Novel regression methods for metacognition

Simon Bang Kristensen

Department of Public Health, Aarhus University, Aarhus, Denmark

Kristian Sandberg

Center of Functionally Integrative Neuroscience, CFIN, Aarhus University

Bo Martin Bibby

Department of Public Health, Aarhus University

Author Note

This article is a preprint and has not been peer-reviewed. Correspondance should be sent to Simon Bang Kristensen, Department of Public Health, Bartholins Allé 2, 8000 Aarhus C, DK-Denmark, simonbk@ph.au.dk.

Abstract

Metacognition is an important component in basic science and clinical psychology, often studied through complex, cognitive experiments. While Signal Detection Theory (SDT) provides a popular and pervasive framework for modelling responses from such experiments, a shortfall remains that it cannot in a straightforward manner account for the often complex designs. Additionally, SDT does not provide direct estimates of metacognitive ability. This latter shortcoming has recently been sought remedied by introduction of a measure for metacognitive sensitivity dubbed $\text{meta-}d'$. The need for a flexible regression model framework remains, however, which should also incorporate the new sensitivity measure. In the present paper, we argue that a straightforward extension of SDT is obtained by identifying the model with the proportional odds model, a widely implemented, ordinal regression technique. We go on to develop a formal statistical framework for metacognitive sensitivity by defining a model that combines standard SDT with $\text{meta-}d'$ in a latent variable model. We show how this agrees with the literature on $\text{meta-}d'$ and constitutes a practical framework for extending the model. We supply several theoretical considerations on the model, including closed-form approximate estimates of $\text{meta-}d'$ and optimal weighing of response-specific meta-sensitivities. We discuss regression analysis as an application of the obtained model and illustrate our points through simulations. Lastly, we present R-software that implements the model. Our methods and their implementation extend the computational possibilities of SDT and $\text{meta-}d$ and are useful for theoretical and practical researchers of metacognition.

Keywords: statistical analysis, modelling, metacognition

Novel regression methods for metacognition

Introduction

Insight into one's actions and abilities is often termed metacognition. Metacognition is a recurrent concept in many psychological fields of study (Metcalf & Shimamura, 1996), and metacognitive disability is a characteristic of clinical diagnoses such as schizophrenia, Alzheimer's and certain injuries of the brain (David, Bedford, Wiffen, & Gilleen, 2012). In cognitive science, these processes and disorders are investigated in a neurological context through cognitive and behavioural experiments, usually involving a complex setup with repeated measurements of test subjects over various stimulus intensities and possibly other changes of experimental settings such as blocked trial runs. The basic structure of most such experiments can be summarised as follows: In a single trial replication, a subject must first complete a task (e.g. identify which of two images were presented) and thereafter report a quantification of her/his confidence in, awareness of, or feeling of control over the performed task. The first assignment is usually dubbed the type 1 task while the reflection over the type 1 task is called the type 2 task. Various quantifications in the type 2 task have been proposed, such as verbal and numeric rating scales of confidence (Dienes, 2007; Dienes, Altmann, Kwan, & Goode, 1995), awareness (Ramsøy & Overgaard, 2004; Sandberg, Timmermans, Overgaard, & Cleeremans, 2010), or as a probability assessment (Dienes, 2007; Fleming & Daw, 2017), and more indirect methods have been proposed, for example allowing the subject to wager money on the correctness of the type 1 assertion (Kiani & Shadlen, 2009; Persaud, McLeod, & Cowey, 2007).

A common assumption when modelling the responses of consciousness experiments is that the subject's decision is based on thresholds in an underlying evidence distribution so that the response changes when the signal crosses a threshold, a model referred to in psychology as Signal Detection Theory (SDT) (Green & Swets, 1966). While the signal detection model remains ubiquitous in modelling cognitive experiments, the question of how to use it in quantifying metacognitive capabilities has been revisited and revised

multiple times, see Fleming and Lau, 2014 for a review. Additionally, later years have seen various expansions and alternatives to the SDT, some examples of which can be found in Maniscalco and Lau, 2016. Such models may be termed 'formal' as they are meant to represent the underlying, e.g. neural, process that the researcher believes to explain the response in a single trial, while they remain abstracted from design structure, data actual and methods of estimation.

The formal model will specify a relationship between the processes controlling performance and insight leading to some index of metacognitive ability. A recently proposed measure in the SDT setting is meta- d ("meta d prime", Maniscalco & Lau, 2012; Rounis, Maniscalco, Rothwell, Passingham, & Lau, 2010), meant to quantify a subject's sensitivity to the difference between a correct and incorrect response. The introduction of meta- d was in itself an attempt to remedy a shortfall of SDT in that the SDT model does not give a direct measure of metacognitive abilities. However, the literature tends to focus on defining meta- d from the observed response frequencies and not based on an underlying model.

An inadequacy shared by SDT and most other formal models is that they are not sufficiently elaborate to directly include the complex designs of most cognitive experiments, particularly it is not clear how to account for multiple stimulus intensities or repeated measurements, and it is not obvious how to include experimental factors and subject characteristics as covariates to adjust comparisons or perform predictions. In neuropsychological experiments, we may require for example to regress cognitive abilities on a structural brain measure while adjusting for age and possibly modulate the effect by brain region. These shortfalls are often addressed by applying the SDT multiple times under various dichotomisations and stratifications with substantial loss of efficiency as a consequence. Moreover, failure to account for sources of variation aside from that of the neurological process can lead to biased estimates and erroneous inference. Morey, Pratte, and Rouder, 2008, for example, showed how three classic results in recognition memory can

be produced as artifacts of ignoring extra-variational factors such as task difficulty.

In the present paper, we analyse models for metacognition from a theoretical perspective focusing on SDT and meta- d and compare them to established statistical models. We define a model integrating SDT and the meta- d measure and show how this leads to expansions to regression models meant to address the above mentioned shortcomings. These approaches are exemplified through simulations to illustrate the importance of accounting for variational sources in an efficient manner.

While our approach is teoretical, we aim to primarily address a practical, fundamental problem of a researcher planning an analysis of a consciousness experiment: How to convert a formal model to an appropriate statistical analysis. A similar idea underlies Wright, Horry, and Skagerberg, 2009 where the authors consider a simpler experiment with only a type 1 task and propose mixed model techniques to enable statistical analysis in type 1 SDT. First, the coupling of a statistical and formal model has the practical advantage that it allows for computation of estimates and standard errors from sound statistical principles that can, ideally, be performed by standard software. Second, and more theoretically, an understanding of the corresponding statistical model may help to make explicit the underlying assumptions of the formal model and standard techniques for model validation can be applied. Furthermore, the statistical model may offer more extensions which can be converted to insights into the subject matter.

The paper is structured as follows. We begin by recounting the basis of Signal Detection Theory with a vu to some earlier versions and expansions of the basic model. We then establish the relationship between generalised linear models (GLM's) for binary data and latent variable models, and we recall two statistical models, the proportional odds (PO) model and a natural extension, the partial proportional odds (PPO) model. We then derive the simple, but often overlooked fact that the SDT model for metacognition is equivalent to the PO model (DeCarlo, 1998), a well-known regression tehniqe for ordinal data.

Next, we introduce meta- d through a statistical model using latent variables and argue that this is in concordance with the definition in for example Barrett, Dienes, and Seth, 2013. We compare the obtained model, dubbed the meta-SDT, to the PPO to show that while the two are conceptually very similar, they differ in an important way that bears consequence to the interpretation of the model parameters. In addition, we discuss model validation and a technique for obtaining approximate, closed-form estimates (i.e. an explicit expression) of the meta parameters. The obtained methods are illustrated and tested through simulations. Finally, we discuss numerical issues connected to the application of regression models for meta-SDT and introduce the R-package `metaSDTreg`.

The novel contributions made by this article include a statistical description of meta- d' that is formal in the sense that it connects a specified model (specified through latent variables) directly to meta- d in the model's parameters. This constitutes a flexible framework that is well suited for expansions of the model. We also supply an optimal weight for combining response-specific meta sensitivities and formulas for approximate closed-form estimates of meta- d , something that is not to our knowledge presented elsewhere. Additionally, we present software to fit the meta-SDT model coded as a package in the open-source statistical software R (R Core Team, 2016).

We end this introduction by introducing some ubiquitous notation. We imagine in the following a cognitive experiment such as that which is presented in Sandberg et al., 2010, but with a two-choice type 1 task. Thus, a subject must for the type 1 task for example report whether a shape was present on an image shown. We denote by $S \in \{0, 1\}$ the true status with $S = 1$ meaning “present” (or “signal”) and $S = 0$ meaning “absent” (or “noise”). The stimulus intensity is varied for example by presenting the image for t milliseconds. In the type 2 task, the subject must quantify perceived awareness of the decision using an ordinal scale $\{1, 2, \dots, L\}$. We will use the subscripts $\mathcal{N}$ and $\mathcal{S}$ to denote quantities in the type 2 model which pertain to a “noise” or “signal” report in the type 1 task, respectively. Subjects are indexed by $i = 1, \dots, N$ and replications are indexed by

$j = 1, \dots, M$. To minimise unwieldy notation we will often omit these subscripts and consider a single replication on a single subject.

Throughout, we will by Φ and ϕ respectively denote the cumulative distribution and density function of the standard normal distribution. We let $\mathbb{P}(\cdot)$ be a probability measure, and let $\mathbb{E}[\cdot]$, $\text{Var}(\cdot)$, and $\text{Cov}(\cdot, \cdot)$ be the associated expectation, variance, and covariance operators, respectively. We further denote by $\mathbb{1}_{\{\cdot\}}$ the indicator function.

Cognitive responses: Signal detection theory

For notational parsimony, the SDT model that we introduce here is the simplest version meant to model experiments with a dichotomous type 2 response. This can be naturally extended to ordinal type 2 assertions by broadening the threshold structure. We end this section by discussing earlier versions as well as possible expansions of the signal distributions.

Signal Detection Theory assumes that responses occur as decisions based on a latent evidence X . We will adhere to the classical conventions of SDT and assume that X follows a standard normal distribution and that a signal shifts the location of the noise distribution by d (“d-prime”), but naturally various alternatives and extensions exist (e.g., DeCarlo, 2010).

The SDT model is shown in Figure 1. Responses are governed by a threshold model on X specifying that a subject responds “signal” if $X > \theta$ and “noise” if $X \leq \theta$. The type 2 response is then “confident” if $X > \tau_S$ (meaning confident in signal) or $X \leq \tau_N$ (confident in noise) and otherwise the assertion is “unsure”. One assumes that $\tau_N < \theta < \tau_S$ under the rationale that for a given response one requires stronger evidence in the given direction to move from “unsure” to “confident”. We refer to θ as the type 1 criterion, to τ_N and τ_S as the type 2 criteria, and to d as the signal sensitivity.

Estimation in SDT is usually based on simple relations between probabilities and

parameters. Thus one would estimate

$$\hat{\theta} = -\Phi^{-1}(f), \quad \hat{d} = \Phi^{-1}(h) - \hat{\theta}, \quad (1)$$

where f is observed F , the type 1 false alarm rate (the probability of incorrectly reporting a signal), and h is the observed H , the type 1 hit rate (correctly reporting a signal).

We refer to (1) as moment estimators. In the single-task SDT (i.e. where there is only a type 1 task) the moment estimators coincide with the maximum likelihood estimates when responses are binomial (Appendix A), but this is generally not true when considering the SDT model in Figure 1 and better estimates (in terms of efficiency) are obtained from the proportional odds model (see below), which is available in any standard statistical software package.

Predecessors and expansions

An important influence for SDT in metacognition is the paper by Galvin, Podd, Drga, and Whitmore, 2003. Galvin and colleagues formulated a response model, where a subject must decide on the type 1 task based on an evidence distribution and then perform the type 2 task using a transformation of the same evidence by some function ω . The idea is to enable overall modeling of type 1 and type 2 response ROC curves, an idea which may be said to go back to an editorial letter by Clarke, Birdsall, and Tanner, 1959.

This is contrasted by the SDT model above which more naturally facilitates models for the type 2 task conditioned on type 1 (e.g. confidence given a “signal” response). We will revisit these conditionals below when we discuss how to quantify metacognitive sensitivity. Note also that the approach of Galvin et al. is very similar to the present SDT model taking $\omega(\cdot) = |\theta - \cdot|$, see Maniscalco and Lau, 2014 for a more in depth discussion. As we will exemplify with the Independence model below, the two approaches may be said to differ in their threshold models.

An obvious alternative to the approach of Galvin et al. and an extension of SDT would be instead to consider a two-dimensional evidence distribution for type 1 and 2

responses. In consciousness and behavioural research, one such model proposed for memory research is the Stochastic Detection and Retrieval model of Jang, Wallsten, and Huber, 2012. A bivariate model for source monitoring is found in DeCarlo, 2003 while another, Bayesian, approach is explored by Fleming and Daw, 2017.

A special case, as we argue below, of the two-dimensional signal distribution is the so-called hierarchical model (Maniscalco & Lau, 2016; Rausch & Zehetleitner, 2017), wherein one assumes that the type 1 task follows a single-task SDT and that the type 2 decision is made in an SDT based on evidence Y , where

$$Y = X + \gamma, \quad (2)$$

for $\gamma \sim N(0, \zeta^2)$ independent of X . The reasoning is that ζ^2 measures the noise or loss of evidence in making the type 2 assertion compared to the type 1 task, where $\zeta^2 = 0$ recovers the SDT.

The relationship between the distributional formal models are shown in Figure 2. This is, however, only the signal distributions and one must also specify the threshold model. The best example of the importance of thresholds is the independence model (Rausch & Zehetleitner, 2017). If we in the independence model were to specify the thresholds in the spirit of Galvin et al., a subject would report a signal if $X > \theta$ and confidence if $Y > \tau$ yielding a rather odd model in which the type 1 and type 2 tasks are independent. However using the ordered rationale of SDT (see Rausch and Zehetleitner, 2017 for details),

$$\text{Confidence} = \begin{cases} 1 & \text{if } Y \leq \tau \text{ and } \text{response} = \text{"Noise"} \\ 1 & \text{if } Y > \tau \text{ and } \text{response} = \text{"Signal"} \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

and thus we obtain a model where the probability of confidence (for Y symmetrically distributed) given X is $\mathbb{P}(Y \leq (1 - 2\mathbb{1}_{\{X > \theta\}})\tau)$, which is not independent of X .

A brief statistical precursor

In the present section, we recount some of the theory concerning generalised linear models (GLM's) and introduce two common regression models for ordinal data.

A GLM can be described as a regression for the mean of a response belonging to an exponential family distribution, where it is assumed that the explanatory variables enter additively after transformation using a so called link function (McCullagh & Nelder, 1989). For example, if the response is Gaussian and the link is taken to be the identity function, the GLM is linear regression while a binary response with a logit link would result in logistic regression.

We will focus on binary outcomes with a particular class of link functions. Suppose to this end that $R \in \{0, 1\}$ is a binary response and that F is a distribution function that is differentiable, strictly increasing on the real numbers, and is such that $F - 1/2$ is an odd function. Given covariates V we might then consider the GLM

$$F^{-1}(\mathbb{E}[R]) = V\beta, \quad (4)$$

where β is a parameter vector. An equivalent formulation of (4) is through a latent variable R^* where we associate the latent to the observed by $\{R = 1\} = \{R^* > 0\}$ and specify the latent variable through a regression equation,

$$R^* = V\beta + \epsilon, \quad (5)$$

where $\epsilon \sim F$. Note that the assumptions on F ensure that ϵ is absolutely continuous, supported on the real numbers and symmetrically distributed. We will refer to (5) as the latent variable representation of the GLM. Also note that the representation consists of two components, the threshold model linking the latent variable to the observed and a

distribution model for the latent variable. The two formulations are indeed equivalent since

$$\begin{aligned}\mathbb{E}[R] &= \mathbb{P}(R = 1) \\ &= \mathbb{P}(R^* > 0) \\ &= \mathbb{P}(-\epsilon \leq V\beta) \\ &= \mathbb{P}(\epsilon \leq V\beta) \\ &= F(V\beta),\end{aligned}$$

which is equivalent to (4).

The majority of statistical models dealing with ordinal data from an ordered scale $\{1, \dots, K\}$ can be characterised as a series of binary models that arise from considering cumulative, possibly conditional, probabilities on the scale. The parameters are constricted across models to increase efficiency and parsimony, see Ananth and Kleinbaum, 1997 for a review. Here we will consider the pervasive proportional odds model.

Proportional odds (PO)

Suppose that we observe an answer $A \in \{1, \dots, K\}$ from an ordinal scale. The *proportional odds model* with a probit link is defined as

$$\Phi^{-1}(\mathbb{P}(A \leq k)) = \alpha_k - V\beta, \quad (6)$$

for $k = 1, \dots, K - 1$ and where $\alpha_1 < \dots < \alpha_{K-1}$. The sign for β is chosen so that the interpretation coincides with that of a standard probit regression when $K = 2$.

The name of the model can be motivated from the following property when using the logit link $\Lambda(x) = 1/(1 + \exp(-x))$. If we consider the odds of observing an outcome no greater than k_1 , this equals $\exp(\alpha_{k_1}) \exp(-\beta V)$ and comparing this to the odds of observing an outcome no greater than k_2 , the ratio of the two odds is the same for any choice of covariates, the ratio being $\exp(\alpha_{k_1} - \alpha_{k_2})$. Moreover, say we wish to calculate the change in odds for the event $\{A \leq k\}$ as we change the explanatory variables from V_1 to

V_2 . This is done using

$$\begin{aligned}\log OR(A \leq k \mid V_1 \rightarrow V_2) &= \alpha_k - V_2\beta - (\alpha_k - V_1\beta) \\ &= -(V_2 - V_1)\beta,\end{aligned}$$

so that the odds ratio is calculated as $\exp(-(V_2 - V_1)\beta)$. The assumption of proportional odds expresses that the OR is independent of k , i.e. there is no interaction between covariates and scale level.

The proportional odds assumption can be checked by examining so called proportional odds plots, that can be constructed by calculating observed frequencies, transforming to the link scale and plotting these against covariate levels. Deviations from parallel lines will be indicative of violations of the assumption. Other approaches to model validation include investigating partial residuals (Harrell, 2015).

From our considerations on latent variables above, it follows that a latent variable representation of the PO model in (6) is given by $\{A \leq k\} = \{A^* \leq \alpha_k\}$, where

$$A^* = V\beta + \epsilon, \tag{7}$$

for $\epsilon \sim N(0, 1)$. Note that the proportional odds assumption in the latent variable context expresses that A^* does not depend on k .

The SDT as a PO model. Consider again the SDT model as outlined above. It is clear that the evidence distribution can be represented as $dS + X$, which is the same distribution as the latent variable in the PO model (7), since $\epsilon \sim X$. Combining this with the fact that the threshold model of the PO model coincides with that in SDT (see Figure 1), it follows directly that SDT is equivalent to the PO model with $K = 4$ and with a single covariate $V = S$ (in the sense of identical likelihoods). The associated ordinal answers A with the latent variable A^* are also shown in Figure 1 and the threshold values of the PO model can be associated with those in SDT by

$$\alpha_1 = \tau_N, \alpha_2 = \theta, \text{ and } \alpha_3 = \tau_S. \tag{8}$$

Additionally, the signal sensitivity d corresponds to the regression parameter β from the PO model.

Partial proportional odds (PPO)

When the proportional odds assumption seems unrealistic or, for other reasons, a more involved parameter structure is deemed appropriate, a natural extension of the PO model can be defined by relaxing the proportionality assumption for a subset of the covariates V_0 ,

$$\Phi^{-1}(\mathbb{P}(A \leq k)) = \alpha_k - V\beta - V_0\delta_k, \quad (9)$$

where δ_k measures the deviation from proportionality. The latent variables associated with the PPO are

$$A_k^* = V\beta + V_0\delta_k + \epsilon, \quad (10)$$

with $\{A \leq k\} = \{A_k^* \leq \alpha_k\}$ for $\epsilon \sim N(0, 1)$. Naturally, $\delta_k = 0$ for every k recovers the PO model.

As an example, we return to our consciousness experiment from the introduction. We define the ordinal variable A as a reordering of the responses, see Figure 1. We might wish to let the effect of S vary depending on the task type, i.e.

$$\Phi^{-1}(\mathbb{P}(A_{ij} \leq k)) = \alpha_k - S_{ij}(\beta + \delta \mathbb{1}_{\{k \neq L\}}).$$

This is the PPO model (9) with $V_{ij} = V_{0,ij} = S_{ij}$ and $\delta_k = \delta \mathbb{1}_{\{k \neq L\}}$. The response to the type 1 task is the event $\{A \leq L\}$ (“noise”) or its complement (“signal”), and this is performed with a discrimination between signal and noise being β . When deciding which awareness rating to assign to the type 1 task, this is done with a signal discrimination $\beta + \delta$. We argue below that β will correspond to the signal sensitivity d known from SDT, while $\beta + \delta$ is similar to, but not the same as, meta- d . If we wished to account for between subject variation, we could modify our model as

$$\Phi^{-1}(\mathbb{P}(A_{ij} \leq k)) = \alpha_k - S_{ij}(\beta + \delta \mathbb{1}_{\{k \neq L\}}) + a_i.$$

Note that the subject intercepts a_i enter as a proportional effect, i.e. it is assumed that a given subject has a response bias that she/he applies equally to both the type 1 and type 2 assertion. Another specification could be to assume that the subject effects were realisations of some population distribution, in which case one would obtain a mixed model. While it is possible to let the subject effects depend also on k , some care is needed as the subject-specific intercepts should still be ordered and reparametrisation is generally needed to ensure ordinality of the scale (Tutz & Hennevogel, 1996). We will term the present model, with no random effects, a fixed effect model.

The meta-SDT model

Below we introduce the main contribution of the present paper, when we define a model for all responses from a meta-cognitive task which includes in its parameters the response-specific meta- d 's which are concordant with the definition in e.g. Barrett et al., 2013. The model relies on a latent variable formulation similar to that in Signal Detection Theory, but also employing the truncated normal distribution $trN_a^b(\mu, \sigma^2)$ (Appendix B).

We discuss first an intuitive definition of meta- d before moving on to the meta-SDT model. This is presented first in the simple case with a two-level type 2 rating and a binary signal intensity. In this case, we study the likelihood to compare to the standard SDT model and to the PPO model. Next, we generalise this model to accommodate an ordinal type 2 task and a general regression model in order to account for more involved designs. We end this section by making a few separate, discursive points concerning the obtained model, including a comparison to the PPO model, model validation and how to obtain a joint meta- d measure from the response-specific meta- d 's. We additionally study how to obtain approximate, closed-form estimates of meta-sensitivity.

Meta- d

In the same manner as d measures the sensitivity to signal in the standard SDT model, meta- d is meant to quantify meta-sensitivity, i.e. the individual's ability to separate

correct from incorrect decisions. Meta- d was introduced in Rounis et al., 2010 and more thoroughly discussed by Maniscalco and Lau, 2012, while a more mathematical approach is found in Barrett et al., 2013 and Maniscalco and Lau, 2014.

Figure 3 shows the distribution of awareness ratings from an experiment for correct and incorrect answers to a shape recognition task. Note that this is something of a muddling of data as the responses are measured at various stimulus intensities and for different subjects. The overall picture, however, shows a pronounced tendency to score an incorrect answer low on the perception awareness scale while a high score is often given to correct assertions and we might take this as an indication of high meta-sensitivity.

Initially, we may wish to separate correct and incorrect assertions conditional on whether the type 1 report was “noise” or “signal”, as a difference between response types may confound the aggregate measure. This leads to the concept of response-specific meta- d . When defining the response-specific meta- d ’s in the SDT model we use two basic properties of SDT: That conditional on the type 1 response, the noise and signal distributions represent correct/incorrect response distributions and that the SDT model implies the response-specific evidence distributions and vice versa (barring some assumptions). This idea is illustrated in Figure 4 for the signal-specific response.

Meta-SDT model formulation

Consider again the response $A \in \{1, 2, 3, 4\}$ on our ordinal scale. Define a conditional threshold model by

$$\begin{aligned} \{A \leq 2\} &= \{X \leq \theta\}, \\ \{A \leq 3 \mid A > 2\} &= \{Z_S \leq \tau_S\}, \\ \{A \leq 1 \mid A \leq 2\} &= \{Z_N \leq \tau_N\}, \end{aligned} \tag{11}$$

for three independent, latent variables X , Z_N and Z_S . The associated distributional model is

$$\begin{aligned} X &\sim N(dS, 1), \\ Z_N &\sim \text{tr}N^\theta(d_N S, 1), \\ Z_S &\sim \text{tr}N_\theta(d_S S, 1). \end{aligned} \tag{12}$$

Note that the right-hand side of Figure 4 depicts the density of Z_S by the fully drawn lines (black / blue for $S = 0$ and gray / red for $S = 1$). We will refer to the model specified by (11) and (12) as the meta-SDT model.

We refer to the parameters d_N and d_S as the response-specific meta- d 's and take the meta-SDT model as their formal definition. This model formulation leads to estimates of meta- d coincident with those obtained from the method in for example Barrett et al., 2013 where τ_S and the signal-specific meta- d are defined as the t_0 and d_0 , respectively, satisfying the equations

$$\begin{aligned} H_S &= \frac{1 - \Phi(t_0 - d_0 S)}{1 - \Phi(\theta - d_0 S)}, \\ F_S &= \frac{1 - \Phi(t_0)}{1 - \Phi(\theta)}, \end{aligned} \tag{13}$$

where H_S is the signal-specific hit rate (the probability of reporting confidence given a correct “signal” assertion) and F_S is the signal-specific false-alarm rate (probability of confidence given an incorrect “signal” response). The equations (13) are recovered in the meta-SDT model by considering $\mathbb{P}(Z_S > \tau_S)$ and varying $S = 1, 0$ in the distribution function for the truncated normal distribution. The noise-specific meta- d is treated analogously.

In Barrett et al., 2013 and Maniscalco and Lau, 2014 θ in the type 2 task, called $\tilde{\theta}$, is allowed to vary from that in the type 1 task. I.e. our approach corresponds to $\tilde{\theta} = \theta$. We return to this in the discussion.

Meta-SDT likelihood

We begin by writing the likelihood of a multinomial distribution in a convenient form. Let $A \in \{1, 2, 3, 4\}$ be a multinomial observation and suppose that the scale is ordered. The log-likelihood for the observation when parametrising the cumulative probabilities by some ψ is then, up to a constant, given by

$$l(\psi) = a_1 \log \mathbb{P}(A \leq 1) + a_2 \log \{\mathbb{P}(A \leq 2) - \mathbb{P}(A \leq 1)\} \\ + a_3 \log \{\mathbb{P}(A \leq 3) - \mathbb{P}(A \leq 2)\} + a_4 \log \{1 - \mathbb{P}(A \leq 3)\}, \quad (14)$$

where $a_k = \mathbb{1}_{\{A=k\}}$. Inserting the probabilities from (6) or (9) would respectively yield the PO or PPO model.

By conditioning, we can write the same likelihood in another manner,

$$l(\psi) = a_1 \log \mathbb{P}(A \leq 1 \mid A \leq 2) + a_2 \log \{1 - \mathbb{P}(A \leq 1 \mid A \leq 2)\} \\ + (a_1 + a_2) \log \mathbb{P}(A \leq 2) \\ + a_3 \log \mathbb{P}(A \leq 3 \mid A > 2) + a_4 \log \{1 - \mathbb{P}(A \leq 3 \mid A > 2)\} \\ + (a_3 + a_4) \log \{1 - \mathbb{P}(A \leq 2)\}, \quad (15)$$

and this form immediately admits the meta-SDT model in (11) and (12). In other words, the meta-SDT is a model for the same ordinal response as SDT and we note that this in fact reduces to the standard SDT when $d_N = d = d_S$. To perform estimation in the meta-SDT model, we would maximise (15) as a function of the parameters.

Inserting the probabilities from (11)–(12) and differentiating (15) with respect to τ_S and d_S we see that the parameter values maximising the likelihood must satisfy the equations in (13). In other words, in the simple meta-SDT model with a dichotomous type 2 response, the maximum likelihood estimates for d_N and d_S coincide with the estimators proposed in Barrett et al., 2013.

Similarly, one may verify that the estimators of θ and d will be given by (1). In other words, the type 1 parameters in the simple case meta-SDT model are estimated in the same manner as in the standard type 1 task SDT model.

The general model formulation

We now formulate the meta-SDT model for a general response scale in a regression framework. While notationally more cumbersome the idea is the same as that considered in (11) and (12).

We consider an experiment in which a subject must perform a mutually exclusive two options choice as the type 1 task and in the type 2 task rate some aspect of the performance on an ordinal scale with $1, \dots, L$ levels. A is still a concatenation of the type 1 and type 2 tasks, so that $A \in \{1, \dots, K\}$ for $K = 2L$, cf. Figure 5.

Define the threshold model by

$$\begin{aligned}
 \{A \leq 1 \mid A \leq L\} &= \{Z_N \leq \tau_N^{L-1}\}, \\
 \{A \leq 2 \mid A \leq L\} &= \{Z_N \leq \tau_N^{L-2}\}, \\
 &\dots \\
 \{A \leq L-1 \mid A \leq L\} &= \{Z_N \leq \tau_N^1\}, \\
 \{A \leq L\} &= \{X \leq \theta\}, \\
 \{A \leq L+1 \mid A > L\} &= \{Z_S \leq \tau_S^1\}, \\
 &\dots \\
 \{A \leq K-1 \mid A > L\} &= \{Z_S \leq \tau_S^{L-1}\},
 \end{aligned} \tag{16}$$

for three independent, latent variables X , Z_N and Z_S . We apply the same distributional model as above but allow a more flexible mean structure,

$$\begin{aligned}
 X &\sim N(V\beta, 1), \\
 Z_N &\sim \text{tr}N^\theta(V_N\beta_N, 1), \\
 Z_S &\sim \text{tr}N_\theta(V_S\beta_S, 1),
 \end{aligned} \tag{17}$$

for covariate matrices V , V_N and V_S and parameter vectors β , β_N and β_S .

The parameters of the model are the thresholds $(\tau_N^1, \dots, \tau_N^{L-1}, \theta, \tau_S^1, \dots, \tau_S^{L-1})$ and the mean parameters $(\beta, \beta_N, \beta_S)$, which together we will denote as ψ . The log-likelihood

function is

$$\begin{aligned}
 l(\psi) = & \log \{ \mathbb{P}(A \leq L) \} \sum_{k=1}^L a_k + \log \{ 1 - \mathbb{P}(A \leq L) \} \sum_{k=L+1}^K a_k \\
 & + \sum_{k=1}^L a_k \log \left\{ \mathbb{P}(A \leq k \mid A \leq L) - \mathbb{P}(A \leq k-1 \mid A \leq L) \right\} \\
 & + \sum_{k=L+1}^K a_k \log \left\{ \mathbb{P}(A \leq k \mid A > L) - \mathbb{P}(A \leq k-1 \mid A > L) \right\}.
 \end{aligned} \tag{18}$$

We postpone some notes on numerical aspects to the section on the `metaSDTreg` package below.

Additional points concerning the meta-SDT model

Below we briefly present a few select points on the meta-SDT model in a discursive form.

Meta-SDT and the PPO model. Let us clarify the relationship between the meta-SDT model and the PPO model. For simplicity we will assume that the confidence scale only has two levels, i.e. $L = 2$, and with only one signal intensity, $S \in \{0, 1\}$. We compare to the PPO model given by

$$\Phi^{-1}(\mathbb{P}(A \leq k)) = \alpha_k - \beta_k S, \tag{19}$$

for $k = 1, 2, 3$. Note that this is in fact equivalent to three separate probit regressions on the cumulative probabilities and thus the estimates are of the form in Appendix A.

By differentiating the likelihood in (15) and studying the resulting estimating equations one may show that the intercept estimates will coincide between the two models. Therefore we will write $\alpha_1 = \tau_N$, $\alpha_2 = \theta$ and $\alpha_3 = \tau_S$. Similarly, examining the estimating equation for d in meta-SDT we note that this coincides with that for β_2 in the PPO model and thus the two models will give the same estimate $\hat{d} = \hat{\beta}_2$.

The difference between the two models lies in the manner in which confidence ratings occur. In the PPO model

$$\mathbb{P}(A = 4 \mid A > 2) = \frac{1 - \Phi(\tau_S - \beta_3)}{1 - \Phi(\theta - d)}. \tag{20}$$

In the meta-SDT model,

$$\mathbb{P}(A = 4 \mid A > 2) = \frac{1 - \Phi(\tau_S - d_S)}{1 - \Phi(\theta - d_S)}. \quad (21)$$

Comparing (20) and (21) we see that the difference lies in the conditional event. In the PPO model, we condition on the subject’s “signal” decision, which is performed with the sensitivity d . In the meta-SDT model, we condition on the hypothetical or unobserved event that the subject responded “signal” using the signal-specific meta- d as sensitivity. This situation arises if the subject’s type 1 and type 2 assertions were both based on a standard SDT model with sensitivity d_S , however the response from the type 1 task in this model is unobserved.

A consequence of this is that in the PPO model the marginal, cumulative probabilities depend only on one sensitivity parameter, while in meta-SDT it depends on two. Conversely, if one considers the conditional probabilities given the type 1 response, the meta-SDT describes this probability using one sensitivity (cf. (21)), while the PPO uses two (cf. (20)). From this, we may characterise the PPO model as focusing on the marginal probabilities where the meta-SDT focuses on the conditional. This may be seen as a consequence of the original definition of meta- d from the response-conditional ROC curves (Maniscalco & Lau, 2012).

Model validation. The meta-SDT model relies on what might be termed a proportional odds assumption when $L > 2$ in that the response-specific type 2 responses each depend on a single latent variable. More precisely, the latent variable does not depend on k .

As stated above, such assumptions are in the PO model often assessed using proportional odds plots, but the nonlinear, noninvertible nature of the response-specific probabilities renders this approach impractical for the meta-SDT model. Note however, that a proportional odds plot may still be used to validate the standard SDT model as this is equivalent to the PO model. A common technique in Signal Detection Theory is to compare the observed response-specific ROC curves to that predicted by the model. This

approach, however, suffers from the disadvantage that it is difficult to interpret deviations from the predicted ROC curve. Here we will discuss a model-based approach to checking the assumptions which can be complemented by a formal test. We focus on the signal-specific response case, but the noise-specific case is symmetric.

In the meta-SDT with signal variable $S \in \{0, 1\}$, the confidence in a “signal” assertion on the scale $\{1, \dots, L\}$ is determined by the latent variable

$$Z_S \sim N(d_S S, 1), \quad (22)$$

which is independent of the confidence scale level k . This is a special case of (17).

We may expand this model to obtain probabilities

$$\begin{aligned} \mathbb{P}(A > k \mid A > L) &= \mathbb{P}(Z_S^k > \tau_S^k) \\ &= \frac{1 - \Phi(\tau_S^k - (d_S + \delta_k)S)}{1 - \Phi(\theta - (d_S + \delta_k)S)} \end{aligned} \quad (23)$$

for $k = L + 1, \dots, K - 1$ where for identifiability we set δ_{L+1} to zero. Thus, the signal-specific meta- d applied by the subject to decide in a confidence of more than k is $d_S + \delta_k$. This is an explicit model of any deviation from proportionality, the deviation being measured by the δ_k 's. We will refer to this model as “saturated”.

A test that may be relevant is a general goodness of fit test to compare the predicted, marginal probabilities to those observed under the assumption that $A \in \{1, \dots, K\}$ has an unstructured multinomial distribution. We discuss this test in the setting of a binary signal, $S \in \{0, 1\}$, but it is easily extended to a situation where the observation of A over factor variables in V , V_N and V_S can be summarised in a contingency table. Suppose that we have predicted marginal probabilities for $\{A = k\}$ for $S \in \{0, 1\}$, say $\hat{\pi}_k(S)$, a total of $2K$ probabilities parametrised by $K - 1$ intercepts and three sensitivities, a total of $K + 2$ parameters. If the observed probabilities, say $p_k(S)$, come from an unstructured multinomial distribution, they must be parametrised using $2K - 1$ parameters. The test is performed by evaluating the likelihood ratio testor

$$-2 \log Q = \sum_{k=1}^K \sum_{S=0}^1 NM p_k(S) \log \left(\frac{p_k(S)}{\hat{\pi}_k(S)} \right) \quad (24)$$

against a χ^2 distribution with degrees of freedom $DF = 2K - 1 - (K + 2) = K - 3$.

Usually this test is recommended only when the expected number of counts per cell is larger than five.

Note that the goodness of fit test in (24) will only be suggestive as the ordered model is not formally nested within the multinomial. However, this can be ammended, and indeed one would obtain a more informative test, if one were to use the expected probabilities arising from the saturated, ordered model described above. This can be achieved by fitting the model in (23) as an expansion of the meta-SDT model. This can also be implemented by applying the meta-STD model for each $l \in \{1, \dots, L - 1\}$ as if one were dealing with a dichotomous type 2 rating cut at l .

Starting from the saturated model in (23) we may formally test the hypothesis of proportionality by testing

$$H : \delta_{L+2} = \delta_{L+3} = \dots = \delta_{K-1} = 0. \quad (25)$$

The test can be performed, for example, by evaluating twice the difference in log-likelihood between the expanded model and the unexpanded in a $\chi^2(DF)$ -distribution with degrees of freedom $DF = L - 2$. Alternatively, calculating the $-2 \log Q$ statistic we may compare the observed probabilities to those expected under the saturated model. We can then further compare to that expected under the meta-SDT model by evaluating the deviance in a $\chi^2(L - 2)$ -distribution. In other words, we do not need the expanded model but only the predicted probabilities.

A joint measure of meta- d . When the response-specific meta-sensitivities d_N and d_S are comparable, it may be desirable to aggregate these into a joint measure of metacognitive sensitivity which we will denote $\tilde{d}$. One may obtain such a measure by applying the meta-SDT model with the restriction $d_N = d_S = \tilde{d}$. Alternatively, $\tilde{d}$ can be obtained as a weighted average of the response-specific meta- d 's.

In Barrett et al., 2013 it is suggested in the simple SDT to concatenate the measures

as

$$\tilde{d} = wd_{\mathcal{N}} + (1 - w)d_{\mathcal{S}}, \quad (26)$$

where the weight is taken to be the proportion of “noise” responses in the type 1 task.

A general approach to the same problem is to construct a weighted estimator of the form in (26) that has minimal variance (i.e. standard error) among the class of such estimators. Using that the variance of the joint measure is

$$\begin{aligned} \mathbb{V}\text{ar}(\tilde{d}) &= w^2\mathbb{V}\text{ar}(\hat{d}_{\mathcal{N}}) + (1 - w)^2\mathbb{V}\text{ar}(\hat{d}_{\mathcal{S}}) \\ &\quad + 2w(1 - w)\mathbb{C}\text{ov}(\hat{d}_{\mathcal{N}}, \hat{d}_{\mathcal{S}}), \end{aligned} \quad (27)$$

one may show that the optimal weight is

$$w_{opt} = \frac{\mathbb{V}\text{ar}(\hat{d}_{\mathcal{S}}) - \mathbb{C}\text{ov}(\hat{d}_{\mathcal{N}}, \hat{d}_{\mathcal{S}})}{\mathbb{V}\text{ar}(\hat{d}_{\mathcal{S}}) + \mathbb{V}\text{ar}(\hat{d}_{\mathcal{N}}) - 2\mathbb{C}\text{ov}(\hat{d}_{\mathcal{N}}, \hat{d}_{\mathcal{S}})}, \quad (28)$$

i.e. the estimators are weighted inverse of their relative contribution to the standard error.

Note also that we may calculate the standard error of the associated optimal estimator of $\tilde{d}$ by inserting w_{opt} in the variance expression in (27).

This technique may be applied as a general and principled approach to obtaining a joint measure of meta-sensitivity, which can also be applied in more complex models than that considered in Barrett et al., 2013. In practice, one could insert the numerical version of the variances and covariance as obtained from the observed information, which we return to in our discussion of the **metaSDTreg** package below.

Approximate closed-form estimates. Here we will argue that when the latent variable distributions in the meta-SDT model are taken to be logistic rather than Gaussian, closed-form estimates for $d_{\mathcal{N}}$ and $d_{\mathcal{S}}$ can be obtained in the simple, dichotomous type 2 rating and binary signal, case. This can be combined with an approximation between the Gaussian and logistic cumulative distribution function to derive closed-form approximate estimates of the response-specific meta-sensitivities in the Gaussian meta-SDT model. A précis of the approximate relation between the logit and probit can be found in for example Demidenko, 2013, Section 7.1.1. Here we will apply a simple scaling

approximation $\Phi(x) \approx \Lambda(cx)$, where Λ is the logistic cumulative distribution function and c is a scaling constant (which is usually taken to be $c = 1.7$).

Applying the approximation to the signal-specific type 2 hit rate,

$$\begin{aligned} H_S &= \frac{1 - \Phi(\tau_S - d_S)}{1 - \Phi(\theta - d_S)} \\ &\approx \frac{1 - \Lambda(c(\tau_S - d_S))}{1 - \Lambda(c(\theta - d_S))}, \end{aligned} \quad (29)$$

we note that (29) is equivalent to

$$cd_S = \log \left(\frac{H_S e^{c\tau_S} - e^{c\theta}}{1 - H_S} \right), \quad (30)$$

which can be used to estimate d_S . Specifically, suppose that h_S is the observed frequency of confident answers given a correct “signal” assertion, and that we have estimators $\hat{\theta}$ for θ and $\hat{\tau}_S$ for τ_S . We would then estimate d_S by

$$\hat{d}_S = \frac{1}{c} \log \left(\frac{h_S e^{c\hat{\tau}_S} - e^{c\hat{\theta}}}{1 - h_S} \right). \quad (31)$$

Similarly, one may derive the approximate estimator for the noise-specific case,

$$\hat{d}_N = -\frac{1}{c} \log \left(\frac{f_N e^{-c\hat{\tau}_N} - e^{-c\hat{\theta}}}{1 - f_N} \right), \quad (32)$$

where f_N is the observed, type 2 noise-specific false alarm rate.

It is of theoretical interest to study the logistic version of the estimates further. If we start from a logistic SDT model, $X \sim \Lambda$, we have that

$$\theta = -\log \left(\frac{F}{1 - F} \right), \quad (33)$$

and

$$d = \log \left(\frac{H}{1 - H} \right) + \theta = \log OR(H, F), \quad (34)$$

i.e. d is the log odds ratio comparing the odds of a hit to that of a false alarm. Defining then the, say signal-specific, type 2 model as before through the truncated logistic distribution, yields

$$\tau_S = -\log \left(\frac{FF_S}{1 - FF_S} \right), \quad (35)$$

where FF_S is simply the probability of reporting confidence in a “signal” assertion when the signal is absent. For the signal-specific sensitivity we obtain the same expression as in (31) with $c = 1$. This can be written,

$$d_S = \log \left(\frac{H_S - e^{\theta - \tau_S}}{1 - H_S} \right) + \tau_S. \quad (36)$$

We recognise the same structure as in (34), but we are not aware of an obvious interpretation of (36). However, if τ_S is much larger than θ so that $\exp(\theta - \tau_S) \approx 0$,

$$d_S \approx \log \left(\frac{H_S}{1 - H_S} \right) - \log \left(\frac{FF_S}{1 - FF_S} \right) = \log OR(H_S, FF_S), \quad (37)$$

i.e. in this case, d_S is approximately a log odds ratio.

A regression model

In the present section we will consider an experiment such as that described in the introduction. The purpose is to discuss how one might approach the analysis using a regression model and additionally to, through simulation in simple cases, make some points concerning the importance of accounting for the different sources of variation in an efficient manner.

We will let A_{ij} be the response of the i 'th ($1, \dots, N$) individual in the j 'th ($1, \dots, M$) replication (trial) of the experimental setting. The subject must for example claim presence or absence of a visual structure in an image to which the subject is exposed for t_{ij} milliseconds. Let $S_{ij} \in \{0, 1\}$ denote the presence of a stimulus in the given trial. Afterwards, the subject must rate the awareness of the type 1 decision using an L -point awareness scale. The combination of the type 1 assertion and the awareness rating is concatenated in the ordinal variable $A_{ij} \in \{1, \dots, K\}$ as previously described (see Figure 5), with $K = 2L$.

We will suppose that individuals prior to the experiment are randomised into one of two groups ($\mathcal{A}$ or $\mathcal{B}$). The main purpose of the study is to estimate the overall group effect on the cognitive and meta-cognitive sensitivities of the subjects.

A simple model taking into account the above described aspects would be the general meta-SDT model in (16) and (17) taking

$$\begin{aligned} X_{ij} &\sim N(d \cdot signal_{ij} + \eta_{ij}, 1) \\ Z_{\mathcal{N},ij} &\sim trN^{\theta}(d_{\mathcal{N}} \cdot signal_{ij} + \eta_{ij}, 1) \\ Z_{\mathcal{S},ij} &\sim trN_{\theta}(d_{\mathcal{S}} \cdot signal_{ij} + \eta_{ij}, 1) \end{aligned} \tag{38}$$

where

$$signal_{ij} = (2 \cdot S_{ij} - 1)t_{ij}, \tag{39}$$

is the directed signal intensity. This ensures and assumes that the intensity of a stimulus affects equally a subjects propensity to report “noise” and “signal”.

The linear predictor in (38) is

$$\eta_{ij} = D \cdot signal_{ij} \mathbb{1}_{\{group_i=\mathcal{B}\}} + a_i, \tag{40}$$

so that D is the effect of group $\mathcal{B}$ while a_i is the subject-specific response bias. These effects are assumed to enter proportionally, i.e. they are independent of k . Naturally, one could also have included a subject effect in the sensitivities, but we have omitted this here for simplicity.

We make a few remarks on the parametrisation. Note that the latent scale is fixed by letting zero be the mean of the underlying normal distribution when the signal intensity is zero ($t_{ij} = 0$) and we are considering an unbiased subject ($a_i = 0$). For such a trial, the subject’s probability of a “signal” assertion is $1 - \Phi(\theta)$, and specifically if the population is unbiased ($\theta = 0$), the probability is $1/2$. This corresponds to the intuition that, in the absense of bias, a subject will answer at random when presented with no stimulus. Also note that the cognitive parameters of a subject in group $\mathcal{A}$ are $(d_{\mathcal{N}}, d, d_{\mathcal{S}})$, while the interaction between the signal and group variable means that a subject in group $\mathcal{B}$ has cognitive parameters $(d_{\mathcal{N}} + D, d + D, d_{\mathcal{S}} + D)$.

Simulations

Simulation study 1: Subject heterogeneity

We simulated the model in (38) with no group effect and with a dichotomous type 2 rating ($L = 2$) and a binary signal intensity ($S_{ij} \in \{0, 1\}$). In this case the sensitivity parameters admit estimation by using moment estimators such as those in (1) and (13) (the method in e.g. Barrett et al., 2013). The subject effects were drawn from a random distribution,

$$a_i \sim N(0, \zeta^2), \quad (41)$$

where ζ measures the between subject variation. In SDT this phenomenon is also referred to as “criterion jitter”, where here the jitter is on the subject level. Note that a large ζ corresponds to a large amount of jittering. Apart from the method of moment (MoM) estimators we additionally applied a meta-SDT regression model to account for the subject response bias. The resulting maximum likelihood estimates (MLE) were obtained using the *metaSDTreg* package. Note that although the data were simulated from a random effect model (41), the analyses applied a fixed effect model. A naïve estimate of ζ can in this case be obtained as the empirical standard deviation of the fixed effect estimates $\hat{a}_1, \dots, \hat{a}_N$.

We simulated the model with $N = 25$ individuals and varying the number of tasks $M = 20, 50, 100, 200$. The simulations were performed for three degrees of subject heterogeneity $\zeta = 0, 0.5, 0.75$. Each scenario was simulated 1000 times. The resulting sensitivity estimates ($\hat{d}, \hat{d}_N, \hat{d}_S$) are shown in Figure 6 (A)-(C), one panel for each parameter.

We observe that in the independent subject case ($\zeta = 0$), the performance of MoM and MLE is interchangeable as would be expected. However, as the between subject variation grows larger, the bias of the MoM estimates becomes more pronounced. This fact is a general lesson from non-linear and generalised linear models: While in the linear case, neglecting to account for a random effect will usually only effect efficiency of the estimator,

the disregard of a random effect in a non-linear model leads to biased, inconsistent and inefficient estimation (e.g. Demidenko, 2013).

We additionally note in connection with Figure 6 that the MLE for these specific choices of parameters converge satisfactorily to the true value for $N = 25$ fixed as the number of replications per subject is increased. This will not generally hold and one would need to also increase the number of subjects (Demidenko, 2013) to ensure convergence to the true value. However, to study the asymptotics of the estimates as N grows to infinity one would need a random effect model as in our present, fixed effect, approach the number of parameters will grow with N leading to inconsistent estimates.

Simulation study 2: Efficient group comparisons

One way to amend the performance within the method of moments is instead to apply the MoM estimators on a subject level, as is often done. This is however inefficient and requires a large number of replications per subject, M , as the subject must have responses in every cell of the $(A, signal)$ table. Often this leads to exclusion of some individuals resulting in another form of bias in the estimation process. We explore this subject level analysis in the simulations below.

We now simulate the model in (38) for the same three degrees of subject heterogeneity as in Study 1 but with a group effect $D = 0.25$ and with a dichotomous type 2 rating and a binary signal intensity. We assume that apart from estimating the group parameter, one wishes to adjust for any subject effect. Thus, for all simulations, $N = 25$ individuals and we vary the number of trials $M = 20, 50, 100, 200$ and heterogeneity $\zeta = 0, 0.5, 0.75$. Again, each scenario was simulated 1000 times We compare the regression model with a subject-specific intercept fitted by maximum likelihood against the MoM estimates obtained by applying the analysis on a subject level and then regressing the obtained sensitivity parameters against *group* to obtain an estimate of D .

The results of Simulation study 2 is plotted in Figure 6, panel **(D)**. As expected, the

error bars are more narrow for the MLE estimates as it is more efficient to obtain the group estimate from an overall model. More importantly, we see that for low number of replicates the MoM estimates are biased. When subject heterogeneity was present, $M = 20$ trials proved insufficient to obtain sufficient sensitivity estimates. Thus, for $\zeta = 0.5$ only one analysis resulted in an estimate of D (and as such, the error bar could not be calculated), while no estimates could be obtained for $\zeta = 0.75$.

The observed bias is due to the exclusion of subjects when a sensitivity parameter can not be determined showing that subject level analyses do not merely lead to imprecision but also to bias.

The metaSDTreg package

The procedure in metaSDTreg proceeds in two steps by first calculating the PPO estimates as detailed above and then maximising (18) with PPO estimates as starting values. The negative Hessian associated with the procedure is the numerical version of $j(\psi) = -\partial^2 l(\psi) / \partial \psi \partial \psi^T$, that is, the observed information of the model, which can be viewed as an empirical estimate of the asymptotic, inverse covariance matrix of $\hat{\psi}$. Thus, it can be used to construct approximate confidence intervals and tests. An alternative would be to use the Fisher information $\mathbb{E}[j(\psi)]$ to perform inference. The Fisher information could also be applied directly in the optimisation (so-called Fisher scoring) with the advantage that it is always non-negative definite in a properly specified model in contrast to the observed information matrix (Demidenko, 2013, section 2.11).

Discussion

In the section below we will delimit our present approach with regards to similar methods and we discuss possible directions for future developments.

Choice of type 1 threshold in the response-specific type 2 model. In the literature on meta- d it is suggested to consider another type 1 threshold than θ for the type 2 models. We would thus have a $\tilde{\theta}_N$ in the noise-conditional type 2 model and $\tilde{\theta}_S$ in the

signal-conditional type 2 model. In our notation, this would mean that Z_N should follow a $trN^{\tilde{\theta}_N}$ -distribution and Z_S a $trN_{\tilde{\theta}_S}$ -distribution. We proceed with the discussion in the signal-specific case.

Naturally, as $\tilde{\theta}_S$ is the type 1 threshold in the unobserved part of the signal-specific SDT (see the section “Meta-SDT and the PPO model” above) it can not be estimated, and one will generally need to fix it as some function of the quantities θ , d and d_S . Barrett et al., 2013 proposes to use the relative type 1 threshold $\tilde{\theta}_S = \theta d_S / d$. Maniscalco and Lau, 2014 suggest the same measure as well as one based on the likelihood ratio.

Interestingly, if we would allow the function to also depend on S , we could take $\tilde{\theta}_S = \theta + (d_S - d)S$ and we would recover the PPO model (cf. equation (20) and (21)). We can extend this argument and apply the approximation $\log z \sim z - 1$ formally to show that if we take $\tilde{\theta}_S$ to be the relative type 1 threshold, then

$$\tilde{\theta}_S = \theta \frac{d_S}{d} \approx \theta + (d_S - d), \quad (42)$$

so that for $S = 1$ the meta-SDT signal-specific hit rate is approximately the same as that in the PPO.

Random and scale effects. In our expansion of the meta-SDT in (17) we focused on location effects, i.e. effects that shift the mean of the latent variables. We could also have expanded the variance structure on the latent variables and allowed this to depend on covariates leading to scale effects. A common scale effect in signal detection theory is found in the Unequal Variance SDT model, which in statistics is known as a heteroscedastic probit model (e.g. DeCarlo, 2010), where the variance of the latent variable is allowed to depend on the presence or absence of a signal.

Additionally, it would be desirable to include random effects, as we have already explored in the simulations. However, estimation procedures for random effects are not implemented in the `metaSDTreg` package at the time of writing, as the likelihood function in this case contains an intractable integral and approximative procedures such as Gaussian quadrature or Laplace approximation are needed (Demidenko, 2013, section 7.1-7.2).

Closed-form estimates. Another aspect that might be worth investigating is the applicability of the closed form estimates in (31) and (32). Specifically, even though the error between the cumulative distribution functions of the probit and logit is small (Demidenko, 2013), it would be interesting to investigate the error on the meta- d estimates. Moreover, it may be possible to improve the approximation for example by focusing on an approximation of the truncated distribution function rather than the distribution function itself.

Apart from the theoretical value of such an investigation, precise, closed-form estimates may prove a robust manner of obtaining starting values for more advanced procedures.

Indices of metacognitive abilities. It has been suggested that when quantifying metacognitive ability using the sensitivities d , d_N and d_S , one should evaluate the metacognitive sensitivities in comparison to the type 1 sensitivity (Maniscalco & Lau, 2012, 2014). Both the absolute difference $d_S - d$ and the relative difference d_S/d have been proposed.

If we think of meta- d as a log-odds ratio, cf. (37) and also (34), this would suggest that one should consider the absolute difference, as this could be interpreted as a log-odds ratio ratio, while the relative difference would have a less straightforward interpretation.

Considering the simulation results in Figure 6 (A) - (C) one might suspect that while the sensitivities are biased separately, the index might retain smaller bias. We found this to hold for the absolute difference for all choices of ζ but less so for the ratio. Of course, we caution against interpreting this as a argument for the absolute difference over the ratio or that bias in general can be avoided by considering absolute differences, as the results are based on a single, one-case, simulation study.

Serial correlation. An aspect that may need to be investigated in future developments is that of serial correlation, although this is generally challenging in non-linear models (e.g. Breinegaard, Rabe-Hesketh, & Skrondal, 2015).

In our approach, we assume that the replications on a single subject may be assumed to be independent given the covariates (in particular the subject’s threshold and sensitivity). However, in many cognitive experiments it is easy to imagine that observations taken close together in time could be more correlated than observations far apart, in which case we say that the correlation is serial. Indeed, one might imagine negative serial correlation to occur if a subject were adverse to give a response having already given the same response multiple times in a row. Conversely in a visual study, the presence of a structure in an image might leave a “residue” that would prompt the subject to claim the structure as present in the following trial, thus leading to positive serial correlation.

Bayesian approaches. Recently there has been some effort in defining hierarchical SDT models and hierarchical models for meta- d in a Bayesian framework. For example, Fleming, 2017 discusses the implementation of a model for meta- d in a Matlab toolbox, which is a front-end to the MCMC sampler JAGS. We are also aware of similar approaches using the MCMC program STAN through the R-package `brms` (Bürkner, 2017), although we do not know of an attempt to study meta- d using this package.

The main difference between Fleming, 2017 and the approach of the present paper is that Fleming, 2017 starts from the multinomial type 2 likelihood as found in Maniscalco and Lau, 2014. Additionally, we are not aware that the approach offers regression capabilities. This is in contrast to the meta-SDT model which is a model for all the responses arising from the cognitive experiment, formulated through latent variables which is well-suited for expansions to include covariate information. Another difference is the manner in which inter-subject variation is modelled, where our approach directly parametrises this variation as effects and where the Bayesian approach in Fleming, 2017 uses hyperpriors to incorporate the extra variation.

There is no reason that the model we proposed in (16)–(17) could not be implemented in a Bayesian manner for example in R through the `brms`-package. This would entail some considerations, which may be said to apply to Bayesian data modelling in

general: namely, the choice of prior distributions and methods for assessing convergence of the Markov chain. Moreover, the importance of these considerations become increasingly critical in non-linear models as these often demand more informative priors to converge (as also discussed in Bürkner, 2017). Thus, one would require prior knowledge of some or even all the parameters in the regression model, which may be difficult to obtain as the interpretation of each parameter would change with the inclusion of other covariates.

As a tentative way of comparing the methods, we applied the out-of-the-box response-specific, group level analysis of Fleming, 2017 to some simulations from Simulation study 1 above. The details and results are presented in Appendix C. These results showing better performances of the MLE procedure over HMeta-d should naturally be interpreted very cautiously as they represent simulations under a very specific scenario. A more realistic delimitation of the performance of the various methods will have to be investigated in a large-scale simulation study, which is beyond the scope of this article.

Summary

We have presented a statistical model for metacognitive sensitivity, the meta-SDT, and have shown how it is well suited for expansion. Specifically, we can derive regression methods for metacognitive experiments. We have also shown that the standard signal detection theory model is equivalent to the proportional odds model, which is a widely used regression technique available in standard statistical software. We have made some theoretical points concerning the meta-SDT model and compared it to the partial proportional odds. Also, we supplied formulas for closed-form approximate estimates of meta-sensitivity and derived an optimal weight to combine response-specific measures. Finally, we described the implementation of the meta-SDT model in the R-package `metaSDTreg`, which is available on CRAN.

The meta-SDT model constitutes a theoretical contribution to signal detection theory and the theory on meta- d . It allows researchers to build flexible and efficient models for

metacognitive experiments, and its implementation is useful for practical researchers of metacognition.

References

- Ananth, C. V., & Kleinbaum, D. G. (1997). Regression models for ordinal responses: A review of methods and applications. *International Journal of Epidemiology*, 26(6), 1323–1333. doi:10.1093/ije/26.6.1323
- Barrett, A. B., Dienes, Z., & Seth, A. K. (2013). Measures of metacognition on signal-detection theoretic models. *Psychological Methods*, 18(4), 535–552. doi:10.1037/a0033268
- Breinegaard, N., Rabe-Hesketh, S., & Skrondal, A. (2015). The transition model test for serial dependence in mixed-effects models for binary data. *Statistical Methods in Medical Research*. doi:10.1177/0962280215588123
- Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, 80(1), 1–28. doi:10.18637/jss.v080.i01
- 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. doi:10.1121/1.1907764
- David, A. S., Bedford, N., Wiffen, B., & Gillean, J. (2012). Failures of metacognition and lack of insight in neuropsychiatric disorders. *Phil. Trans. R. Soc. B*, 367(1594), 1379–1390. doi:10.1098/rstb.2012.0002
- DeCarlo, L. T. (2003). Source monitoring and multivariate signal detection theory, with a model for selection. *Journal of Mathematical Psychology*, 47(3), 292–303. doi:10.1016/S0022-2496(03)00005-1
- DeCarlo, L. T. (1998). Signal detection theory and generalized linear models. *Psychological methods*, 3(2), 186. Retrieved November 9, 2016, from <http://psycnet.apa.org/journals/met/3/2/186/>
- DeCarlo, L. T. (2010). On the statistical and theoretical basis of signal detection theory and extensions: Unequal variance, random coefficient, and mixture models. *Journal of Mathematical Psychology*, 54(3), 304–313. doi:10.1016/j.jmp.2010.01.001

- Demidenko, E. (2013). *Mixed Models: Theory and Applications with R*. John Wiley & Sons.
- Dienes, Z. (2007). Subjective measures of unconscious knowledge. In *Progress in Brain Research* (Vol. 168). doi:10.1016/S0079-6123(07)68005-4
- Dienes, Z., Altmann, G. T. M., Kwan, L., & Goode, A. (1995). Unconscious knowledge of artificial grammars is applied strategically. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 21(5), 1322–1338.
doi:10.1037//0278-7393.21.5.1322
- Fleming, S. M. (2017). HMeta-d: Hierarchical Bayesian estimation of metacognitive efficiency from confidence ratings. *Neuroscience of Consciousness*, 3(1).
doi:10.1093/nc/nix007
- Fleming, S. M., & Daw, N. D. (2017). Self-evaluation of decision-making: A general Bayesian framework for metacognitive computation. *Psychological Review*, 124(1), 91–114. doi:10.1037/rev0000045
- Fleming, S. M., & Lau, H. C. (2014). How to measure metacognition. *Frontiers in Human Neuroscience*, 8. doi:10.3389/fnhum.2014.00443
- 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.
- Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. Signal detection theory and psychophysics. Oxford, England: John Wiley.
- Harrell, F. (2015). *Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis*. Springer.
- Jang, Y., Wallsten, T. S., & Huber, D. E. (2012). A stochastic detection and retrieval model for the study of metacognition. *Psychological Review*, 119(1), 186–200.
doi:10.1037/a0025960
- Johnson, N. L., Kotz, S., & Balakrishnan, N. (1994). *Continuous Univariate Distributions* (2nd). John Wiley & Sons.

- Kiani, R., & Shadlen, M. N. (2009). Representation of Confidence Associated with a Decision by Neurons in the Parietal Cortex. *Science*, 324(5928), 759–764.
doi:10.1126/science.1169405
- Maniscalco, B., & Lau, H. (2012). A signal detection theoretic approach for estimating metacognitive sensitivity from confidence ratings. *Consciousness and Cognition*, 21(1), 422–430. doi: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). DOI: 10.1007/978-3-642-45190-4_3. Springer Berlin Heidelberg.
- Maniscalco, B., & Lau, H. (2016). The signal processing architecture underlying subjective reports of sensory awareness. *Neuroscience of Consciousness*, 2016(1).
doi:10.1093/nc/niw002
- McCullagh, P., & Nelder, J. (1989). *Generalized Linear Models, Second Edition*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. bibtex: mccullagh1989generalized LCCB: 99013896. Taylor & Francis.
- Metcalfe, J., & Shimamura, A. P. (Eds.). (1996). *Metacognition: Knowing about Knowing* (F First Paperback Edition edition). Cambridge, Mass: A Bradford Book.
- Morey, R. D., Pratte, M. S., & Rouder, J. N. (2008). Problematic effects of aggregation in z ROC analysis and a hierarchical modeling solution. *Journal of Mathematical Psychology*, 52(6), 376–388. doi:10.1016/j.jmp.2008.02.001
- Persaud, N., McLeod, P., & Cowey, A. (2007). Post-decision wagering objectively measures awareness. *Nature Neuroscience*, 10(2), 257–261. doi:10.1038/nm1840
- R Core Team. (2016). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from <https://www.R-project.org/>

- Ramsøy, T. Z., & Overgaard, M. (2004). Introspection and subliminal perception. *Phenomenology and the Cognitive Sciences*, 3(1), 1–23.
doi:10.1023/B:PHEN.0000041900.30172.e8
- Rausch, M., & Zehetleitner, M. (2017). Should metacognition be measured by logistic regression? *Consciousness and Cognition*, 49, 291–312.
doi:10.1016/j.concog.2017.02.007
- Rounis, E., Maniscalco, B., Rothwell, J. C., Passingham, R. E., & Lau, H. (2010). Theta-burst transcranial magnetic stimulation to the prefrontal cortex impairs metacognitive visual awareness. *Cognitive Neuroscience*, 1(3), 165–175.
doi:10.1080/17588921003632529
- Sandberg, K., Timmermans, B., Overgaard, M., & Cleeremans, A. (2010). Measuring consciousness: Is one measure better than the other? *Consciousness and Cognition*, 19(4), 1069–1078. doi:10.1016/j.concog.2009.12.013
- Tutz, G., & Hennevogl, W. (1996). Random effects in ordinal regression models. *Computational Statistics & Data Analysis*, 22(5), 537–557.
doi:10.1016/0167-9473(96)00004-7
- Wright, D. B., Horry, R., & Skagerberg, E. M. (2009). Functions for traditional and multilevel approaches to signal detection theory. *Behavior Research Methods*, 41(2), 257–267. doi:10.3758/BRM.41.2.257

Figure 1. The SDT model with ordinal responses A . The black / blue curve is the evidence density under noise ($S = 0$) while gray / red is the evidence distribution under signal ($S = 1$). X is the underlying evidence governing the responses and A^* governs the concatenated, ordinal answers A .

Figure 2. Interrelationship of signal distributions in some common signal models. In the two-dimensional model, the simultaneous distribution of X and Y is assumed to follow a bivariate normal distribution (note that the variance structure is too general to be identifiable when used to generate latent variables in a threshold model). The other models are obtained as special cases by applying standard results for the multivariate normal distribution. For example, letting the components be independent and equivariable we would obtain the Independent model. Alternatively, taking the mean parameters to be the same and the variance structure additive yields the hierarchical model, of which SDT is the special case for $X = Y$ occurring as ζ goes to zero.

Figure 3. Distribution of Perception Awareness Scale ratings for incorrect and correct assertions.

Figure 4. Intuitive definition of meta- d : The left plot depicts the SDT model.

Conditioning on a “signal” response ($X > \theta$), we obtain the right hand plot where the dashed lines are those expected if the type 2 decision is based on the SDT. If we conversely start from the signal-specific type 2 data in the right-hand plot we may ask which SDT model has generated these responses (the fully drawn lines). The signal-specific meta- d , d_S , is then the sensitivity in this SDT model. Note that the fully drawn lines in the right-hand plot is the density of Z_S in (12).

Figure 5. Response scales in the meta-SDT model for $L = 6$.

Figure 6. (A)-(C) Simulations to illustrate the effect of heterogeneity on estimates of sensitivities. For each of the three parameters, the left column is simulations from a scenario with no subject heterogeneity ($\zeta = 0$) while the middle column has an intermediate between subject variation ($\zeta = 0.5$) and the right column has a high degree of heterogeneity ($\zeta = 0.75$). *(A)* depicts estimates of d , *(B)* estimates of d_N and *(C)* shows estimates of d_S . *(D)* Estimates of group effect D for three levels of heterogeneity. For all graphs, the points represent simulation means and error bars are twice the empirical standard deviation of the simulated estimates (i.e. the observed standard error), while the horizontal dashed line is the true value.

Appendix A

ML Estimation in probit regression

In the present appendix we derive the maximum likelihood estimators of a probit model with a binary explanatory variable. The results may be standard especially to readers familiar with the theory of exponential families, but we take a more hands on approach and study the likelihood directly.

Model and likelihood. Suppose that we have N Bernoulli observations

$$Y_i \sim b(1, p_i), \text{ for } i = 1, \dots, N, \quad (43)$$

and say that we wish to model the probability parameter in a GLM with probit link, that is

$$\Phi^{-1}(p_i) = a + bx_i, \quad (44)$$

where $x_i \in \{0, 1\}$. Write $\eta_i = a + bx_i$. We wish to estimate a and b based on data such as that given in Table A1.

		y	
		0	1
x	0	n_{00}	n_{01}
	1	n_{10}	n_{11}

Table A1

Data for the Bernoulli observations Y and the binary, explanatory variable x .

The log likelihood of the model is given by

$$\begin{aligned} l(a, b) &= \sum_{i=1}^N y_i \log \Phi(\eta_i) + \sum_{i=1}^N (1 - y_i) \log (1 - \Phi(\eta_i)) \\ &= n_{00} \log (1 - \Phi(a)) + n_{01} \log \Phi(a) \\ &\quad + n_{10} \log (1 - \Phi(a + b)) + n_{11} \log \Phi(a + b). \end{aligned} \quad (45)$$

Likelihood equations. Differentiating the expression in (45) we obtain

$$\begin{aligned} \frac{\partial}{\partial a} l(a, b) &= n_{01} \frac{\phi(a)}{\Phi(a)} - n_{00} \frac{\phi(a)}{1 - \Phi(a)} \\ &\quad + n_{11} \frac{\phi(a+b)}{\Phi(a+b)} - n_{10} \frac{\phi(a+b)}{1 - \Phi(a+b)}. \end{aligned} \quad (46)$$

$$\frac{\partial}{\partial b} l(a, b) = n_{11} \frac{\phi(a+b)}{\Phi(a+b)} - n_{10} \frac{\phi(a+b)}{1 - \Phi(a+b)}. \quad (47)$$

It is straightforward to verify that the likelihood equation $\frac{\partial}{\partial b} l(a, b) = 0$ is satisfied for

$$b = \Phi^{-1} \left(\frac{n_{11}}{n_{10} + n_{11}} \right) - a. \quad (48)$$

Taking the right hand side of (48) as $\hat{b}$ and plugging this into (46) we see that the corresponding equation for a is

$$\begin{aligned} \phi(a) \frac{n_{01} - (n_{01} + n_{00})\Phi(a)}{\Phi(a)(1 - \Phi(a))} \\ + \phi(a+b) \frac{n_{11} - (n_{11} + n_{10})\Phi(a+b)}{\Phi(a+b)(1 - \Phi(a+b))} = 0. \end{aligned} \quad (49)$$

Noting that

$$n_{11} - (n_{11} + n_{10})\Phi(a+b) = 0, \quad (50)$$

the likelihood equation for a reduces to

$$\phi(a) \frac{n_{01} - (n_{01} + n_{00})\Phi(a)}{\Phi(a)(1 - \Phi(a))} = 0, \quad (51)$$

which is seen to be solved for

$$a = \Phi^{-1} \left(\frac{n_{01}}{n_{00} + n_{01}} \right). \quad (52)$$

In conclusion,

$$\begin{aligned} \hat{a} &= \Phi^{-1} \left(\frac{n_{01}}{n_{00} + n_{01}} \right). \\ \hat{b} &= \Phi^{-1} \left(\frac{n_{11}}{n_{10} + n_{11}} \right) - \hat{a}. \end{aligned} \quad (53)$$

Appendix B

The truncated normal distribution

Recall that a stochastic variable V is said to follow a truncated normal distribution with lower threshold a , upper threshold b , mean μ and variance σ^2 if

$$\mathbb{P}(V \leq v) = \frac{\Phi\left(\frac{v-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)}{\Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)}, \quad v \in [a, b]. \quad (54)$$

For such a variable we write $V \sim trN_a^b(\mu, \sigma^2)$. Note that if $V \sim trN_a^b(\mu, \sigma^2)$, then $V \sim U \mid a \leq U \leq b$ for some variable $U \sim N(\mu, \sigma^2)$ which motivates the distribution's name. We have that

$$\mathbb{E}[V] = \mu + \sigma \frac{\phi\left(\frac{b-\mu}{\sigma}\right) - \phi\left(\frac{a-\mu}{\sigma}\right)}{\Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)}, \quad (55)$$

see for example Johnson, Kotz, and Balakrishnan, 1994.

Since $trN_{-\infty}^{\infty}(\mu, \sigma^2) \sim N(\mu, \sigma^2)$, we use the notational convention of omitting a threshold when it is trivial. More precisely, we write $trN_a(\mu, \sigma^2)$ for $trN_a^{\infty}(\mu, \sigma^2)$ and $trN^b(\mu, \sigma^2)$ for $trN_{-\infty}^b(\mu, \sigma^2)$.

Appendix C

Additional simulations

We simulated 100 cases as in Simulation study 1, but now only varying the between subject variation as $\zeta = 0$ and $\zeta = 0.75$. For each simulation, we applied the meta SDT regression model as in Simulation study 1. Additionally, we analysed the data using the response-specific, group level analysis from the *HMeta-d* toolbox. All priors and MCMC initialisation parameters were unaltered from the “out of the box” settings in the toolbox. The results are shown in Figure C1.

The results show that in this specific case the MLE outperforms the HMeta-d method both in terms of bias and efficiency (the width of the confidence intervals), especially at moderate sample sizes. It seems that the HMeta-d model fares better in the presence of heterogeneity compared to the case $\zeta = 0$ which may be explained by the presence of hyperpriors in the model which are in a sense unused when $\zeta = 0$. As already stated, this should be interpreted very cautiously as it is the result under one specific set of assumptions. In particular, as the HMeta-d model relies on a number of informative priors one would expect the performance to depend strongly on their validity in the simulated data.

Figure C1. (A)-(C) Simulations to tentatively compare to the performance of the *HMeta-d* toolbox. For each of the three parameters, the left column is simulations from a scenario with no subject heterogeneity ($\zeta = 0$) and the right column has a high degree of heterogeneity ($\zeta = 0.75$). (A) estimates of d , (B) estimates of d_N and (C) estimates of d_S . For all graphs, the points represent simulation means and error bars are twice the empirical standard deviation of the simulated estimates (i.e. the observed standard error), while the horizontal dashed line is the true value.