Journal of Memory and Language (in press, May 2017)

A Bayesian Approach to the Mixed-Effects Analysis of Accuracy Data in Repeated-Measures Designs

Yin Song[a], Farouk S. Nathoo[a,*], Michael E.J. Masson[b,*]

[a] *Department of Mathematics and Statistics, University of Victoria*
[b] *Department of Psychology, University of Victoria*

[*] Corresponding authors. Email addresses: nathoo@uvic.ca (F.S. Nathoo), mmasson@uvic.ca (M.E.J. Masson)

Abstract

Many investigations of human language, memory, and other cognitive processes use response accuracy as the primary dependent measure. We propose a Bayesian approach for the mixed-effects analysis of accuracy studies using mixed binomial regression models. We present logistic and probit mixed models that allow for random subject and item effects, as well as interactions between experimental conditions and both items and subjects in either one- or two-factor repeated-measures designs. The effect of experimental conditions on accuracy is assessed through Bayesian model selection and we consider two such approaches to model selection: (a) the Bayes factor via the Bayesian Information Criterion approximation and (b) the Watanabe-Akaike Information Criterion. Simulation studies are used to assess the methodology and to demonstrate its advantages over the more standard approach that consists of aggregating the accuracy data across trials within each condition and over the contemporary use of logistic and probit mixed models with model selection based on the Akaike Information Criterion. Software and examples in R and JAGS for implementing the analysis are available at `https://v2south.github.io/BinBayes/`.

Keywords: Accuracy Studies, Bayesian Analysis, Behavioural Data, Model Selection, Repeated-Measures

Many types of behavioural data generated by experimental investigations of human language, memory, and other cognitive processes processes entail the measurement of response accuracy. For example, in studies of word identification, error rates in word-naming or lexical-decision tasks are analyzed to determine whether manipulated variables or item characteristics influence response accuracy (e.g., Chateau & Jared, 2003; Yap, Balota, Tse, & Besner, 2008). Similarly, in experiments on memory topics such as false memory and the avoidance of retroactive and proactive interference on recall, response errors or probability of accurate responding are the critical measures of performance (e.g., Arndt & Reder, 2003; Jacoby, Wahlheim, & Kelley, 2015).

The common treatment of accuracy or error-rate data has, and to a large extent continues, to consist of aggregating data across trials within each condition for each subject to generate the equivalent of a proportion correct or incorrect score, ranging from 0 to 1. These scores are then analyzed using repeated-measures analysis of variance (ANOVA) or, in the simplest cases, a t test. Although this standard approach, hereafter termed 'standard aggregating approach', has serious problems that have repeatedly been pointed out to researchers, it continues to be used. Here we illustrate a solution to these problems offered by Bayesian data analysis. We first (re)summarize the problems of the standard aggregating approach. Then we summarize one approach to this problem that has gained traction over the last decade (non-Bayesian Generalized Linear Mixed Models), followed by a brief review of some of the general pros and cons of Bayesian approaches. The rest of the paper then presents a Bayesian statistical modeling framework for repeated-measures accuracy data, simulation studies evaluating the proposed

methodology, and an application to actual data arising from a single-factor repeated-measures design.

The Standard Aggregating Approach

To assess the validity of our characterization of how researchers typically analyze accuracy, error, or other classification data, we examined articles published in recent issues of four of the leading journals in the field of cognitive psychology: the *Journal of Memory and Language (JML)*, the *Journal of Experimental Psychology: Learning, Memory, and Cognition (JEP)*, *Cognition*, and *Cognitive Psychology*. All articles appearing in issues with a publication date of January to August 2016 (up to the October 2016 issue for *JML* because later issues of that journal were available at the time the survey was conducted) were considered. Articles in which accuracy was analyzed using a transformed measure such as d', receiver operating characteristic curves, or parameters of computational models based on simulation of accuracy data were not included. A total of 180 articles across the four journals reported data expressed as proportions or the equivalent (e.g., accuracy, error, classification responses). Among these articles, 69 were on a topic related to language processing and the remaining 111 addressed other issues in memory and cognition. For each article, we determined whether the authors used standard methods of analyzing data that included aggregating performance across items or across subjects or whether generalized linear mixed models were used in which individual trials were the units of analysis. We included in the standard-analysis category any standard univariate method of analysis, such as analysis of variance, t tests, correlation, and regression in which data were aggregated over items or over subjects. The application

of analysis of variance using subjects and items as random effects in separate analyses and reporting F_1 and F_2 was also classified as using a method of aggregation. This approach, used widely since Clark's (1973) seminal paper on item variability, relies on an analysis that aggregates across defined subsets of trials (items for F_1 and subjects for F_2), rather than analyzing data at the level of individual trials. Our assessment indicated that for articles on language-related topics, 37 (54%) applied some form of the standard aggregating approach (of these, 15 used methods that reported effects aggregated over subjects and effects aggregated over items; i.e., F_1 and F_2). For articles on other topics of memory and cognition, 99 (89%) relied on the standard aggregating approach (two of these reported F_1 and F_2 analyses). Overall, then, 76% of recently published articles in these four leading cognitive psychology journals analyzed accuracy or other binomial data in the historically standard way, which involves aggregating performance across items for at least a subset of the analyses. The remaining articles used generalized linear mixed models to analyze the data[1], which does not aggregate across items and which we discuss in detail below.

The shortcomings of what continues to be a widely applied method of analyzing accuracy data, and binomial data in general (i.e., aggregating across items), have been known for some time (Cochran, 1940) and have been reiterated in recent accounts of alternative approaches (e.g., Dixon, 2008; Jaeger, 2008; Quen & van den Bergh, 2008). For instance, the proportions generated

[1]In four of the articles reporting linear mixed model regression analyses of accuracy, it was not clear whether logistic regression was used or whether raw accuracy was the dependent measure.

from binary observations (correct versus incorrect) need not be normally distributed, which violates one of the fundamental assumptions of ANOVA and t tests. Moreover, the variance of accuracy scores will depend on the mean, with larger variance when the mean is closer to .5 and variance vanishing to zero as the mean approaches 0 or 1. This dependency implies that if effects are present (i.e., means vary across conditions), the assumption of homogeneity of variance on which ANOVA depends is likely to be violated. By aggregating data across trials, error variance is likely to be reduced, leading to an elevation of type I error probability in null-hypothesis significance testing (Quen & van den Bergh, 2008). Finally, because proportion correct is bounded by 0 and 1, confidence intervals created from such data may well extend outside that range when the relevant mean approaches one of those limits, meaning that probability mass is being assigned to impossible values (Dixon, 2008; Jaeger, 2008).

A common strategy that is adopted to avoid these problems is the application of a data transformation such as the square-root arcsine transformation. Unfortunately, this approach makes the interpretation of the analysis more difficult as the hypothesis tests then correspond to the means of the transformed data and not to the actual accuracy data. Jaeger (2008) also shows that these transformations do not fix the problem when the mean proportions are close to 0 or 1.Furthermore, transforming the data after aggregating across items precludes the investigation of item effects.

Generalized Linear Mixed Models

A viable solution to these difficulties with the standard aggregating approach to analyzing accuracy data involves using generalized linear mixed-

models of logistic regression (Dixon, 2008; Jaeger, 2008; Quen & van den Bergh, 2008). In this setting a hierarchical model based on two levels is specified for the data, where, at the first level the response variables are assumed to be generated from a Bernoulli distribution. At the second level of the model the accuracy or error rates are converted to a logit scale (the logarithm of the odds of success or failure): $\text{logit}(p) = \ln(p/(1-p))$ and the variability in the log-odds across subjects, items, and conditions is based on a mixed effects model. We emphasize here that p is not computed from the data and does not correspond to the proportion of accurate responses aggregated over items for a given condition and subject; rather, p is an unknown parameter representing the probability of an accurate response for a given subject, item, and experimental condition. Rather than aggregating data over trials to obtain a single estimate of the proportion correct in a given condition for each subject, the individual binary accuracy trial scores are the unit of measurement. This level of granularity allows the assessment of possible random effects for both subjects and items. That is, effects of a manipulation may not be consistent from subject to subject or item to item and a mixed-effects analysis can characterize the extent of these differences. Variance in effects across items can thus be assessed, which addresses the concern raised by Clark (1973) about the "language-as-a-fixed-effect fallacy" (Jaeger, 2008; Quen & van den Bergh, 2008).

The proposed use of mixed-effects logistic regression for the analysis of accuracy data can be implemented either with or without significance tests. In the latter case, information criteria such as the Akaike Information Criterion (AIC) can be used for model selection. In the former case, these analyses

continue to rely on the basic principles of null-hypothesis significance testing (NHST) for making decisions about whether independent variables are producing effects on performance. A number of recent reports in the psychological literature have highlighted potential deficiencies associated with NHST (e.g., Kruschke, 2013; Rouder, Speckman, Sun, Morey, & Iverson, 2009; Rouder, Morey, Speckman, & Province, 2012; Wagenmakers, 2007). We will briefly mention only a few of those difficulties here.

First, the probability value associated with a significance test reflects the probability of obtaining an observed result, or one more extreme, on the assumption that the null hypothesis is true. An inference must then follow, establishing one's belief that the null hypothesis is false on those occasions where the obtained probability value is very low. Many researchers mistakenly interpret that probability as the likelihood that the null hypothesis is true, given the observed data (e.g., Haller & Krauss, 2002). That inference is not available through NHST, but it can, for example, be generated by a Bayesian analysis. Second, NHST is, by design, capable of providing evidence in favour of only the alternative hypothesis. When evidence does not allow rejection of the competing null hypothesis, no strong conclusion can be reached. Another potential advantage of the Bayesian approach is that it allows the strength of evidence in favor of either a null or an alternative hypothesis to be quantified. Although such reasoning can, and has been, accommodated under some non-Bayesian approaches, it follows naturally from the Bayesian perspective and Bayesian methods provide one principled approach to doing this. Finally, although not an inherent problem of non-Bayesian approaches but rather a potential pitfall of their misuse, when using NHST researchers

are susceptible to the problems caused by optional stopping during data collection. One may be tempted, for example, to collect additional data if a NHST applied to data currently in hand generates a p value that is just shy of significance. It has been clearly demonstrated that this approach to data collection substantially raises the probability of a type I error (e.g., Wagenmakers, 2007), whereas Bayesian analysis is not susceptible to this problem and will only yield increasingly accurate results as more data accumulate (Berger & Berry, 1988; Wagenmakers, 2007), assuming the methods are used adequately. Again, we emphasize that this is not an inherent problem of non-Bayesian approaches but one that can and often does arise when these procedures are misused.

Bayesian Approaches

As a solution to this problem with NHST and to advance the use of mixed-effects analyses of binomial data, we propose the use of a Bayesian version of mixed-effects models of logistic and probit regression for the repeated-measures case. The nature of these modified versions of regression analysis is discussed in detail below. Although Bayesian analysis of generalized linear mixed-models has been developed extensively in the statistical literature over the past decade, our interest is specifically on the application to behavioural data arising from accuracy studies where a method combining the use of Bayesian analysis with generalized linear mixed-models for binomial data has not been considered previously. We investigate two options for selecting between null and alternative hypotheses (models) using Bayesian analysis: the Bayes factor computed using a Bayesian Information Criterion (BIC) approximation, and the Watanabe-Akaike Information Criterion (WAIC). We

provide R software to implement these approaches in addition to computing posterior distributions.

Although the Bayesian approach offers an exciting avenue for the analysis of memory and language data, there is a large body of work that debates the pros and cons of a Bayesian analysis. Efron (1986) discussed the potential problems with the Bayesian approach and provided examples where the frequentist approach provides easier solutions. The use of Bayesian approaches can lead to an increase in conceptual complexity of some aspects of the data analysis. Users must take the time to acquire the necessary background in order to use Bayesian methods appropriately. In addition, an important issue that has received a lot of attention in the literature is the choice of the prior distribution, which can have an influence on the results. Informative prior distributions can be chosen to reflect prior knowledge on certain parameters of a model, though formulating such priors can be very difficult and is often impractical. In our work, we favour the use of weakly informative priors that have high or infinite prior variance so that the prior plays a limited role in the inference. For Bayesian logistic regression the issue of priors and the development of weakly informative priors is discussed extensively in Gelman et al. (2008).

The use of logistic mixed models is an alternative to methods that involve aggregation over items or subjects. As demonstrated in the literature, this alternative is a more effective methodology for the analysis of repeated-measures accuracy studies. In the memory and language literature this has been considered primarily from a classical, non-Bayesian perspective (Dixon 2008; Jaeger 2008; Quen & van den Bergh 2008). In parallel, there is cur-

rently a shift towards the use of Bayesian methods for the analysis of cognitive studies (Wagenmakers 2007; Rouder et al. 2009; Rouder et al. 2012). To date, much of this shift has focused on the analysis of continuous response variables. The goal of this project is to provide tools for memory and language researchers to combine the advantages of both logistic/probit mixed models and Bayesian methods for the analysis of repeated-measures studies of accuracy. In doing so, we allow for the evaluation of posterior distributions for the effects of interest, and we also offer two possible approaches for Bayesian model selection (a) the Bayes factor based on the BIC approximation; and (b) the more recently proposed WAIC. These two approaches are motivated based on different utilities, the former assesses model performance using the marginal likelihood while the latter assesses model performance by estimating a measure of out-of-sample prediction error.

Given that we offer two different approaches for Bayesian model selection it is naturally of interest to consider comparisons between them. We therefore make these comparisons by evaluating the operating characteristics of both approaches using simulation studies. In order to place these comparisons within the larger field of methods that can be applied to repeated-measures accuracy data we also evaluate two other approaches. One is a Bayesian analysis based on item aggregation with model selection determined by the Bayes factor. The other is logistic mixed modeling within the classical (non-Bayesian) setting with model selection based on the standard Akaike Information Criterion (AIC). The latter is arguably the current non-Bayesian state-of-the-art. We use these evaluations to inform a discussion on the pros and cons of the Bayesian approach and its combination with logistic/probit

mixed modelling.

In the first simulation study, comparisons are made within the context of testing for a fixed effect of the experimental conditions. In that study, we find that the proposed Bayesian approach and the current non-Bayesian state-of-the-art perform equally well in the sense that the two exhibit identical power curves after calibrating for type I error. In the second simulation study, comparisons are made within the context of testing for a random effect of the experimental conditions. More specifically, we consider the scenario where the effect of the experimental manipulation varies across items. In that case, we find that the fully Bayesian approach based on the WAIC exhibits uniformly higher power than logistic mixed modeling within the classical (non-Bayesian) setting with model selection based on the standard AIC.

The primary contributions of our work are: (a) Facilitating the combination of binomial mixed-effects modeling and Bayesian inference for repeated-measures analysis of accuracy studies, (b) simulation studies evaluating this approach relative to the standard aggregating approach and classical logistic mixed models, and (c) making available easy-to-use R software with examples facilitating such analysis.

Overview

The remainder of the article proceeds as follows. In the next section we present the statistical modeling framework and discuss Bayesian approaches for evaluating the effect of experimental conditions on accuracy. We then present two simulation studies evaluating the proposed methodology in comparison to a benchmark consisting of a Bayesian analysis (using the BayesFactor R package based on Rouder et al., 2012) of data aggre-

gated in the standard way. Comparisons to logistic mixed modelling within the non-Bayesian setting with model selection based on the AIC are also made. This is followed by the description of an application to actual data, where we demonstrate how our methodology can be applied to a single-factor repeated-measures design. Two-factor repeated-measures designs can also be accommodated and we provide an example of this analysis on a webpage `https://v2south.github.io/BinBayes/` that is associated with this paper. The final section concludes with a discussion and practical recommendations for the analysis of accuracy studies.

Method

Statistical Models for Repeated-Measures Accuracy Studies using Single-Factor Designs

We first consider a repeated-measures design involving K subjects, I experimental conditions corresponding to a single factor, and J items, and where generally the number of experimental conditions will be much smaller than the number of items. For each subject the data consist of a binary response measuring accuracy on each of the J items. The response on each item occurs in a particular experimental condition and these conditions are randomly assigned across items. The structure of the data are illustrated in Table 1 for the case where there are $I = 3$ experimental conditions with labels b, g, or r, and these labels are indicated as subscripts on each of the binary measurements, the latter taking values either 0 or 1. It should be noted that our framework can accommodate cases where items are seen by the same subject in multiple conditions. In that case the data could be represented

by adding additional columns for each item in Table 1. Furthermore, the studies considered here may use counterbalancing, so that for some subjects a particular item will be assigned to a particular condition whereas for other subjects that same item will be assigned to a different condition. We also do not require that a response is obtained on every item for each subject, so it is possible for different subsets of items to be presented to different subjects.

Table 1: Example data structure: $I = 3$ conditions, K subjects, J items where conditions are indicated as subscripts b, g, or r of each binary data value.

	Item 1	Item 2	...	...	Item J
Subject 1	1_b	1_r	1_g	0_b	1_b
Subject 2	0_g	0_b	0_r	1_g	0_g
$\vdots$	0_r	0_r	0_g	1_r	0_g
$\vdots$	0_r	0_g	0_r	1_b	0_r
Subject K	0_b	0_b	0_b	1_r	0_b

In the model specifications that follow, the notation $X \sim F$ where X is a random variable and F is a probability distribution, such as the standard normal distribution with mean 0 and standard deviation 1, $N(0, 1)$, denotes that the random variable X is drawn from the distribution F.

Similarly, if Y is also a random variable, the notation $X|Y \sim F_Y$ denotes that the conditional distribution of X given the value of the random variable Y is F_Y. For example, $X|Y = y \sim N(y, 1)$ denotes that, given that $Y = y$, the conditional distribution of X, given the knowledge that $Y = y$, is a normal distribution with mean y and standard deviation 1. In addition, if

$X_1, \ldots, X_n$ is a collection of random variables, the notation $X_i \overset{iid}{\sim} F$, $i = 1, \ldots, n$ is used to denote that these random variables are independent and identically distributed from the distribution F (i.e., independently selected from the same distribution, F).

We let Y_{ijk} denote the binary response obtained from subject k when item j is assigned to condition i. The approach, hereafter termed 'standard aggregating approach', often applied in the analysis of such data begins by averaging the response variables across the items for each condition to obtain accuracy scores corresponding to each subject and condition. Referring to Table 1, this averaging results in three response variables for each row of the table, one for each of the three conditions. A repeated-measures ANOVA is then applied to the aggregated data. In contrast, trial-based analyses like those pursued here (and in Dixon, 2008) avoid averaging over items and model each binary score using a Bernoulli distribution. We let p_{ijk} denote the probability of an accurate response ($Y_{ijk} = 1$) when item j is assigned to condition i and subject k. The model assumes

$$Y_{ijk}|p_{ijk} \overset{ind}{\sim} \text{Bernoulli}(p_{ijk}),$$

and we emphasize that our modeling approach is specified at the level of binary observations (i.e., Y_{ijk} is either 0 or 1) and the accuracy probability p is a parameter of the corresponding Bernoulli distribution with $p = Pr(Y = 1)$. Each of these binary observations is specific to a particular subject, item, and level of the experimental condition. The analysis we propose evaluates a number of models each corresponding to different assumptions on how this probability varies across items, subjects, and conditions. We note that the experimental design is such that we expect lack of independence of the data

within subject (the rows of Table 1) and also within item (the columns of Table 1). As a result all of the models that we consider include random effects for both subjects and items (Clark, 1973) to account for possible dependence.

These models rely on first transforming the accuracy probability (p_{ijk}) using a link function $g(p) : [0, 1] \rightarrow \mathbb{R}$. That is, the link function, g, takes an value of p from 0 to 1 and maps it onto an element of the set of real numbers. We consider two common link functions $g(p) = \log \frac{p}{1-p}$ which corresponds to a logistic model, and $g(p) = \Phi^{-1}(p)$ which corresponds to a probit model, where $\Phi(\cdot)$ denotes the cumulative distribution function of the standard normal distribution. In the latter model it is useful to think of the probability values, p, as being converted to a corresponding Z-score. The two link functions are depicted in the left panel of Figure 1.

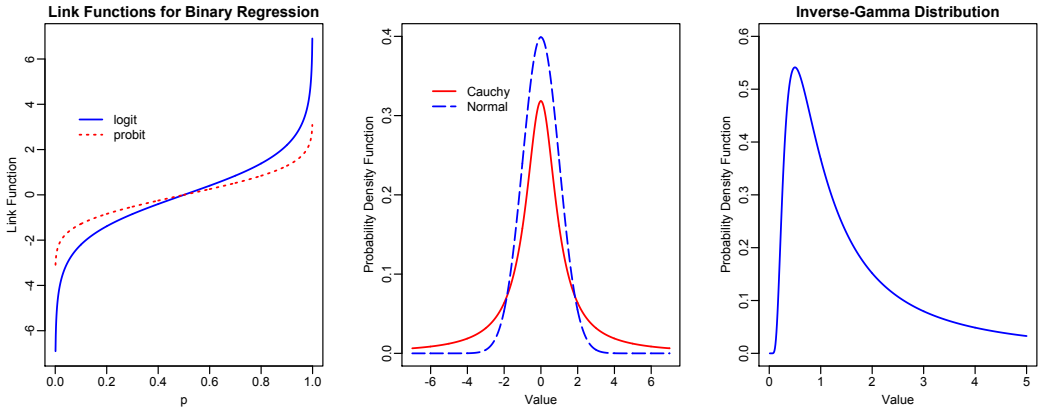

Figure 1: Left Panel - the logistic and probit link functions; Centre Panel - the probability density function of a Cauchy distribution and a normal distribution; Right Panel - the probability density function of an inverse-Gamma distribution.

The hierarchical Bayesian models for accuracy are presented below. The different models represent different possible effects. Following the recom-

mendations of Gelman (2008), Cauchy prior distributions are assigned to the fixed effects in all of the models. The centre panel of Figure 1 provides an illustration of the Cauchy prior distribution in relation to the normal distribution. As is typical in Bayesian generalized linear mixed models, the random effects are assigned normal distributions and the variance components corresponding to the random effects are assigned inverse-gamma distributions. The latter distribution is illustrated in the right panel of Figure 1 and is a convenient choice as it means that the algorithm used to fit the model to the data has an analytical form that is easy to work with.

1. *(LM_0 - Logit/PM_0 - Probit) Baseline model with random subject and item effects with no effect of the experimental condition:*

$$g(p_{ijk}) = \beta_0 + a_j^{(R)} + b_k^{(R)}$$

$$a_j^{(R)} \overset{iid}{\sim} N(0, \sigma_a^2), \quad b_k^{(R)} \overset{iid}{\sim} N(0, \sigma_b^2), \quad \beta_0 \sim \text{Cauchy}(0, \sigma_\beta = 10)$$

$$\sigma_a^2 \sim \text{Inverse-Gamma}(\kappa_{\sigma_a}, \tau_{\sigma_a}), \quad \sigma_b^2 \sim \text{Inverse-Gamma}(\kappa_{\sigma_b}, \tau_{\sigma_b})$$

   Here β_0 is the model intercept, $a_j^{(R)}$ is the item random effect with variance σ_a^2, and $b_k^{(R)}$ is the subject random effect with variance σ_b^2. The hyper-parameters $\kappa_{\sigma_a}, \tau_{\sigma_a}, \kappa_{\sigma_b}, \tau_{\sigma_b}$ are fixed to values that make the inverse-gamma prior distributions weakly informative, with infinite variance. We reiterate that this model assumes that there is no effect of the experimental condition on the probability of an accurate response and this is the only model where this assumption is made. This model corresponds to the null hypothesis.

2. *(LM_F - Logit/PM_F - Probit) Fixed effect for the experimental condition:*

$$g(p_{ijk}) = \beta_0 + \alpha_i + a_j^{(R)} + b_k^{(R)}$$

with $\alpha_1 = 0$, $\alpha_i \overset{iid}{\sim} \text{Cauchy}(0, \sigma_\alpha = 2.5)$, $i = 2, \ldots, I$, and with all other priors identical to model 1. The constraint $\alpha_1 = 0$ is imposed for model identification and as a result one arbitrarily selected experimental condition is considered a baseline condition and the remaining fixed effects α_i represent the effect of condition i relative to that baseline. The value of the scale parameter $\sigma_\alpha = 2.5$ in the Cauchy prior distribution for the fixed effects is different from the corresponding value in the prior for the intercept $\sigma_\beta = 10$ based on the work of Gelman et al. (2008) who recommend these choices for Bayesian logistic regression as a weakly informative prior. The Cauchy prior is centered around zero so that there is no preference for either a positive or negative effect. Using cross-validation, Gelman et al. (2008) show that this class of priors outperform Gaussian and Laplace priors and we use it here for both logistic and probit regression. This model extends the first model by assuming that the probability of an accurate response depends on the experimental condition through a fixed effect α_i.

3. *(LM_{R_S} - Logit/PM_{R_S} - Probit) The effect of experimental condition varies across subjects:*

$$g(p_{ijk}) = \beta_0 + \alpha_i + a_j^{(R)} + b_k^{(R)} + (\alpha b)_{ik}^{(R)}$$

where, in this case, the effect of the experimental condition is represented by both the fixed effects α_i and the random effects $(\alpha b)_{ik}^{(R)}$ which

represent an interaction between subject and condition, implying that the effect of condition varies across subjects. As before, constraints are imposed so that one arbitrarily selected condition is taken as a baseline condition, $(\alpha b)_{1k}^{(R)} = 0$, and the remaining random effects are assumed to be normally distributed $(\alpha b)_{ik}^{(R)} \overset{iid}{\sim} N(0, \sigma_{\alpha b}^2)$, $i = 2, \ldots, I; k = 1, \ldots, K$. An inverse-gamma prior distribution is assumed for the corresponding variance component, $\sigma_{\alpha b}^2 \sim$ Inverse-Gamma$(\kappa_{\sigma_{\alpha b}}, \tau_{\sigma_{\alpha b}})$ with hyper-parameters $\kappa_{\sigma_{\alpha b}}, \tau_{\sigma_{\alpha b}}$ fixed to values that make the inverse-gamma prior distribution weakly informative, with infinite variance.

4. *(LM_{R_i} - Logit/PM_{R_i} - Probit) The effect of experimental condition varies across items:*

$$g(p_{ijk}) = \beta_0 + \alpha_i + a_j^{(R)} + b_k^{(R)} + (\alpha a)_{ij}^{(R)}$$

with the constraint, $(\alpha a)_{1j}^{(R)} = 0$, imposed so that one arbitrarily selected condition is taken as a baseline condition and the remaining random effects are assumed to be normally distributed,
$(\alpha a)_{ij}^{(R)} \overset{iid}{\sim} N(0, \sigma_{\alpha a}^2)$, $i = 2, \ldots, I; j = 1, \ldots, J$. An inverse-gamma prior distribution is assumed for the corresponding variance component, $\sigma_{\alpha a}^2 \sim$ Inverse-Gamma$(\kappa_{\sigma_{\alpha a}}, \tau_{\sigma_{\alpha a}})$ with $\kappa_{\sigma_{\alpha a}}, \tau_{\sigma_{\alpha a}}$ fixed to values that make the inverse-gamma prior distribution weakly informative, with infinite variance. All other prior distributions are identical to model 2. In this case the effect of the experimental condition is represented by both the fixed effects α_i and the random effects $(\alpha a)_{ij}^{(R)}$ which represent an interaction between item and condition (items potentially vary in the extent to which they exhibit effects of the experimental conditions).

5. *($LM_{R_{S,i}}$ - Logit/$PM_{R_{S,i}}$ - Probit) The effect of experimental condition varies across items and subjects:*

$$g(p_{ijk}) = \beta_0 + \alpha_i + a_j^{(R)} + b_k^{(R)} + (\alpha a)_{ij}^{(R)} + (\alpha b)_{ik}^{(R)}$$

with distributions for random effects (condition-by-item and condition-by-subject effects) and hyper-priors set as in models 3 and 4. This is the most general of the models considered for a single-factor repeated measures design.

Each of the five models presented above represents different assumptions on the effect of the experimental conditions while explicitly modeling the binary response through the Bernoulli distribution and accounting for between-subject and between-item variability with random effects. Considering the two possible choices for the link function, logit or probit, there are ten possible models and these models are summarized in Table 2.

Analysis of Two-Factor Designs

In the case of a design with two experimental factors we let Y_{hijk} denote the binary response obtained from subject k when item j is assigned to condition i of the first factor and condition h of the second factor, $k = 1, \ldots, K$, $j = 1, \ldots, J$, $i = 1, \ldots, I$, $h = 1, \ldots, H$. We assume

$$Y_{hijk}|p_{hijk} \overset{ind}{\sim} \text{Bernoulli}(p_{hijk})$$

where p_{hijk} is the corresponding probability of an accurate response. Different models correspond to different assumptions on how this probability varies across items, subjects, and the levels of the two experimental factors. The

Table 2: The full set of Bernoulli mixed models for single-factor designs representing different assumptions about effects on the accuracy probability.

Model	**Link**	**Condition Effect**
LM_0	Logistic	Null
LM_F	Logistic	Fixed
LM_{R_s}	Logistic	Varies across subjects
LM_{R_i}	Logistic	Varies across items
$LM_{R_{s,i}}$	Logistic	Varies across subjects and items
PM_0	Probit	Null
PM_F	Probit	Fixed
PM_{R_s}	Probit	Varies across subjects
PM_{R_i}	Probit	Varies across items
$PM_{R_{s,i}}$	Probit	Varies across subjects and items

most general model we consider for a two-factor design takes the form

$$g(p_{hijk}) = \beta_0 + \gamma_h + \alpha_i + (\gamma\alpha)_{hi} + a_j^{(R)} + b_k^{(R)} + (\gamma a)_{hj}^{(R)} + (\alpha a)_{ij}^{(R)} + (\gamma b)_{hk}^{(R)} + (\alpha b)_{ik}^{(R)}$$

where α_i, γ_h, and $(\gamma\alpha)_{hi}$ are fixed effects corresponding to the first factor, the second factor, and their interaction respectively. As before $a_j^{(R)}$ and $b_k^{(R)}$ are random effects for items and subjects while the random effects $(\gamma a)_{hj}^{(R)}$, $(\alpha a)_{ij}^{(R)}$, $(\gamma b)_{hk}^{(R)}$, $(\alpha b)_{ik}^{(R)}$ allow the effects of the two experimental factors to vary across items and subjects. Just as with the singe-factor case, we assume Cauchy priors for the fixed effects and adopt the identification constraints $\alpha_1 = \gamma_1 = (\gamma\alpha)_{1i} = (\gamma\alpha)_{h1} = 0$ so that the first level of both factors are taken to be baseline conditions. The random effects are assigned normal priors as before and the corresponding variance components are assigned weakly informative inverse-gamma hyper-priors.

For simplicity, we do not consider models where the interaction between the two experimental factors itself interacts with either items or subjects. Although allowing such terms increases the flexibility of the model, the associated parameters can be only weakly estimatable in practice and will therefore exhibit a low degree of Bayesian learning (prior-to-posterior movement) in particular with binary data. We note that there is some controversy in the literature with regards to the inclusion of high-order random effects in mixed models. For example, Barr, Levy, Scheepers, & Tily (2013) suggest that linear mixed models generalize best when the maximal random effects structure supported by the design is employed; however, Bates, Kliegl, Vasishth, & Baayen (2015) indicate problems with this suggestion, including convergence problems of numerical algorithms for fitting mixed models and the potential lack of model interpretability. Our choice to exclude models

with random effects that represent three-way interactions is inline with the discussion in Bates et al.

Many different models can be obtained by removing certain terms in the model equation above, and either the logit or probit link can be employed. For a single-factor design, the set of ten possible models is summarized in Table 2, where, for example, LM_{R_i} (PM_{R_i}) denotes the logistic (probit) model where the effect of the experimental condition is represented as a random effect that varies across items, and we assume the presence of a fixed effect in any model that contains a corresponding random effect for the experimental condition. In the case of two factors, the set of models obtainable by removing appropriate terms from the general model equation above is considerably larger. For the logistic (probit) link we refer to specific models using the notation $LM_xN_yI_z$ ($PM_xN_yI_z$), where $x \in \{0, F, R_s, R_i, R_{s,i}\}$ denotes the model structure for the first factor as defined for single-factor designs in Table 2, $y \in \{0, F, R_s, R_i, R_{s,i}\}$ similarly denotes the model structure for the second factor, and $z \in \{0, 1\}$ is used to specify the presence or absence of an interaction between the two factors, with $z = 1$ ($z = 0$) indicating presence (absence). For example, $LM_{R_{s,i}}N_{R_{s,i}}I_1$ denotes the most general model specified in the equation above with a logit link, $PM_0N_0I_0$ denotes the null probit model where all terms corresponding to the effects of the two factors have been removed, and $LM_{R_s}N_{R_{s,i}}I_1$ denotes a logistic model where the effect of the first factor varies across subjects, the effect of the second factor varies across subjects and items, and the interaction between the factors is included.

Model Fitting and Software

The posterior distribution of the model parameters (fixed effects, random effects, and variance components) associated with each model can be computed using standard Markov chain Monte Carlo (MCMC) sampling algorithms. These procedures can be implemented in the R (Ihaka & Gentleman, 1996) and JAGS (Plummer, 2003), programming languages in conjunction with the R package 'rjags' (Plummer, 2013) which provides an interface between the two. We have developed an R function 'BinBayes.R' that allows these models and algorithms to be used in a relatively straightforward manner requiring only very basic knowledge of the R language. The software along with a detailed illustration of its use for single-factor and two-factor repeated-measures designs, sample data, and examples are available for download at: `https://v2south.github.io/BinBayes/`.

Bayesian Model Comparison

For a given dataset, we are able to summarize the posterior distribution for any of the models for single-factor and two-factor designs. An arguably more important task is the comparison of models, as each model represents different assumptions regarding the effect of the experimental condition on the probability of response accuracy. For example, and in reference to Table 2, a comparison of models LM_0 and LM_F corresponds to testing for a fixed effect of the experimental condition, whereas a comparison of models LM_0 and LM_{R_i} corresponds to testing for an effect of the experimental condition that allows for this effect to vary across items. As the link function is a modelling choice, logit and probit models can also be compared (e.g. LM_F and PM_F) to determine which is more appropriate for the data at hand.

The traditional approach for model comparison in the Bayesian framework is based on the Bayes factor. Given two models denoted by $M0$ and $M1$ the Bayes factor comparing $M0$ to $M1$ is defined as

$$BF_{01} = \frac{Pr(\mathbf{y}|M0)}{Pr(\mathbf{y}|M1)}$$

where $\mathbf{y}$ denotes the data, and $Pr(\mathbf{y}|M)$ denotes the probability of the data under M. A value of $BF_{01} > 1$ can be viewed as evidence in favour of model $M0$ over $M1$ in the sense that the probability of the data is higher under $M0$. Kass and Raftery (1995) provide a comprehensive review of the Bayes factor including information about its interpretation where it is suggested that a value of $BF_{01} \geq 3$ corresponds to positive evidence in favour of model $M0$ over $M1$, whereas, decisive evidence corresponds to $BF_{01} > 150$.

In general, the Bayes factor can be difficult to compute and a great deal of research in the area of statistical computing has been dedicated to this problem (see e.g. Chib & Jeliazkov, 2001; Meng & Wong, 1996; Meng & Schilling, 2003; Chen, 2005; Raftery, Newton, Satagopan & Krivitsky, 2006). A number of Monte Carlo algorithms can be applied for the computation of the Bayes factor; however, for the Bernoulli mixed models under consideration in this article, we have found that stable estimation of the Bayes factor is extremely time consuming (e.g. several hours to days on a fast laptop for datasets of standard to large size). As a more practical alternative that is easy to compute in just a few seconds, we use the Bayesian information criterion (BIC) defined for a given model M by

$$BIC(M) = -2 \log \hat{L} + p \log n,$$

where $\hat{L}$ is the maximized likelihood function for model M, p is the number

of parameters in the model, and n is the sample size. In the case of generalized linear mixed models for repeated-measures designs, both the number of parameters p and the sample size n are not straightforward to define (Jones, 2011; Spiegelhalter 2002). We will assume that p excludes the random effects but includes the corresponding variance components and the number of fixed effects. Indeed, this definition for p is the default used in the computation of the BIC in the R package lme4 (Bates, Machler, Bolker, & Walker, 2014). For the sample size, we assume that $n = K$, the number of subjects. The issue of effective sample size for BIC is considered in detail by Berger, Bayarri, & Pericchi, (2014); see also Nathoo & Masson, (2016).

Given the value of BIC for two competing models the Bayes factor is approximated by $BF_{01} \approx \exp\{(BIC(M1) - BIC(M0))/2\}$ where this expression assumes $Pr(M0) = Pr(M1)$ a priori and the accuracy of approximation will increase with the sample size. The approximation is based on a unit information prior for the model parameters (see e.g. Kass & Raftery, 1995; Masson, 2011; Nathoo & Masson, 2016; Wagenmakers, 2007). Alternatively, given a set of competing models such as those listed in Table 2, the BIC for each model can be computed in order to rank the models, with lower values corresponding to preferred models. Typically we require a difference in the BIC scores of two models to be at least $|\Delta BIC| = 2$ in order to claim that there is positive evidence in favour of one model over the other, which corresponds to a Bayes factor of $BF \approx 2.72$.

Rather than comparing models based on Bayes factors, an alternative second approach that can be used to compare models is based on an evaluation of how well each model can predict new data or heldout data. By heldout

data, we mean a subset of the data that is not used in the process of fitting the model but is subsequently used to evaluate the predictive ability of the model. In this context, cross-validation is a common approach for estimating the out-of-sample prediction error which can then be used to compare models (Gelman, Hwang, & Vehtari, 2014). As cross-validation requires splitting the data into multiple partitions and then repeatedly fitting the model to subsets of the data, which is computationally demanding, alternative measures of predictive accuracy that in some sense approximate cross-validation have been proposed for model selection. One such approximation that has been applied extensively for model selection is the AIC (Akaike, 1998) which is computed using the maximum likelihood estimator. The computation of the AIC, like the BIC, requires a value for the number of model parameters p which is not clearly defined in the case of hierarchical models as described above. An alternative, fully Bayesian approximation to cross-validation that avoids this problem, is the WAIC, which has been proposed by Watanabe (2010) and takes the form

$$WAIC = -2\sum_{k=1}^{K} \log E_\theta[p(\mathbf{y}_k|\theta)|\mathbf{y}_1, \ldots, \mathbf{y}_K]$$

$$+2\sum_{k=1}^{K} Var_\theta[\log p(\mathbf{y}_k|\theta)|\mathbf{y}_1, \ldots, \mathbf{y}_K]$$

where $\mathbf{y}_k$ is the data collected from subject k, θ denotes the set of all unknown parameters, $p(\mathbf{y}_k|\theta)$ denotes the probability mass function of $\mathbf{y}_k$ conditional on θ, and the expectation $E_\theta[\cdot]$ and variance $Var_\theta[\cdot]$ are taken with respect to the posterior distribution of the model parameters. The first part of the formula for WAIC provides a measure of the fit of the model against the

data and the second part is meant to capture the inherent complexity of the model, where the reasoning is that more complex models are a priori less likely (i.e., it is a form of Occam's razor).

In practice, these are computed using an MCMC algorithm that is used to fit the model. The WAIC is thus easily computed as a by-product of fitting the model. As discussed in Gelman et al. (2014), the WAIC has the desirable property of averaging over the posterior distribution of the model parameters, that is, considering many possible values of the model parameters weighted by their posterior density, rather than conditioning on just a single value of the parameter, the maximum likelihood estimator as is done with AIC. This is desirable because it captures the estimated uncertainty a rational researcher should have into the estimates of the model given the assumptions the researcher was willing to make about the prior and model structure.

More importantly, the WAIC works well with so called singular models, that is, models with hierarchical structures where the number of parameters increases with sample size. It is thus particularly well suited for the random effect models we are considering in the present article, and unlike the penalties used by the BIC and AIC, the penalty used in the WAIC has been rigorously justified for model comparison in this setting (Watanabe, 2010).

Although a generally applicable calibration for differences in the WAIC scores of two models is currently lacking, a reasonable heuristic is to apply the criteria often used for the standard AIC (Burnham & Anderson, 1998; Dixon, 2008). One such criterion is to require that $|\Delta WAIC| = 2$ in order to claim that there is positive evidence in favour of one model over the other, though other criteria may also be used. The operating characteristics of this

decision rule are evaluated in comparison with the BIC and AIC through our simulation studies.

We emphasize that BIC and WAIC, although both Bayesian in their formulation, are constructed with different utilities in mind. The BIC is based on the notion of posterior model probabilities and the Bayes factor, whereas the WAIC is an estimate of the expected out-of-sample-prediction error. Given the differing utilities it is certainly possible that the two criteria will disagree, and we view the approaches as complimentary. Detailed comparisons are made in the next section. In addition to computing the posterior distribution for a given model, our R function BinBayes.R will also compute the BIC and WAIC for any of the models in Table 2 or for models associated with the two-factor designs discussed above.

Simulation Studies

In order to evaluate the methodology described in the previous section we conducted two simulation studies. The studies examined the type I error and the power of specific decision rules based on the Bernoulli mixed models and the BIC or the WAIC for evaluating the effect of experimental conditions. These are compared with the type I error and the power of the standard aggregating approach after aggregating over items. Although we make our comparisons based on frequentist criteria, we note that it is not uncommon to evaluate Bayesian methods using such criteria (see e.g. Carlin & Louis, 2008; Gelman, Carlin, Stern, Dunson, Vehtari, & Rubin, 2013). In the non-Bayesian context, simulation studies somewhat similar to those presented here are conducted in Dixon (2008). For the standard aggregating approach,

we let $\tilde{Y}_{ik}$ denote the proportion of accurate responses at the i^{th} condition in the experiment for subject k. This is obtained by averaging that subject's binary response variables Y_{ijk} over the items at condition i. The statistical model considered in this case is

$$\tilde{Y}_{ik} = \mu_i + b_k + \epsilon_{ik}, \;\; \epsilon_{ik} \overset{iid}{\sim} N(0, \sigma_\epsilon^2), \;\; i = 1, \ldots, I; \;\; k = 1, \ldots, K \qquad (1)$$

where $b_k \overset{iid}{\sim} N(0, \sigma_b^2)$ are subject-specific random effects that are assumed independent of the model errors ϵ_{ik}, and μ_i is the fixed effect of the experimental condition. To evaluate the effect of the experimental condition, this model is compared with the simpler model that assumes no effect of the experimental condition and thus has $\mu_i = \mu, i = 1, \ldots, I$. For the standard aggregating approach we will again make this evaluation within the Bayesian paradigm and compute the Bayes factor using the BayesFactor package (Rouder et al., 2012) in the R programming language. The BayesFactor package implements Monte Carlo techniques for computing the Bayes factor for a class of Gaussian error models, of which (1) is a special case. Finally, we compare the three Bayesian approaches for model selection (the first Bayesian approach is the standard aggregating approach under a Bayesian framework) with the standard AIC criterion applied to logistic mixed models.

In each study, we simulated binary response accuracy data of the type depicted in Table 1, with $K = 72$ subjects, $J = 120$ items, and $I = 4$ experimental conditions manipulated as a repeated-measures factor. In the first study we simulate data where the effect of the experimental condition does not depend on items or subjects, and in the second study this effect is not held constant, but instead varies across the items. In each setting, 1,500 simulation replicates were used to estimate each point of the power curve

for comparing the null and alternative models as the effect size varies. To create power curves when the BIC approximation to the BF was used, our decision rule was to reject the null model whenever $BF > \exp(1)$ (which occurs when $\Delta BIC = BIC_{null} - BIC_{alt} > 2$) and the same rule was used for both the WAIC and AIC. For the standard item aggregated approach using the BF, the null model was rejected whenever the Bayes factor favouring the alternative model was greater than $BF > \exp(1)$. Additionally, we also created power curves for each of the four methods where the *decision rule* was chosen for each so that the type I error rate was fixed at 0.05. Fixing the type I error rate for all four methods to have the same value has the advantage of making the power curves more directly comparable.

Simulation Study I

We generated the simulated data using the logistic mixed model LM_F which contains a fixed effect for the experimental condition. The simulated data could also be generated from a probit mixed model; however, we simulated from the logistic model as it is more commonly used in practice. In this case, the effect of the experimental condition does not depend on subjects or items. The fixed effect is represented by α_i, and we set $\alpha_1 = \alpha_2 = \alpha_3 = 0$ and $\alpha_4 = C$, where C ranged over a series of values from $C = 0$ to $C = 1.5$. This particular pattern of fixed effects where the effect of only a single condition varies was chosen so that the results are easier to summarize and visualize (i.e., in a two-dimensional plot). The intercept was set at $\beta_0 = 3.22$ corresponding to a baseline accuracy rate of 96% (representative of performance in speeded word identification experiments, for example). We also considered and will summarize later results from simulations in which performance

is not near ceiling or floor. The variance components were set according to $\sigma_b = 1.045$ for the standard deviation of the subject random effects, and we considered three values for the standard deviation of the item random effects $\sigma_a = 1.5, 3$, or 5. These choices for the simulation parameter values are guided by the model estimates obtained from the single-factor real-data example analyzed in the next section.

For each value of the fixed effect C and the item standard deviation σ_a, we simulated $1,500$ datasets, and for each dataset we fit model LM_F which contains the fixed effect of the experimental condition, as well as model LM_0 which has no effect for the experimental condition (the null model). The BIC approximation to the BF, AIC, and WAIC for both models were computed, and in addition, the standard model Eq. (1) was applied after aggregating over items, and the Bayes factor comparing the models with and without a fixed effect for condition was computed. In all four cases, the decision rules described in the previous section were applied to create power curves for the different model selection criteria.

The results of the simulation study are presented in Figure 2 which shows the result of assessing the significance of the experimental manipulation by comparing models with and without fixed effects for each approach. In the case where the decision rules were set so that the type I error rate was fixed at 0.05 for all three Bayesian approaches as well as the AIC (second column of Figure 2), a clear pattern emerges. When the between-item variability is at its lowest level, with $\sigma_a = 1.5$, all four approaches have power curves that are virtually identical. In this case there appears to be no advantage to applying the Bernoulli mixed model over the standard aggregating approach. How-

ever, as the between-item variability increases, the Bernoulli mixed models with BIC approximation to BF, AIC, or WAIC outperform the standard aggregating approach with BF and have uniformly higher power. Interestingly, the BIC approximation to the BF, AIC, and WAIC have identical power curves when they are calibrated to have the same type I error rate.

In practice, outside of a simulation study, it may not be possible to calibrate the decision rule so as to ensure a specific value for the type I error rate. The first column of Figure 2 depicts the power curves when the decision rules $\Delta AIC > 2$, $\Delta WAIC > 2$, and $BF > \exp(1)$ are applied. In this case, it must be understood that a comparison of the power curves is not an 'apples-to-apples' comparison as the type I error rates are not the same. For all values of σ_a the type I error rate for the Bernoulli mixed model with $BF > \exp(1)$ ($\Delta BIC > 2$) is 0. This produces a conservative rule that has uniformly lower power than the Bernoulli mixed model with $\Delta AIC > 2$ and $\Delta WAIC > 2$. As a tradeoff, the latter have a higher type I error rate and we note that the power curves for WAIC and AIC are virtually identical in this case. The standard aggregating approach with rejection of the null when $BF > \exp(1)$ also has a type I error rate of 0 in all cases. It should be noted that although both the BIC approximation to the BF and the standard item aggregated approach with BF have a type I error rate of 0, the corresponding adjusted power curves differ due to the fact that the two approaches require different decision rules in order to fix the type I error rates at 0.05.

The power curve associated with the standard aggregating approach is always below that of the WAIC and AIC, and is above the power curve of the BIC approximation to the BF when σ_a is set to its lowest value of

1.5. As the value of σ_a increases, the power of the Bernoulli mixed model with BIC approximation to the BF begins to improve relative to the standard aggregating approach. In the case where $\sigma_a = 5$ both the BIC approximation to the BF and standard aggregating approach have a type I error rate of 0, but the power of the Bernoulli mixed model with BIC apprximation to the BF is generally higher than that of the standard aggregating approach, particularly for higher values of the effect size C.

Simulation Study II

We next consider the situation where the effect of the experimental condition varies across the items and where the objective is again to evaluate the data for the existence of a condition effect. We generated data sets from the model LM_{R_i} which contains random effects for both subjects and items, a fixed effect for the experimental condition, and a random effect representing the interaction between experimental condition and items. The parameter values for the simulation were again guided by model estimates obtained from the single-factor dataset considered in the next section. For simulating data, we set the intercept to be $\beta_0 = 3.21$ corresponding to a baseline accuracy rate of 96%, the variance components for the subject and item random effects were set based on $\sigma_b = 1.04$ and $\sigma_a = 0.44$ respectively, with these values being based on estimates obtained from fitting the model to the real data example discussed in the next section. The fixed effect for the experimental condition was again set based on $\alpha_1 = \alpha_2 = \alpha_3 = 0$ and $\alpha_4 = C$, and we considered three possible values, $C = 0, 0.25, 0.5$. The effect of the experimental condition varied across items through the random effect $(\alpha a)_{ij}^{(R)} \overset{iid}{\sim} N(0, \sigma_{\alpha a}^2)$ and the magnitude of this variability depended on the parameter $\sigma_{\alpha a}$, which

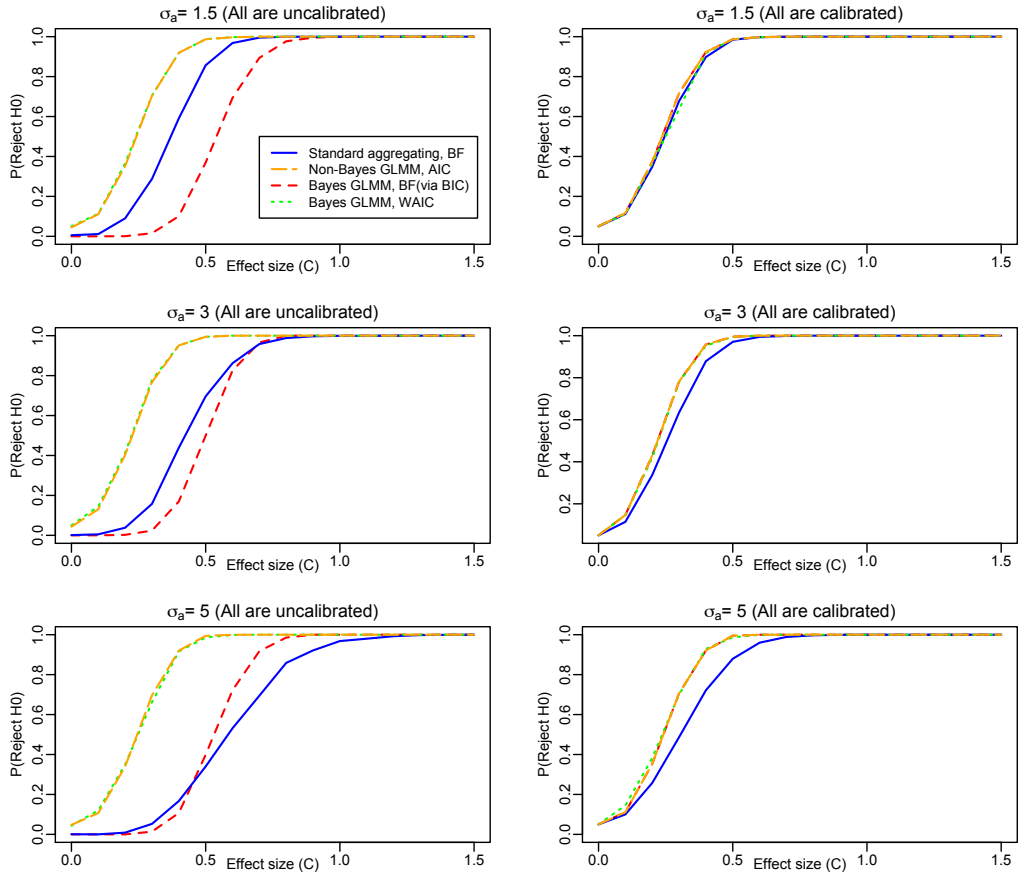

Figure 2: Results from simulation study I. The left column corresponds to the decision rules $\Delta BIC > 2$ (for Bayes GLMM, BF via BIC), $\Delta AIC > 2$ (for non-Bayes GLMM, AIC), $\Delta WAIC > 2$ (for Bayes GLMM, WAIC), and $BF >$ exp(1) (for standard aggregating BF), whereas in the right column the decision rules were chosen to ensure that all four methods had a type I error rate of 0.05. The rows correspond to different values of the between item variability $\sigma_a = 1.5, 3, 5$. Values of C represent the strength of the effect of the experimental conditions.

we varied over a series of values ranging from 0 to 2. The values of C and $\sigma_{\alpha a}$ are varied factorially so that overall there can be a random effect even without a fixed effect ($C = 0$) and there can be a fixed effect without a random effect ($\sigma_{\alpha a} = 0$). When both $C = 0$ and $\sigma_{\alpha a} = 0$ there is no effect present.

For each value of C and $\sigma_{\alpha a}$ we again simulated $1,500$ datasets, and for each dataset we fit the model LM_{R_i} which contains an effect for the experimental condition that varies across the items, as well as the model LM_0 which has no effect for the experimental condition (the null model). The comparison of models LM_{R_i} and LM_0 then corresponds to an overall test for an effect of the experimental condition, where this effect can be either random and varying across items, fixed, or both. After aggregating over items, the standard aggregating approach is applied as in the previous section. We note that the standard aggregating approach is not sufficiently flexible to allow the condition effect to depend on items, and so as before, the effect of experimental condition was evaluated through the model Eq. (1) and the Bayes factor comparing the models with and without a fixed effect for condition was computed. In all cases the decision rules used in the first simulation study were applied again.

The results are depicted in Figure 3. In this case the effect of the experimental condition is represented by both the fixed effect C and the standard deviation of the random condition-by-item interaction $\sigma_{\alpha a}$. We clarify that in this figure the power curves vary on the x-axis by $\sigma_{\alpha a}$; whereas, they vary on the x-axis by C in Figure 2. As we are testing for an overall effect of the experimental conditions, a type I error can occur only when $C = 0$ and

$\sigma_{\alpha a} = 0$, that is, when both the fixed and random components of the experimental effect are absent, which is possible only in the left-most data point in the top panels of Figure 3, where both $C = 0$ and $\sigma_{\alpha a} = 0$. The second column of Figure 3 compares the power curves when the decision rules were chosen so as to fix the probability of a type I error at 0.05. In these cases, when $C = 0$ (so that the fixed effect of condition is absent but the random condition-by-item effect is not) the Bernoulli mixed model with any of the BIC approximation to the BF, AIC, or WAIC outperforms the standard aggregating approach rather significantly with respect to power. In this case the latter approach does not contain parameters in the model for a random effect, and can only detect this through the fixed effect for condition, though it is interesting to note that the model is able to do this once $\sigma_{\alpha a}$ is sufficiently large. This phenomenon can be thought of as a type of false alarm, because the approach detects the existence of a fixed effect when in fact the real effect that is present is a random effect. It also appears that WAIC has higher power than both the AIC and BIC approximation to the BF in this case. *Thus we see here an advantage in terms of power for the fully Bayesian approach compared with the non-Bayesian approach when the same generalized linear mixed models are applied and it is only the model selection criterion that varies.*

When C is increased so that $C = 0.25$ (so that there is now both a fixed effect of condition as well as a random condition-by-item effect) the power of all four approaches increases; however, the Bernoulli mixed model still generally outperforms the standard aggregating approach with a fairly large difference in power at most values of $\sigma_{\alpha a}$. Considering the same model

comparisons based on generalized linear mixed models (LM_{R_i} versus LM_0) *we again see that WAIC has higher power than both AIC and BIC approximation to the BF in this case.* When $C = 0.5$ the fixed effect is now sufficiently large that all four methods have very high power to detect an effect of experimental condition and the power curves are generally flat as $\sigma_{\alpha a}$ varies.

Turning to the first column of Figure 3 where the methods are not calibrated to have the same type I error rate, we see that the Bernoulli mixed model with decision rule $BF > \exp(1)$ ($\Delta BIC > 2$) results in an extremely conservative procedure, where the type I error is 0 but the power is generally lower (particularly with $\sigma_{\alpha a}$ small) than the other procedures, with the exception of AIC which is also very conservative and has a power curve matching that of the BIC approximation to the BF when $C = 0.25$ and $C = 0.5$. Interestingly, when $C = 0$ (so that the fixed of the experimental condition is absent but the random condition-by-item effect is not) the power curve for the AIC is higher than that of both the BIC approximation to the BF and the standard aggregating approach but is uniformly lower than that of the WAIC. The power of the standard aggregating approach improves as the values of C increases, but its power is uniformly dominated by the Bernoulli mixed model with decision rule $\Delta WAIC > 2$ in all cases. We emphasize again that this is not an 'apples-to-apples' comparison as the two approaches are not calibrated to have the same type I error rate. The type I error rate of both approaches is indicated in the left-most data point in the first row and first column of Figure 3 when $\sigma_{\alpha a} = 0$, and we note that the Bernoulli mixed model with decision rule $\Delta WAIC > 2$ has a slightly higher type I error than the other approaches. Nevertheless, its power to detect an exper-

imental effect is much higher for most values of $\sigma_{\alpha a}$, and this is particularly evident within a neighbourhood of $\sigma_{\alpha a} = 0.5$. *Thus it is interesting to note that fully Bayesian WAIC has higher power than AIC when the same logistic mixed models are being compared.*

In both simulation studies I and II we have assumed that the true value of the intercept in the logistic model is $\beta_0 = 3.22$. This is a fairly large value corresponding to a baseline accuracy rate of approximately 96%. We have also conducted additional simulation studies where the baseline accuracy rate was not so extreme, based on setting the true value to $\beta_0 = 0$, corresponding to a baseline accuracy rate of approximately 50%. The results for this more moderate accuracy rate are depicted for study I in Figure 4 which corresponds to the third row of Figure 2, and for study II in Figure 5 which corresponds to the first row of Figure 3.

Figure 4 and Figure 5 indicate that the comparison of the power curves at a baseline accuracy rate of 50% yields results that are quite similar to those already presented where the baseline accuracy rate was 96%. The primary difference is that the relative performance of the standard aggregating approach appears to drop in the case where the baseline accuracy rate is 50%. A baseline rate of 50% was also the value considered in Dixon (2008) where the intercept was taken to be zero in the simulations that evaluated generalized linear mixed models.

Example Application: Single-Factor Design

We now present an example application illustrating the use of Bernoulli mixed models for the analysis of response accuracy for a single-factor design.

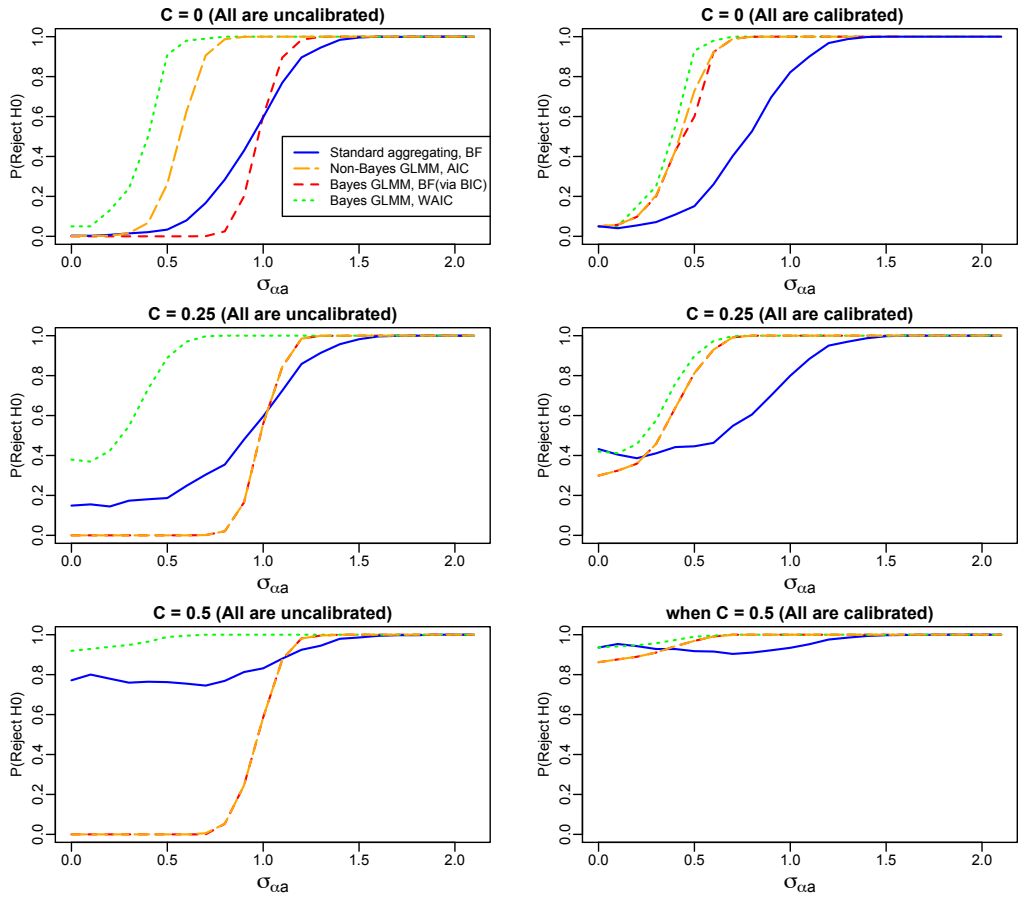

Figure 3: Results from simulation study II. The left column corresponds to the decision rules $\Delta BIC > 2$ (for Bayes GLMM, BF via BIC), $\Delta AIC > 2$ (for non-Bayes GLMM, AIC), $\Delta WAIC > 2$ (for Bayes GLMM, WAIC), and $BF > \exp(1)$ (for standard aggregating BF), whereas in the right column the decision rules were chosen to ensure that all four methods had a type I error rate of 0.05. Note that a type I error can occur only when $C = 0$ and $\sigma_{\alpha a} = 0$ (since otherwise the null is false) so the calibration of the type I error rates is based on the $C = 0$ and $\sigma_{\alpha a} = 0$ case for all three panels in the right column.

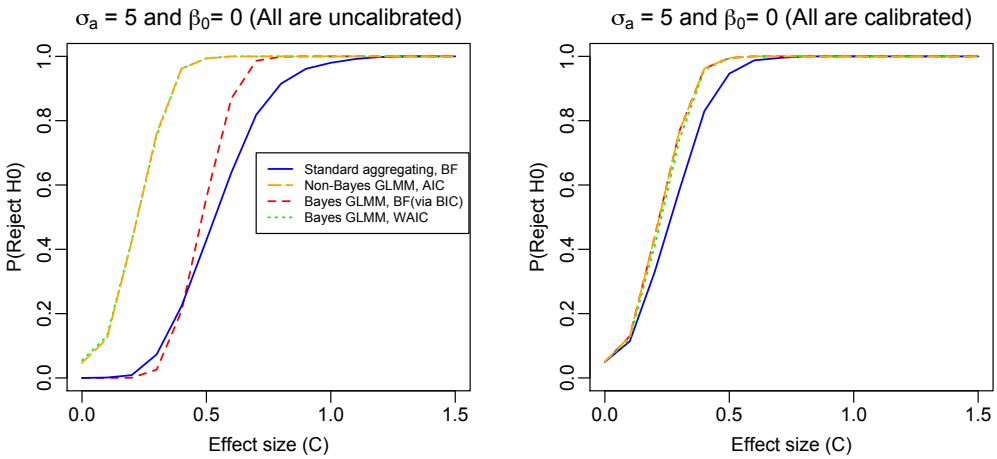

Figure 4: Results from simulation study I with $\beta_0 = 0$ corresponding to a baseline accuracy of 50%. The left panel corresponds to the decision rules $\Delta BIC > 2$ (for Bayes GLMM, BF via BIC), $\Delta AIC > 2$ (for non-Bayes GLMM, AIC), $\Delta WAIC > 2$ (for Bayes GLMM, WAIC), and $BF >$ exp(1) (for standard aggregating BF), whereas in the right panel the decision rules were chosen to ensure that all four methods had a type I error rate of 0.05. These settings correspond to the third row of Figure 2 where the baseline accuracy rate is 96%.

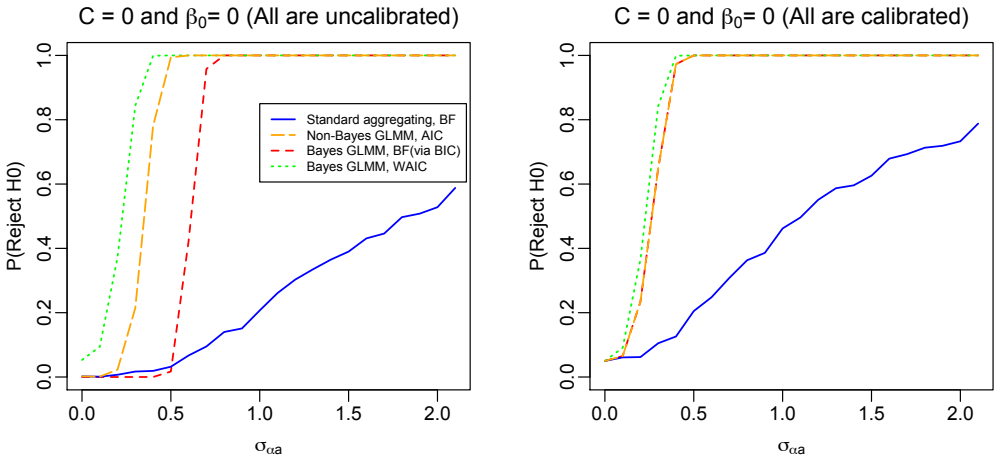

Figure 5: Results from simulation study II with $\beta_0 = 0$. The left figure corresponds to the decision rules $\Delta BIC > 2$ (for Bayes GLMM, BF via BIC), $\Delta AIC > 2$ (for non-Bayes GLMM, AIC), $\Delta WAIC > 2$ (for Bayes GLMM, WAIC), and $BF > \exp(1)$ (for standard aggregating BF), whereas in the right figure the decision rules were chosen to ensure that all four methods had a type I error rate of 0.05. These settings correspond to the first row of Figure 3 where the baseline accuracy rate is 96%.

The analysis presented here can be reproduced using the software, data, and examples provided at: `https://v2south.github.io/BinBayes/`. The data considered here were taken from a study that investigated the influence of a semantic context on the identification of printed words shown either under clear (high contrast) or degraded (low contrast) conditions. The semantic context consisted of a prime word presented in advance of the target item. On critical trials, the target item was a word and on other trials the target was a nonword. The task was a lexical-decision procedure where the subject was instructed to classify the target on each trial as a word or a nonword, and the response was either accurate or inaccurate. Our interest was confined to trials with word targets. The prime word was either semantically related or unrelated to the target word (e.g., granite-STONE vs. attack-FLOWER), and the target word was presented either in clear (high contrast) or degraded (low contrast) form. Combining these two factors produced four conditions: related-clear (RC), unrelated-clear (UC), related-degraded (RD), unrelated-degraded (UD). The two factors are treated as a single factor with four levels for this example. For the current analysis, the accuracy of the response was the dependent measure.

The study comprised $K = 72$ subjects, $I = 4$ conditions, and $J = 120$ items, and the total number of binary observations was $KJ = 8,640$. The overall rate of response accuracy was 95.4%. We took the UD (unrelated-degraded) level of the experimental condition as the baseline condition and fit each of the ten Bernoulli mixed models listed in Table 2, and the resulting WAIC and BIC scores were obtained for each model. The model comparisons are presented in Table 3.

According to the WAIC, the optimal model is the logistic model with random subject and item effects, and where the effect of the experimental condition depends on items. According to the BIC the optimal model is the logistic model with random subject and item effects, and where the effect of condition is fixed and does not depend on items. Taken together both criteria point to the existence of an effect for the experimental condition; however, the fixed condition effect has the highest (approximate) posterior model probability (BIC), whereas the model with random condition effects depending on item is expected to make the best out-of-sample predictions (WAIC). Both model selection criteria taken together seem to provide evidence in support of the logistic link as opposed to the probit link. This is primarily the case with BIC as the WAIC scores are more neutral towards the link function but do show some support in favor of the logistic link.

The BIC scores can be used to compute an approximate Bayes factor for comparing models. For example, comparing LM_0 (null condition) versus LM_F (fixed condition) we obtain a Bayes factor of

$$BF \approx \exp\{(BIC(LM_0) - BIC(LM_F))/2\} = 4.48$$

which indicates substantial evidence in favour of the model with a fixed condition effect when compared to the model with no condition effect.

Because it was chosen as the optimal model by the WAIC, we summarize the posterior distribution of model LM_{R_i} in more detail. In this case, the effect of experimental condition varies across items. The condition UD (unrelated-degraded) is taken as the baseline condition so that the item-dependent condition effects associated with the remaining three conditions are then interpreted relative to UD.

Table 3: The BIC and WAIC values for each of the ten binomial mixed models presented in Table 2 after application to the study data. Note: the lowest (i.e., best) scores are in bold.

Model	**Link**	**Condition Effect**	**WAIC**	**BIC**
LM_0	Logit	Null	2827	2928
LM_F	Logit	Condition: Fixed	2803	**2925**
LM_{R_s}	Logit	Condition*Subject: Random	2804	3010
LM_{R_i}	Logit	Condition*Item: Random	**2800**	2997
$LM_{R_{s,i}}$	Logit	Condition*Subject + Condition*Item: Random	2802	3078
PM_0	Probit	Null	2827	2935
PM_F	Probit	Condition: Fixed	2803	2934
PM_{R_s}	Probit	Condition*Subject: Random	2806	3015
PM_{R_i}	Probit	Condition*Item: Random	2803	3007
$PM_{R_{s,i}}$	Probit	Condition*Subject + Condition*Item: Random	2806	3088

The posterior distributions for the item-dependent condition effects represent the information obtained about these effects from the data, the model, and Bayes rule. These are depicted as box-plots in Figures 6, 7, and 8, for the conditions UC, RD, and RC, respectively. These plots relate to the best linear unbiased predictors (BLUPs) plots for standard generalized linear mixed models. Overall, we see that all three conditions increase the probability of response accuracy relative to the baseline UD. This increase is roughly the same for RD and UC; however, it appears that RC leads to higher accuracy rates compared to all other conditions. Examining all three conditions, we see that the variability in the condition effects across the items is not substantial; however, it does appear that some items have relatively lower or higher effects, particularly in the RD condition. Examining Figure 8, there is also one item in the RC condition that appears to have a substantially lower effect relative to all other items. Thus, it seems that the variability in the effect size across items picked up by the WAIC is driven primarily by just a few items. Our analysis enables the user to identify such items through the posterior distribution.

Bayesian interval estimates can be obtained by selecting those values of highest posterior probability density including the mode. This is sometimes called the 95% highest posterior density interval when the posterior probability associated with the interval is 0.95. With respect to the variance components for the random effects, the between-item effect standard deviation σ_a has a 95% highest posterior density (HPD) interval of $(0.863, 1.261)$, the between-subject effect standard deviation σ_b has a 95% HPD interval of $(0.280, 0.625)$, and the standard deviation for the condition-by-item random

effects $\sigma_{\alpha a}$ has a 95% HPD interval of $(0.109, 0.587)$.

Conclusions and Recommendations

We have introduced a collection of Bernoulli mixed models that can be used for the analysis of accuracy studies in the Bayesian framework. The set of models represents a number of different assumptions for the effect of the experimental condition on accuracy. These assumptions range from a null model to a model where the experimental effect varies across both items and subjects. The models we consider are based on random effects that are assumed normally distributed. Although one may consider generalized linear mixed models where the random effects are not normally distributed (see e.g. Nathoo & Ghosh, 2013) it is generally the case that inference in generalized linear mixed models is fairly robust to misspecification of the random effects distribution (e.g. McCulloch & Neuhaus, 2011).

For the models we have considered here, we have generally assumed the presence of a fixed effect in any model that contains a corresponding random effect. Florian Jaeger has suggested the alternative of assessing the significance of effects by comparing to a model without the fixed effect but, critically, with the subject- or item-based random slopes for the fixed effects. This comparison makes sure that the assessment of significance (e.g., through model comparison) does not confound the effect of the fixed effect predictor

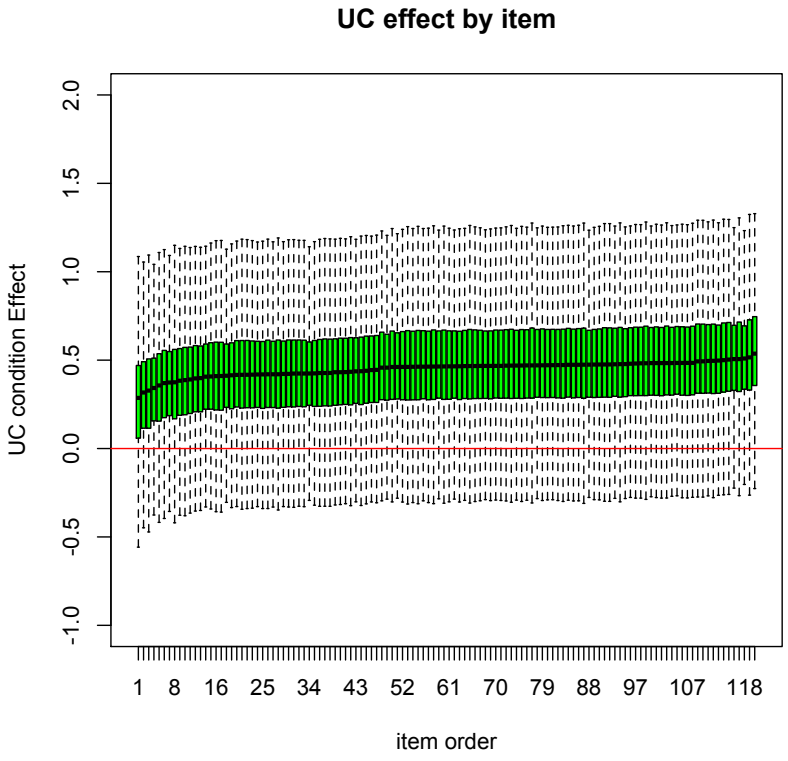

Figure 6: The posterior distributions for the effects of the unrelated-clear (UC) condition on the probability of response accuracy. In this case there is a separate effect for each of the 120 items used in the experiment and the figure depicts the posterior distribution for condition UC across the items as a boxplot summarizing Markov chain Monte Carlo samples drawn from the posterior distribution. Items are ordered according to their estimated effect sizes. The effects depicted are on the logit scale. The black line represents the posterior median for each item, the green region represents the set of values between the first and third quartile of the posterior distribution, and the dotted bars extend out to the extreme values.

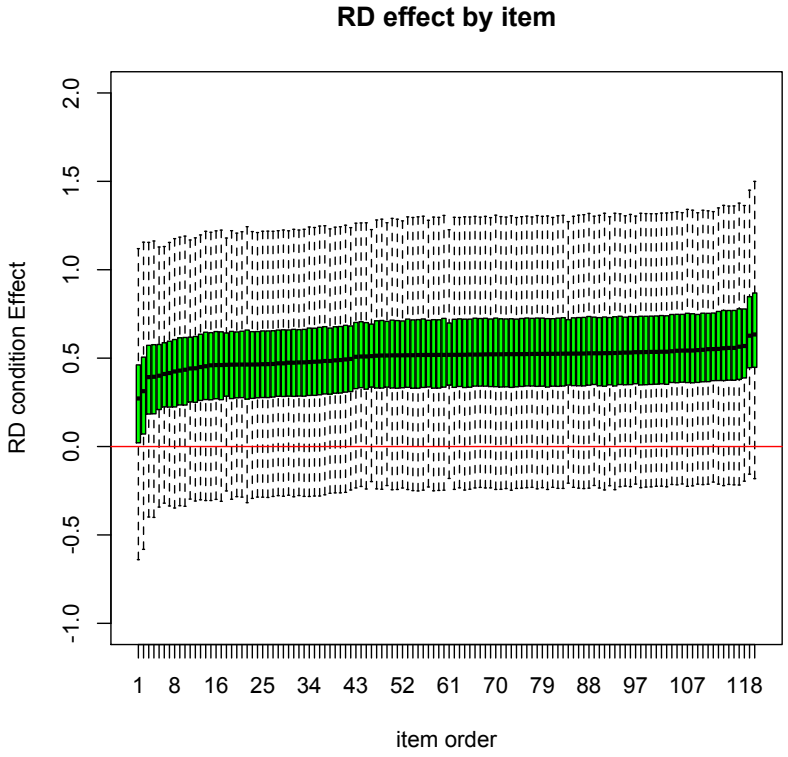

Figure 7: The posterior distributions for the effects of the related-degraded (RD) condition on the probability of response accuracy. In this case there is a separate effect for each of the 120 items used in the experiment and the figure depicts the posterior distribution for condition RD across the items as a boxplot summarizing Markov chain Monte Carlo samples drawn from the posterior distribution. Items are ordered according to their estimated effect sizes. The effects are depicted on the logit scale. The black line represents the posterior median for each item, the green region represents the set of values between the first and third quartile of the posterior distribution, and the dotted bars extend out to the extreme values.

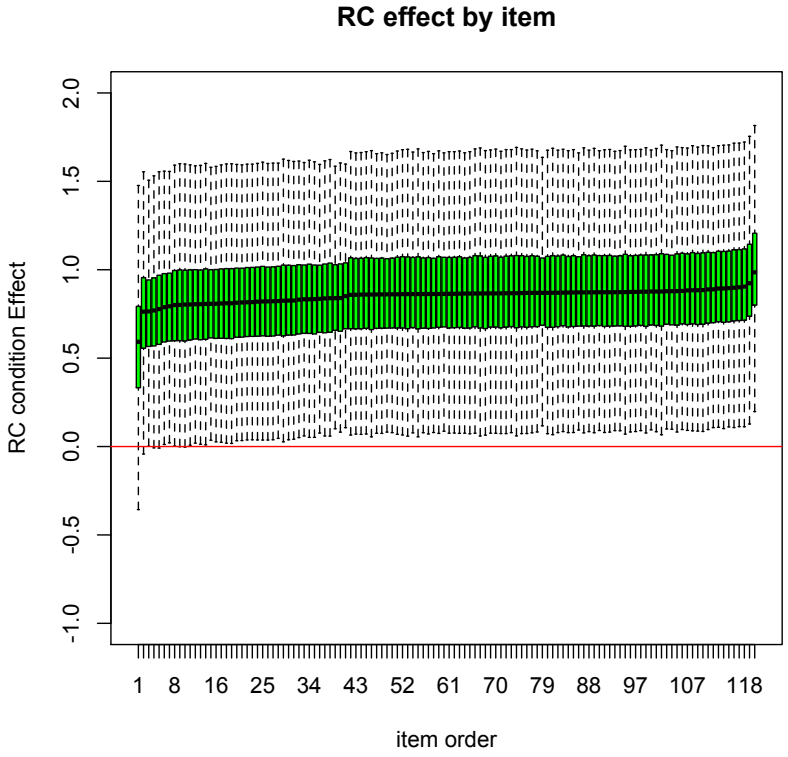

Figure 8: The posterior distributions for the effects of the related-clear (RC) condition on the probability of response accuracy. In this case there is a separate effect for each of the 120 items used in the experiment and the figure depicts the posterior distribution for condition RC across the items as a boxplot summarizing Markov chain Monte Carlo samples drawn from the posterior distribution. Items are ordered according to their estimated effect sizes. The effects are depicted on the logit scale. The black line represents the posterior median for each item, the green region represents the set of values between the first and third quartile of the posterior distribution, and the dotted bars extend out to the extreme values.

with the variance captured by the random slopes for that predictor. The extent to which such confounding can cause problems for the model comparisons considered here is unclear but presents a potentially interesting avenue for investigation.

To compare possible models, we have investigated the AIC and both the BIC as a large-sample approximation to the Bayes factor and the WAIC as a large-sample approximation to Bayesian cross-validation. These have been compared to the standard, item-aggregated approach using simulation studies. An alternative approach for model selection that we did not consider in our simulation studies is that of nested comparison of models via a chi-square over differences in model deviances. This approach, although arguably a current standard for model comparison across generalized linear mixed models (in the field), is less flexible than the approaches we considered as it requires the models being compared to be nested and is not appropriate for testing random effects because the asymptotic chi-square distribution under the null is not valid when the null hypothesis lies on the boundary of the parameter space (as it does when testing random effects where the null corresponds to setting a variance component to zero, see e.g., Lin, 1997).

Overall, we recommend the use of the WAIC as it appears to have power that is higher or at least as high as the BIC approximation to the BF, AIC, and the standard aggregating approach with BF when applied with industry-standard decision rules. The BIC approximation to the BF can be used alongside WAIC and the results treated as complimentary when Bayes factors are of interest. While not typical, there could be specific cases where the choice of random effects structure and link function could be driven by the-

oretical considerations. In these cases, these considerations could be used to narrow the class of possible models and then combined with Bayesian model selection.

For the simulation settings considered here, we found that the performance of the Bernoulli mixed model approach improved relative to the standard repeated-measures approach as both the between-item variability σ_a (study I) and the item-condition variability $\sigma_{\alpha a}$ (study II) increased. In the latter case, the observed difference in performance when comparing our approach to the standard item-aggregated approach was rather substantial for a fairly large range of values for $\sigma_{\alpha a}$. In general, we see that as the variability across items increases, the application of the proposed approach becomes increasingly more valuable and likely to detect effects, relative to the standard aggregating approach to evaluating the effects of independent variables on accuracy.

With respect to comparisons between AIC and the fully Bayesian WAIC, when the same logistic mixed models were applied the two approaches had identical power curves when testing for fixed effects of the experimental condition in study I. Interestingly, the WAIC had higher power than the AIC when testing for random effects of the experimental conditions in study II. As with any simulation study, the conclusions drawn are specific to the particular simulation settings adopted. The conclusions we have drawn regarding the power of the Bayesian approach relative to its competitors can be guaranteed to hold only under the conditions assumed for the simulation studies. Nevertheless, these initial findings are very instructive.

We note that the WAIC is based on the priors included in our model

specifications, whereas the BIC approximation to the BF is based on the unit information prior. The differences in performance seen when comparing these approaches is due in part to the differences in priors. It was clear that the BIC when used with the decision rule based on $\Delta BIC = 2$ results in a conservative approach. Aside from comparisons based on standard decision rules, we also made comparisons after controlling for type I error, in which case the performance of the BIC approximation to the BF improved with respect to power. Practically speaking, outside of a simulation study, it is currently not possible to control the type I error of a decision rule when using information criteria for model comparison; however, we are currently investigating an approach for doing so based on the parametric bootstrap as an avenue for future work.

We again note that there is a large body of work that debates the pros and cons of a Bayesian analysis. The use of Bayesian approaches requires researchers to digest new ideas that can be conceptually difficult. Users must take the time to acquire the necessary background in order to use Bayesian methods appropriately and we refer the interested reader to the introductory textbook of Gelman et al. (2014a). We also note that our approach, which requires MCMC sampling, is more computationally intensive than either the standard aggregating approach or the generalized linear mixed model fit by maximum likelihood with AIC used for model comparisons. We have demonstrated that WAIC has power that is higher or at least as high as AIC under certain settings, and we believe that this justifies the additional computation required. In addition, the advantages of a Bayesian analysis in the ability to construct posterior distributions for parameters of interest, and

the availability of a user-friendly software implementation for both single-factor and two-factor designs make our approach an exciting alternative to standard item aggregation approaches and contemporary mixed logit/probit procedures for the analysis of repeated-measures accuracy studies.

Software Implementation

The software implementing the methodology presented in this paper, along with sample datasets and two examples illustrating its use is available at: `https://v2south.github.io/BinBayes/`.

References

Akaike, H., 1998. Information theory and an extension of the maximum likelihood principle. In: Selected Papers of Hirotugu Akaike. Springer, pp. 199–213.

Arndt, J., Reder, L. M., 2003. The effect of distinctive visual information on false recognition. Journal of Memory and Language 48 (1), 1–15.

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.

Bates, D., Kliegl, R., Vasishth, S., Baayen, H., 2015. Parsimonious mixed models. arXiv preprint arXiv:1506.04967.

Bates, D., Mächler, M., Bolker, B., Walker, S., 2014. Fitting linear mixed-effects models using lme4. arXiv preprint arXiv:1406.5823.

Berger, J., Bayarri, M., Pericchi, L., 2014. The effective sample size. Econometric Reviews 33 (1-4), 197–217.

Berger, J. O., Berry, D. A., 1988. The relevance of stopping rules in statistical inference. Statistical decision theory and related topics IV 1, 29–47.

Burnham, K., Anderson, D., 1998. Model selection and inference: a practical informationtheoretic approach: 60-64.

Carlin, B. P., Louis, T. A., 2008. Bayesian methods for data analysis. CRC Press.

Chateau, D., Jared, D., 2003. Spelling–sound consistency effects in disyllabic word naming. Journal of Memory and Language 48 (2), 255–280.

Chen, M.-H., 2005. Computing marginal likelihoods from a single mcmc output. Statistica Neerlandica 59 (1), 16–29.

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.

Cochran, W. G., 1940. The analysis of variance when experimental errors follow the poisson or binomial laws. The Annals of Mathematical Statistics 11 (3), 335–347.

Dixon, P., 2008. Models of accuracy in repeated-measures designs. Journal of Memory and Language 59 (4), 447–456.

Efron, B., 1986. Why isn't everyone a bayesian? The American Statistician 40 (1), 1–5.

Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A., Rubin, D., 2013. Bayesian data analysis. Chapman Hall, London.

Gelman, A., Carlin, J. B., Stern, H. S., Rubin, D. B., 2014a. Bayesian data analysis. Vol. 2. Chapman & Hall/CRC Boca Raton, FL, USA.

Gelman, A., Hwang, J., Vehtari, A., 2014b. Understanding predictive information criteria for bayesian models. Statistics and Computing 24 (6), 997–1016.

Gelman, A., Jakulin, A., Pittau, M. G., Su, Y.-S., 2008. A weakly informative default prior distribution for logistic and other regression models. The Annals of Applied Statistics, 1360–1383.

Haller, H., Krauss, S., 2002. Misinterpretations of significance: A problem students share with their teachers. Methods of Psychological Research 7 (1), 1–20.

Ihaka, R., Gentleman, R., 1996. R: a language for data analysis and graphics. Journal of computational and graphical statistics 5 (3), 299–314.

Jacoby, L. L., Wahlheim, C. N., Kelley, C. M., 2015. Memory consequences of looking back to notice change: Retroactive and proactive facilitation.

Jaeger, T. F., 2008. Categorical data analysis: Away from anovas (transformation or not) and towards logit mixed models. Journal of memory and language 59 (4), 434–446.

Jones, R. H., 2011. Bayesian information criterion for longitudinal and clustered data. Statistics in medicine 30 (25), 3050–3056.

Kass, R. E., Raftery, A. E., 1995. Bayes factors. Journal of the american statistical association 90 (430), 773–795.

Kruschke, J. K., 2013. Bayesian estimation supersedes the t test. Journal of Experimental Psychology: General 142 (2), 573.

Lin, X., 1997. Variance component testing in generalised linear models with random effects. Biometrika, 309–326.

Masson, M. E., 2011. A tutorial on a practical Bayesian alternative to null-hypothesis significance testing. Behavior research methods 43 (3), 679–690.

McCulloch, C. E., Neuhaus, J. M., 2011. Misspecifying the shape of a random effects distribution: why getting it wrong may not matter. Statistical science, 388–402.

Meng, X.-L., Schilling, S., 2002. Warp bridge sampling. Journal of Computational and Graphical Statistics 11 (3), 552–586.

Meng, X.-L., Wong, W. H., 1996. Simulating ratios of normalizing constants via a simple identity: a theoretical exploration. Statistica Sinica, 831–860.

Nathoo, F. S., Ghosh, P., 2013. Skew-elliptical spatial random effect modeling for areal data with application to mapping health utilization rates. Statistics in medicine 32 (2), 290–306.

Nathoo, F. S., Masson, M. E., 2016. Bayesian alternatives to null-hypothesis significance testing for repeated-measures designs. Journal of Mathematical Psychology 72, 144–157.

Plummer, M., 2013. rjags: Bayesian graphical models using mcmc. R package version 3.

Plummer, M., et al., 2003. Jags: A program for analysis of Bayesian graphical models using gibbs sampling. In: Proceedings of the 3rd international workshop on distributed statistical computing. Vol. 124. Technische Universit at Wien Wien, Austria, p. 125.

Quené, H., Van den Bergh, H., 2008. Examples of mixed-effects modeling with crossed random effects and with binomial data. Journal of Memory and Language 59 (4), 413–425.

Raftery, A. E., Newton, M. A., Satagopan, J. M., Krivitsky, P. N., 2006. Estimating the integrated likelihood via posterior simulation using the harmonic mean identity.

Rouder, J. N., Morey, R. D., Speckman, P. L., Province, J. M., 2012. Default Bayes factors for anova designs. Journal of Mathematical Psychology 56 (5), 356–374.

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., Iverson, G., 2009. Bayesian t tests for accepting and rejecting the null hypothesis. Psychonomic bulletin & review 16 (2), 225–237.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., Van Der Linde, A., 2002. Bayesian measures of model complexity and fit. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 64 (4), 583–639.

Wagenmakers, E.-J., 2007. A practical solution to the pervasive problems of p values. Psychonomic bulletin & review 14 (5), 779–804.

Watanabe, S., 2010. Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. The Journal of Machine Learning Research 11, 3571–3594.

Yap, M. J., Balota, D. A., Tse, C.-S., Besner, D., 2008. On the additive effects of stimulus quality and word frequency in lexical decision: Evidence for opposing interactive influences revealed by rt distributional analyses. Journal of Experimental Psychology: Learning, Memory, and Cognition 34 (3), 495.

Author Note

This work was supported by discovery grants to Farouk Nathoo and Michael Masson from the Natural Sciences and Engineering Research Council of Canada. Farouk Nathoo holds a Tier II Canada Research Chair in Biostatistics. Research was enabled in part by support provided by WestGrid (www.westgrid.ca) and Compute Canada (www.computecanada.ca). The authors thank Dr. Belaid Moa for assistance with the implementation of the simulation studies on the Westgrid high-performance computing clusters. We are also grateful to Peter Dixon, Florian Jaeger, Adam Krawitz, and Anthony Marley for very helpful comments on earlier versions of this article.