



The `bhsdtr` package: a general-purpose method of Bayesian inference for signal detection theory models

Borysław Paulewicz¹ · Agata Blaut²

Published online: 24 March 2020
© The Psychonomic Society, Inc. 2020

Abstract

We describe a novel method of Bayesian inference for hierarchical or non-hierarchical equal variance normal signal detection theory models with one or more criteria. The method is implemented as an open-source R package that uses the state-of-the-art Stan platform for sampling from posterior distributions. Our method can accommodate binary responses as well as additional ratings and an arbitrary number of nested or crossed random grouping factors. The SDT parameters can be regressed on additional predictors within the same model via intermediate unconstrained parameters, and the model can be extended by using automatically generated human-readable Stan code as a template. In the paper, we explain how our method improves on other similar available methods, give an overview of the package, demonstrate its use by providing a real-study data analysis walk-through, and show that the model successfully recovers known parameter values when fitted to simulated data. We also demonstrate that ignoring a hierarchical data structure may lead to severely biased estimates when fitting signal detection theory models.

Keywords Signal detection theory · Bayesian inference · Hierarchical models

Introduction

Many tasks used in psychology studies are essentially classification tasks. In a memory study, for example, participants may be required to decide if a given test item is old or new, or, in a perceptual study, an object may be either a letter or a digit. If a task requires classification, it is always possible that conclusions based on accuracy or percent correct are invalid because the ability to discriminate between stimulus classes (i.e., sensitivity) is confounded with bias, which is a tendency to classify stimuli as belonging to a particular class. In principle, any effect that manifests itself in differences in classification accuracy may reflect differences in sensitivity, bias, or both.

Signal detection theory (Peterson et al., 1954; Tanner & Swets, 1954) provides a simple and popular solution to this common problem: according to Google Scholar, the seminal

book by Green and Swets (1966) which introduced SDT to psychology researchers was cited more than 15,000 times before the year 2020. Despite the fact that the theory solves a common and important problem and is even described in cognitive psychology handbooks, there are reasons to believe that it may be heavily underutilized (Stanislaw & Todorov, 1992).

Because the SDT model is non-linear, variability in its parameters due to factors such as participants or items has to be accounted for. When they are not accounted for, e.g., by aggregating data across participants or items, the estimates of SDT parameters are asymptotically biased (Rouder & Lu, 2005). As we explain later in this paper, none of the available methods of inference for hierarchical SDT models that we are aware of addresses this problem in its full generality, meaning that none of the available methods allow for fixed and random effects in both the sensitivity and the criteria parameters while restricting the parameters in accordance with SDT assumptions. Later in the paper; we explain why, in our view, the `bhsdtr` package for R (R Core Team, 2017), which we have made publicly available at <https://github.com/boryspaulewicz/bhsdtr>, provides a correct implementation of the general hierarchical linear regression structure defined on SDT parameters. This package repository also contains the

✉ B. Paulewicz
bpaulewicz@swps.edu.pl

¹ Faculty of Psychology in Katowice, SWPS University of Social Sciences and Humanities, Katowice, Poland

² Institute of Psychology, Jagiellonian University, Krakow, Poland

annotated source code that was used to perform all the analyses and produce all the figures presented in this paper.

In what follows, after introducing the most common version of the SDT model, we describe its generalization, which can accommodate data from rating experiments. Note that our brief introduction to the SDT theory is meant as a refresher—the reader interested in a more comprehensive treatment is advised to consult any of the three most popular contemporary books on this subject, i.e., McNicol (2005), Wickens (2002), or Macmillan and Creelman (2004), listed here in order of increasing mathematical sophistication.

After describing the generalized SDT model, we explain why, if a method of inference for SDT models were to be of general use in psychology studies, it is important that it is based on a model equipped with a correct hierarchical linear regression structure. The `bhsdtr` package meets this requirement thanks to a novel parametrization; we describe this and explain how reliance on relatively standard parametrizations leads to problems in the three other available implementations. We end the first part of this paper with a formal definition of the model as implemented in `bhsdtr`. The second part contains an overview of the package and a tutorial in which we demonstrate how to use our method in practice, as well as a demonstration of bias resulting from ignoring the effects of grouping factors.

Equal variance normal signal detection theory model with additional criteria

According to signal detection theory, each stimulus in a classification task gives rise, by some unspecified cognitive process, to a unidimensional internal evidence value s sampled from a distribution that depends on the stimulus class. For historical reasons, the two stimulus classes are often referred to as “noise” and “signal”, and task performance is described in terms of hits (when the participant

responds “signal” to signals), correct rejections (responding “noise” to noise stimuli), omissions (responding “noise” to signals), and false alarms (responding “signal” to noise stimuli), but this terminology is appropriate only when the model is applied to tasks that require detection, which is far from always being the case. To emphasize the general applicability of SDT models, we will use the classical terminology only at the beginning of our paper, and later we will mostly talk about two arbitrary kinds or classes of stimuli, indexed by the numbers 1 and 2.

In the most widely used version of the model, shown in Fig. 1, the two evidence distributions are normal with the same variance, which is usually fixed at unity to make the model identifiable. The distance d' between the means of the evidence distributions represents sensitivity. Because normal distributions are unbounded, s is always ambiguous, and so a criterion c placed on the evidence axis has to be used to reach a binary decision. A participant is assumed to decide that a stimulus belongs to the first class (e.g., an old item or “noise”) if $s < c$, or that it belongs to the second class (e.g., a new item or “signal”) if $s \geq c$. The location of the decision criterion may be interpreted in terms of classification or response bias.

Perhaps the simplest way of using this model is to fit it to observed response counts and use the estimated d' values in place of the percent correct ($p(c)$) scores; if the model is correct, the resulting performance measure is not contaminated by bias. The model may not be correct, which in our view is the most important reason to focus more on the generalized version shown in Fig. 2 below.

This generalized model is applicable to studies in which participants are asked to rate their binary classification decisions on confidence or some other performance- or stimulus-related dimension. The ratings and the binary classification decisions can be provided together (e.g., “I am almost certain that it was a digit”), or in an arbitrary order.

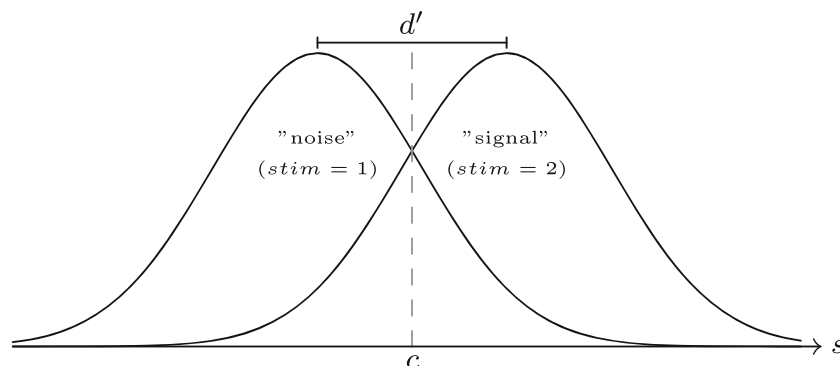


Fig. 1 Equal-variance normal signal detection theory model. The s -axis represents the evidence space; the left-most (right-most) density curve represents the distribution of internal evidence for the noise (signal) stimuli (stimulus class 1 and 2, respectively); c is the decision criterion, and the distance d' between the distribution means represents sensitivity

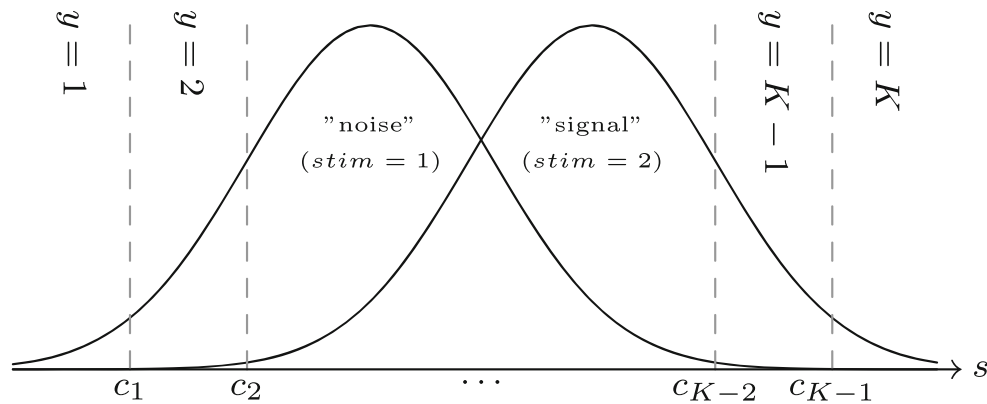


Fig. 2 Equal-variance normal signal detection theory model with additional criteria. This is the same model as in Fig. 1, but there is more than one criterion. The response y is a single number that combines the binary classification decision and the rating. Larger y values correspond to increasing certainty that the stimulus is of type “signal” (stimulus class 2)

Ratings are accommodated by introducing additional criteria and modeling a combined response y , which represents both the binary classification decision and rating as a single number. The value of y increases with the strength of evidence in favor of the second stimulus class. For example, if confidence is rated on a four-point scale, then $y = 1$ when a participant decides that a stimulus belongs to the first class with certainty 4, and $y = 5$ when a participant decides that a stimulus belongs to the second class with certainty 1. More formally, if K is the number of possible combined responses, then a participant is assumed to give response k if $s \in (c_{k-1}, c_k]$, where c are the decision criteria, with c_0 and c_K fixed at $-\infty$ and $+\infty$, respectively.

There is a good reason to collect the ratings and use the generalized SDT model from Fig. 2, even when neither the ratings nor the placement of criteria are relevant to the research problem. When $K = 2$ (no ratings), the SDT model fits perfectly (it is saturated), regardless of whether it is a reasonably good approximation to reality, because the data and the model have the same dimensionality. This makes the generalization to the $K > 2$ case particularly important, as the formal assumptions of the model (e.g., equal or unequal variance) can only be tested when $K > 2$.¹

The model is usually tested by comparing the empirical and the implied receiver operating characteristic (ROC) curves, which represent the relationship between the hit rate ($p(H)$), the false-alarm rate ($p(F)$) and d' . An example of the empirical ROC curve is shown later in the paper (see for example Fig. 6). Each curve in the implied ROC space represents all the possible pairs of $p(H)$ and $p(F)$

values that correspond to some unique d' value. If the observed points corresponding to the pairs of hit and false-alarm rates for the same d' value do not lie on the same implied isosensitivity curve, the distributional assumptions may be false. This may be easier to see when using the zROC plot, which shows the relationships between $z(H)$ and $z(F)$, because the z transform turns the original ROC curves into straight lines. In particular, if the equal variance normal SDT model is true then all the zROC slopes equal 1.

A false SDT model cannot be trusted to serve its main purpose of deconfounding sensitivity from bias and it is in fact not uncommon that results are obtained that seem to falsify an SDT model; for example, the survey studies by Swets (1986) and Swets and Pickett (1982) indicate that the slopes of empirical zROC curves are often different from unity. That is why we believe that the additional complexity introduced by collecting and modeling the ratings is more than justified, unless—for some reason—it defies the purpose of the study.

As we later explain, in a typical case there are also good reasons to complicate the model even further by supplementing it with a regression structure. When SDT models are used in psychology studies, researchers are usually interested not in the values of the SDT parameters themselves but in the relationships between SDT parameters and additional measured or manipulated variables; a good example is the dependence of d' on stimulus strength. Also, in a great majority of psychology studies in which classification tasks are used, the data have a hierarchical structure, i.e., there are repeated measures for each participant or item, and participants or items are only samples from the target population. A general-purpose method of inference for SDT models should accommodate both kinds of situations.

¹In contrast to the formal assumptions, a psychological interpretation of the SDT parameters can be tested even when ratings are not available, e.g., by means of selective influence (Sternberg, 2001)

The importance of a correct hierarchical regression structure

If data have a hierarchical structure, but variability due to participants, items, or other grouping factors is not accounted for, estimates of average (fixed) effects are not guaranteed to be unbiased and conclusions are not guaranteed to generalize to the target population.

The not uncommon practice of analyzing data aggregated over grouping factors represents an extreme case of ignoring hierarchical data structure. The invalidity of this approach in the context of SDT was clearly illustrated with the results of simulational studies by Morey et al. (2008); however, strictly speaking, such demonstrations are irrelevant to proving the invalidity in question. Because SDT is a non-linear model, by definition, when estimates of its parameters are based on data aggregated over sampled factors—e.g., d' estimated for hits and false alarms averaged over participants or items—the expected values of these estimates (e.g., what the calculated d' actually estimates) are not in general true population averages (true d' in some condition). In fact, such estimates are asymptotically biased, which means that increasing the sample size will not make the bias disappear and any inference about a target population based on such estimates is simply not valid. To give a concrete example, consider two unbiased participants, one with $d'_1 = 2$ and one with $d'_2 = 4$. Their expected average accuracy is given by $(\Phi(d'_1/2) + \Phi(d'_2/2))/2 = .91$, which corresponds to $d' = 2.68$, whereas their true average d' is 3.

Aggregating data over grouping factors leads to two kinds of estimate bias. One is a bias in point estimates of d' and c parameters, as illustrated by the last example. The other is a bias in interval estimates. Ignoring a hierarchical data structure in an SDT analysis is analogous to using a fixed effects ANOVA for repeated measures data, which is a major violation of the modeling assumptions. When a hierarchical data structure is ignored, the data are no longer independent given the model because the data points for the same level of a grouping factor are informative about each other. For the same reason, it is important to model the random effects' correlations when these may be non-negligible. In the worst case, because of non-linearity the point estimates will systematically differ from the true values and, because the variability is underestimated, the interval estimates will be too narrow, leading to conclusions that are both invalid and apparently strongly supported by the data.

Except for toy examples like the one we have described earlier, we are not able to say much about the estimate bias magnitude for the d' or c parameters in quantitative terms because predicting estimate bias magnitude is in general a difficult problem. However, using a real dataset we will

demonstrate later in the paper how overly aggregating the data may easily lead to invalid conclusions.

As we repeatedly stress in this paper, aggregation is not the only way of ignoring a hierarchical data structure. Sometimes *non*-aggregated data are analyzed by using separate estimates for every participant \times item \times condition combination, but uncertainty due to distributions of participant or item effects is not accounted for by means of a hierarchical model structure. In such cases, conclusions—at least with respect to the uncertainties in estimates of population-average (fixed) effects—are guaranteed to be valid only for the given sample, not the target population.

Furthermore, when the SDT parameters are estimated separately for each participant and condition, and only later regressed on predictors of interest, a number of additional issues may arise. Firstly, the standard errors or credible intervals associated with the regression coefficients do not reflect the uncertainty in the SDT parameter estimates because the latter are treated as mere data points. The precision of parameter estimates often varies between participants, items, or conditions, but when the estimates are treated as data points, no use is made of this information. Secondly, regressing parameters on numerical predictors makes their estimates dependent on the common regression structure, and so also on each other, which can improve the quality of the estimates, just as assuming that random effects are samples from a common distribution may improve their estimates.

The three most popular contemporary books on SDT modeling that we have mentioned in the introduction differ in how much emphasis their authors place on the issue of point and interval estimate bias. For example, McNicol (2005) notes that the d' estimates based on aggregated hit and false-alarm rates are biased and recommends aggregating the z transformed rates: this does solve the problem of bias in point (but not interval) estimates of group average d' and c in the binary classification case, but not when the ratings are collected; it also does little in the way of accounting for the uncertainty associated with the distribution of (and correlation between) random effects, since this latter problem can only be solved by supplementing an SDT model with a hierarchical regression structure.

Although almost every example of SDT analysis that Macmillan and Creelman (2004) consider in their book is based on data aggregated over participants, they do provide an analysis of estimate bias based on the results obtained by Macmillan and Kaplan (1985) and Hautus (1997). However, this analysis is restricted to a simple case of a small number of participants and no regression structure, which, as the authors themselves admit, is a limited scenario. When considering more realistic situations in which *the*

effects in d' or c are estimated, the authors recommend using a generalized linear regression modeling framework, as described by DeCarlo (1998).

Wickens (2002) seems to be the only of these books that emphasizes both the point estimate bias and the interval estimate bias resulting both from averaging and from separating the hierarchical regression analysis from SDT model fitting. Wickens also shows that that the latter kind of bias may be severe in certain simple situations.

Hierarchical signal detection theory in a constrained parameter space

Both d' and c have the virtue of being directly interpretable in terms of sensitivity and bias. However, d' is assumed to be non-negative and when there is more than one criterion the elements of the c vector are order restricted ($c_{i+1} > c_i$).

The non-negativity of d' seems to deserve some explanation, since the authors who write about SDT models differ in how clear they are on this issue. For example, Macmillan and Creelman (2004) state this assumption explicitly, Wickens (2002) describes d' as a measure of distance and, when considering negative d' values, interprets them as arising from sampling error or response reversal, whereas, to our knowledge, McNicol (2005) does not discuss the possibility of negative d' values at all.

In signal detection theory d' is a measure of distance and as such is a non-negative quantity by definition (Luce, 1963). Usually, the true negative d' values are not like non-negative d' values only smaller—they are *qualitatively* different. To see why, observe that participants in a typical study cannot perform the task below chance level unless—for some reason which is outside the scope of the SDT model, such as misunderstanding the task instructions—they reverse their responses. Consequently, it makes no sense to say that $d' = -1$ represents a sensitivity which is lower by two standard deviations than the sensitivity represented by $d' = 1$. A more natural interpretation is that in both cases the sensitivity is exactly the same, and there is some other reason, such as response reversal, for the difference in sign.

Macmillan and Creelman (2004) as well as Wickens (2002) seem to agree that the most common reason why negative d' values are observed is that some participants have true near-zero positive d' and their observed hit rates are lower than their observed false-alarm rates purely because of the sampling error. However, the prior on d' as well as the distribution of d' random effects represent the uncertainty in the *true* d' values, whereas the sampling error is handled by the distributional assumptions of the SDT model itself.

If the reason that the negative d' values are observed is either sampling error or response reversal in a small number of participants, then the normal prior on d' does not correctly express this possibility because normal prior on d' represents the assumption that there is a natural continuity of true d' values that extend below and above zero. A more appropriate way to model rare cases of response reversal would be to represent random d' effects as an uneven (assuming response reversal is an exception) mixture of two distributions of participants, which are essentially the same except that for one of the distributions the values of the response variable are reversed. On the other hand, in some situations true negative d' values may even be common; for example, certain experimental conditions may systematically cause the participants to reverse their responses, but these are special cases which require special treatment; in particular, such situations call for a non-trivial generalization of the SDT model.

Since, due to the sampling error, $p(H) < p(F)$ is not impossible for an arbitrary true negative d' value, an unbounded prior on d' inevitably forces the posterior distribution to have a non-zero mass on the negative d' values. This makes all the posterior samples of all the model parameters that correspond to the negative posterior d' samples problematic. Whether this is an important problem in a particular case depends on a number of factors, the main two being how informative the data are about the true d' value and how close the true d' is to 0. Regardless of how large the dataset is as a whole, when there is substantial variability in d' or d' effects between the participants, the participant-specific estimates become the data points since the fixed effects are estimated as averages of random effects, and the number of participants is often much more limited than the number of raw data points that can be collected; This limits the amount of information available in typical datasets. Using a truncated normal distribution as a prior for d' does not solve this problem either because it does not allow for unbounded fixed and random effects; we are also not aware of any reason why the true d' random effects could be approximated well by a truncated normal distribution.

All this leads to the conclusion that the constraints on d' (positivity) are as important as the constraints on c (ordering), but these constraints make it impossible to supplement an SDT model with a hierarchical linear regression structure. Such a structure can only be defined on unconstrained parameters because in hierarchical linear regression (1) random effects are assumed to be normally distributed and normal distribution is unbounded, and (2) effects such as differences between conditions or regression slopes are allowed to assume arbitrary real values. The only way to solve this problem is to re-parametrize the model so

that the parameters are no longer constrained, but the model is essentially the same.

Limitations of available implementations of hierarchical SDT models with ratings

In the following summary of two hierarchical SDT implementations, we provide examples of the problems that may arise when the SDT model is not correctly reparametrized. One is the Gibbs sampler proposed by Morey et al. (2008) and the other is the Hierarchical Meta- d' model (HMeta- d') proposed by Fleming (2017). After describing the problems associated with those two software implementations, we will briefly explain how our method compares to what the `brms` package has to offer.

The HMeta- d' model is a hierarchical version of the meta- d' model (Maniscalco & Lau, 2012), which in turn is a generalization of the SDT model that allows for a separate “meta-sensitivity” to account for possible discrepancies between a binary stimulus classification (referred to as a type 1 task) and the associated rating task (referred to as a type 2 or meta-cognitive task). We consider HMeta- d' here because it reduces to an SDT model with ratings when the type 1 and type 2 sensitivities are equal.

The Gibbs sampler created by Morey et al. (2008) allows for at most two grouping factors to have independent normally distributed random effects on the evidence distribution means. Unlike d' , each evidence distribution mean considered in isolation is an unconstrained parameter, but the mean of the second evidence distribution is by definition greater than ($d' > 0$) or equal to ($d' = 0$) the mean of the first. The model does not enforce the non-negativity of d' because this would introduce non-conjugate priors, which are problematic in a Gibbs sampler. The outermost criteria are fixed at 0 and 1, and the ordering restriction is enforced by assuming that the likelihood is 0 whenever $c_{i+1} \leq c_i$. As the authors explain, because a grouping factor can have independent random effects on the evidence distribution means, it can have an effect on all the criteria: shifting both means by the same amount in the same direction is equivalent to keeping the sensitivity intact, while shifting the criteria relative to the evidence distributions. However, the *individual* criteria cannot be affected differently by the same grouping factor.

In HMeta- d' the hierarchical structure is limited to normally distributed random intercepts of one grouping factor. In this model the d' parameter is also allowed to assume negative values, but the most problematic aspect of this implementation is again the representation of the criteria. For reasons that are outside the scope of this paper, in the HMeta- d' model the main criterion is interpreted as qualitatively different from the rest of the criteria. The main

criterion, as a random effect, is assumed to be a sample from a normal distribution, whereas the criteria above (below) the main criterion are assumed to be samples from a normal distribution which is bounded below (above) by the value of the main criterion. The mean of the distribution of the upper criteria random effects is the same as the mean of the lower criteria random effects, only the sign is reversed. These unordered “proto-criteria” are sorted to obtain the actual criteria values, but because sorting is not injective the space of the actual criteria is only loosely related (i.e., not isomorphic) to the space of the unrestricted criteria vectors that are associated with the hierarchical structure. In this way the HMeta- d' model enforces some but not all the necessary order constraints, and it only accounts in a limited way for the variability in the criteria due to the grouping factors.

Some extensions of the SDT model can be fitted correctly using the excellent `brms` package, as described in Bürkner (2017) or in a tutorial on ordinal models by Bürkner and Vuorre (2019). The `brms` package is a flexible tool that shares some deep design similarities with our method. Both our package and `brms` belong to a growing family of software tools that aim to provide a somewhat simplified and domain- or application-specific interface for one of the general purpose Bayesian inference engines, in this case the Stan modeling language (Carpenter et al., 2016). The `brms` is a highly flexible, well-documented package that offers an elegant interface for fitting a large class of hierarchical models. Among the models that can be fitted using this package are the ordinal regression models. The hierarchical SDT model with ratings is essentially a generalization of hierarchical ordinal regression since ratings are an ordinal-scale variable.

However, that does not mean that a general hierarchical SDT model with ratings can be fitted using the `brms` package. There are three categorical distributions available at present in `brms`, each with its own set of link functions: the cumulative model, the adjacent category model and the sequential model. As the author of the package explains (Bürkner, 2017), the only model that respects the ordering of the thresholds is the cumulative model, but in the cumulative model—just like in the Morey et al. (2008) model—the predictors can only have constant effects across categories. Moreover, just like in the other two implementations, in the cumulative model d' is unbounded.

None of the three available methods that we consider here forces the d' parameter to be non-negative. The `brms` package is the only one of the three methods that allows for an arbitrary hierarchical linear regression structure with possibly correlated random effects in an SDT model with ratings. The Morey et al. (2008) model allows for at most two grouping factors, but it cannot account for random effects' correlations. The HMeta- d' model allows for at

most one grouping factor, but it can only be associated with variability in the intercept, which means that the within-subject (or within-item) effects cannot be modeled at all out of the box.

If a participant has one criterion shifted to the right, the criterion above it will usually also be shifted to the right, and so random effects in criteria are likely to be strongly correlated. Lastly, none of the three methods are able to account for the fact that participants or items may differ not only in by how much and in what direction all of their criteria are shifted relative to the population mean, but also in the relative positions of individual criteria; In the cumulative model as implemented in the `brms` package and in the Morey et al. (2008) model, random criteria effects are constant across categories. In the HMeta- d' model the variability in the criteria due to the grouping factors seems to be reduced to the average distance of the unordered proto-criteria from the middle criterion. It follows that when an SDT model is fitted to the data from a rating experiment, the data cannot be assumed to be independent given any of the three models. As pointed out by Wickens (2002), violation of the assumption of the independence of data given the model may lead to severely biased interval estimates.

Note that the estimates of the SDT parameters are *interdependent*, meaning that point or interval estimate bias in one parameter may lead to point or interval estimate bias in all the other parameters. Conversely, if an estimate of one parameter is improved, the estimates of the other parameters may improve as well. In the three available methods of fitting hierarchical SDT models with ratings discussed so far, there are independent reasons for point or interval estimate bias in every SDT parameter considered in isolation. Such biases are difficult to predict theoretically and difficult to detect because obtaining empirical evidence of their existence or magnitude is only possible when the true model is known.

We should clarify that not all forms of bias are absent when using the SDT model as implemented in the `bhsdtr` package. In a Bayesian model the posterior distribution is influenced by the data and the priors, and the prior-induced bias is often unavoidable. Since in the `bhsdtr` package the default priors are weakly informative, they may have a non-negligible impact when the number of data points is small and the variance of the priors is not large. Another concern is that when the true d' values are near-zero, our model will automatically infer that the true d' value is non-negative, thus discarding part of the error, but only the negative part. On the other hand, if the true d' is zero, given enough data the Highest Posterior Density (HPD) interval estimates can be expected to include zero. This is also an efficient solution, given the positivity assumption, since it only discards part of the error, and we see no reason why it would eliminate the asymptotic unbiasedness. This feature also

does not preclude the researcher from choosing a prior that favors the near-zero values. When the positivity assumption is acceptable, which usually seems to be the case according to Wickens (2002) and Macmillan and Creelman (2004), the alternative leads to the problems that we have already described. In any case, whenever there is reason to believe that the chosen priors lead to misleading conclusions, the researcher can perform a sensitivity analysis.

Hierarchical signal detection theory in an unconstrained parameter space

The general hierarchical linear regression structure can be defined on SDT parameters only if the latter are derived from unconstrained parameters. In the `bhsdtr` package, d' is derived from $\delta = \ln(d')$, thus random effects on d' can be modeled by assuming that δ is normally distributed. Modeling the logarithm of d' has an additional advantage: as noted by Macmillan and Creelman (2004), d' has ratio-scaling properties and it makes sense to say that one d' value is twice as large as another. Because both d' and δ have a natural interpretation in terms of the differences in sensitivity but they are related by a non-linear transformation, inspecting both the d' values and the δ values may reveal that an interactive effect is illusory.²

The problem of representing the criteria by unconstrained parameters can be solved by mapping the R^{K-1} space of unconstrained criteria vectors to the K dimensional probability simplex space using the softmax function, and mapping the simplex space to the space of order-restricted criteria vectors by means of the inverse normal CDF:

$$c_i = \Phi^{-1}\left(\frac{\sum_{k=1}^i (e^{\gamma_k})}{\sum_{j=1}^K (e^{\gamma_j})}\right) \quad (1)$$

where Φ is the CDF of the standard normal distribution and $\boldsymbol{\gamma} \in R^K$, with γ_K fixed at 0 for identifiability. The idea is illustrated in Fig. 3 below:

Note that the normal distribution centered at the midpoint is merely a mapping device, not a third evidence distribution, and that, for reasons that will soon become clear, it is wider than the two evidence distributions. The mapping expressed by Eq. 1 is an isomorphism between the R^{K-1} space and the space of order-restricted criteria vectors. Its inverse is given by $\gamma_i = \ln\left(\frac{\int_{c_{i-1}}^{c_i} f(s) ds}{\int_{c_{K-1}}^{\infty} f(s) ds}\right)$, where f is the standard normal probability density function. The elements of the $\boldsymbol{\gamma}$ vector correspond to relative distances between pairs

²Whenever a nonzero additive effect is observed on the original (logarithmic) scale of the dependent variable, it will look like an interaction on the logarithmic (original) scale. In such cases, the effect of one variable can be predicted without knowledge of the other variable, and the observed interaction may be illusory.

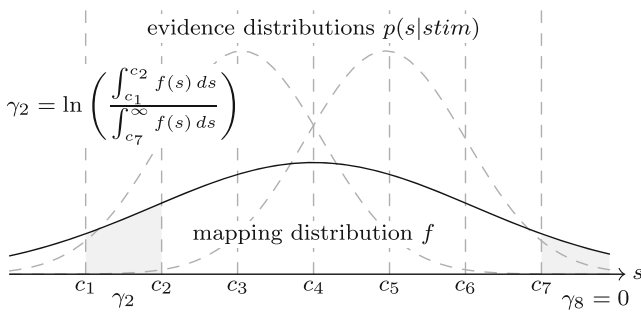


Fig. 3 Mapping between the unconstrained γ vector and the criteria. The dashed lines represent the SDT model with additional criteria; the distribution in the middle is the mapping distribution used to translate between the γ and the c vectors

of adjacent criteria because their exponents represent the relative magnitudes of areas under the standard normal curve, delineated by the pairs of adjacent criteria: $e^{\gamma_i} / e^{\gamma_j} = (\Phi(c_i) - \Phi(c_{i-1})) / (\Phi(c_j) - \Phi(c_{j-1}))$. When $K = 2$, only γ_1 is free to vary, and its value directly represents the direction and magnitude of bias: γ_1 is 0 when the criterion is placed at the midpoint between the evidence distributions; the more negative (positive) γ_1 is, the more the criterion is shifted to the left (right) of the midpoint.

In our model, it is often a good idea to multiply all the criteria by a value greater than 1, which is equivalent to making the mapping distribution wider. This tends to even out values of γ by preventing the outermost areas under the mapping distribution curve from becoming very small relative to areas delineated by adjacent pairs of non-outermost criteria. This is especially important when the criteria are widely spread, as can happen for moderate to large d' values. This feature is implemented in the `bhsdtr` package by introducing a scaling factor.

Once d' and c are derived from the unconstrained δ and γ parameters, the SDT model can be supplemented with a hierarchical linear regression structure. To avoid having to deal with an even more complicated index notation,³ below we present only the simple case of one grouping factor.

$$\begin{aligned} \delta &= X^{(\delta)}\beta^{(\delta)} + Z^{(\delta)}\theta^{(\delta)} \\ d'_i &= e^{\delta_i} \\ \gamma_{i,\cdot} &= X_{i,\cdot}^{(\gamma)}\beta^{(\gamma)} + Z_{i,\cdot}^{(\gamma)}\theta^{(\gamma)} \\ c_{i,k} &= s \Phi^{-1}\left(\frac{\sum_{l=1}^k (e^{\gamma_{i,l}})}{\sum_{m=1}^K (e^{\gamma_{i,m}})}\right) \end{aligned}$$

$$\begin{aligned} p(y_i = k | stim_i = 1) &= \Phi(c_{i,k} + d'_i/2) - \Phi(c_{i,k-1} + d'_i/2) \\ p(y_i = k | stim_i = 2) &= \Phi(c_{i,k} - d'_i/2) - \Phi(c_{i,k-1} - d'_i/2) \end{aligned}$$

³The reader familiar with hierarchical models may be surprised by our use of superscript parenthesized Greek letters to express hierarchical relationships. We chose this convention because it allowed us to use subscripts to denote elements of vectors and matrices while minimizing the number of nested sub- or superscripts.

Here $i = 1 \dots N$ is the observation number, X is the fixed effects model matrix for the respective parameter, Z is the random effects model matrix, β and θ are the fixed and random effects, c is an $N \times K - 1$ matrix, s is the criteria scaling factor, and y is the combined response. Note that d'_i is a scalar, but $\gamma_{i,\cdot}$ is in general a vector, and so $\beta^{(\gamma)}$ and $\theta^{(\gamma)}$ are matrices. The j -th rows of the $\beta^{(\gamma)}$ and $\theta^{(\gamma)}$ matrices represent fixed and random effects on the j -th element of the γ vector.

Following Sorensen and Vasishth (2015), we make use of the Cholesky decomposition of the correlation matrices because it improves efficiency and admits a convenient prior on random effects correlations:

$$\begin{aligned} \text{vectorized}(\theta^{(\gamma)}) &= \text{diag}(\tau^{(\gamma)})L^{(\gamma)}z^{(\gamma)} \\ \theta^{(\delta)} &= \text{diag}(\tau^{(\delta)})L^{(\delta)}z^{(\delta)} \\ z_i^{(\delta)} &\sim \text{Normal}(0, 1) \\ z_j^{(\gamma)} &\sim \text{Normal}(0, 1) \end{aligned}$$

where each τ is a vector of standard deviations of random effects and each L is a Cholesky decomposition of a random effects correlation matrix, i.e., $C = LL'$. Thus, θ is multivariate normal with the covariance matrix $\text{diag}(\tau)L$.

Finally, as recommended by Gelman et al. (2006), we use weakly informative proper priors because they provide regularization and help stabilize computation. The fixed effects $\beta^{(\delta)}$ and $\beta^{(\gamma)}$ are given independent normal priors, the random effects standard deviations $\tau^{(\delta)}$ and $\tau^{(\gamma)}$ are given independent half-Cauchy priors, and each L is given an independent lkj prior:

$$\begin{aligned} \beta_i^{(\delta)} &\sim \text{Normal}(\mu_i^{(\delta)}, \sigma_i^{(\delta)}) \\ \beta_{k,l}^{(\gamma)} &\sim \text{Normal}(\mu_{k,l}^{(\gamma)}, \sigma_{k,l}^{(\gamma)}) \\ \tau_i^{(\delta)} &\sim \text{half-Cauchy}(0, \zeta_i^{(\delta)}) \\ \tau_{k,l}^{(\gamma)} &\sim \text{half-Cauchy}(0, \zeta_{k,l}^{(\gamma)}) \\ L^{(\delta)} &\sim \text{lkj}(v^{(\delta)}) \\ L^{(\gamma)} &\sim \text{lkj}(v^{(\gamma)}) \end{aligned}$$

Specifying the prior distributions

A Bayesian model is not complete without providing fixed values of all the parameters that define prior distributions. Specifying the priors on sensitivity effects does not pose any special difficulties. The sensitivity of an unbiased classifier given percent correct is given by $2\Phi^{-1}(p(c))$. When $p(stim = 1) = .5$, the greater the bias, the lower the accuracy, meaning that an unbiased sensitivity is a lower bound on sensitivity given percent correct. Let us assume that the majority of participants are expected to achieve percent correct within the .51 to .99 range, with negligible

bias. Since $\ln(2\Phi^{-1}(.51)) = -2.99$ and $\ln(2\Phi^{-1}(.99)) = 1.54$, a reasonable weakly informative prior on δ is normal with mean $(1.54 - 2.99)/2$ and standard deviation $(1.54 + 2.99)/2$, which is the default prior on delta effects in the `bhsdtr` package.

Specifying the priors on criteria effects can be challenging because the criteria are order-restricted. On the other hand, specifying the priors on γ effects is challenging because of the complexity of the mapping expressed by Eq. 1. In our view, this is a major limitation of our implementation and we are currently working on improving it. By default, in the `bhsdtr` package each entry in the $\sigma^{(\gamma)}$ and $\zeta^{(\gamma)}$ matrices is set to $\ln(100)$ and the criteria scaling factor is fixed at 2. The prior on random effects' standard deviations are parametrized by ζ , which represents half-width at half-maximum of the half-Cauchy distribution. In our opinion, a not unreasonable starting point is to set ζ at a value that is greater than or equal to the most likely value of the random effects' standard deviation.

Finally, by default $\nu^{(\delta)} = \nu^{(\gamma)} = 1$, which implies a uniform prior on random effects' correlation matrices. Because the greater the value of ν , the more emphasis is put on zero off-diagonal correlations, the researcher can force the correlations to be near-zero by choosing a large ν value.

Overview of the software implementation

The `bhsdtr` package implements the model described in the previous section in the Stan modeling language because

```
devtools::install_git('git://github.com/boryspaulewicz/bhsdtr')
library(bhsdtr)
```

The essential steps of a typical data analysis process will usually involve preparing the data, creating the model code, fitting the model, assessing the fit, and possibly converting the unconstrained δ and γ parameters to d' and c .

Preparing the data

The `bhsdtr` package contains a dataset, `gabor`, from an unpublished study in which on each trial the participants had to classify a briefly presented Gabor patch as tilted to the left or to the right using the arrow keys. The participants were also asked to rate the stimuli on a 4-point Perceptual Awareness Scale (Ramsøy & Overgaard, 2004) presented at the bottom of the screen. The Gabor patch was immediately followed by a mask. The PAS ratings ranged from “no experience” to “absolutely clear image” and were provided either before (RATING-DECISION order condition) or after

it uses a state-of-the-art adaptive Hamiltonian Monte Carlo sampling algorithm which often handles high-dimensional correlated posteriors better than a Gibbs sampler.

Our package is essentially a collection of documented functions: The `aggregate_responses` function aggregates data as much as possible for efficiency, but without distorting the hierarchical structure. The `make_stan_model` function creates a model definition in the Stan language. The Stan code produced by the `make_stan_model` function can be fitted as is or modified by the user if needed, e.g., to change the prior distributions or to drop the equal variance assumption. The `make_stan_data` function creates regression model matrices and other data structures required by the model created using the `make_stan_model` function. Finally, the `plot_sdt_fit` function can be used to visually assess the fit of the model by creating publication-ready ROC curve plots or response distribution plots with posterior predictive intervals calculated for the chosen α level.

Usage example: installing the package and testing the model on real data

To make full use of the `bhsdtr` package functionality, three non-standard R packages are required, namely `rstan`, `plyr`, and `ggplot2`. We recommend using the `devtools` package to install the `bhsdtr` package directly from the GitHub repository. This will automatically install any missing required packages:

(DECISION-RATING order condition) the arrow keys were pressed. On each trial the Gabor patch was equally likely to be presented for 32 ms or 64 ms. Order was a between-subject variable and duration was a within-subject variable. There were 47 participants and 48 trials per condition.

In the study in question, the response was originally encoded using separate variables for accuracy and rating, so the first step was to create an appropriate response variable using the `combined_response` function. This function requires three variables, one encoding the stimulus class, one encoding the rating (as an integer), and one binary variable encoding the decision accuracy.

```
gabor$resp = combined_response
              (gabor$stim, gabor$rating,
              gabor$acc)
```

This step is required only if the ratings are available and a combined response variable is not already present in the

data. In the single criterion case, the combined response variable is simply the binary classification decision. To fit a single-criterion SDT model to this dataset, the code above would have to be replaced with the following:

```
gabor$resp = combined_response
             (gabor$stim ,
              accuracy = gabor$acc)
```

Next, the data has to be aggregated using the `aggregate_responses` function, but only to an extent that preserves all the random effects. This function requires as arguments a data frame containing all the relevant variables, the name of the stimulus class variable, the name of the combined response variable, and the vector of the names of all the variables that are to be preserved in the resulting aggregated dataset (apart from the stimulus class variable and the combined response variable), i.e., those encoding the grouping factors and those representing the independent variables used in the regression part of the model:

```
adata = aggregate_responses(gabor ,
                             'stim' , 'resp' , c('id' ,
                                                  'duration' , 'order'))
```

The main purpose of the aggregation step is to improve the efficiency of sampling from the posterior distribution. When data are aggregated in this way, the likelihood for each condition \times participant combination has to be computed only once rather than as many times as there are trials per condition per participant. Note that if there are other grouping factors present in the data (e.g., items, replications, etc.), and the user decides to model the effects of these factors, then these factors have to be specified at this stage to preserve the hierarchical data structure. The `aggregate_responses` function creates a list with three components. The `data` component is a data frame containing additional preserved variables, the `stimulus` component is the stimulus class variable, and the `counts` component is an $N \times K$ matrix of combined response counts, where N is the number of data points and K is the number of possible combined response values.

Creating the model code

A model is fitted using the `stan` function from the `rstan` package. The `stan` function returns a `stanfit` object, which can be used to obtain the summary statistics or the posterior samples as described in the `rstan` package documentation. This function requires a special list of data structures used by the model as well as a model specification expressed in the Stan language.

Every model has some fixed effects structure since, even when there are no predictors, the model parameters can

be expressed as regressed on a vector of ones (i.e., an intercept). However, many models also have a hierarchical structure and, if that is the case, this hierarchical structure has to be specified when using the `make_stan_model` function. This is done by providing a list of lists of R model formulae. Each list of model formulae is composed of at least three elements and specifies the correlated random effects of one grouping factor. The `group` element specifies the sampled factor; the `delta` and `gamma` elements specify which effects are assumed to vary between the levels of this sampled factor. When `make_stan_model` is used without any arguments, it specifies a model without any random effects. Fixed effects model matrices are specified by providing a list with at least two model formulae, named `delta` and `gamma`, to the `make_stan_data` function that is described later in this paper. Non-default priors can be specified by adding optional elements to the random and fixed effects specification lists, as described in the `make_stan_data` function documentation.

In the study in question, there was only one grouping factor, i.e., the participants. Because duration was a within-subject variable, in principle its effect could vary between the participants for all the SDT parameters. However, a preliminary data analysis indicated that the 32-ms difference in duration seemed to affect only the sensitivity. Thus, it was assumed that δ may depend on duration and order (`delta = ~ -1 + duration:order`), but γ may only be affected by order (`gamma = ~ order`). Because duration was a within-subjects variable, its effect on δ was assumed to vary between the participants (`group = ~ id [...] delta = ~ -1 + duration`), but the only random effect associated with γ was the participant specific intercept (`group = ~ id [...] gamma = ~ 1`):

```
fixed = list(delta = ~ -1 + duration:order ,
              gamma = ~ order)
random = list(list(group = ~ id ,
                   delta = ~ -1 + duration ,
                   gamma = ~ 1))
model = make_stan_model(random)
```

The `make_stan_data` function creates fixed and random effect model matrices based on the respective model formulae using dummy contrast coding. Note that the implicit intercept was suppressed for the δ model matrix (the `-1` term on the right-hand side of the model formula). In this way, δ was estimated for every duration \times order condition. The resulting separate intercepts and slopes parametrization makes it easier to calculate arbitrary contrasts on posterior samples. A more standard parametrization was used for the γ parameter because it was initially assumed that the criteria depend only on order, and so there was only one contrast of interest for every element of the γ vector. On

the other hand, in such cases nested parametrization (with separate intercepts and slopes for every condition) may be more convenient if a researcher is interested in the actual criteria, as we will later explain when introducing the `gamma_to_crit` function. This example also illustrates how the separation of the δ and γ regression structures makes it possible to test a broad class of linear models representing the dependence of the SDT parameters on additional variables.

Fitting the model

In order to fit the model, a separate data structure used by the Stan sampler has to be created using the `make_stan_data` function. The obligatory arguments to this function are an aggregated data object created by the `aggregate_responses` function and a fixed effects' specification. Importantly, if random effects are modeled,

the same specification of random effects has to be provided to the `make_stan_model` and `make_stan_data` functions.

```
sdata = make_stan_data(adata, fixed,
                      random)
fit = stan(model_code = model,
          data = sdata,
          pars = c('delta_fixed', '
                  gamma_fixed',
                  'delta_sd_1',
                  'gamma_sd_1',
                  'delta_random_1',
                  'gamma_random_1',
                  'Corr_delta_1',
                  'Corr_gamma_1',
                  'counts_new'),
          iter = 8000,
          chains = 4)
```

Table 1 Model fit summary statistics

	Mean	SE_{mean}	SD	2.5%	97.5%	No. eff. samples	\hat{R}
delta_fixed[1]	-0.11	0.00	0.15	-0.42	0.17	4339	1.00
delta_fixed[2]	1.12	0.00	0.09	0.94	1.29	4327	1.00
delta_fixed[3]	-0.39	0.00	0.20	-0.79	-0.03	5615	1.00
delta_fixed[4]	1.28	0.00	0.11	1.07	1.49	5417	1.00
gamma_fixed[1,1]	-0.14	0.00	0.06	-0.27	-0.02	3982	1.00
gamma_fixed[1,2]	-0.22	0.00	0.11	-0.43	-0.01	7356	1.00
gamma_fixed[2,1]	-0.70	0.00	0.18	-1.06	-0.36	3438	1.00
gamma_fixed[2,2]	0.49	0.00	0.29	-0.08	1.06	3845	1.00
gamma_fixed[3,1]	-0.54	0.00	0.22	-0.96	-0.11	3385	1.00
gamma_fixed[3,2]	0.83	0.01	0.35	0.14	1.51	3381	1.00
gamma_fixed[4,1]	0.28	0.00	0.25	-0.20	0.77	3143	1.00
gamma_fixed[4,2]	0.41	0.01	0.40	-0.39	1.20	3299	1.00
gamma_fixed[5,1]	-0.21	0.01	0.30	-0.79	0.38	3521	1.00
gamma_fixed[5,2]	0.82	0.01	0.49	-0.14	1.78	3596	1.00
gamma_fixed[6,1]	-0.78	0.00	0.24	-1.25	-0.30	3521	1.00
gamma_fixed[6,2]	0.80	0.01	0.39	0.03	1.57	3461	1.00
gamma_fixed[7,1]	-0.32	0.00	0.17	-0.65	0.01	3804	1.00
gamma_fixed[7,2]	0.43	0.00	0.28	-0.12	0.97	4138	1.00
delta_sd_1[1]	0.66	0.00	0.11	0.48	0.91	6078	1.00
delta_sd_1[2]	0.44	0.00	0.06	0.34	0.56	5500	1.00
gamma_sd_1[1]	0.16	0.00	0.07	0.02	0.32	2723	1.00
gamma_sd_1[2]	0.82	0.00	0.11	0.63	1.07	6362	1.00
gamma_sd_1[3]	1.08	0.00	0.12	0.86	1.33	4776	1.00
gamma_sd_1[4]	1.27	0.00	0.14	1.03	1.57	4430	1.00
gamma_sd_1[5]	1.55	0.00	0.17	1.26	1.91	4616	1.00
...

This is part of the table that was obtained using the `summary` function from the `rstan` package on the `stanfit` object produced by the `stan` function

Note that since more than one grouping factor is allowed, the names of all the hierarchical parameters are indexed (e.g., `delta_sd_1`, `delta_random_1`, `Corr_delta_1`). The name `counts_new` refers to posterior predictive samples that are required by the `plot_sdt_fit` function. Names starting with `Corr` refer to random effects correlation matrices, which are computed from Cholesky decompositions.

Assessing the model fit

As can be seen in the previous code fragment, four chains of 8000 iterations each were run simultaneously; the first half of the posterior samples, which served as a warm-up period for tuning the parameters of the sampling algorithm, was discarded. Part of the resulting Stan output is presented in Table 1 below.

Given the complexity of the model, the chains exhibited good mixing and seemed to have converged; there were enough effective samples for the fixed effect parameters to estimate 95% credible intervals well and none of the Gelman-Rubin statistics crossed the conventional 1.01 threshold, suggesting negligible sensitivity to the initial values.

Figures 4 and 5 below contain normal quantile-quantile plots of δ and γ random effects. The plots indicate that the distributions of random δ and γ effects can be approximated by normal distributions and that—at least in this particular example—these parameters seem to be good candidates for representing variability in the sensitivity and criteria parameters due to the grouping factors.

Once enough good-quality posterior samples are obtained for the parameters of interest, the inference process can be carried out by calculating credible intervals or HPD intervals for any function of the parameters, or Bayes

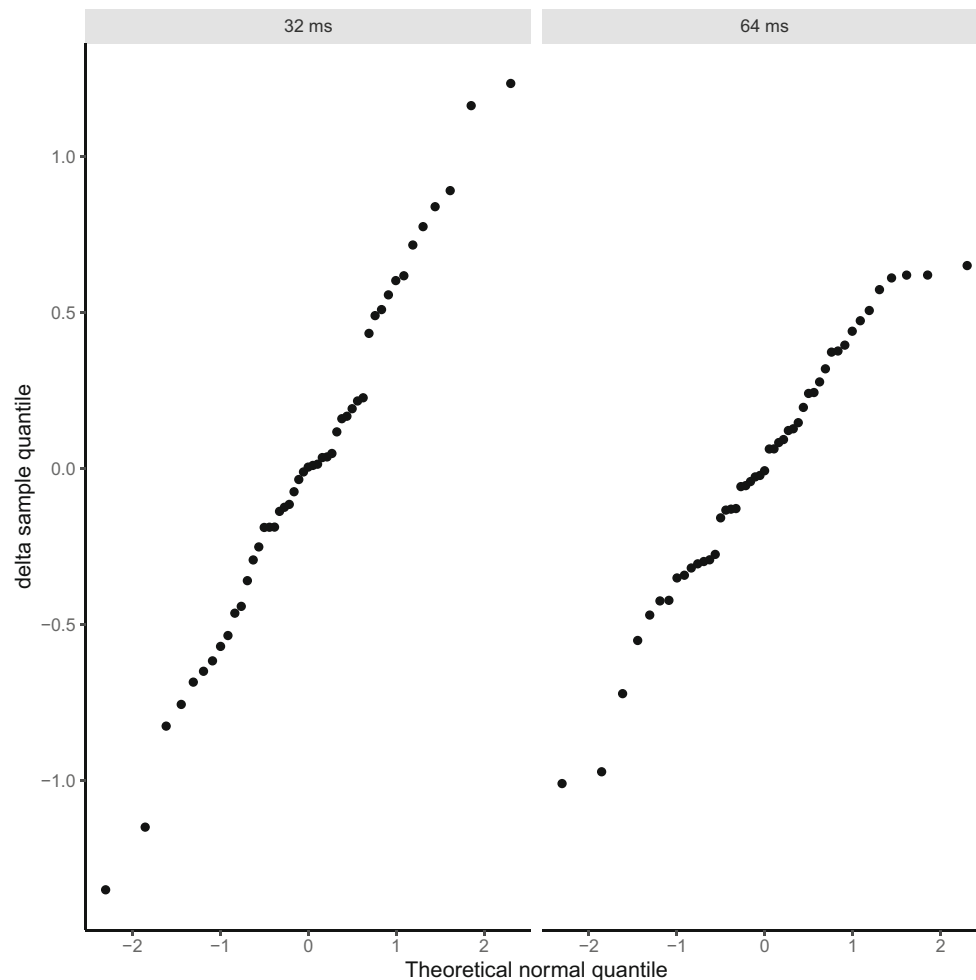


Fig. 4 Normal quantile-quantile plots of δ random effects. If the data points are normally distributed, they should form an approximately straight line. Here the delta random effects in the 64 ms condition seem to deviate from the straight line in a way that indicates that the tails of their distribution may be different than the tails of the normal distribution

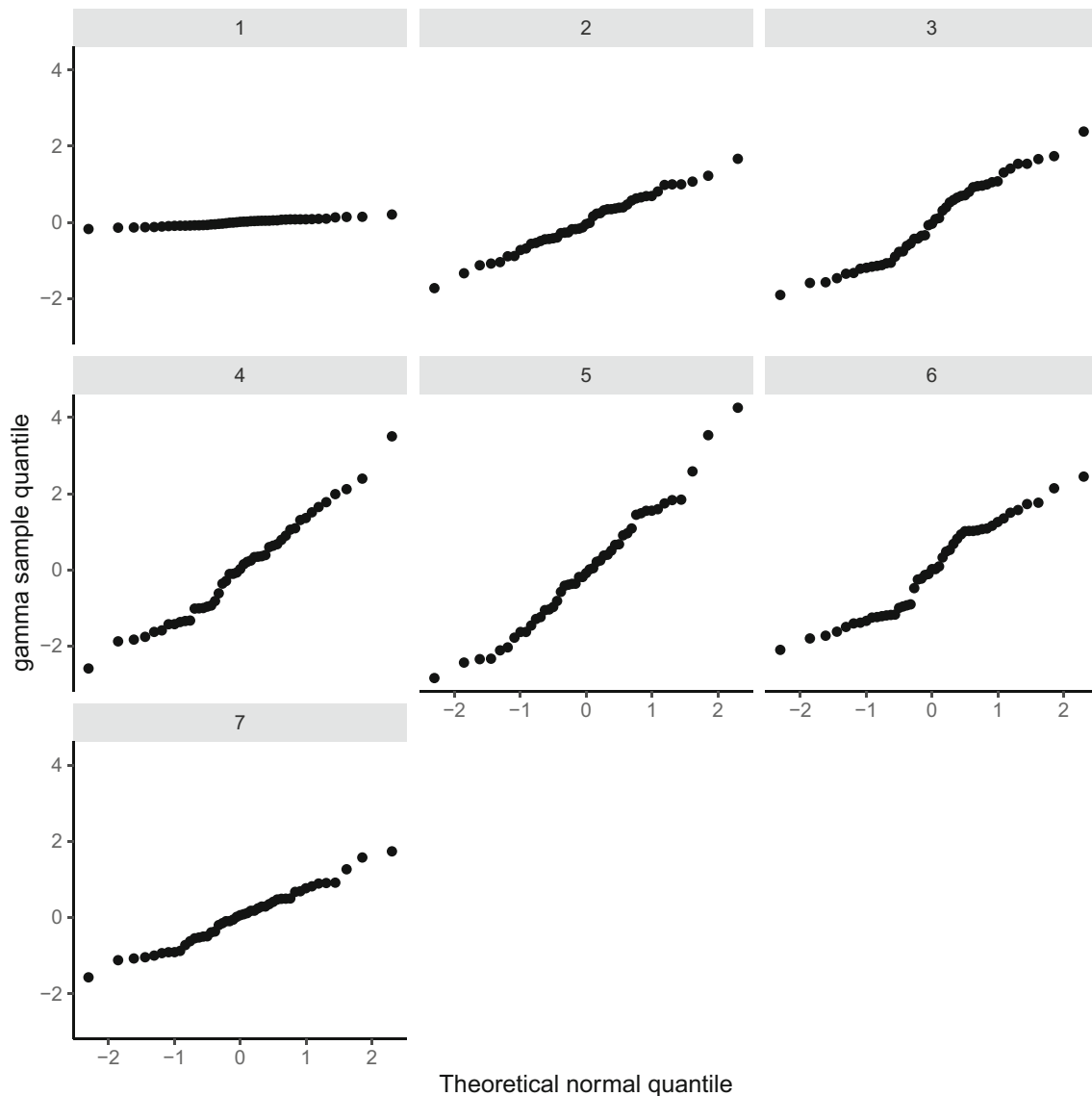


Fig. 5 Normal quantile-quantile plots of γ random effects. See Fig. 4 for a description

factors for each parameter separately. However, even when the Stan output summary does not indicate sampler convergence issues, before drawing any further conclusions the researcher should first check if the model fits the data. The `plot_sdt_fit` function can be used for this purpose:

```
plot_sdt_fit(fit, adata, c('order',
                          'duration'))
```

This function requires at least three arguments: a stanfit object, an aggregated data list produced by the `aggregate_responses` function that was used to produce the stanfit object, and a vector of names of variables that will determine how the data will be partitioned before plotting. We recommend assessing the fit at the individual level, but we did not include the participant identification

number in the list of conditioning variables because the resulting plot would take up too much space.

As can be seen in Fig. 6, which shows the ROC curves produced by the last code fragment, the model seemed to fit the data well in all but one condition (RATING-DECISION, 64-ms duration, lower right panel), in which three out of seven relevant⁴ points were outside the two-dimensional 95% posterior predictive regions.

Another way to assess model fit visually is by inspecting the conditional response distributions ($p(y|stim)$), such as those shown in Fig. 7, which was also created using

⁴The point in the upper-right corner of a ROC curve is always in the (1, 1) position

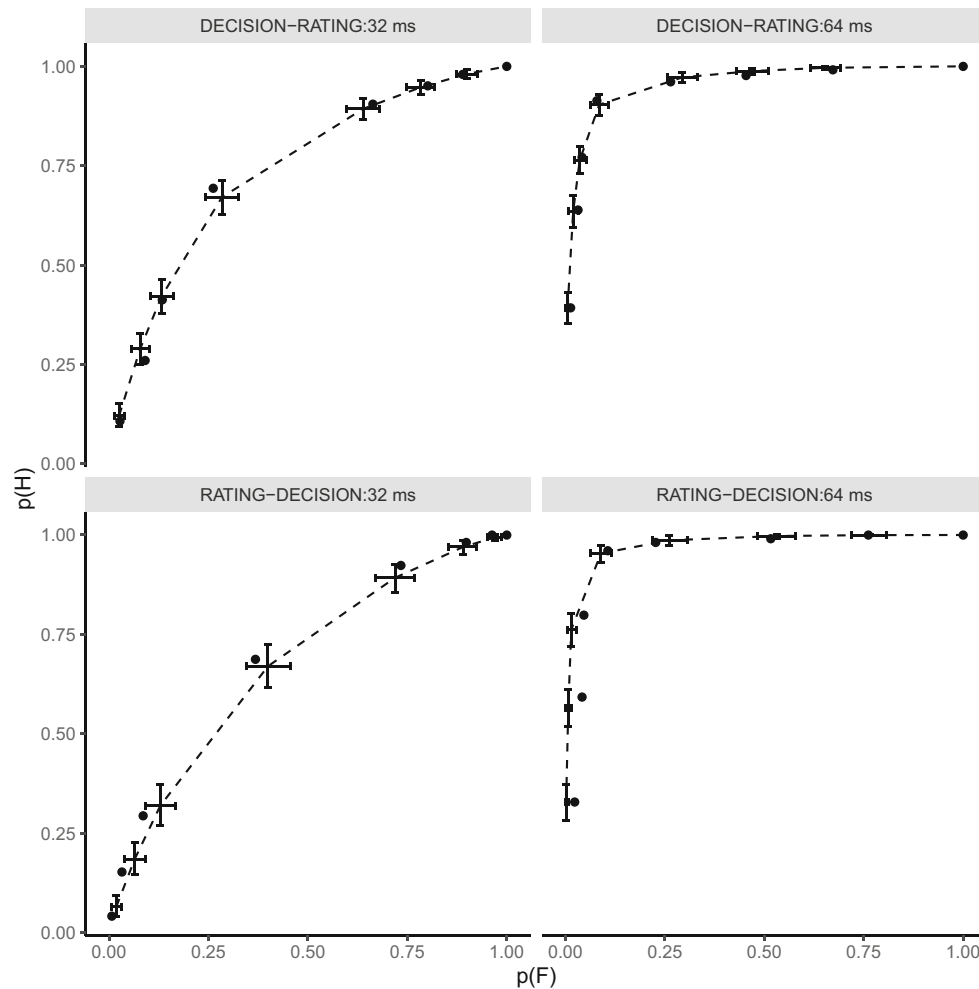


Fig. 6 ROC curve fit. The *dashed lines* represent the implied ROC curves, the *points* represent the observed ($p(H)$, $p(F)$) points, and the *horizontal and vertical 95% credible intervals* represent the posterior uncertainty in the estimates of those points

the `plot_sdt_fit` function by adding the `type = 'response'` argument.

Both plots can be informative about the reasons why a model does not fit the data. In this particular case, the plot seems to suggest that it may be a good idea to inspect the fit at the individual level and see if there are some participants with unusual $p(y|stim = 1)$ distributions in the RATING-DECISION \times 64 ms condition. On the other hand, it is also possible that the lack of fit is mainly a consequence of the assumption that duration had zero effect on γ , or that more substantial modifications are necessary, such as dropping the equal variance assumption.

Converting unconstrained δ and γ parameters to sensitivities and criteria

Posterior δ and γ samples have to be transformed in order to work with the d' and c parameters. Because δ (γ)

and d' (c) are related by an isomorphism, they contain exactly the same information and the translation between the two representations is always possible, although certain inferential tasks are not automatized in the current version of our package. In particular, translating between the two representations is straightforward only when fixed effects represent average parameter values in separate conditions, not when they represent differences between conditions, regression slopes, or interactive effects. For that reason, if there are no numerical predictors in the model, we recommend always using a separate intercepts parametrization for the δ and γ parameters. This is obtained when the effects are represented by the R model formulae of the form $\sim -1 + f_{-1}:f_{-2}:\dots:f_{-n}$, where the -1 term suppresses the common intercept and the f_{-i} terms represent nominal variables (i.e., factors). In this, way all the SDT parameters will be estimated for each condition separately, d' can be recovered from δ for every condition using the exponential function, and all the criteria can be

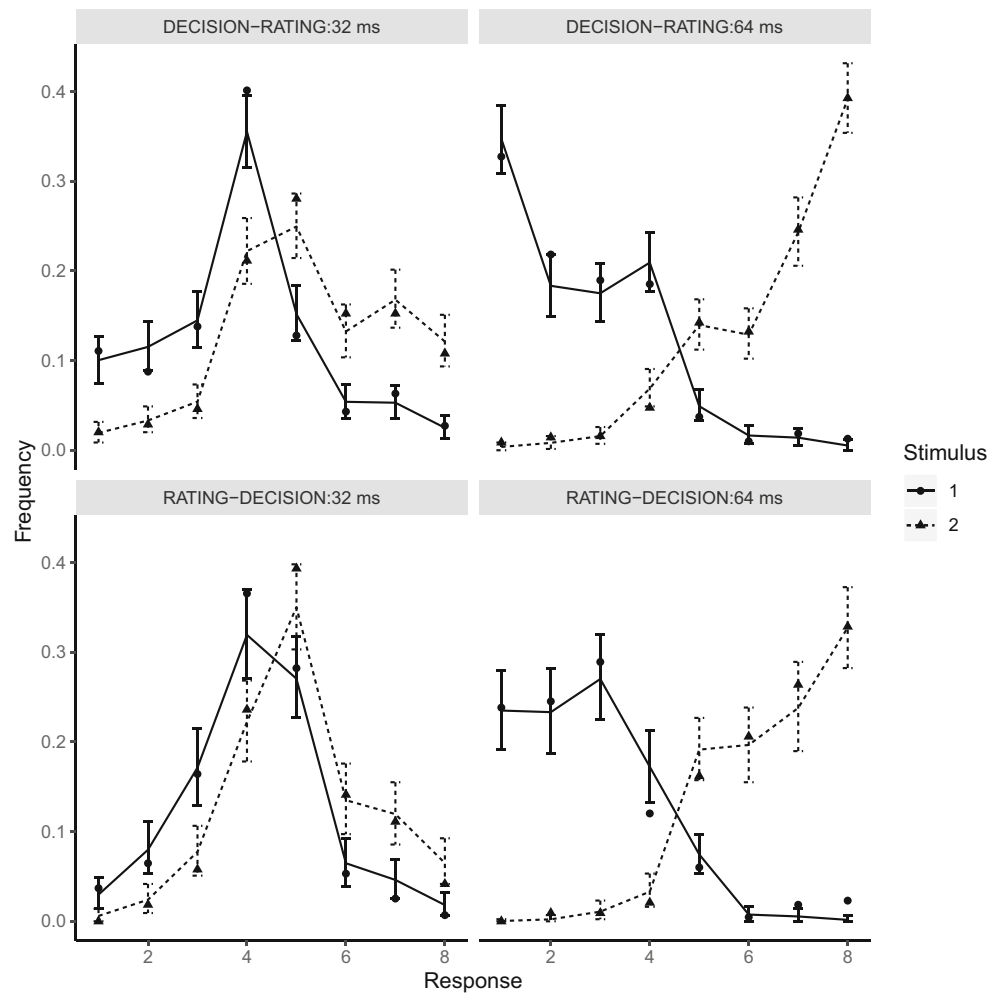


Fig. 7 Response distribution fit. The *solid* and *dashed lines* represent the implied response (γ) distributions, the points represent the observed proportions of responses, the 95% credible intervals represent the posterior uncertainty, and the type of line (dashed or solid) represents the stimulus class

recovered for every condition using the `gamma_to_crit` function described later in this paper. Since this is a Bayesian model, arbitrary contrasts, including the contrasts that correspond to interactive effects, can be calculated using the posterior samples.

In our example, because nested parametrization was used for the δ fixed effects model matrix, all four `delta.fixed` parameters can easily be transformed to sensitivities by applying the exponential function. It is important to remember that because the logarithm is a non-linear transformation, the δ to d' conversion step should be done first before applying any other transformations to the posterior samples; in general, the logarithm of a point and interval d' estimate is not equal to the point and interval estimate calculated after transforming the posterior δ samples to the d' samples. The same is true of the γ (c) parameters.

In this case, the first column of the `gamma.fixed` matrix (the intercept) corresponds to the values of the γ vector in the DECISION-RATING condition, but the second column corresponds to the *effect* of order on γ . For this reason the posterior criteria samples can be obtained using the `gamma_to_crit` function only for the first column of the `gamma.fixed` matrix. This is because the second column represents the difference in γ between conditions and translating from γ to c will not give the correct difference in c . In order to recover the criteria for the second condition, the γ posterior samples for this condition would have to be computed first. This could be achieved by adding the posterior samples of γ effects to the posterior samples of γ in the first condition, and then converting the obtained γ posterior samples for the second condition to the criteria posterior samples using the mapping in Eq. 1.

Testing the model on simulated data

We simulated the data from a hypothetical exact replication of the previously described experiment using the point estimates from the previous fit as known realistic parameter values. The true hierarchical model was fitted to the simulated data. Mixing performance was similar to the real data case. All the model parameters were correctly recovered in a sense that the true values were outside the 95% credible intervals no more than 5% of the time. Note that unless there is an error in the software, in this case the credible intervals and point estimates are automatically correct since they are based on the model fitted to the data obtained from itself.

As we have emphasized, the models that lack the necessary hierarchical structure may easily show reliable effects where none exist, or they may fail to detect true differences. To illustrate this problem, an SDT model that differed from the true model only in that it did not have

any hierarchical structure was fitted to the same simulated dataset. Since the non-hierarchical model was much simpler and the data consisted of only eight vectors of response counts, the mixing of the chains was excellent. This simple simulational example shows how misleading the results of such analyses can be.

The 95% credible intervals and the point estimates calculated for the fixed effects based on each model are compared in Figs. 7 and 8 below. The estimates were centered on the true values to simplify the presentation, thus the true values are represented by the horizontal line at 0.

The bias in the estimates based on the simplified model is even more apparent in Fig. 9 which shows the effects for the d' and c parameters. Note that the δ and γ parameters do not share a common scale, but the more familiar d' and c parameters do (i.e., the common standard deviation of the evidence distributions). The c posterior samples for the second condition were calculated by adding the effect of order on γ to γ in the first condition and then using Eq. 1.

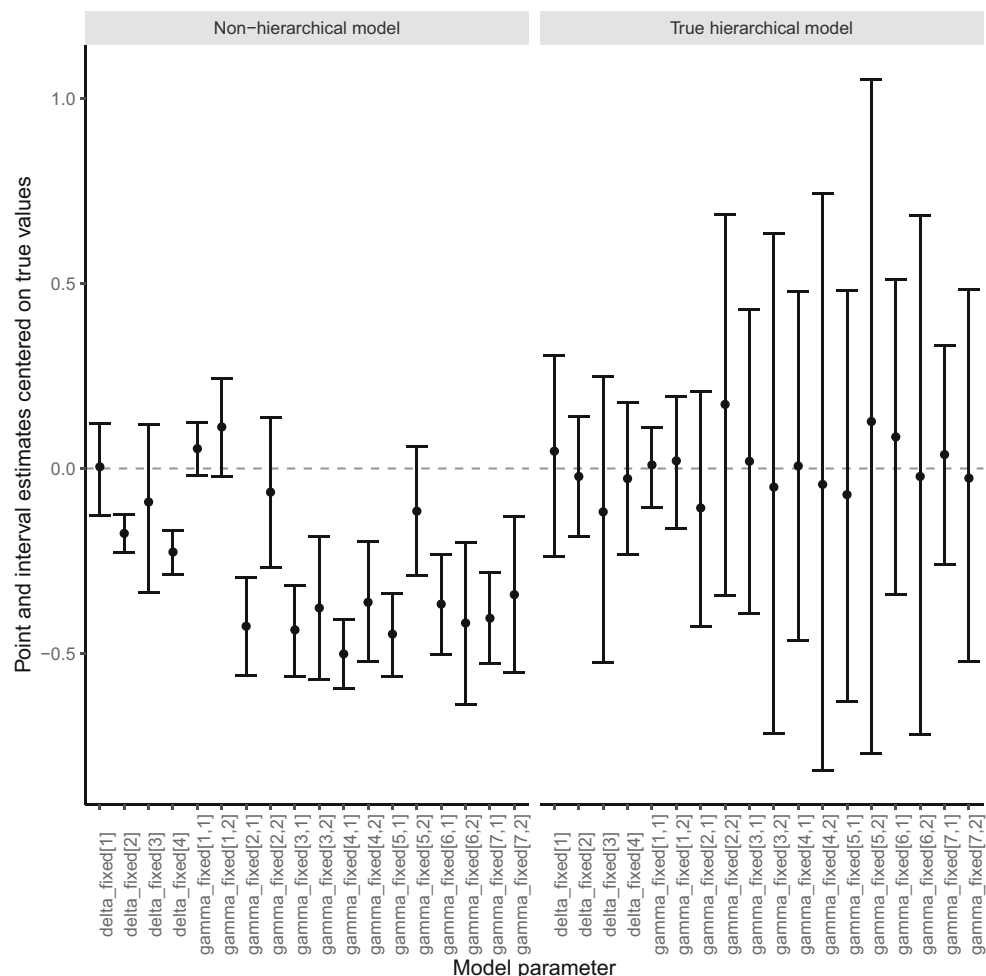


Fig. 8 Comparison of the point and 95% interval posterior δ and γ estimates based on the true hierarchical and the simplified non-hierarchical models. The posterior estimates were centered at the true values to allow for easy inspection of the direction and magnitude of estimate bias. The dashed horizontal line centered at 0 represents the true values of the parameters. The estimates based on the true model are correct

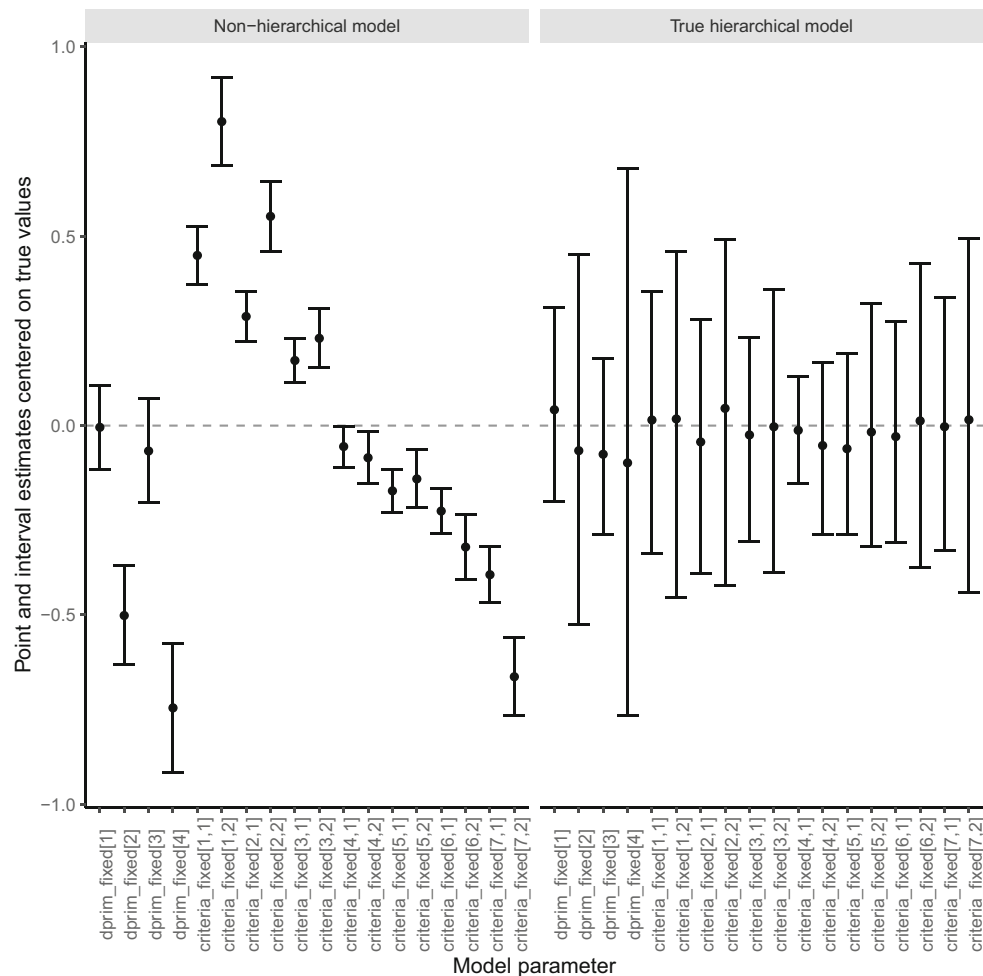


Fig. 9 Comparison of the point and 95% interval posterior d' and c estimates based on the true hierarchical and the simplified non-hierarchical models. The posterior samples for the effect of order on γ were added to γ in the first condition to obtain the posterior γ samples in the second condition. Condition specific posterior γ samples were converted to c using Eq. 1. The dashed horizontal line centered at 0 represents the true values of the parameters

As can be seen, the true model correctly recovered the known parameter values, but the estimates based on the simplified, non-hierarchical model were severely biased; the credible intervals were not only much shorter than the correct intervals, but also failed to contain most of the true values. In fact, *most of the point d' and c estimates based on the non-hierarchical model differed from the true values by several standard deviations of their posterior distributions.*

One of the main reasons that the ROC curves are calculated when an SDT model is fitted to the data from psychology experiments is to determine whether the model is approximately true. However, as can be seen in Fig. 10 below, in this case the observed ROC curves seemed to fit the false simplified model's predictions quite well.

This is a clear example of what we have previously described as the worst-case scenario of estimate bias: the point estimates are severely biased, the interval estimates are much more narrow than they should be, and the ROC

curve plot indicates that the model fits the data well. All this gives a false impression of validity of conclusions that systematically differ from the truth.

As striking as this example may be, there is nothing special about the dataset that was used to produce it; To our knowledge the observed fixed and random effects are within the range of values observed in other studies: the average d' values ranged from 0.9 to 3.6 and the average criteria ranged from -2.9 to 2.5 .

Limitations of the current implementation and some future directions

Certain aspects of our implementation are experimental. This is especially true of the criteria scaling factor and the default parameter values that specify the prior distributions for fixed and random effects. Without fitting the model to

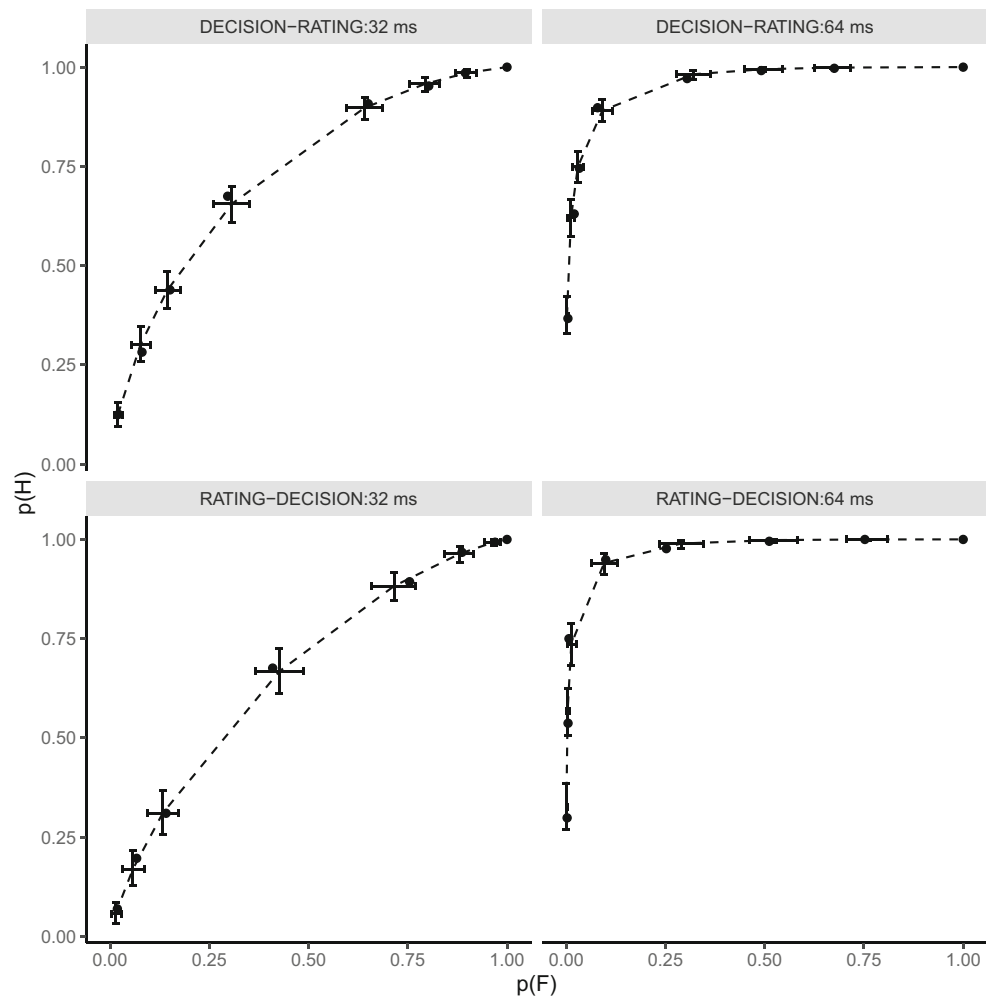


Fig. 10 ROC curve fit for the non-hierarchical model. See the description of Fig. 6

a large number of different datasets it is impossible to say with any degree of certainty if the default values for the parameters that define the priors on δ and γ effects are good starting points in the majority of typical cases. When they are not, the model may not converge, or the proportion of effective samples may be low. This is a common problem when fitting complicated Bayesian models. The only advice that we can provide at this stage is to always carefully inspect the posterior samples, test if the model fits the data, and use informed judgment to see if the obtained results make theoretical sense.

The correlations between the γ random effects as well as the correlations between the δ random effects are accounted for in our model, but the correlations between the δ and the γ random effects are not. The results of the tests with real datasets that we have done so far seem to indicate that implementing this feature is not urgent; this may be due to the fact that the criteria are midpoint-centered, but more

extensive testing with many different datasets is necessary to see how serious this limitation is.

Perhaps a more pressing matter is the possibility to fit the unequal variance SDT model as it seems to be one of the main alternative models tested against the equal variance SDT model. The results of the two surveys by Swets and Pickett (1982) and Swets (1986) seem to indicate that unequal variance may be common, although these results are mostly based on aggregated data; as shown by Morey et al. (2008) ROC curves based on aggregated data may falsely indicate lack of variance equality. This is a more difficult task than implementing the correlations between the two kinds of SDT parameters, because a new kind of parameter has to be introduced with all the associated hierarchical linear regression structure and an appropriate link function. A related problem is that when the variances of the two evidence distributions are not equal, there is more than one notion of the midpoint between the

evidence distribution means. Consequently, the correlations between the γ and the δ random effects may have to be introduced simultaneously with the unequal variance model, resulting in an increase in model complexity and the associated demand for a large number of data points and participants to obtain interval estimates that are narrow enough for effects of typical size to be reliably detected.

Thanks to Dobromir Rahnev's initiative, a substantial collection of datasets that can be modeled using the `bhsdtr` package was recently made publicly available, as described in Rahnev et al. (2019). This presents a great opportunity for extensive testing and makes it possible to obtain well-calibrated default priors for all the parameters in the future.

Conclusions

The importance of SDT to psychology stems from the fact that given weak assumptions about an underlying decision process, it promises to deconfound sensitivity from bias in arbitrary classification tasks—a problem almost as common in psychology studies as the usage of classification tasks. To the best of our knowledge, at present the `bhsdtr` package provides the only method of Bayesian inference for SDT models with or without ratings that can be recommended as a default choice in typical applications. That is because it is the only method that allows for fixed and random effects in all the parameters of an SDT model with additional criteria. Our parametrization forces the sensitivity to be non-negative and the criteria to be order-restricted, while the isomorphisms between the d' and c parameters and the unconstrained δ and γ parameters make it possible to supplement the SDT model with the general hierarchical linear regression structure. There is no limit to the number of grouping factors except for the one imposed by available computational resources; correlations of random effects of the same grouping factor are accounted for, all the SDT parameters can be modeled by linear regression within the same model, and all the effects on all the SDT parameters estimable within the levels of the grouping factors can have associated random effects. Finally, if the need arises to relax a built-in restriction, experienced users can extend the model in arbitrary ways by using automatically generated human-readable Stan code as a template.

The GitHub package repository (<https://github.com/boryspaulewicz/bhsdtr>) contains the annotated source code and data that were used to perform all the analyses and produce all the figures presented in this paper.

Acknowledgements This work was supported by the National Science Centre, Poland SONATA grant given to BP (2013/09/D/HS6/02792).

References

- Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, 80(1), 1–28.
- Bürkner, P.-C., & Vuorre, M. (2019). Ordinal regression models in psychology: A tutorial. *Advances in Methods and Practices in Psychological Science*, 2(1), 77–101.
- Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., . . . , Riddell, A. (2016). Stan: A probabilistic programming language. *Journal of Statistical Software*, 20, 1–37.
- DeCarlo, L. T. (1998). Signal detection theory and generalized linear models. *Psychological Methods*, 3(2), 186.
- Fleming, S. M. (2017). HMeta-d: Hierarchical Bayesian estimation of metacognitive efficiency from confidence ratings. *Neuroscience of Consciousness*, 3(1), nix007.
- Gelman, A. et al. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Analysis*, 1(3), 515–534.
- Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York: Wiley.
- Hautaus, M. J. (1997). Calculating estimates of sensitivity from group data: Pooled versus averaged estimators. *Behavior Research Methods, Instruments, & Computers*, 29(4), 556–562.
- Luce, R. D. (1963). Detection and recognition. In Luce, R. D., Bush, R. R., & Galanter, E. (Eds.) *Handbook of mathematical psychology*, (Vol. 1, pp. 245–307). New York: Wiley.
- Macmillan, N. A., & Creelman, C. D. (2004). *Detection theory: A user's guide*. Psychology Press.
- Macmillan, N. A., & Kaplan, H. L. (1985). Detection theory analysis of group data: Estimating sensitivity from average hit and false-alarm rates. *Psychological Bulletin*, 98(1), 185.
- Maniscalco, B., & Lau, H. (2012). A signal detection theoretic approach for estimating metacognitive sensitivity from confidence ratings. *Consciousness & Cognition*, 21(1), 422–430.
- McNicol, D. (2005). *A primer of signal detection theory*. Psychology Press.
- 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.
- Peterson, W., Birdsall, T., & Fox, W. (1954). The theory of signal detectability. *Transactions of the IRE Professional Group on Information Theory*, 4(4), 171–212.
- R Core Team (2017). R: A language and environment for statistical computing. Vienna, Austria. <https://www.R-project.org/>.
- Rahnev, D., Desender, K., Lee, A.Ian.L. F., Adler, W. T., Aguilar-Lleyda, D., Akdogan, B., et al. (2019). The confidence database. PsyArXiv. Retrieved from psyarxiv.com/h8tju. [10.31234/osf.io/h8tju](https://doi.org/10.31234/osf.io/h8tju).
- Ramsøy, T. Z., & Overgaard, M. (2004). Introspection and subliminal perception. *Phenomenology and the Cognitive Sciences*, 3(1), 1–23.
- Rouder, J. N., & Lu, J. (2005). An introduction to Bayesian hierarchical models with an application in the theory of signal detection. *Psychonomic Bulletin & Review*, 12(4), 573–604.
- Sorensen, T., & Vasishth, S. (2015). Bayesian linear mixed models using Stan: A tutorial for psychologists, linguists, and cognitive scientists. arXiv:1506.06201.
- Stanislaw, H., & Todorov, N. (1992). Documenting the rise and fall of a psychological theory: Whatever happened to signal detection theory? (Abstract). *Australian Journal of Psychology*, 41, 128.

- Sternberg, S. (2001). Separate modifiability, mental modules, and the use of pure and composite measures to reveal them. *Acta Psychologica, 106*(1), 147–246.
- Swets, J. A. (1986). Form of empirical ROCs in discrimination and diagnostic tasks: Implications for theory and measurement of performance. *Psychological Bulletin, 99*(2), 181.
- Swets, J. A., & Pickett, R. M. (1982). *Evaluation of diagnostic systems: Methods from signal detection theory*. New York: Academic Press.
- Tanner, W. P. Jr., & Swets, J. A. (1954). A decision-making theory of visual detection. *Psychological Review, 61*(6), 401.
- Wickens, T. D. (2002). *Elementary signal detection theory*. USA: Oxford University Press.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.