Simulating Sensitivity to Reach Powerful Conclusions: A Comparison of Aggregated and Hierarchical Bayesian Models of Signal Detection

Emily N. Mech

Department of Psychology, University of Illinois at Urbana-Champaign

Psychology 555: Detection and Discrimination

Dr. Aaron Benjamin

April 25, 2022

Simulating Sensitivity to Reach Powerful Conclusions: A Comparison of Aggregated and Hierarchical Bayesian Models of Signal Detection

Signal detection theory is a powerful framework used to model decisions made under uncertainty (Green & Swets, 1966). The framework can be applied to many tasks, although it has been commonly used to model participant responses indicating whether they detected a signal (e.g., a flash of light) or participant responses choosing between two alternative choices (e.g., is this stimulus old or new?). The basic premise behind signal detection theory is that decisions made under uncertainty can be modeled with a noise distribution and a signal distribution. The distance between the noise and signal distribution gives the metric d', which describes how discriminable the signal is from the noise. However, the real contribution of signal detection theory is that it can consider *how* people may employ different decision criteria to make a decision separately from sensitivity (e.g., Stanislaw & Todorov, 1999). For example, when told to prioritize making an accurate decision over a fast decision, participants may use one criterion for making decisions, and this may differ from a criterion used when prioritizing speed over accuracy. Signal detection theory has led to many breakthroughs in theoretical and applied domains of decision making such as perception (e.g., Gescheider, 1997; Wixted, 2020), recognition memory (e.g., Glanzer, Adams, Iverson, & Kim, 1993), diagnostics (e.g., Swets, 1988; Swets, Dawes, & Monahan, 2000), weather forecasting (cf. DeCarlo, 1998), and even police lineup procedures (e.g., Wixted, 2020).

Signal detection theory has been implemented with many different analysis approaches (e.g., Stanislaw & Todorov, 1999; DeCarlo, 1998; Rouder & Lu, 2005). To get the measure of sensitivity or discriminability, d', participants' responses can be characterized as hits (correctly respond *old* when the stimulus is old), misses (respond *new* when the stimulus is old), correct

rejections (respond *new* when the stimulus is new), or false alarms (respond *old* when the stimulus is new). From these response classifications, participants' hit and false alarm rates can be calculated, and taking the difference between the normalized hit and false alarm rates gives the measure, d'. This method allows one to calculate d' for each participant. Averaging over all participants gives the population level d' which allows conclusions to be made about experimental conditions. However, this approach does not explicitly account for variability in different participants' decisions due to, for example, differences in memory ability. It also does not consider item variability due to differences in, for example, item difficulty or bias.

While averaging over participants and items should reduce the "noise" in the sensitivity estimate associated with these sources of variability, there are several problems that could arise if there is a failure to explicitly account for participant and item level variability. First, by treating participants and items as fixed effects, an implicit assumption is made that the participants and items in the given study constitute the entire population of participants and items to which one would like to generalize. In the context of language research, Clark (1973) argued this very point in the context of the words that experimenters used as stimuli and termed it the "language-as-fixed-effect fallacy." If participants and items are not treated as random variables, there is not statistical evidence allowing one to generalize the experimental findings beyond the participants and items in a given experiment. However, this generalization is precisely what most research aims to do, and often researchers make this claim without the statistical backing for it (Clark, 1973).

Second, estimating discriminability while aggregating over both participants and items can lead to conclusions that problematically do not apply to individual participants or items in the study. For example, Estes (1956) raised this issue in the context of learning curves, noting

that obtaining the mean, population level learning curve does not necessarily provide the information necessary for describing individual's learning. In the context of signal detection theory, Rouder and Lu (2005) demonstrated that when there is item variability present, estimating sensitivity without accounting for the item variability leads to a systematic underestimation of d'. Although potentially less devastating when estimating linear models, failing to account for participant and item level variable can be particularly detrimental in nonlinear contexts such as those typically under investigation with signal detection theory (Rouder & Lu, 2005).

Beyond systematically mis-estimating effects when participant and item level variability is present in the data, aggregation can also impact claims made from assessing the curvature of zROC curves (Morey, Pratte, & Rouder, 2008). As zROC curves have been widely used in the literature to inform theories of recognition memory, artifacts due to aggregation would be particularly problematic for choosing between competing theories and forming conclusions. The problem is exacerbated by the fact that the source of variability could differ. For example, items in a recognition memory experiment may vary in memorability or bias or both. Failing to account for this variation by aggregating over items limits conclusions to describing how variability affects responses from a bird's eye view (Morey, Pratte, & Rouder, 2008).

Finally, failing to account for participant and item level variability present in the data can lead to an inflation of the Type I error rate of the test (e.g., Barr, Levy, Scheepers, & Tily, 2013). As researchers, we set an alpha level to 0.05 (or lower) and make conclusions understanding that the false positive rate is controlled at the specified alpha level. However, if variability due to participants and items is not properly accounted for, the false positive rate of the test could creep

much higher than the alpha level, and dangerously, this would not be easily diagnosable to the researcher.

However, while many implementations of signal detection theory require aggregating over participants and items, there has been a recent push to implement signal detection theory within a hierarchical framework which allows the variability due to participants and items to be modeled as random effects (Rouder & Lu, 2005; Rouder, Lu, Speckman, Sun, & Jiang, 2005; Rouder, Sun Speckman, Lu, & Zhou, 2003; Rouder et al., 2007; DeCarlo, 2010). By partialing out the estimated variance due to participants and items, hierarchical models (also termed mixed models) solve many of the problems created by aggregation. Additionally, hierarchical models can be implemented with frequentist or Bayesian estimation methods. While frequentist approaches dominate the field of psychology, as computational power has become more accessible, Bayesian estimation has started to become more widely used. Bayesian estimation has several benefits relative to frequentist estimation such as a more intuitive interpretability, conditioning only on data that has been observed, the possibility to incorporate prior knowledge, and the ability to collect evidence in favor of the null hypothesis (Wagenmakers, Lee, Lodewyckx, & Iverson, 2008; Kruschke & Liddell, 2017).

Rouder and colleagues published a series of studies fitting hierarchical Bayesian models to recognition decisions and response times (Rouder & Lu, 2005; Rouder, Lu, Speckman, Sun, & Jiang, 2005; Rouder, Sun, Speckman, Lu, & Zhou, 2003; Rouder et al., 2007). Across these studies, they demonstrated that using hierarchical models resolves the problem of systematic underestimation of sensitivity that arises in aggregated techniques. Additionally, they have shown that for small sample sizes as can be typical for some experiments invoking signal detection theory, Bayesian estimation is more accurate than maximum likelihood estimation

(Rouder et al., 2003). Finally, they have also shown that hierarchical Bayesian models provide accurate fits to the data when the sampling strategy is to collect a large number of participants with limited numbers of trials (Rouder et al., 2005).

The current study follows the logic of Rouder and colleagues' approach to assess the impact of aggregation and Bayesian estimation on the d' parameter. Three different models under the umbrella of equal variance signal detection theory were implemented. The goal of the current work is to conclude whether there are differences in the Type I error rate across implementations that vary in aggregation and estimation method.

The Present Study

In the present study, data was simulated following the design of a planned experiment. In the planned experiment, participants will be emailed an online survey that asks about their preferences for a variety of topics. After responding to the survey, two participants at a time will come into the lab to learn about each other's preferences (see Coronel & Federmeier, 2016 for example materials). The experimenter will provide prompts of specific preferences to discuss based on each partner's answers to the pre-study survey. After discussing the preferences prompted by the experimenter, participants will concurrently read sentences containing statements about either their own or their partner's preferences. After reading each sentence, participants will respond whether the preference stated in the presented sentence was correct or incorrect. As the stimuli will be created from the pre-study survey, the presented sentences will contain preferences that are both correct and incorrect as well as known and unknown to the partner. For example, if the participants (e.g., Emily and Melinh) discussed that Emily's favorite pizza topping was artichokes, after reading the sentence, "Emily's favorite pizza topping is

pepperoni.," both participants should respond that the preference in the sentence was incorrect as this information was known by both participants.

Data simulated for this 2 (correct: yes, no) x 2 (known: yes, no) design contained variability due to both participants and items. In the simulated data, condition effects due to the correctness of the presented item and whether the preference was known were set to 0. An equal variance signal detection model was implemented with three methods: 1. Manually calculating point estimates of condition discriminability (Stanislaw & Todorov, 1999), 2. Estimating condition discriminability with a Bayesian generalized linear model (DeCarlo, 1998), 3. Estimating condition discriminability with a Bayesian generalized linear mixed model (Rouder & Lu, 2005). These different implementations were chosen such that they varied in the estimation method and the treatment of participant and item level variability. As all condition effects were simulated to be null, critical comparison of the different implementations allows conclusions about how the treatment of participant and item level variability by the different methods impacts false positive rates and condition sensitivity estimates. I predict that the Bayesian generalized linear mixed model will have lower false positive rates than both the manual point estimate method and Bayesian generalized linear model method as it is the only method that does not aggregate over participant and item level variability.

Method

Data Simulation

All data simulations were generated in R (R Core Team, 2020) with the faux (DeBruine, 2021) package. The data generating simulation followed the extended binomial method shared in DeBruine and Barr (2021). Using this method, data was generated from a random normal multivariate distribution based on a specified number of participants and items, specified

participant and item level variability and correlations, as well as specified condition effects. To simulate data from the planned 2 (correct, incorrect) x 2 (known, unknown) experimental design, the number of participants was always set to 40, and the number of total items was specified as 100 (known: 50, unknown: 50). By-item and by-participant correlations were also set to 0, and all participant and item level variability was set to 1.

For each participant, the gaussian response for each item was calculated by adding the intercept (grand mean, 0), the condition effect for correctness of the preference (0), the condition effect for knowledge of the preference (0), and the interaction effect of the conditions (0). This effect was then transformed with an inverse logit function to get the probability of answering with a "correct" response (50%). Simulated responses for each item were then obtained by sampling from a random Bernoulli distribution with the calculated probability of answering with a "correct" response. Each trial was then labeled as either a hit, miss, correct rejection, or false alarm.

Data Analysis

The data were analyzed with three different implementations of equal variance signal detection theory. All models and estimates were generated in R (R Core Team, 2020), and Bayesian models were generated with the brms package (Bürkner, 2017). The analysis approach follows the model fitting procedure specified in Vuorre (2017a). The first implementation involved manually calculating a point estimate for the discriminability of the preferences based on whether they were correct or incorrect directly from the data (Stanislaw & Todorov, 1999). First, items were aggregated over, and the number of hits, misses, correct rejections, and false alarms was counted for each participant. From these counts, the Z-transformed hit rate and false alarm rate were calculated. D' was calculated by subtracting the Z-transformed false alarm rate

from the Z-transformed hit rate for each participant. The population d' was obtained by averaging d' across participants.

A Bayesian generalized linear model was fit for the second implementation of equal variance signal detection theory (DeCarlo, 1998). This approach allows predictors to be estimated for d' but does not separately model variability due to participants and items. The generalized linear model was fit by specifying the Bernoulli distribution with a probit link function. Participant responses were estimated from predictors coding whether the preference was correct or incorrect, whether the preference was known or unknown, as well as the interaction. The posterior distribution was estimated by sampling four MCMC chains for 2000 iterations, and the priors were set to the default. The intercept estimate can be interpreted as the criterion, and the predictor estimates as the d' for that condition.

Finally, a Bayesian generalized linear mixed model was fit as the third implementation of equal variance signal detection theory (EVSDT) (Rouder & Lu, 2005). This implementation was identical to the generalized linear model with the exception that participant and item level intercept and slope variability were accounted for in the model.

Power Simulation

Simulating data and analyzing it with three different methods allows one to draw conclusions about which method is preferred for recovering the effects present in the data. However, as with experimental work, to trust the conclusions drawn from such a comparison, the results must be replicated many times. Therefore, following the approach prescribed by Debruine and Barr (2021), multiple experiments were simulated with the same data and analysis parameters with the goal of understanding the power of each model to recover the effects present in the data. Each experiment consisted of simulating data drawn from the same distribution

described above and analyzing the same data with each of the EVSDT methods. To draw a conclusion about the results of the analysis for each experiment, the condition effects (Implementation 1: correctness; Implementations 2 & 3: correctness, knowledge, interaction effect) were tested with a one sample t-test (Implementation 1) or a one sample hypothesis test comparing the posterior probability under the hypothesis against the alternative of a null (0) effect (Vuorre, 2017b). One hundred experiments were simulated with this procedure.

The validity of the conclusions of each method was then tested by averaging the number of significant results found for each condition across the one hundred experiments. As the real condition effects in the simulated data were set to 0, a model has successfully recovered the effect from the data if the estimate of each condition effect is null (non-significant). However, as there is variability included in the data simulation process, there will sometimes be false positive results. In this experiment, the alpha level was set to 0.05. Any implementation that maintains an alpha level of .05 or below across the 100 experiments we will conclude as having successfully recovered the condition effects in the data.

Results

The d' for determining whether the preference presented was correct or incorrect as calculated from each analysis method for each of the one hundred replications of the experiment is plotted in Figure 1. Descriptively, it should be noted that the d' calculated for the same condition for the same data (same replication) is not equivalent across methods. However, it should be noted that across replications, the d' for each method hovers around 0, which is the true condition effect present in the data.

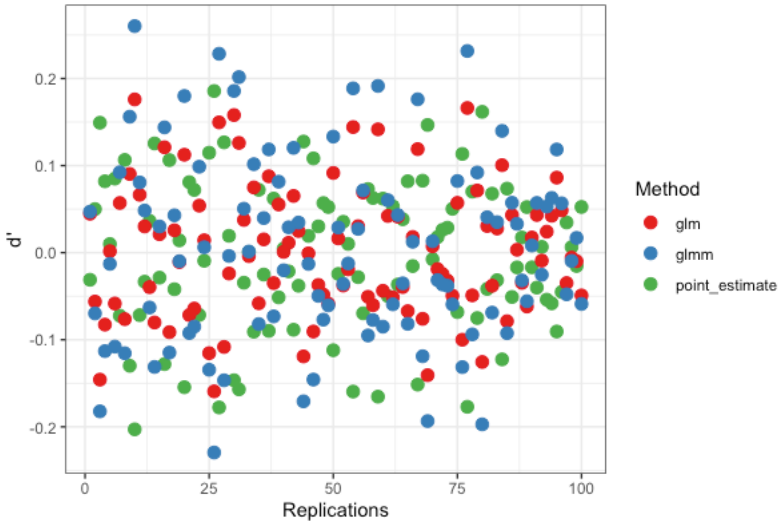

Figure 1. d' calculated by each method for each experiment.

To assess how many incorrect conclusions would be drawn for each method across replications, the number of d' results for the correct vs. incorrect decision that significantly differed from 0 were counted and averaged to get the false positive rate for the method[1]. Manually calculating point estimates of d' for correct and incorrect preferences led to an average d' of -0.003 and a false positive rate of 0.06. Estimating d' with a generalized linear model resulted in an average d' of 0.002 and a false positive rate of 0.46. Finally, estimating d' with a generalized linear mixed model resulted in a mean d' of 0.003 and a false positive rate of 0.04.

Condition	d'	False Positive Rate	Method
Correctness	0.0034	0.04	GLMM
Correctness	0.0016	0.46	GLM
Correctness	-0.0029	0.06	Point Estimate
Interaction	-0.0152	0.04	GLMM
Interaction	-0.0137	0.35	GLM
Knowledge	0.0174	0.02	GLMM
Knowledge	0.0131	0.59	GLM

Table 1. Mean d' estimates across and the associated false positive rate.

[1] Only the d' results for the correct vs. incorrect parameter are reported in text as this was calculated in each method. See Table 1 for d' results for the knowledge and interaction parameters for the generalized linear model and the generalized linear mixed model.

Discussion

While each method was able to estimate d' parameters that were close to the true effects in the simulated data, the absence of random effects of participants and items catastrophically inflated the false positive rate across experiments for the generalized linear model. The point estimate method, as it was calculated directly from the data, maintained the expected alpha level as did the generalized linear mixed model that estimated d' and accounted for the variability of participants and items. This simulation demonstrates how truly egregious it was to estimate d' without accounting for variability due to participants and items.

There could be several reasons why the generalized linear model performed so badly. First, as the d' estimates for the generalized linear model for each experiment were similar to the d' calculated by the other methods, the variance used to calculate the significance of the effect can be blamed for erroneously reaching a conclusion of a significant result. Second, several analysis choices specific to the given implementation could be to blame. Both generalized linear models were fit with default priors. For the generalized linear mixed model, this did not have a meaningful effect as it could have used the participant variable as a prior and scale itself appropriately (Kruschke & Lidell, 2017). However, as the non-hierarchical model did not include participants as a parameter, the set of default priors may have been particularly catastrophic for the results. Additionally, the number of iterations was limited to 2000 for each of the generalized linear models. Perhaps if the non-hierarchical model was allowed to iterate longer, it could have reached the correct conclusion more often. Other limitations of the current approach are related to the number of experimental replications. While the use of 100

replications here provides a glimpse into the power of each of the methods, to be truly confident in the results, a greater number (~10,000) of replications should be simulated[2].

It should also be considered that in this study, different methods under the framework of equal variance signal detection theory were implemented. However, this approach could be easily extended to unequal variance methods to test the validity of the conclusions when the signal and noise (or choice) distributions do not have the same variance. Additionally, the benefit of this approach is that the impact of different experimental parameters on the results can be easily tested. For example, one could test whether it is better to optimize the number of participants or items for a given study based on this approach. Further, in this study, the amount of participant and item variability was kept constant. However, it could be the case that the different methods would perform better or worse depending on the level of variability present in the data. This is could also be simulated in follow-up studies. In conclusion, this study suggests that while different estimation methods can be implemented to test questions in signal detection theory, it is critical to account for participant and item level variability to maintain an acceptable false positive rate across experiments.

[2] It took over nine hours for 100 replications to complete. Given the time constraints of the project deadline, I decided that 10,000 replications would be an endeavor for another time.

References

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of memory and language*, *68*(3), 255-278.

Bürkner, Paul-Christian. 2017a. "Brms: An R Package for Bayesian Multilevel Models Using Stan." *Journal of Statistical Software* 80 (1): 1–28. https://doi.org/10.18637/jss.v080.i01.

Clark, H. H. (1973). The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of verbal learning and verbal behavior*, *12*(4), 335-359.

Coronel, J. C., & Federmeier, K. D. (2016). The N400 reveals how personal semantics is processed: Insights into the nature and organization of self-knowledge. *Neuropsychologia*, *84*, 36-43.

DeBruine, L. M., & Barr, D. J. (2021). Understanding mixed-effects models through data simulation. *Advances in Methods and Practices in Psychological Science*, *4*(1), 2515245920965119.

DeCarlo, L. T. (1998). Signal detection theory and generalized linear models. *Psychological methods*, *3*(2), 186.

Estes, W. K. (1956). The problem of inference from curves based on group data. *Psychological bulletin*, *53*(2), 134.

Gescheider, G. A. (1997). Psychophysical measurement of thresholds: differential sensitivity. *Psychophysics: the fundamentals*, 1-15.

Glanzer, M., Adams, J. K., Iverson, G. J., & Kim, K. (1993). The regularities of recognition memory. *Psychological review*, *100*(3), 546.

Green, D. M., & Swets, J. A. (1966). Signal detection theory and psychophysics. New York:

Wiley.

Kruschke, J.K., Liddell, T.M. Bayesian data analysis for newcomers. *Psychon Bull Rev* **25,** 155–177 (2018). https://doi.org/10.3758/s13423-017-1272-1

McCarley, J. S., & Benjamin, A. S. (2013). Bayesian and signal detection models. In *The Oxford handbook of cognitive engineering*.

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.

R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Rouder, J. N., Lu, J., Speckman, P., Sun, D., & Jiang, Y. (2005). A hierarchical model for estimating response time distributions. *Psychonomic Bulletin & Review*, *12*(2), 195-223.

Rouder, J. N., Lu J., Sun, D., Speckman, P., Morey, R.D., & Naveh-Benjamin, M., 2007. "Signal Detection Models with Random Participant and Item Effects." *Psychometrika* 72 (4): 621–42. https://doi.org/10.1007/s11336-005-1350-6.

Rouder, J. N., Sun, D., Speckman, P. L., Lu, J., & Zhou, D. (2003). A hierarchical Bayesian statistical framework for response time distributions. *Psychometrika*, *68*(4), 589-606.

Stanislaw, H., & Todorov, N. (1999). Calculation of signal detection theory measures. *Behavior research methods, instruments, & computers*, *31*(1), 137-149.

Swets, J. A. (1988). Measuring the accuracy of diagnostic systems. *Science*, *240*(4857), 1285-1293.

Swets, J. A., Dawes, R. M., & Monahan, J. (2000). Psychological science can improve

diagnostic decisions. *Psychological Science in the Public Interest*, *1*(1), 1-26.

Vuorre (2017, Oct. 9). Sometimes I R: Bayesian Estimation of Signal Detection Models. Retrieved from: https://mvuorre.github.io/posts/2017-10-09-bayesian-estimation-of-signal-detection-theory-models/

Vuorre (2017, March 21). Sometimes I R: Bayes Factors with brms. Retrieved from https://mvuorre.github.io/posts/2017-03-21-bayes-factors-with-brms/

Wagenmakers, E. J., Lee, M., Lodewyckx, T., & Iverson, G. J. (2008). Bayesian versus frequentist inference. In *Bayesian evaluation of informative hypotheses* (pp. 181-207). Springer, New York, NY.

Wixted, J. T. (2020). The forgotten history of signal detection theory. *Journal of experimental psychology: learning, memory, and cognition*, *46*(2), 201.