A sample size study and comparison of Averaged, Pooled, and Bayesian Hierarchical estimators of group sensitivity (*d′*) in a yes/no detection task

Don Li, Michael Hautus

The University of Auckland

2019

Corresponding author: Don Li

School of Psychology, The University of Auckland City Campus, Private Bag 92019, Auckland 1142, New Zealand

Email: yli877@aucklanduni.ac.nz

# **Abstract**

We investigate the problem of estimating group-level sensitivity (*d′*) from a yes/no detection experiment. With multiple subjects, this type of protocol yields a dataset with a nested variance structure. That is, there is within- and between- subject variability. We present a Bayesian hierarchical model that attempts to estimate group sensitivity by explicitly modelling these two levels of variability. The point estimates from the hierarchical model are compared to point estimates obtained by averaging individual estimates of *d′* or by pooling trials across subjects. The hierarchical model tended to be more efficient than the averaged or pooled estimates when more than eight subjects and eight trials per subject were available. In addition, the hierarchical model was resistant to large variation in the criterion (response bias), while the pooled and averaged estimators were adversely affected. R code and an associated graphical application are provided for implementing the hierarchical model.

Supplementary material can be found at https://github.com/Don-Li/hierarhical\_binomial\_SDT\_group\_sensitivity

# **An overview of Signal Detection Theory**

Signal Detection Theory (SDT; Tanner & Swets, 1954) is a theory of decision making under uncertainty. Our interest is in the application of SDT to binary responses. The experimental protocol that generates this data is such that, the subject is presented with a sequence of intermixed *stimulus-present* or *noise* (stimulus-absent) trials. The subject is required to detect the presence of the stimulus by responding whether the stimulus was present on a given trial. The combination of trial types and responses generate a 2 x 2 decision matrix. If the subject reports the detection of the stimulus on a stimulus-present trial, the outcome is referred to as a *hit*. However, if the subject reports the absence of the stimulus on a stimulus-present trial, the outcome is referred to as a *miss*. If detection is reported on a noise trial, then the outcome is a *false alarm*; whereas if detection is not reported on a noise trial, then the outcome is a *correct rejection*.

In SDT, the stimulus/noise input is represented on a one-dimensional continuum of signal magnitude. It is assumed that the presence of absence of a stimulus generates a signal value. If the signal value is greater than some decision criterion, the subject will respond that the stimulus has been detected. In the Gaussian equal-variance SDT model, it is assumed that the signal values in each trial are perturbed by normally distributed noise where the variance of the perturbation is the same for both stimulus-present and noise trials. Thus, uncertainty in decision making arises due to the possibility of stimulus-present signals being perturbed below the decision criterion (generating misses) and stimulus-absent signals possibly being perturbed above the decision criterion (generating false alarms).

Within the SDT framework, the performance of an individual on a detection task is characterised by sensitivity to the stimulus classes and response bias. A common measure of sensitivity is *d′*, which is computed (Macmillan & Creelman, 2004), where is inverse standard normal distribution function and and are hit and false alarm rates, respectively. In addition to sensitivity, a subject may show systematic bias to one response class (e.g., a tendency to report the detection of the stimulus). We will use the SDT measure, *c* (criterion; Macmillan & Creelman, 2004) to represent bias; . The appeal of SDT as a framework to characterise sensitivity is that sensitivity and response bias are (independent) linear terms in the response measures (hits and false alarms).

## **The variance structure of experiments**

In the present study, we are concerned with the problem of summarising the performance of a group of subjects. For example, given two products, we may be interested to what extent consumers can detect the difference between two products, A and B. To answer this question, one would sample a collection of participants, conduct the appropriate yes/no task, and summarise the sensitivity (*d′*) to the products for the group.

Experimental protocols generally yield data with nested variance structures. In the case of a simple yes/no detection task with some number of subjects, we have within- and between- subject variability. We therefore have two measures of sample size; the number of trials per subject and the number of subjects.

Multiple methods exist for summarising the sensitivity of a group of subjects. One method is to average the individual estimates of *d′* across subjects, and to treat the averaged *d′* as an estimate of the sensitivity of the group (*average d′*). Averaging *d′* has appeal because of the Central Limit Theorem. An alternative method is to pool the trials across subjects and compute *d′* using hits and false alarms summed across all the subjects (*pooled d′*). Pooling trials assumes that we essentially have a single super-subject that all the responses come from. Therefore, pooling trials assumes that within- and between- subject variability is interchangeable. However, in the case of small sample sizes, we may wish to make this assumption of interchangeable variance in order to improve the power of our analysis. Thus, our intuition would lead us to presume that the difference in quality of the summaries obtained from averaging *d′* or computing *d′* using pooled trials would depend on the sample size. Indeed, Hautus (1997) showed that pooled *d′* is a less biased estimator of group *d′* than averaged *d′* when there are few trials per subject and the criterion variability is low[[1]](#footnote-1).

The present study extends the results of Hautus (1997) by considering an alternative method for estimating group sensitivity. That is, we explicitly model the relationship between individual-level parameters (i.e., *d′* and *c* for each subject) and group-level parameters (i.e., group-level *d′* and *c*) and use these two levels of parameters to mutually inform one another. Although there are different approaches to this kind of multilevel modelling, the present paper will use a Bayesian hierarchical model. Therefore, our estimates of group sensitivity will incorporate both within- and between- subject variability.

However, a consequence of using both within- and between- subject variability is that the quality of our hierarchical model estimates will depend on the sample size – both in the number of trials per subject and the number of trials. Although the hierarchical model may represent a more “correct” way to the nested variance structure of these experiments, there may be situations where point estimates from other methods may be more accurate. Indeed, McNeish (2016) showed that in meta-analyses and multilevel modelling, Bayesian techniques can perform worse than Frequentist techniques when sample sizes are small and the priors are inaccurately specified (e.g., overly diffuse/uninformative). Thus, we seek to compare the quality of group estimates from our hierarchical model to those obtained by pooling trials or averaging *d′* across subjects for different experimental contexts (e.g., different sample sizes, different amounts of criterion variability, etc.).

A hierarchical model for yes/no data has been presented previously by Rouder and Lu (2005). However, that presentation was framed in terms of a tutorial for Bayesian modelling, rather than an in-depth study of the quality of group-level estimates. Thus, the present paper combines the estimator quality study of Hautus (1997) with the modelling techniques presented by Rouder and Lu (2005).

# **A Bayesian hierarchical model**

In a Bayesian analysis, the target of the analysis is the posterior distribution or some summary of the posterior distribution. Recall that a Bayesian analysis requires a prior distribution of the model parameters and a likelihood model for the data , for data and parameter vector . The posterior is then a compromise between prior information and the data: . In the context of the performance of a single subject in the yes/no task, is a vector of *d′* and *c*, and each parameter would have its own prior. We recommend Gelman, Carlin, Stern, Dunson, and Rubin (2014) for a comprehensive text on Bayesian analysis. An implementation of software available for the Bayesian analysis of single subject SDT data has been developed by Lee (2008).

Thus, the Bayesian analysis treats the model parameters as random variables. We can recursively use this property to treat the prior parameters as random variables. In doing so, the priors for each subject come from a hyperprior distribution, each with hyperprior parameters. In other words, suppose that we have subjects so that the posterior for each subject has the form . Thus, each subject has their own sensitivity and criterion parameters. We then assume that the *d′* and *c* for each subject come from a common distribution. Taking sensitivity as an example, we may write , where the sensitivity of each of the subjects come from a shared hyperprior distribution, with the hyperparameter . would correspond to the group-level sensitivity.

Transitioning back to generic terms, let denote the hyperprior distribution of . The hierarchical model then has the form: . Thus, the posterior distribution of the data given the subject-level parameters and the group-level parameters depends on the likelihood of the data given the subject-level parameters, the subject-level parameters given the group-level parameters, and the prior distribution of the group-level parameters. For a tutorial on hierarchical modelling that focusses on yes/no detection data, we recommend Rouder and Lu (2005).

Another way to conceptualise the hierarchical model is that it comprises two components. The first is an *observation model*, this part of the model comprises the likelihood and it links the parameters of each subject (*d′* and *c*) to the observed data (hits and false alarms). This link is made through the likelihood function. The second part is a *latent variable model* that links the parameters of each subject to the group-level parameters. That is, the parameters of each subject are treated as deviations from the group-level parameters. This link is made through the conditional distribution .

## **Likelihood and parameterisation**

The typical approach for constructing a likelihood model for SDT data is to use a binomial distribution for hits and false-alarms (Dorfman & Alf, 1968; Lee, 2008; Rouder & Lu, 2005). Recall that the binomial distribution is the distribution of identically distributed Bernoulli trials[[2]](#footnote-2) with probability of success . Thus, and , where and denote hits and false alarms and denotes the subject.

However, the parameters of interest are the SDT parameters, *d′* and *c*. Therefore, we have a choice of how to parameterise our Bayesian model. We could choose to put priors on the binomial probability parameters – beta priors could be a good choice. In this case, *d′* and *c* would be obtained by using the SDT equations. Alternatively, we may choose to use a probit transformation to put the binomial probability parameters on the real numbers (e.g., Rouder & Lu, 2005). A third approach is to parameterise the model directly in terms of *d′* and *c*, then invert the SDT equations to get the binomial probabilities to compute the likelihood (e.g., Lee, 2008). Our preference is for the third approach. In an SDT analysis, *d′* and *c* are the targets of the analysis, so researchers may be more familiar with the typical mean and variabilities of these measures rather than the probability or probit forms. This may make it easier to specify priors that better match the plausibility of the model parameters.

## **Priors and hyperpriors**

In our hierarchical model, we specify prior distributions for the subject-level SDT parameters, *d′* and *c*. For each of the parameters, we use normal distributions. That is, for subjects, and , where and are the mean and standard deviations for the corresponding parameters.

The motivation for using normal distributions for our priors is that our simulation results showed that the sample parameters were relatively Gaussian, especially for larger sample sizes (number of trials). Therefore, in using Gaussian priors, we aim to improve the robustness of our model to small sample sizes. Further, Lee (2008) used Gaussian priors on *d′* and *c*, but their justification was based on the inverse transformation of uniform priors for the binomial response probabilities.

In addition, we also specify hyperprior distributions from which the priors are drawn. These are the group-level parameters. Because and are both normally distributed, we require two hyperpriors for each parameter – one for the mean parameter and one for the standard deviation. For the mean parameters, we use Cauchy priors; for the standard deviations we use half-Cauchy priors. Using *d′* as an example, and , where and are the location and scale parameters for the Cauchy distribution corresponding to the prior parameter. In total, we have eight hyperprior parameters for the model – four for *d′* and four for *c*.

Our choice of the Cauchy and half-Cauchy hyperpriors are motivated by robustness. The Cauchy distribution has relatively long tails compared to other bell-shaped distributions. This allows us to specify regions of the parameter space where the posterior parameters are likely to be, while at the same time allowing for outliers. The half-Cauchy on the variance parameters is based on the recommendations of Gelman (2006). See Rouder and Lu (2005) for an alternative model with inverse-gamma priors for variance parameters and normal priors for means.

In our view, it is important to specify priors/hyperpriors that accurately specify the relative likelihood of the posterior values. Previous presentations of hierarchical modelling of yes/no data has used uninformative (Lee, 2008) or very diffuse priors (Rouder & Lu, 2005). However, the present study is a study of sample size. As a result, we will examine data that is more extreme. When the sample size is small, the posterior density is very sensitive to the hyper/prior distribution. When we have, for instance, two subjects and four trials, diffuse or non-informative priors can lead to estimates that are many times worse than an accurately specified and relatively informative prior (see subsequent analyses). In the case of large sample sizes, the prior contributes relatively less to the posterior density than the data, so the consequences of prior diffusiveness are less severe. We wish to reiterate here that noninformative priors or overly diffuse priors can lead to severe biases in hierarchical modelling when the sample size is small (Gelman, 2006; McNeish, 2016).

Because a Bayesian analysis depends on the priors to the extent of the sample size, we will investigate two versions of our model with hyperprior parameters that vary in diffusiveness. The value of these hyperprior parameters are described in Table 1. For the narrower-hyperprior model, most of the density for the group *d′* mean is between 0 and 3, which is a reasonable range for group *d′*. This corresponds to proportion correct values of 50% to 93.32%, assuming no response bias. In general, the values of group *d′* less than zero are unlikely unless the stimuli are designed in such a way to elicit a mislabelling of stimuli (e.g., visual illusions). However, when the true values of group *d′* are extremely high, it becomes difficult to determine accurately the individual and group *d′* without proportionally large sample sizes. Because the experimenter has some degree of control over the experimental stimuli and thus some control over the upper limit of *d′*, extremely large may also be unlikely.

The relatively long tails of the Cauchy distribution permit more extreme values with some reasonable probability. Overall, we designed the prior values in Table 1 (narrower-hyperprior model) to cover a wide range of possible experiments so that our subsequent sample size recommendations can be as general as possible without being overly diffuse. However, our general recommendation is to use all available information to make the hyperpriors as informative as possible without being overconfident.

As a comparison to the narrower-hyperprior model, the wider-hyperprior model has very diffuse hyperpriors to the point where they are relatively uninformative. Our main analyses will focus on the narrower-hyperprior model, while the wider-hyperprior model will be used as a contrast to show catastrophic errors in the presence of small sample sizes.

Table 1. Hyperpriors for two versions of the hierarchical model.

|  |  |  |  |
| --- | --- | --- | --- |
| Parameter | Parameter name | Narrower prior | Wider prior |
|  | Group *d′* mean location | 1 | 1 |
|  | Group *d′* mean scale | 1 | 10 |
|  | Group *d′* SD location | 0.5 | 0.5 |
|  | Group *d′* SD scale | 2 | 15 |
|  | Group *c* mean location | 0 | 0 |
|  | Group *c* mean scale | 1 | 10 |
|  | Group *c* SD location | 0.5 | 0.5 |
|  | Group *c* SD scale | 2 | 15 |



Figure 1. Densities for the hyperpriors for the “narrower” (Prior 1) and “wider” (Prior 2) models. See Table 1 for parameter values.

## **Overall model specification**

Overall, the hierarchical Bayesian model that we apply for the yes/no task is as follows:

Let and denote the number of hits and false alarms for the subject. The hits and false alarms are binomially distributed: and where denotes the probability of a hit or false alarm (Dorfman & Alf, 1968). The are derived from a probit transformation of an individual’s *d′* () and criterion (): and .

We assume that values of *d′* and *c* across subjects are normally distributed:  and where and are the mean and standard deviations for the corresponding parameters – these are the group-level (prior) distributions for *d′* and *c*. Each of the and in turn have their own distributions (hyperpriors)[[3]](#footnote-3): , , , and , where and are the location and scale for the corresponding parameter – these are the hyperparameters for the hierarchical model. A graphical representation of this model is shown in Figure 2. For estimating group *d′*, the target distribution is the posterior distribution for the group *d′* ().

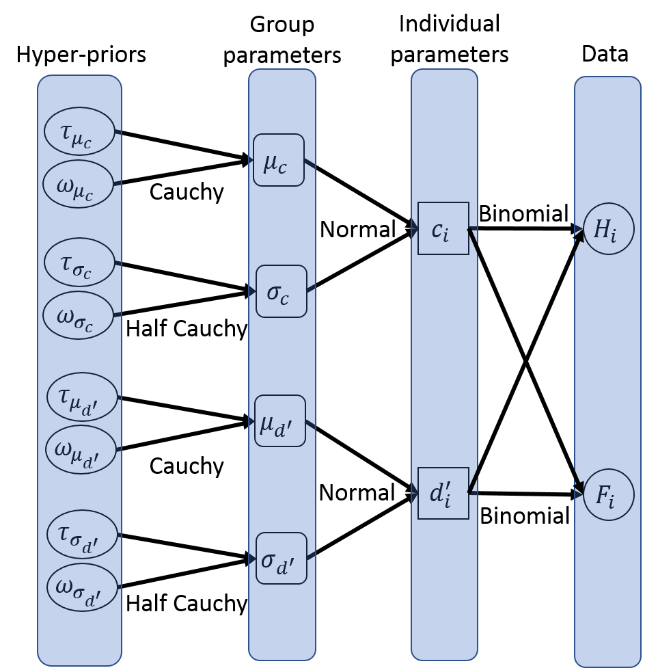


Figure 2. A hierarchical Bayesian model. Refer to the text for details.

## **Extreme proportions**

A feature of the Bayesian model is that corrections are not required for conducting inference in the presence of extreme proportions (Lee, 2008). This is because the Bayesian analysis approaches the problem of estimation by constructing a generative model – i.e., for a given combination of parameters and model, what is the likelihood that we would have obtained the observed data.

Although a correction is not strictly required in this context, it does not imply that a Bayesian hierarchical model would generate better estimates of group *d′* than a technique that requires corrections. This is because corrections have biasing properties that can improve estimates depending on the correction value/method. For example, adding a large constant to each cell in the SDT outcome matrix will bias estimates of *d′* towards zero, which would improve estimates if the true *d′* is also close to zero.

# **Method**

## **Evaluating estimators**

For the hierarchical model, we will derive a point estimate of group sensitivity (*d′*) by taking the posterior median of the *d′* mean hyperprior (). We use the median over other estimates of central tendency (e.g., mean or mode) because of robustness. Because we have an extreme range of data and we simulate a massive quantity of datasets, we cannot guarantee that we generate the optimal number of posterior samples for all the simulated datasets. For example, unusual datasets, especially those with a small sample size, may require additional modifications to the modelling procedure in order to optimally sample from the posterior. In our previous testing, we found that the median tended to perform better than the posterior mean for small sample sizes.

We are concerned with two aspects of estimator quality, bias and efficiency. The bias of an estimator refers to the difference between the expected value of an estimator and the true value of the parameter being estimated. In contrast, the efficiency of an estimator refers to the average difference between an estimator and the true value of the parameter being estimated. Our primary interest will be in the efficiency of these estimators, which we will index using the square root of the average squared deviation between an estimator and the true value (root-mean-squared error; RMSE)[[4]](#footnote-4). Corresponding results for bias will be shown in the Appendix.

## **Simulation study**

A simulated experiment was defined by five parameters: group *d′* mean (), group *d′* standard deviation (), group *c* mean (), group *c* standard deviation (), number of trials (; stimulus-present and noise trials), and the number of subjects (). The values of each parameter were as follows: , , , and . The group *d′* standard deviation was set to be 1/5 of the group *d′* mean, and the group *c* mean was set to be zero. A design matrix was generated by parametrically varying the group *d′* mean, group *c* standard deviation, number of trials, and the number of subjects; this yielded 315 different permutations.

In each simulated experiment, the designated combination of parameters was input into the data-generating model. This generated a group-level sampling distribution for *d′* and *c*, parameterised by , , , and . subjects were then generated by sampling individual-level *d′* and *c* values from their corresponding sampling distribution. These subject parameters were then used to generate the frequency of hits () for stimulus-present trials and the frequency of false alarms () for stimulus-absent trials, for each subject. The collection of hits and false alarms for each subject comprised the results for a single replication of a simulated experiment.

The data-generating model was as follows:

The *d′* for the subject was gamma distributed, parameterised by the group *d′* mean () and the group d′ standard deviation (), where denotes the total number of subjects. The *c* for the subject was normally distributed, parameterised by the group *c* mean () and the group *c* standard deviation (), The choices for the data-generating model were inherited from Hautus (1997).

Probabilities of hits and false-alarms for the subject were obtained using the probit transformation (): and The observed hits and false-alarms for the subject were subsequently generated from a binomial distribution: and where denotes the number of trials for that trial type.

Note here that the generative process specifies that the subject-level *d′* is gamma distributed, whereas our hierarchical model specifies a normal distribution. This is because even though the true *d′* for a subject is positive-bounded, the binomial response process can generate hits and false alarms that are consistent with negative *d′* values. From previous testing, our model had to be consistent with the possible range of sample *d′*, especially for small sample sizes. Otherwise, our posterior estimates would be positively biased.

## **Computational details**

For each condition of the design matrix (i.e., each permutation of experimental parameters), 10,000 simulated datasets were created.

For the Bayesian analysis, model fitting was done using the Stan software package (Carpenter et al., 2017). For each simulated dataset, 4 Markov Chain Monte Carlo (MCMC) chains were run for 5000 iterations, with 2000 iterations of warm-up. For 315 conditions and 10,000 generated datasets per condition, this resulted in a total of 12.6 million MCMC *chains* being generated, totalling 63 billion MCMC iterations in total.

# **Results**

Our analyses will focus on the narrower-hyperprior model (see Table 1). Unless otherwise specified, the results shown will pertain to the model with that set of priors. In this section, we will show only the results for the RMSE. For posterity, bias results are included in the Appendix.

Figures 3, 4, and 5 show the RMSE of the Bayes, averaged and pooled estimators of group *d′* for different values of the true group *d′*, *c* standard deviation, number of trials, and the number of subjects. In general, each estimator was more efficient (lower RMSE) as a function of the number of trials and the number of subjects. The efficiency of the Bayes estimator was relatively unaffected by the standard deviation of *c*. In contrast, the averaged and pooled estimators tended to be less efficient as the standard deviation of *c* increased, especially when the true group *d′* was also large. Furthermore, the pooled estimator was particularly inefficient at large values of *c* standard deviation, because the asymptotic RMSE (as a function of subjects and trials) was consistently higher than that of the Bayes and averaged estimators.

In general, the Averaged estimator tended to be at least as efficient as the Bayes estimator when the number of subjects and the number of trials were small. However, with more subjects, the Bayes estimator becomes more efficient than the averaged estimator at low trial numbers, especially when the standard deviation of *c* was higher.



Figure 3. The root mean squared error (RMSE) for each estimator (columns) as a function of the standard deviation of *c* (rows), trial number, and subject number. The true group *d′* is 0.5.



Figure 4. The root mean squared error (RMSE) for each estimator (columns) as a function of the standard deviation of c (rows), trial number, and subject number. The true group d′ is 1.5.



Figure 5. The root mean squared error (RMSE) for each estimator (columns) as a function of the standard deviation of *c* (rows), trial number, and subject number. The true group *d′* is 3.

Table 2 shows the estimators with the lowest RMSE as a function of the number of subjects, the true group *d′*, and the standard deviation of *c*. RMSE values within 5% of each other are treated as equivalent. A decision tree is shown in Figure 6 to summarise the results of Table 2 and Figures 3 to 5. The decision tree was created using a non-parametric classification tree, where the target of the classifier was the estimator that had the lowest RMSE. Because this classifier does not allow for ties between estimator that have RMSEs with similar values, it should only be used as a visual summary of the present results.

Overall, the RMSE results show that the Bayes estimator tended be more efficient when the standard deviation of *c* was moderate-to-high (≥ 0.6), especially when there were more than four subjects and the true group *d′* was moderate-to-high (≥ 0.5). When the standard deviation of *c* was low, the averaged and pooled estimators tended to be more efficient when there were fewer than eight trials. In addition, when the standard deviation of *c* was low, and the true group *d′* was high (> 1.5), the Bayes estimator tended to be more efficient with more than 16 subjects. Finally, when the standard deviation of *c* was low and the true group *d′* was around 1.5, the Bayes estimator tended to be more efficient than the pooled and averaged estimators.

Table 2. The estimator (A = averaged, P = pooled, B = Bayes) with the lowest RMSE (within 5% of the estimator with the lowest RMSE) as a function of the number of subjects, number of trials, the true group *d′*, and the standard deviation of *c*. Grey cells indicate configurations where the Bayes estimator was one of the best estimators.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  | Trials | | | | | | |
| Subjects | Group *d′* | SD(*c*) | 4 | 8 | 16 | 32 | 64 | 128 | 256 |
| 2 | 0.5 | 0.2 | A | A/B | B | P/B | A/P/B | A/P/B | A/P/B |
| 2 | 0.5 | 0.6 | A | A | A/P/B | P/B | P/B | A/P | A/P/B |
| 2 | 0.5 | 1 | A | A | A/P | A/P | A/P | A/P | A |
| 2 | 1.5 | 0.2 | A | A | B | B | B | B | B |
| 2 | 1.5 | 0.6 | B | B | B | B | B | B | B |
| 2 | 1.5 | 1 | B | B | B | B | B | B | B |
| 2 | 3 | 0.2 | B | P | A/P | A | A | A/P | A/P |
| 2 | 3 | 0.6 | B | P | A/P | A | A | A | A |
| 2 | 3 | 1 | B | P/B | A/P/B | A/B | A | A | A |
|  |  |  |  |  |  |  |  |  |  |
| 4 | 0.5 | 0.2 | A | A | A/P/B | A/P/B | A/P/B | A/P/B | A/P/B |
| 4 | 0.5 | 0.6 | A | A | A/P | P | A/P | A/P/B | A/B |
| 4 | 0.5 | 1 | A | A | A/P | A/P | A/P | A/B | A/B |
| 4 | 1.5 | 0.2 | A/P | A | A | B | B | P/B | A/P/B |
| 4 | 1.5 | 0.6 | P | A/P | A | A/B | A/B | A/B | B |
| 4 | 1.5 | 1 | B | B | B | B | B | B | B |
| 4 | 3 | 0.2 | P | P | P | A | A | A | A/P |
| 4 | 3 | 0.6 | P | P | P | A | A | A | A |
| 4 | 3 | 1 | P | P/B | B | B | A/B | A | A |
|  |  |  |  |  |  |  |  |  |  |
| 8 | 0.5 | 0.2 | A | A | A/P | A/P/B | A/P/B | A/P/B | A/P/B |
| 8 | 0.5 | 0.6 | A | A/P | A/P | A/P | A/P/B | A/B | A/B |
| 8 | 0.5 | 1 | A | A/P | A/P | A | A/B | A/B | B |
| 8 | 1.5 | 0.2 | P | A/P | A/P | A/P/B | A/P/B | A/P/B | A/P/B |
| 8 | 1.5 | 0.6 | P | P | A/B | A/B | A/B | A/B | A/B |
| 8 | 1.5 | 1 | B | B | B | B | B | B | B |
| 8 | 3 | 0.2 | P | P | P | A/P | A | A | A/B |
| 8 | 3 | 0.6 | P | P | P | A/B | A | A | A |
| 8 | 3 | 1 | P/B | B | B | B | B | B | B |
|  |  |  |  |  |  |  |  |  |  |
| 16 | 0.5 | 0.2 | A | A | A/P | A/P/B | A/P/B | A/P/B | A/P/B |
| 16 | 0.5 | 0.6 | A | A/P | A/P | A/P/B | A/B | A/B | A/B |
| 16 | 0.5 | 1 | A/P | A/P | A | A/B | A/B | B | B |
| 16 | 1.5 | 0.2 | P | P | A/P/B | A/P/B | A/P/B | A/B | A/B |
| 16 | 1.5 | 0.6 | P | B | B | A/B | A/B | A/B | A/B |
| 16 | 1.5 | 1 | B | B | B | B | B | B | B |
| 16 | 3 | 0.2 | P | P | P | P/B | A | A | A/B |
| 16 | 3 | 0.6 | P | P | B | B | B | A/B | A/B |
| 16 | 3 | 1 | B | B | B | B | B | B | B |
|  |  |  |  |  |  |  |  |  |  |
| 32 | 0.5 | 0.2 | A | A/P | A/P/B | A/P/B | A/P/B | A/P/B | A/P/B |
| 32 | 0.5 | 0.6 | P | A/P | A/B | A/B | A/B | A/B | A/B |
| 32 | 0.5 | 1 | P | A/P/B | A/B | B | B | B | B |
| 32 | 1.5 | 0.2 | P | P | A/P/B | A/B | A/B | A/B | A/B |
| 32 | 1.5 | 0.6 | P | B | B | B | A/B | A/B | A/B |
| 32 | 1.5 | 1 | B | B | B | B | B | B | B |
| 32 | 3 | 0.2 | P | P | P/B | B | B | A/B | A/B |
| 32 | 3 | 0.6 | P | B | B | B | B | B | A/B |
| 32 | 3 | 1 | B | B | B | B | B | B | B |



Figure 6. A classification tree summarising the results of Table 2. The histograms at the terminal nodes show the proportion of each estimator (A = averaged, B = Bayes, P = pooled) that had the lowest RMSE for that tree partition.

Regarding bias, the bias of each estimator was a function of the number of subjects and the number of trials per subject. For the Bayes estimator, there was non-negligible bias present up to the maximum assessed number of trials (256) and subjects (32). With eight subjects, the Bayes estimator generally converged towards the true value of group *d′* with negligible bias. The pooled estimator had a negative bias that was proportional to the standard deviation of *c*, and pooling across more subjects adversely affected the rate of convergence of the pooled estimates to the true group *d′*. The averaged estimator converged towards the true group *d′*, but the rate of convergence was slower when the group *d′* and standard deviation of *c* were higher. See the Appendix for details.

## **Different priors**

Figure 7 shows the RMSE for the Bayes estimator with different hyperpriors (Table 1). When the hyperpriors were wide (wider-hyperprior model), the Bayes estimates were severely inefficient compared to the narrower-hyperprior model when the sample size (trials and subjects) was small. With 32 subjects, the difference in bias and RMSE between the two models was negligible regardless of sample size.



Figure 7. Root mean squared error, for the narrow- and wide-prior models, as a function of the true group d′, standard deviation of c, number of trials, and the number of subjects.

# **Discussion**

Bayesian posterior distributions represent a compromise between the prior and the data. Thus, the accuracy of the estimates from a hierarchical model will depend on an interaction between the number of subjects, the number of trials, and the true group *d′*. Consequently, the exact values of our conclusions depend on the priors that we used. We cannot guarantee to what extent our current conclusions generalise to point estimates generated from other models or the same model with different priors. However, in our main analysis, the hyperpriors were arranged to be relatively diffuse, while maintaining most of the density for the group *d′* around 0 to 3 (Figure 1; Table 1). Thus, we do expect some robustness of the results of our main analysis for data that represent interpolations or slight extrapolations with respect to the datasets that we have simulated in the present paper.

In our main analysis, we examined the point estimates obtained from a hierarchical model, where location and scale parameters of the *d′* hyperprior were 1. For this model (Figure 1; narrower-hyperprior model) and this set of priors, the Bayes estimator tended to be more efficient than the pooled and averaged estimators when the standard deviation of *c* was moderate-to-high (≥ 0.6). When the standard deviation of *c* was low, the Bayes estimator tended to be more efficient when there were more than eight trials per subject, and either 1) the true group *d′* was between 0.5 and 1.5, or 2) the true group *d′* was larger than 3 and there were more than 16 subjects.

One of the key results was that the performance of the Bayes estimator was relatively invariant with respect to the standard deviation of *c*. This is because the sample variance of *c* is accounted for in the Bayesian hierarchical model. In contrast, the efficiency of the pooled and averaged estimators tended to decrease as the standard deviation of *c* increased.

In practice, the true group *d′* and *c* are unknown. Thus, Figure 8 shows a reanalysis of the classification tree derived from Table 2 (Figure 6), where the true group *d′* and standard deviation of *c* are not included in the classification tree model. From the set of parameters examined in the present study, the Bayes estimator tended to be more efficient with more than eight subjects and eight trials. Thus, our recommendation from Figure 8 is that if we have *absolutely no information* about the plausible values of the true group *d′* and standard deviation of *c*, *we should prefer the Bayes estimate when the number of subjects is greater than eight and the number of trials is greater than 16*. Note that caution should be applied when generalising this result beyond the mixture sampling distribution that generated the dataset used in Figure 8.

We reiterate again here that our recommendations are based on the hyperpriors that we have arranged. For example, if the stimuli are arranged to elicit mislabelling, such as in the case of visual illusions, we would expect the group *d′* to be negative. Consequently, our hyperpriors would be inappropriate and the averaged or pooled estimators would be more efficient and/or less biased. However, if one can use domain knowledge and *a priori* knowledge about the dataset to construct more accurate priors than the relatively generic priors we have arranged here, it is likely that our recommendations can be relaxed and that the hierarchical model may be more efficient with fewer subjects or fewer trials.



Figure 8. Reanalysis of Table 2 (cf. Figure 6) with the true group d′ and standard deviation of c excluded from the classification tree. The histograms at the terminal nodes shows the number of instances that each estimator had the lowest RMSE for that combination of parameters.

## **The effect of wide hyperpriors**

In a secondary analysis (Figure 7), we compared the efficiency of the point estimates for group *d′* from two versions of the hierarchical model (Figure 1; Table 1) with a wider or narrower hyperprior. The general effect of having a wider prior was that the small-sample (trials and subjects) efficiency was worse. However, with more trials and more subjects, the differences in efficiency became negligible due to the increased influence of the likelihood on the posterior.

In terms of bias, there is additional nuance in the results (Appendix 1; Figure A2). There was an interaction between the number of subjects, the informativeness of the prior, and the true group *d′*. With only two subjects, the posterior median of the narrower-hyperprior model tended to be less biased for small-to-moderate number of trials, but slightly more biased when the number of trials was large. This represents a situation where we have very high certainty about the subject-level parameters (many trials), but less certainty about the group-level parameters (few subjects). Although wider hyperpriors may reduce bias in these situations, the efficiency is still relatively poorer than with narrower hyperpriors (Figure 7). Therefore, in these situations of large trial numbers but few subjects, we recommend shifting the location of the hyperpriors rather than defaulting to wider hyperpriors.

## **Tools for implementing an analysis**

To implement an analysis using a version of the hierarchical model (Figure 2), our supplementary material includes a set of R functions with annotated examples for fitting the hierarchical model, conducting diagnostics, and extracting the averaged, pooled, and Bayes estimates. In addition, the supplementary material contains a Shiny application written in R that provides a graphical interface for importing a dataset and fitting the hierarchical model. This application also includes graphical tools for the selection of hyperprior parameters from an imported dataset, as well as graphics for quick visualisation of the marginal posteriors for the group-level parameters and each of the individual-level parameters.

## **Conclusion**

We examined the efficiency of point estimates of group *d′* from a Bayesian hierarchical model (Figure 2) with Cauchy hyperpriors on the group *d′* and group *c*. For the model and the set of priors used here, the Bayes estimates tended to be more efficient than estimates obtained from averaging *d′* across subjects or by computing *d′* by pooling hits and false alarms across trials, given that some sample size requirements for the number of subjects and the number of trials are met. The general result was that the Bayes estimate should be preferred when more than eight subjects and eight trials per subject are available.

# **Appendix**

## **Bias results**

Figure A1 shows the estimates of *d′* for the pooled, averaged, and Bayes estimators for each combination of group *d′* and *c* standard deviation as a function of the number of trials and the number of subjects. To simplify the plot, only one series was plotted for the averaged estimator (averaged over two subjects) – Note that in expectation, the estimate of group *d′* obtained from averaging is the same regardless of the number of subjects averaged over, due to the linear properties of averaging.

From Figure A1, the bias of these estimators is deduced by the difference between each estimator and the true group *d′*. Because the hyperpriors will affect bias for the Bayes estimator, the conclusions that we wish to draw should be considered across different values of *d′*.

The bias of the Bayes estimator appeared to be a function of the number of subjects and the number of trials. With only two subjects, there was non-negligible bias present up to the maximum assessed number of trials (256) and subjects (32). However, increases in sample size along both these dimensions decreased the bias of the Bayes estimator. With eight subjects, the Bayes estimator generally converged towards the true value of group *d′* with negligible bias.

The bias for the averaged and pooled estimators replicate those shown previously (Hautus, 1997). The pooled estimator had a negative bias that was proportional to the standard deviation of *c*, and pooling across more subjects decreased the rate of convergence of the pooled estimates to the true group *d′*. The averaged estimator converged towards the true group *d′*, but the rate of convergence was slower when the group *d′* and standard deviation of *c* were higher.

The bias for the narrower- and wider-hyperprior models (see Table 1) are shown in Figure A2. In general, the wider-hyperprior model tended to be more biased when the sample size (subjects and trials) were small. The exception appeared to be when the number of subjects was small (2), there were many trials, and the true group *d′* was large. In these cases, the hyperpriors for the narrower-hyperprior may be too strong for when there are only two subjects. However, with many subjects, our *a priori* uncertainty about the posterior group *d′* is better informed, so we may better specify a relatively informative prior instead of defaulting to a very diffuse prior.



Figure A1. Estimates of group *d′* for the pooled, averaged, and Bayes estimators. To simplify the graphs, only one series is plotted for the averaged estimator (averaged across two subjects). Narrower-hyperprior model (see Table 1).



Figure A2. Estimates of *d′* from the Bayesian hierarchical model with different priors, as a function of the number of subjects, the number of trials, the true group *d′*, and the standard deviation of *c*. Wider-hyperprior model compared to the narrower-hyperprior model (see Table 1).

# **References**

Berger, J. O. (1985). *Statistical decision theory and Bayesian analysis*. Springer New York. http://public.ebookcentral.proquest.com/choice/publicfullrecord.aspx?p=3086710

Bi, J., & Ennis, D. M. (1998). A Thurstonian variant of the beta-binomial model for replicated difference tests. *Journal of Sensory Studies*, *13*(4), 461–466. https://doi.org/10.1111/j.1745-459X.1998.tb00100.x

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., & Riddell, A. (2017). Stan: A Probabilistic Programming Language. *Journal of Statistical Software*, *76*(1). https://doi.org/10.18637/jss.v076.i01

Dorfman, D. D., & Alf, E. (1968). Maximum likelihood estimation of parameters of signal detection theory—A direct solution. *Psychometrika*, *33*(1), 117–124. https://doi.org/10.1007/BF02289677

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Analysis*, *1*(3), 515–534. https://doi.org/10.1214/06-BA117A

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., & Rubin, D. B. (2014). *Bayesian data analysis* (Third edition). CRC Press.

Hautus, M. J. (1997). Calculating estimates of sensitivity from group data: Pooled versus averaged estimators. *Behavior Research Methods, Instruments, & Computers*, *29*(4), 556–562.

Lee, M. D. (2008). BayesSDT: Software for Bayesian inference with signal detection theory. *Behavior Research Methods*, *40*(2), 450–456. https://doi.org/10.3758/BRM.40.2.450

Macmillan, N. A., & Creelman, C. D. (2004). *Detection Theory: A User’s Guide*. Psychology Press.

McNeish, D. (2016). On using Bayesian methods to address small sample problems. *Structural Equation Modeling: A Multidisciplinary Journal*, *23*(5), 750–773. https://doi.org/10.1080/10705511.2016.1186549

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. https://doi.org/10.3758/BF03196750

Tanner, W. P., Jr., & Swets, J. A. (1954). A decision-making theory of visual detection. *Psychological Review*, *61*(6), 401–409. https://doi.org/10.1037/h0058700

1. Criterion variability in this situation can be thought of as a source of between-subject variability that, within each subject, is independent of the sensitivity. However, when we assume that the within- and between- subject variability in the criterion is interchangeable, our estimates of group sensitivity become biased towards zero (Hautus, 1997). [↑](#footnote-ref-1)
2. This model is misspecified when the hits and false-alarms are over/under-dispersed compared to binomial variance. In these situations, we may wish to use other likelihood functions, such as the beta-binomial distribution (e.g., Bi & Ennis, 1998). Implementing different likelihood models in the hierarchical model would follow the same logic as that described for the binomial likelihood. [↑](#footnote-ref-2)
3. We used Cauchy and Half-Cauchy distributions for our hyperpriors, based on the recommendations of Gelman (2006). Cauchy distributions can yield relatively diffuse priors due to their long tails, which makes our model more robust to outliers, while still being semi-informative. See Rouder and Lu (2005) for an alternative model with inverse-Gamma priors for variance parameters and normal priors for means. [↑](#footnote-ref-3)
4. One may argue that it is unusual to investigate the “Frequentist” properties of Bayesian models. However, the evaluation of Bayesian point estimates has a long history in decision theory (e.g., Berger, 1985). In our application, we are using a Bayesian model to generate a point estimate. Thus, the point estimates are estimators as would any other statistic be and consequently has properties of bias, efficiency, and consistency (Rouder & Lu, 2005). [↑](#footnote-ref-4)