Received: 27 September 2022 | Accepted: 29 May 2023

DOI: 10.1002/jeab.872

TECHNICAL REPORT

Journal of the Experimental Analysis of Behavior

SEAB SOCIETY FOR THE EXPERIMENTAL ANALYSIS OF BEHAVIOR

Multilevel analysis of matching behavior: A comparison of maximum likelihood and Bayesian estimation

Michael John Ilagan[1] | Pier-Olivier Caron[2] | Milica Miočević[1]

[1]McGill University, Montréal, QC, Canada

[2]Université TÉLUQ, Montréal, QC, Canada

Correspondence
Pier-Oliver Caron, Département des Sciences humaines, Lettres et Communication, Université TÉLUQ, 5800, Rue Saint-Denis, Bureau 1105, Montréal, QC, Canada H2S 3L5.
Email: pier-olivier.caron@teluq.ca

Funding information
Fonds d'Aide à la Recherche-Université TÉLUQ awarded to Dr. Caron and SSHRC Insight Grant awarded to Dr. Miočević

Editor-in-Chief: Mark Galizio
Handling Editor: Brent Alsop

Abstract

While trying to infer laws of behavior, accounting for both within-subjects and between-subjects variance is often overlooked. It has been advocated recently to use multilevel modeling to analyze matching behavior. Using multilevel modeling within behavior analysis has its own challenges though. Adequate sample sizes are required (at both levels) for unbiased parameter estimates. The purpose of the current study is to compare parameter recovery and hypothesis rejection rates of maximum likelihood (ML) estimation and Bayesian estimation (BE) of multilevel models for matching behavior studies. Four factors were investigated through simulations: number of subjects, number of measurements by subject, sensitivity (slope), and variance of the random effect. Results showed that both ML estimation and BE with flat priors yielded acceptable statistical properties for intercept and slope fixed effects. The ML estimation procedure generally had less bias, lower RMSE, more power, and false-positive rates closer to the nominal rate. Thus, we recommend ML estimation over BE with uninformative priors, considering our results. The BE procedure requires more informative priors to be used in multilevel modeling of matching behavior, which will require further studies.

KEYWORDS
Bayesian estimation, matching behavior, matching law, maximum likelihood, multilevel model, pooled data, statistical analysis

There is a long tradition within behavior analysis of studying behavior in a concurrent schedule in which a subject has to choose between two options. Each option is associated with its own differential reinforcer rate, generally a variable-interval or a random-interval schedule of reinforcement. Over numerous sessions, the subject can choose repeatedly between alternatives. In this procedure, a behavioral pattern known as matching behavior emerges. This is well described by the generalized matching law (Baum, 1974) in log form:

$$\log\left(\frac{b_1}{b_2}\right) = a\log\left(\frac{r_1}{r_2}\right) + \log c + \epsilon, \tag{1}$$

where b refers to response rate and r to reinforcer rate; the indices (1 and 2) specify the two options, and ϵ represents the residual term. The parameter a refers to sensitivity to reinforcement, or the degree to which an organism adjusts its response ratio according to the reinforcer ratio, and c refers to the bias, or the behavioral preference for one response (the numerator) over the other (the denominator). Equation 1 is a simple regression analysis, where $x = \log\left(\frac{r_1}{r_2}\right)$, $y = \log\left(\frac{b_1}{b_2}\right)$, and a and $\log c$ are the slope and intercept, respectively.

Matching behavior is a within subject-oriented model; that is, parameters of the generalized matching law are specific to the subject (Herrnstein, 1970). While trying to infer laws of behavior for the subjects' species (e.g., rats, pigeons, primates, humans), accounting for both the within-subjects and the between-subjects variances is often overlooked (Caron, 2019). Matching behavior is generally analyzed by subjects independently (within-subjects variance only) or by pooling subjects (between-subjects variance only). Both types of analyses yield potentially biased and insufficiently detailed and informative conclusions, as the use of direct aggregation of data across subjects or averaged parameters do not consider both within- and between-subjects variances.

It has been advocated recently to use multilevel modeling (MLM) to analyze matching behavior (Caron, 2019). Multilevel analysis structures data into different levels (Gelman & Hill, 2006)—for instance, behavior nested within subjects. Multilevel modeling is preferred when combining several subjects' behavior because it accounts simultaneously for the within-subjects and between-subjects variance. The parameters of all subjects are considered to infer group characteristics using both the behavioral (within) and the subject (between) levels simultaneously. For a repeated-measures design, the number of subjects is the Level 2 sample size, whereas the number of data points per subject is the Level 1 sample size. If a single model is estimated from all points (complete pooling, Level 2 only), the variation between subjects is unaccounted for. If each model is estimated without pooling (Level 1 only), the estimate is uninformed by the between-subject variations in the parameter estimates. Multilevel modeling provides a mathematically sound framework to recognize both levels.

The extension of Equation 1 to a multilevel model can be written as

$$\log\left(\frac{b_{1,s}}{b_{2,s}}\right) = a_s \log\left(\frac{r_{1,s}}{r_{2,s}}\right) + \log c_s + \epsilon, \tag{2}$$

where the indices s emphasize the subjects' response and reinforcer rates as well as that sensitivity (a_s) and bias ($\log c_s$) are coefficients specific to individuals that are sampled from the same population. Multilevel analysis assumes that the sensitivity and bias parameters follow a bivariate normal distribution. Their population means are μ_a and $\mu_{\log c}$, and their corresponding variances and covariance are σ_a^2, $\sigma_{\log c}^2$, and $\rho\sigma_a\sigma_{\log c}$ such that

$$\begin{pmatrix} a_s \\ \log c_s \end{pmatrix} \sim \mathcal{N}\left[\begin{pmatrix} \mu_a \\ \mu_{\log c} \end{pmatrix}, \begin{pmatrix} \sigma_a^2 & \rho\sigma_a\sigma_{\log c} \\ \rho\sigma_a\sigma_{\log c} & \sigma_{\log c}^2 \end{pmatrix}\right], \tag{3}$$

where ρ is the correlation between sensitivity and bias (e.g., higher value of sensitivity could be related to higher value of bias) and σ_a and $\sigma_{\log c}$ are the standard deviations of sensitivity and bias. From Equation 3, slopes (a_s) and intercepts ($\log c_s$) are random coefficients that are free to vary across subjects and differ according to their given means and variance–covariance matrix.

Using MLM in behavior analysis has its own challenges though. Adequate sample sizes are required (at both levels) to estimate multilevel models without bias. Multilevel models are often estimated with maximum likelihood (ML) or Bayesian methods; both estimation approaches have their own advantages and disadvantages.

Maximum likelihood estimation is asymptotically unbiased but is known to behave less desirably with smaller sample sizes, particularly with a low number of clusters (Cousineau & Laurencelle, 2016). The recommendation from the methodological literature on smaller cluster sizes is to use the Kenward–Roger (2009) adjustment, a correction to the degrees of freedom and covariance matrix adjustment that reduces the underestimation of fixed-effect standard errors when the number of clusters is small (McNeish, 2017; McNeish & Stapleton, 2016). Even with the correction, sample size matters. Even though there is no gold standard, there have been a few guidelines suggesting 30 clusters (Level 2) with a cluster size of 30 (Level 1; Kreft, 1996), a minimum of 20 clusters (Snijders & Bosker, 2012), or 50 clusters with a cluster size of 20 for cross-level interactions (Hox, 2010). These sizes appear reasonable for fields that typically collect large samples, such as education and economics, where 30 clusters (classes, companies) measured quarterly is reasonable. However, these recommendations are not directly applicable to single-case studies where few subjects are measured daily (or more) for long periods (Kazdin, 2021). There have been no thorough investigations of statistical properties of relevant parameters in multilevel models with fewer than 20 clusters, which is considered a small number (Arend & Shäfer, 2019; Austin & Leckie, 2018; Maas & Hox, 2005).

Bayesian methods relieve the researcher from relying on maximum likelihood asymptotics, at the cost of using priors and relying on Markov chain Monte Carlo (MCMC) sampling. Parameter estimates in MLM are known to be biased and have less precise variance estimates when the number of subjects (Level 2) is small. Studies on multilevel modeling of single-case data, specifically AB phase designs for each subject, showed that five and seven subjects yielded relative bias close to 5% and yielded more precise estimates relative to analyses with three subjects (Moeyaert et al., 2017). Using more informative priors with Bayesian techniques led to more precise estimates for the fixed effects and random effects, even for three subjects. However, matching behavior studies are different in two respects: (a) they do not use AB designs and (b) the variance in the average response rate ratio explained by the linear model is typically very high, higher than 80% in most studies in experimental settings (Davison & McCarthy, 1988). Those specific aspects were not thoroughly investigated in prior research. Thus, multilevel models for matching behavior are expected to yield high precision and power for parameters of interest, even with uninformative priors.

Even though models with small sample sizes may converge toward a solution and produce estimates, they may be biased and may mislead research (McNeish, 2017; McNeish & Stapleton, 2016). It is of interest for behavior analysts to better understand the statistical properties of parameters in multilevel models under conditions typically encountered in behavior-matching research. Before going further, the next section will briefly introduce ML estimation and Bayesian estimation (BE).

ESTIMATION METHODS IN TWO-LEVEL MODELING OF MATCHING BEHAVIOR

Maximum likelihood estimation

When a model is specified, there are multiple possible values the parameter of interest can take. In estimation, it is the goal of the researcher to find the "best" candidate to characterize the observed data. To find a best candidate for the parameter, what is "good" must be defined first.

One general definition of "good" applicable to many models relies on the likelihood function (see Myung, 2003, for a tutorial on ML estimation). Precisely, the likelihood of a candidate value of the parameter is the probability it assigns to the data.[1] When a candidate value makes the observed data likely, the candidate has high likelihood. By this criterion, the "best" candidate, called the ML estimate, is the one that yields the highest probability for the observed data. For instance, under the conventional assumptions in linear regression, the ML estimate for the regression slopes coincides with the ordinary least squares estimate.

As is often the case in MLM studies, when the sample size is small, traditional ML estimation is compromised in two aspects (McNeish, 2017; McNeish & Stapleton, 2016). First, the variances of random effects are underestimated. Second, the fixed-effect standard errors and the degrees of freedom used for hypothesis testing are incorrect, thus inflating the Type I error rate. As McNeish (2017) explained, the MLM literature recommends small-sample methods that help ML estimation avoid these consequences: restricted maximum likelihood (REML) is used to properly estimate random effects, and the Kenward–Roger correction (Kenward & Roger, 1997) is used to give more accurate standard errors and degrees of freedom to t-test statistics. When the sample is large, REML and traditional ML yield similar results.

Under ML estimation, the parameter itself, though unknown, is nonrandom. There is no provision to find a distribution over the candidate values of the parameter without the Bayesian perspective.

Bayesian estimation and inference

In the Bayesian framework, probability is viewed as a degree of belief, which does not rest on the frequentist assumption of repeating an event many times. It is possible to express one's belief without observing any data, which is why the Bayesian view of probability is sometimes labeled as subjective. Consequently, parameters have distributions: The prior distribution of a parameter encodes the best guess and uncertainty regarding the parameter value prior to observing the data, and the posterior distribution is obtained by updating the prior distribution with the observed data using Bayes' theorem.

Bayesian estimation starts with the specification of a prior distribution for all freely estimated parameters. The single predictor regression model in Equation 1 consists of the intercept a and the slope $\log c$. The specification of prior distributions requires selecting a distributional form (e.g., normal, gamma) and hyperparameters that govern its shape. For intercepts and slopes in regression models it is typical to select normal prior distributions such as

$$a \sim \mathcal{N}\left(\mu_{0a}, \sigma^2_{0a}\right) \tag{4}$$

and

$$\log c \sim \mathcal{N}\left(\mu_{0\ \log c}, \sigma^2_{0\ \log c}\right), \tag{5}$$

where the mean hyperparameters of the normal priors, μ_{0a} and $\mu_{0\log c}$, encode the best guess for the value of the parameter and the variances of the normal priors, σ^2_{0a} and $\sigma^2_{0\log c}$, encode the level of confidence surrounding the best guess. Larger variance hyperparameters correspond to less informative (more diffuse) prior distributions. The prior distributions are then updated with observed data to obtain the posterior distribution.

When model parameters vary by subject, Bayesian methods offer an intuitive way to specify the hierarchical nature of the model. The parameters in Equation 3 and the Level 1 residual variance are unknown parameters that are assigned prior distributions.

Although the result of a Bayesian analysis is a posterior *distribution*, making inferences requires summarizing the posterior in terms of the familiar quantities such as point estimates, interval estimates, and hypothesis tests from frequentist analysis. We consider the following appropriate Bayesian analogues. For point estimates, one possible Bayesian analogue is the mean of the posterior distribution, called the posterior mean. For interval estimates, the Bayesian analogue is the interval covering the middle 95% of the posterior distribution, called the 95% equal-tail credible interval. Unlike the frequentist confidence interval, the credible interval can be interpreted probabilistically: There is a 95% probability that the true value lies within the interval. For hypothesis testing, the Bayesian analogue rejects a hypothesis that the parameter is equal to a given value if and only if this parameter value is not included in the 95% credible interval.

Such Bayesian analogues require the calculation of means and percentiles from the resulting posterior distributions. However, not all posterior distributions have an

[1] Although it may appear confusing to use the term probability in the definition of the likelihood function, we are simply following conventions from textbooks on Bayesian statistics (Gelman & Hill, 2006, p. 388).

analytic derivation (Hoff, 2009). In this case, the posterior distribution (and consequently, its summary statistics) is approximated using MCMC sampling instead, and MCMC sampling takes draws that in the limit could be considered draws from the posterior distribution (Hoff, 2009). The initial iterations in MCMC sampling are likely not sampling from the posterior distribution, and the analyst needs to diagnose at which point the chains converged to the target (posterior) distribution and discard all draws prior to convergence. There are numerous convergence diagnostics (Cowles & Carlin, 1996). In the social sciences literature, the most commonly encountered convergence diagnostics are those offered by the majority of software packages, which are the potential scale reduction factor (Gelman & Rubin, 1992), Geweke's diagnostic (1992), and trace plots of draws plotted against the iteration number for each parameter (Brooks, 1998). In the current study, we use the potential scale reduction factor (referred to as R-hat) and the effective sample size (percentage of independent draws from the posterior; Stan Development Team, 2020).

The purpose of the current study is to compare parameter recovery and hypothesis rejection rates of ML estimation and BE in contexts of matching-behavior studies using a simulation study.

SIMULATION STUDY

All the simulations[2] were carried out in R (R Core Team, 2022), using the package *MASS* (Ripley, 2016) to generate the data; *lme4* (Bates et al., 2015), *pbkrtest* (Halekoh & Højsgaard, 2014), and *lmerTest* (Kuznetsova et al., 2017) for ML estimation of multilevel models; and *Rstan* (Stan Development Team, 2020) for BE.[3]

Generating data

Four factors were investigated: number of subjects (n_1), number of measurements by subject (n_2), sensitivity (expected slope), and variance of both random effects. The numbers of subjects were set to four, six, and eight, and the numbers of measurement were 20, 40, and 60. Other parameters were taken from studies on matching behavior. We used the data sets from Davison and Hogsden (1984; henceforth DH), an experimental setting (Part 4 is similar to usual matching studies), and Rivard et al. (2014; henceforth RFKB), an applied setting to set expected slope, expected intercept and their variances, and the error (Level 1) standard deviation. In both cases, the bias (intercept) was set to 0 throughout the simulation (as it is found to be close to 0 in most studies). Expected slope values were set to .875 (taken from the DH data set) and 1 (taken from the RFKB data sets). The variance was the same for slope and intercept, and their covariance was set to 0. Variances were set to either .03 (taken from the DH data set) for both parameters or .17 (taken from the RFKB data sets) for both parameters. The residuals were independent draws from a normal distribution with 0.3 standard deviation, which corresponds to the percentage of variance accounted for (equivalent to $R^2 = .836$) by the generalized matching law and was approximately equal in both data sets. These chosen parameters closely match seminal results in Baum (1979) and Wearden and Burgess (1982)—that is, a slope of .867, an intercept of .010 and an explained variance of .899. Log reinforcer ratios, which were the predictor in the linear regression, were also randomly generated from a standard normal distribution for each subject in the sample. Each of the $3 \times 3 \times 2 \times 2 = 36$ scenarios was replicated 1,000 times.

Fitting models

In line with Moeyaert et al. (2017), the covariance between the random intercept and the random slope was fixed to be zero for both ML and BE.

The ML estimation procedure was carried out via REML, as implemented in the R package *lmerTest* (Kuznetsova et al., 2017). The Kenward–Roger correction, as implemented in the R package *pbkrtest* (Halekoh & Højsgaard, 2014) was used to calculate the degrees of freedom and standard error of the fixed-effects tests. The corresponding hypothesis tests used the same packages.

For BE, we used the R package *Rstan* (Stan Development Team, 2020) for MCMC sampling. For the intercept and slope fixed effects, we used the default flat priors unbounded, but for the variance parameters, the priors were flat but bounded below by zero. Four MCMC chains were used.

Analyzing data

In each scenario, we conducted estimation and hypothesis testing on MLM under both ML estimation and BE. The outcome measures were rejection rates, bias and efficiency, and nonconvergence rates. Each outcome measure was averaged over only the replicates where the estimation converged. For ML estimation, the software explicitly indicated whether convergence was met. In BE, the sampler was considered converged if the potential scale reduction factor (Gelman & Rubin, 1992) value was below 1.05 and the effective sample size was at least 10% of the 10,000 MCMC iterations for every parameter.

[2] The files can be found online, https://osf.io/xdu3y/?view_only=acc9f244c2dd45389207721b5bf69365.

[3] For users that are more familiar with *lmer*, but wish to fit the model in the Bayesian framework, see these two tutorials for *rstanarm* (https://mc-stan.org/users/documentation/case-studies/tutorial_rstanarm.html) and *brms* (https://ourcodingclub.github.io/tutorials/brms/).

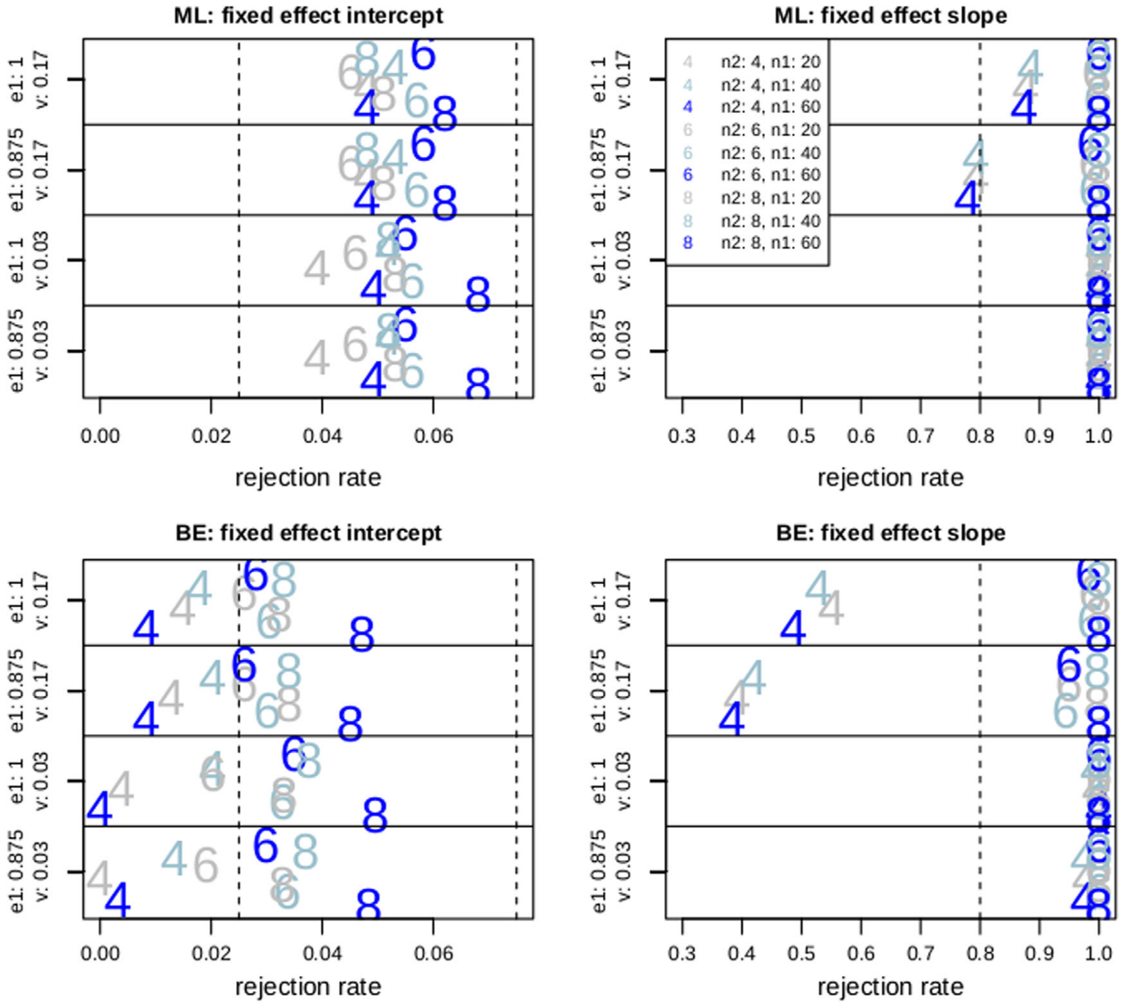

FIGURE 1 Rejection rates. Panels in the top row are for maximum likelihood estimation (ML) estimation, and panels in the bottom row are for Bayesian estimation (BE). The left column represents the intercept fixed effect, and the right column represents slope fixed effect: e0 = fixed effect intercept, e1 = fixed effect slope, and v = random effect variances.

For testing hypotheses, the BE analogues to ML estimation, as described in the previous section, were applied. Rejection rates were evaluated for two hypothesis tests of interest: the intercept fixed effect (is the expected intercept zero?) and slope fixed effect (is the expected slope zero?). We evaluated the bias, defined as the average deviation from the estimand, $E\left(\widehat{\theta}-\theta\right)$; relative bias, the bias divided by the estimand, $\frac{E\left(\widehat{\theta}-\theta\right)}{\theta}$; and the root mean square error (RMSE), $\sqrt{E\left(\widehat{\theta}-\theta\right)^2}$, which is equivalent to square of the bias plus the variance, of the intercept, slope, and variance parameters. Between bias and relative bias, we used bias for the expected intercept, as its true value was zero, but we used relative bias for the rest of the parameters.

The chosen values of the intercept and slope determined whether a rejection rate was a Type I error rate or power. For the expected intercept, the true parameter was always zero, so the rejection rate always represented the Type I error. For the expected slope, the true parameter was always nonzero, so rejection rate always represented power. Type I error rate was considered acceptable if it was within Bradley's (Bradley, 1978) liberal robustness criterion, which was within the range of 2.5% to 7.5%. For power, acceptable values were at least 80% (Cohen, 1992).

RESULTS

The simulation results reported below are organized first by outcome measure: rejection rate, (relative) bias, and RMSE. Then within each outcome measure, they are organized by parameter of interest: intercept fixed effect, slope fixed effect, and random effects. Both ML estimation and BE results are reported.

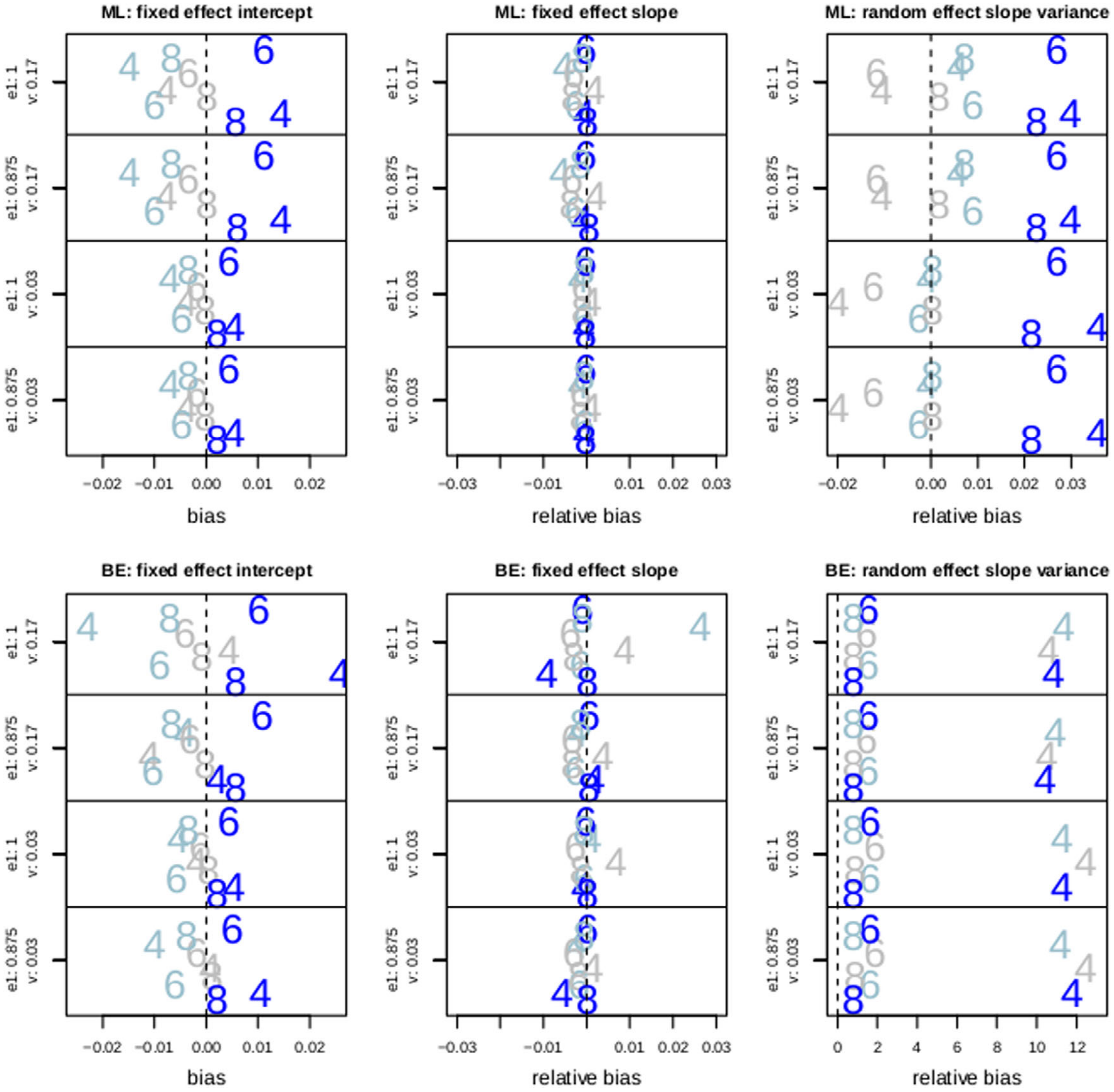

FIGURE 2 Bias and relative bias. See note to Table 1 for abbreviations.

All figures are structured similarly. The panels in the top row are for ML estimation, and panels in the bottom row are for BE. From left to right, columns represent the intercept fixed effect, slope fixed effect, and variance of slope random effect (except for Figure 1, see fn4). Each panel is made up of four boxes, demarcated by horizontal lines. Each box corresponds to a combination of true slope (either .875 or 1) and true random-effect variances (either .03 or 0.17). Within each box, the height of a point has no meaning: The points are drawn at different heights merely for visual clarity. As shown in the legend, colors and characters denote different combinations of sample sizes n_1 (which may be 20, 40, or 60) and n_2 (which may be 4, 6, or 8). Dashed vertical lines mark the 2.5% and 7.5% boundaries for acceptable Type I error rates (where the nominal rate is 5%) and the minimum acceptable 80% rejection rate for power. Also, a dashed vertical line marks zero bias and zero relative bias.

Rejection rates

Figure 1 shows the results for the rejection rates.[4] The left column of Figure 1 depicts the intercept fixed effect. Rejection rate is Type I error rate. For ML estimation (top row), all Type I error rates were within the acceptable range. For BE (bottom row), the test was conservative, going below the nominal rate for fewer subjects.

For the slope fixed effect shown in the right column of Figure 1, rejection rate is power. For scenarios where $n_2 = 4$ and the variance of the random effects were larger, power was noticeably smaller. However, for ML estimation (top row), these scenarios were still mostly within the acceptable range, whereas they were underpowered for BE (bottom row). For the rest of the scenarios, both ML estimation and BE had power well above 80%.

[4] Rejection rates for random effects were also investigated for ML. Results are in the supplementary material.

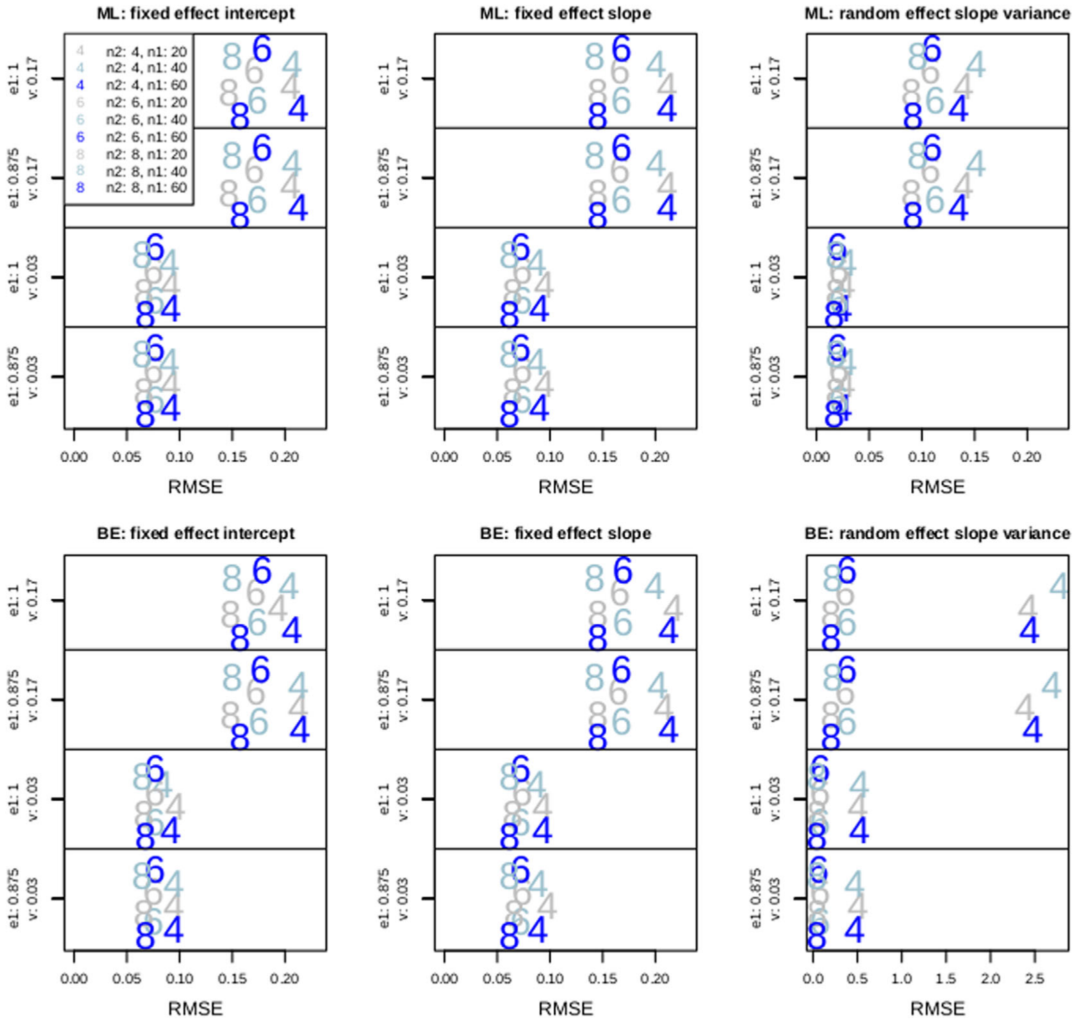

FIGURE 3 Root mean square error. Panels in the top row are for ML, and panels in the bottom row are for BE. From left to right, columns represent the intercept fixed effect, slope fixed effect, and slope random effects: RMSE = root mean squared error, e0 = fixed effect intercept, e1 = fixed effect slope, v = random effect variances, and v1 = slope random effect variances.

Bias and relative bias

Figure 2 shows plots for the bias or relative bias. The format is similar to that for Figure 1, with bias or relative bias on the horizontal axis. Bias is shown for the intercept fixed effect (left column), whereas relative bias is shown for the slope fixed effect (middle column) and the slope random-effect variance (right column). The random effect is shown for only the slope, as the results bear a similar pattern for the intercept random effect.

For the fixed-effect intercept estimate in the left column of Figure 2, bias did not exceed 0.03 in absolute value for all scenarios in both ML estimation (top row) and BE (bottom row) analyses.

For the fixed-effect slope estimate shown in middle column of Figure 2, the relative bias values were close to zero.

In the right column of Figure 2, there is a stark contrast between the ML estimation (top row) and BE (bottom row) results for the slope random-effect estimate. Due to Bayesian analysis having much higher relative bias, the plots could not be placed in the same range. Relative bias for ML estimation was no worse than 4% in absolute value, whereas it was greater than 100% in many BE scenarios.

Root mean square error

Figure 3 shows plots for RMSE. The format is similar to that for the previous figures, with RMSE taking the horizontal axis. Only the random-effect variances for the slope are shown, as the results are similar for the intercept. Overall, RMSE was higher for scenarios with larger random effects and with fewer subjects (i.e., smaller n_2).

For the fixed-effect intercept estimate (left column of Figure 3), RMSE was similar between ML estimation (top row) and BE (bottom row). For the fixed-effect slope estimate in middle column of Figure 3, RMSE was similar for ML estimation (top row) and BE (bottom row). For the slope random-effect estimate in right column of

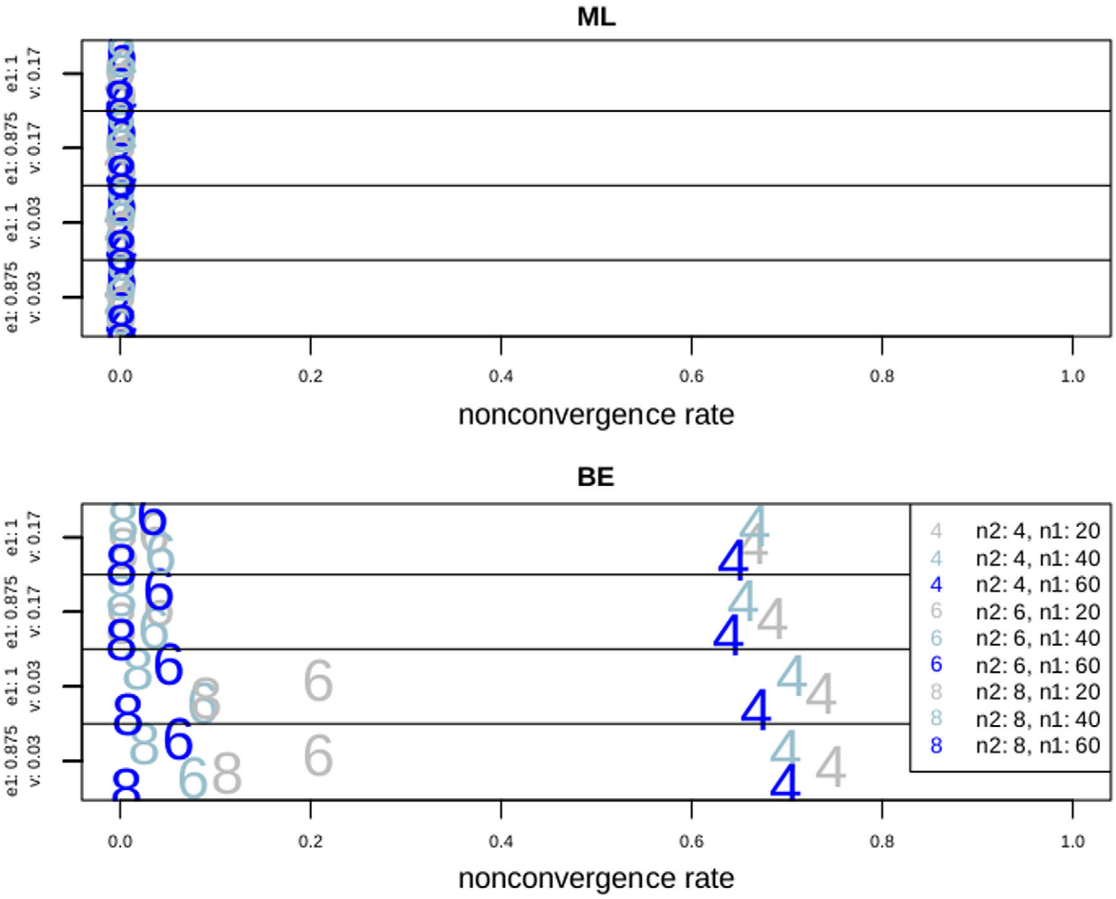

FIGURE 4 Nonconvergence. Panels in the top row are for ML, and panels in the bottom row are for BE.

Figure 3, there is a stark contrast between the ML estimation (top row) and BE (bottom row) results. Due to BE having much larger RMSE, the plots could not be placed in the same range.

Nonconvergence

Figure 4 shows the plots for nonconvergence. Similar to the previous figures, the top plot corresponds to ML estimation, and the bottom plot corresponds to BE. However, nonconvergence is not associated with any parameter of the MLM, so there is only one column. The horizontal axis is the proportion of replicates that did not meet the convergence criteria. There is a stark contrast between the ML estimation and the BE results. Nonconvergence did not exceed 1% in ML estimation, but in BE nonconvergence occurred at a much higher rate, reaching over 60% for scenarios with fewer subjects (i.e., smaller n_2).

DISCUSSION

The purpose of the current study was to compare parameter recovery and hypothesis rejection rates of ML estimation and BE in contexts of matching-behavior data. Simulations were carried out with 36 scenarios from which the parameters were derived from real studies. In each scenario, estimation and hypothesis testing on the MLM bivariate normal parameters under both ML and BE were conducted. The outcome measures were rejection rates (fixed effect, slope fixed effect, and the random effects), bias, relative bias, RMSE, and nonconvergence rates.

The current simulations show that the rejection rates intercept of ML estimation was adequate, whereas BE was conservative. Power to test the slope was good in both ML estimation and BE. However, ML estimation was better with small sample sizes compared with BE because it reaches 80% of statistical power in all scenarios, whereas it was lower (around 40% to 60%) when the number of subjects was four cases for BE.

Regarding bias and relative bias, there was no systematic bias for the intercept in ML estimation and BE. However, BE overestimates of the random slope effect were so important, especially in the $v = .17$ scenarios, that a differential had to be used to compared ML estimation and BE. The referential scales in Figure 2, third column range from −.01 to .03 for ML estimation, whereas for BE it ranges from 0 to 13. Overall, BE tended to overestimate the variance of the slope random effect.

The current results showed no particular trends in RMSE. As expected, RMSE was higher for conditions with higher error variance. The similarity in RMSEs indicates that the differences in bias between ML estimation and BE were not substantial for the fixed effects. The only noticeable difference between ML estimation and BE occurred for the RMSE of the random slope effect, where, like previously, panels necessitated different scales: ML estimation had the smaller scale ranging from 0 to .20, whereas BE scale ranged from 0 to 4.

Finally, ML estimation had no convergence issues, whereas BE had some convergence issues, especially with low cluster sizes like $n_2 = 4$. Upon closer inspection, it appeared that the convergence issues were due to difficulty in estimating the random effects.

A limitation of the current study was the range of parameters being investigated. The data were derived from only two data sets that are examples of a wider range of data sets that researchers might use for matching analyses. Another limitation was that only uninformative default priors were investigated for BE. Given the numbers of replication, conditions, and the time for the chains to convergence, we had to prioritize completing the simulations in a reasonable time. The next step would be to investigate a wider range of prior specifications and parameter values, especially for the residual variances. Alternative computational modeling approaches to simulate matching behavior could be used in future studies.

Conclusions

For scientific results to be trustworthy, statistical methods must be appropriate for the data structure and the method ought to have good performance under representative conditions. Traditional methods in the matching-behavior literature fall short of accounting for within- and between-subjects levels of variation. As Caron (2019) argued, at least conceptually, MLMs were more suited to do so. However, good performance cannot be taken for granted, as simulation studies with MLMs (e.g., McNeish & Stapleton, 2016) were based on larger sample sizes. The current study provides evidence that MLMs can indeed perform acceptably to perfectly well for matching-behavior data.

Overall, the results were promising. In most scenarios, for intercept and slope fixed effects, both ML estimation and BE with flat priors yielded acceptable statistical properties. However, ML estimation generally had less bias, lower RMSE, more power, and rejection rates closer to the nominal rate. Thus, we recommend ML estimation over BE considering our results.

Nonetheless, BE can be improved to be used to analyze matching-behavior data. Our simulation study was limited to using flat priors, which did especially poorly for estimating random effects. This is consistent with findings from a systematic review of studies comparing Bayesian and frequentist methods at small sample sizes (Smid et al., 2020); specifically, ML with small sample corrections can outperform Bayesian methods with diffuse priors, especially in the estimation of variance parameters. Statistical properties can be improved by informative accurate priors, and the wider literature on MLM offers guidance on informative priors (e.g., Moeyaert et al., 2017). Further studies are needed to better understand the influence of informative priors on MLM for the analysis of matching behavior and to develop thorough guidelines for researchers.

CONFLICT OF INTEREST STATEMENT

The authors report no conflict of interest.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available in the supplementary material of this article.

ETHICS APPROVAL

No human or animal subjects were used to produce this article.

ORCID

Michael John Ilagan https://orcid.org/0000-0001-6340-9346
Pier-Olivier Caron https://orcid.org/0000-0001-6346-5583
Milica Miočević https://orcid.org/0000-0001-8487-3666

REFERENCES

Arend, M. G., & Shäfer, T. (2019). Statistical power in two-level models: A tutorial on Monte Carlo simulation. *Psychological Methods*, *24*(1), 1–19. https://doi.org/10.1037/met0000195

Austin, P. C., & Leckie, G. (2018). The effect of number of clusters and cluster size on statistical power and Type I error rates when testing random effects variance components in multilevel linear and logistic regression models. *Journal of Statistical Computation and Simulation*, *88*(16), 3151–3163. https://doi.org/10.1080/00949655.2018.1504945

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Baum, W. M. (1974). On two types of deviation from the matching law: Bias and undermatching. *Journal of the Experimental Analysis of Behavior*, *22*(1), 231–242. https://doi.org/10.1901/jeab.1974.22-231

Baum, W. M. (1979). Matching, undermatching, and overmatching in studies of choice. *Journal of the Experimental Analysis of Behavior*, *32*(2), 269–281. https://doi.org/10.1901/jeab.1979.32-269

Bradley, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology*, *31*(2), 144–152. https://doi.org/10.1111/j.2044-8317.1978.tb00581.x

Brooks, S. P. (1998). Markov chain Monte Carlo method and its application. *The American Statistician*, *47*(1), 69–100. https://doi.org/10.1111/1467-9884.00117

Caron, P.-O. (2019). Multilevel modeling of matching behavior. *Journal of the Experimental Analysis of Behavior*, *111*(2), 183–191. https://doi.org/10.1002/JEAB.510

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*(1), 155–159. https://doi.org/10.1037//0033-2909.112.1.155.

Cousineau, D., & Laurencelle, L. (2016). A correction factor for the impact of cluster randomized sampling and its applications.

Psychological Methods, *21*(1), 121–135. https://doi.org/10.1037/met0000055

Cowles, M. K., & Carlin, B. P. (1996). Markov chain Monte Carlo convergence diagnostics: A comparative review. *Journal of the American Statistical Association*, *91*(434), 883–904. https://doi.org/10.1080/01621459.1996.10476956

Davison, M., & Hogsden, I. (1984). Concurrent variable-interval schedule performance: Fixed versus mixed reinforcer durations. *Journal of the Experimental Analysis of Behavior*, *41*(2), 169–182. https://doi.org/10.1901/jeab.1984.41-169

Davison, M., & McCarthy, D. (1988). *The matching law: A research review*. Erlbaum.

Gelman, A., & Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press.

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, *7*, 457–472. https://doi.org/10.1214/ss/1177011136

Geweke, J. (1992). Evaluating the accuracy of sampling-based approaches to calculating posterior moments. In J. M. Bernado, J. O. Berger, A. P. Dawid, & A. M. F. Smith (Eds.), *Bayesian Statistics 4* (pp. 169–193). Clarendon Press.

Halekoh, U., & Højsgaard, S. (2014). A Kenward-Roger approximation and parametric bootstrap methods for tests in linear mixed models: The R package pbkrtest. *Journal of Statistical Software*, *59*(9), 1–32. https://doi.org/10.18637/jss.v059.i09

Herrnstein, R. J. (1970). On the law of effect. *Journal of the Experimental Analysis of Behavior*, *13*(2), 243–266. https://doi.org/10.1901/jeab.1970.13-243

Hoff, P. D. (2009). *A first course in Bayesian statistical methods*. Springer.

Hox, J. J. (2010). *Multilevel analysis: Techniques and applications*. Routledge.

Kazdin, A. E. (2021). Single-case experimental designs: characteristics, changes, and challenges. *Journal of the Experimental Analysis of Behavior*, *115*(1), 56–85. https://doi.org/10.1002/jeab.638

Kenward, M. G., & Roger, J. H. (1997). Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics*, *53*(3), 983–997. https://doi.org/10.2307/2533558

Kenward, M. G., & Roger, J. H. (2009). An improved approximation to the precision of fixed effects from restricted maximum likelihood. *Computational Statistics & Data Analysis*, *53*(7), 2583–2595. https://doi.org/10.1016/j.csda.2008.12.013

Kreft, I. (1996). *Are multilevel techniques necessary? An overview, including simulation studies*. ERIC. https://eric.ed.gov/?id=ED371033

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*(13), 1–26. https://doi.org/10.18637/jss.v082.i13

Maas, C. J. M., & Hox, J. J. (2005). Sufficient sample sizes for multilevel modeling. *Methodology*, *1*(3), 86–92. https://doi.org/10.1027/1614-2241.1.3.86

McNeish, D. M. (2017). Small sample methods for multilevel modeling: A colloquial elucidation of REML and the Kenward-Roger correction. *Multivariate Behavioral Research*, *52*(5), 661–670. https://doi.org/10.1080/00273171.2017.1344538

McNeish, D. M., & Stapleton, L. M. (2016). The effect of small sample size on two-level model estimates: A review and illustration. *Educational Psychology Review*, *28*(2), 295–314. https://doi.org/10.1007/s10648-014-9287-x

Moeyaert, M., Rindskopf, D., Onghena, P., & Van den Noortgate, W. (2017). Multilevel modeling of single-case data: A comparison of maximum likelihood and Bayesian estimation. *Psychological Methods*, *22*(4), 760–778. https://doi.org/10.1037/met0000136

Myung, J. (2003). Tutorial on maximum likelihood estimation. *Journal of Mathematical Psychology*, *47*(1), 90–100. https://doi.org/10.1016/S0022-2496(02)00028-7

R Core Team. (2022). *R: A language and environment for statistical computing* [Computer software]. In R Foundation for Statistical Computing. http://www.Rproject.org/

Ripley, B. (2016). *Package 'MASS'* [Computer software]. https://cran.r-project.org/web/packages/MASS/MASS.pdf

Rivard, M., Forget, J., Kerr, K., & Bégin, J. (2014). Matching law and sensitivity to therapist's attention in children with autism spectrum disorders. *The Psychological Record*, *65*(1), 79–88. https://doi.org/10.1007/s40732-014-0015-1

Smid, S. C., McNeish, D., Miočević, M., & van de Schoot, R. (2020, 2020/01/02). Bayesian versus frequentist estimation for structural equation models in small sample contexts: A systematic review. *Structural Equation Modeling: A Multidisciplinary Journal*, *27*(1), 131–161. https://doi.org/10.1080/10705511.2019.1577140

Snijders, T. A. B., & Bosker, R. J. (2012). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. Sage.

Stan Development Team. (2020). *RStan: the R interface to Stan* [Computer software]. http://mc-stan.org/

Wearden, J. H., & Burgess, I. S. (1982). Matching since Baum (1979). *Journal of the Experimental Analysis of Behavior*, *38*(3), 339–348. https://doi.org/10.1901/jeab.1982.38-339

How to cite this article: Ilagan, M. J., Caron, P.-O., & Miočević, M. (2023). Multilevel analysis of matching behavior: A comparison of maximum likelihood and Bayesian estimation. *Journal of the Experimental Analysis of Behavior*, *120*(2), 253–262. https://doi.org/10.1002/jeab.872