0(\tau_+)}{1-\Phi_0(\theta')} + \int_{\tau_+}^{\infty} \mathrm{d}\theta' \Phi_{\theta,\eta}(\theta')\,, && (21)\\ H_- &= \int_{\tau_-}^{\infty} \mathrm{d}\theta' \Phi_{\theta,\eta}(\theta') \frac{\Phi_0(\tau_-)}{\Phi_0(\theta')} + \int_{-\infty}^{\tau_-} \mathrm{d}\theta' \Phi_{\theta,\eta}(\theta')\,, && (22)\\ F_- &= \int_{\tau_-}^{\infty} \mathrm{d}\theta' \Phi_{\theta,\eta}(\theta') \frac{\Phi_{d,\sigma}(\tau_-)}{\Phi_{d,\sigma}(\theta')} + \int_{-\infty}^{\tau_-} \mathrm{d}\theta' \Phi_{\theta,\eta}(\theta')\,. && (23) \end{aligned}$$

Figures S3 and S4 show the behaviour of d', $\tilde{d}'_{\mathrm{b}}$ and $\tilde{d}'_{\mathrm{SSE}}$ on example criterion jitter models with respectively equal and unequal variances. In both examples d' is approximately independent of decision and confidence thresholds, and only slightly less than the distance d between the two evidence distributions. The meta-d' measures are approximately equal to d' for all decision and confidence threshold values, reflecting well the fact that there is no enhancement or degradation of the evidence in between the type I and II responses.

References

Ashby, F., & Maddox, W. (1993). Relations between prototype, exemplar, and decision bound models of categorization. Journal of Mathematical Psychology, 37(3), 372 - 400. doi: 10.1006/jmps.1993.1023

Benjamin, A. S., Diaz, M., & Wee, S. (2009). Signal detection with criterion noise: Applications to recognition memory. Psychological Review, 116(1), 84115. doi: 10.1037/a0014351

Mueller, S., & Weidemann, C. (2008). Decision noise: An explanation for observed violations of signal detection theory. Psychonomic bulletin and review, 15(3), 465494.

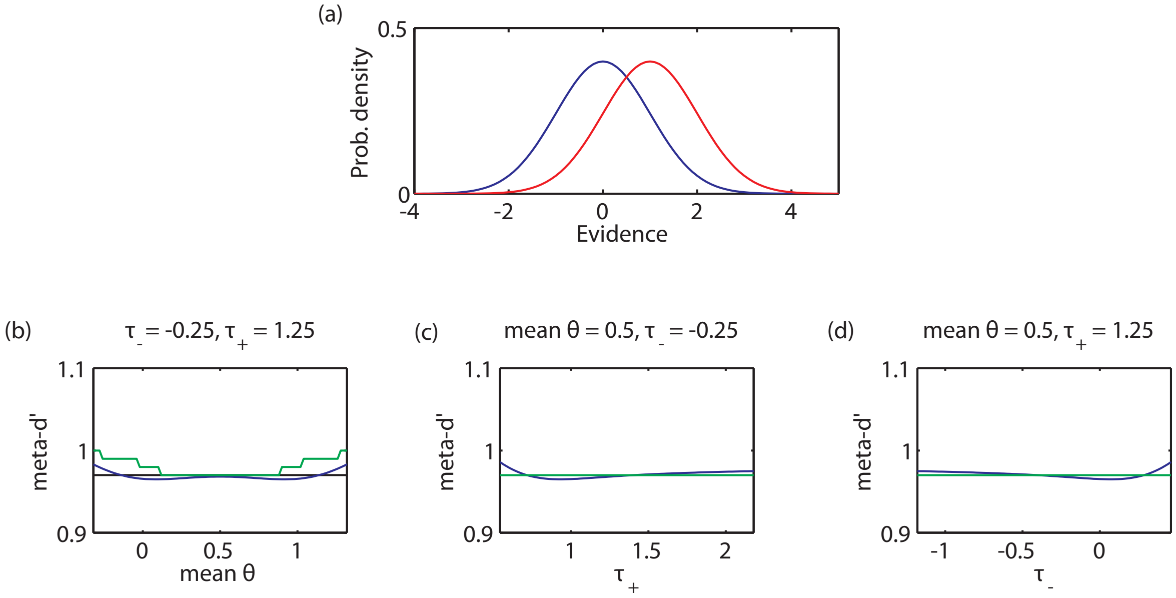

Figure S3. Meta-d' on a model with type I criterion jitter and equal variances ($d = 1$, $\sigma = 1$, $\eta = 0.25$). (a) Evidence distributions for the type I and II responses; stimulus absent in blue, and stimulus present in red. Bottom row: meta-d' for varying (b) mean decision threshold θ, (c) upper confidence threshold τ_+, and (d) lower confidence threshold τ_-. Blue curves show $\tilde{d}'_{\mathrm{b}}$, green curves show $\tilde{d}'_{\mathrm{SSE}}$. The black line in (b) shows d', which varies in this case due to the jitter. In (c) and (d) $d' = \tilde{d}'_{\mathrm{SSE}}$. In (b-d) the threshold is varied across the full range satisfying the inclusion criterion (29). While the jitter causes d' to be slightly reduced as compared to the distance between the two evidence distributions, meta-d' remains approximately equal to d'.

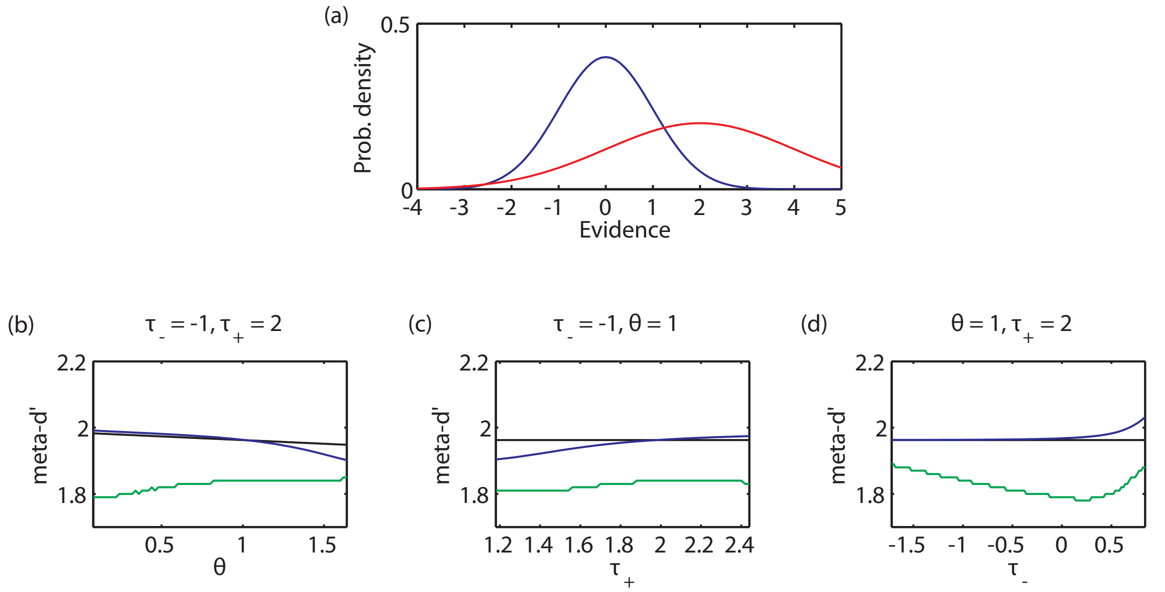

Figure S4. Meta-d' on a model with type I criterion jitter and unequal variances ($d = 2$, $\sigma = 2$, $\eta = 0.25$). (a) Evidence distributions for the type I and II responses, stimulus absent in blue and stimulus present in red. Bottom row: behaviour of meta-d' for varying (b) θ, (c) τ_+, and (d) τ_-. Blue curves show $\tilde{d}'_{\mathrm{b}}$, green curves show $\tilde{d}'_{\mathrm{SSE}}$, and black curves show d', which varies in this case due to the jitter. In each panel, the threshold being varied is taken across the full range that satisfies the inclusion criteria.