Investigating diagnostics for generalized linear mixed models

by

Adam J. Karren

B.S., Brigham Young University, 2019

A REPORT

submitted in partial fulfillment of the requirements for the degree

MASTER OF SCIENCE

Department of Statistics
College of Arts and Sciences

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2022

Approved by:

Major Professor
Dr. Nora M. Bello

Copyright

Abstract

Generalized linear mixed models (GLMMs) are extensions of linear mixed models that enable non-normal distributional assumptions on the response of interest. Effective diagnostic metrics and tools to assess GLMM fit and performance are limited. The objective of this study was to develop and explore potential diagnostics to assess GLMM fit and performance to ultimately inform model choice, specifically for discrete count responses. We conducted a simulation study whereby a count response variable was generated by three realistic data generation processes (DGP) under a 2x2 factorial treatment structure in randomized complete blocks. Simulated data were fitted with competing models, including various GLMM specifications and normal approximations with and without transformations. For each DGP, model performance was assessed for accuracy of estimation of treatment means, as well as for Type I error and power for inference on differential treatment effects. Models were evaluated and compared using the Pearson Chi-Square over degrees of freedom statistic for overdispersion and information criteria. Further, we developed an array of potential diagnostic metrics based on model point predictions and used them to assess the ability of competing models to recreate selected features of count data, specifically skewness and dispersion. Overall, the diagnostic metrics evaluated failed to identify the corresponding true model for each DGP. Meanwhile, regardless of DGP, a Poisson-Unit GLMM outperformed other model specifications in fitting selected features of count data. An entomological dataset was used for proof-of-concept application. Further study is warranted to best inform GLMM specification for count response variables.

Table of Contents

List of Figures v
List of Tables vi
Acknowledgements ix
Introduction 1
Methodology 6
2.1 Simulation Scenarios 6
2.1.1 Data Generation Process 1: Poisson-Unit 6
2.1.2 Data Generation Process 2: Poisson-Gamma 7
2.1.3 Data Generation Process 3: Additive Means 8
2.2 Specification of Competing Models 10
2.2.1 Linear Mixed Models 11
2.2.2 Generalized Linear Mixed Models 12
2.3 Model Comparison 13
2.3.1 Existing Performance Statistics 13
2.3.2 Additional Performance Statistics Considered 15
2.4 Data Application 18
2.5 Software implementation 19
Results 20
3.1 Simulation Study: Poisson-Unit DGP 20
3.2 Simulation Study: Poisson-Gamma DGP 24
3.3 Simulation Study: Additive Means DGP 27
3.4 Data application 31
Discussion 33
Conclusion 43
References 44
Tables 50
Appendix A - Supplemental Tables 78

List of Figures

Figure 1: Histograms for illustration of simulated datasets with equal and unequal treatment means generated under the Poisson-Unit DGP (panel A and D), Poisson-Gamma DGP (panel B and E) and Additive Means DGP (panel C and F). Simulation scenarios with 50 blocks are shown. 10

Figure 2: Simple visualization of major axis regression 16

Figure 3: Fitted major axis regression (MAR) lines of predicted values $\widehat{\boldsymbol{y}}_{\boldsymbol{ijk}}$ over observed values $\boldsymbol{y}_{\boldsymbol{ijk}}$ under selected competing models, namely LMM-hm (-- - -⁎- - --), GLMM-PsU-Q (--●--) , and GLMM-NB-Q (square). Observed values $\boldsymbol{y}_{\boldsymbol{ijk}}$ were generated under a Poisson-Gamma DGP with 10 blocks and unequal treatment means. Identity line represented by -·-·-. 27

List of Tables

Table 1: Frequencies of model failures to converge, by simulation scenario 50
Table 2: Empirical Type I Error & Statistical Power, Poisson-Unit DGP 51
Table 3: Average, minimum and maximum values of the Pearson chi-sq/df statistic under simulation scenario of unequal treatment means with 10 or 50 blocks and a Poisson-Unit DGP 51
Table 4: Summary of estimates of expectations of treatments and 95% confidence interval coverage for competing models fitted to data generated under the Poisson-Unit DGP 52
Table 5: Average, minimum, and maximum values of Akaike's and Bayesian Fit Criteria under simulation scenarios of unequal treatment means with 10 or 50 blocks and a Poisson-Unit DGP 55
Table 6: Average, minimum, and maximum values for point estimates of MAR coefficients, both intercept and slope, in simulation scenarios of unequal treatment means with 10 or 50 blocks under a Poisson-Unit DGP 56
Table 7: Average, minimum, and maximum values for 97.5th Percentile Ratios from simulation scenarios with 10 or 50 blocks and unequal treatment means under a Poisson-Unit DGP .. 57
Table 8: Average, minimum, and maximum values for Skewness Ratios from simulation scenarios with 10 or 50 blocks and unequal treatment means under a Poisson-Unit DGP .. 58
Table 9: Average, minimum, and maximum values for Coefficient of Variation Ratios from simulation scenarios with 10 or 50 blocks and unequal treatment means from a Poisson-Unit DGP 59
Table 10: Empirical Type I Error & Statistical Power, Poisson-Gamma DGP 60
Table 11: Summary of estimates of expectations of treatments and 95% confidence interval coverage for competing models fitted to data generated under the Poisson-Gamma DGP .. 61
Table 12: Average, minimum, and maximum values of Akaike's and Bayesian Fit Criteria under simulation scenarios of unequal treatment means with 10 or 50 blocks and a Poisson-Gamma DGP 64

Table 13: Average, minimum, and maximum values for point estimates of MAR coefficients, both intercept and slope, in simulation scenarios of unequal treatment means with 10 or 50 blocks under a Poisson-Gamma DGP 64
Table 14: Empirical Type I Error & Statistical Power, Additive Means DGP 65
Table 15: Summary of estimates of expectations of treatments and 95% confidence interval coverage for competing models fitted to data generated under the Additive Means DGP ... 66
Table 16: Average, minimum and maximum values of the Pearson chi-sq/df statistic under simulation scenario of unequal treatment means with 10 or 50 blocks and an Additive Means DGP 69
Table 17: Average, minimum, and maximum values of Akaike's and Bayesian Fit Criteria under simulation scenarios with 10 or 50 blocks and an Additive Means DGP 70
Table 18: Average, minimum, and maximum values for point estimates of MAR coefficients, both intercept and slope, in simulation scenarios of unequal treatment means with 10 or 50 blocks under an Additive Means DGP 71
Table 19: Average, minimum, and maximum values for 97.5th Percentile Ratios from simulation scenarios with 10 or 50 blocks and unequal treatment means under an Additive Means DGP 72
Table 20: Average, minimum, and maximum values for Skewness Ratios from simulation scenarios with 10 or 50 blocks and unequal treatment means under an Additive Means DGP 73
Table 21: Average, minimum, and maximum values for Coefficient of Variation Ratios from simulation scenarios with 10 or 50 blocks and unequal treatment means from an Additive Means DGP 73
Table 22: Pearson chi-sq/df statistic, AIC, and BIC, proof-of-concept data application 74
Table 23: Estimates and 95% confidence interval bounds for MAR coefficients from the proof-of-concept data application 74
Table 24: 97.5th Percentile, Skewness, and Coefficient of Variation Ratios from the proof-of-concept data application 75
Table 25: F-test results for treatment effects of competing models from the proof-of-concept data application at α=0.05 75

Table 26: Treatment mean estimates of LMM-hm, LMM-ht, LMM-l, GLMM-PsU, GLMM-NB, and GLMM-Gm from the proof-of-concept data application 76

Acknowledgements

With sincerest gratitude, I recognize Dr. Nora M. Bello for her mentorship, patience, encouragement, correction, and insight as both my advisor and major professor throughout my time as a master's student at Kansas State University. My experience at K-State has been shaped by my interactions with her, as has my development into a Master of Statistics. She has pushed me to refine my craft in every way and reach my greatest potential as a statistician. This project could not have been a success without her tutelage.

I also acknowledge the vital insights and contributions offered by my Graduate Advisory Committee, comprised of Dr. Trevor Hefley and Dr. Matthew Kramer. Dr. Walt Stroup, though not a formal member of the committee, graciously offered his expertise to the project as well. His invaluable input was essential to the progression of this project.

I would also like to recognize Dr. Constantine Daskalakis who supervised my internship work with Thomas Jefferson University in the summer of 2020. While attempting to characterize the progression of COVID-19 in Washington, USA, we noticed discrepancies between which time-series models were recommended by fit criteria and which time-series models diagnostics dictated ought to be used. Work on the internship project during a summer otherwise punctuated by a global pandemic contributed to the inspiration of this Master's research.

This research and project would not have been possible without the willingness of Kansas State University Department of Statistics and Graduate School to not only offer admission to the Master's program, but to also provide funding in the form of a Graduate Teaching Assistantship and in the Dr. Lynn Ying-Shiang Lin Statistics Graduate Scholarship. I also thank Dr. Lynn Ying-Shiang Lin directly for the scholarship which made continuation on this research possible in periods where this Assistantship was not available. I'm grateful to the University for their

willingness to offer me a place in the K-State family and I'm grateful to my students these last four semesters for their eagerness to engage with challenging material and their willingness to endure sometimes cumbersome technological challenges prompted by the pandemic.
I could not have been in a position to engage in Master's-level research without the support of the fine professors at Brigham Young University who oversaw my undergraduate education, and I offer special thanks to the staff and faculty in the Brigham Young University Department of Statistics.

On the precious, personal level, I offer my undying love and appreciation to my family. Although statistics, goodness-of-fit, and generalized linear mixed models are effectively a foreign language to my parents, my siblings, and my sisters-in-law, their love, support, encouragement, and confidence in me and my capabilities have carried me through the ordinary and pandemic-prompted stresses of life. Additionally, I recognize the emotional and spiritual support of the members of the Manhattan Young Single Adult Branch of The Church of Jesus Christ of Latter-day Saints. And, finally, I thank God for His unfailing support and unending blessings in helping me get to this point in my life.

Introduction

Generalized Linear Mixed Models (GLMMs) are extensions of linear models that integrate discrete or otherwise non-normal distributional assumptions on the response of interest, with a mixed models framework that accommodates correlation patterns in the data (Wolfinger and O'Connell 1993). Since their advent, GLMMs have expanded the versatility of statistical methods to deal with non-normal responses common across scientific disciplines, meanwhile correcting some of the problems resulting from an over-reliance on Fisher's linear-models-based Analysis of Variance (ANOVA) (Wolfinger & O'Connell, 1993). The limitations of linear models are particularly apparent when its normal distributional assumptions do not align with the behavior of discrete data. In particular, count data tend to be skewed to the right, especially for low counts, thus violating normality assumptions essential to linear mixed models (Stroup, 2015). Additionally, for count data, it often the case that the variance is some function of the mean, thereby violating the ANOVA assumption of constant variance (Stroup, 2015). Variance-stabilizing transformations (e.g., log function or square-root-plus-3/8) have historically been used to alleviate this issue (Miller, 1997; Stroup, 2013, 2015). Admittedly, the robustness of linear models and ANOVA for non-normal datasets when sufficiently large sample sizes are available is well documented (Larrabee et al., 2014; Miller, 1997). However, linear models applied to non-normal data can result in non-sensical mean estimates, e.g. negative estimates for means or confidence interval boundaries (Larrabee et al., 2014; Stroup, 2015). Additional issues include losses in statistical power and biased estimates of treatment means, even if sensical (Larrabee et al., 2014; Stroup, 2015), especially when a sufficiently large sample size is not possible due to budgetary or practical constraints.

As a general methodological framework, GLMM are applicable to a wide variety of data types under a range of distributional assumptions, not necessarily the normal distribution, meanwhile retaining estimation accuracy and Type I error control without compromising power (Stroup, 2015). Recent computational and algorithmic developments (Pinheiro & Chao, 2006; Wolfinger & O'Connell, 1993) have facilitated implementation of GLMM and their incorporation into mainstream statistical practice through multiple software developments (R Core Team, 2021; SAS Institute Inc, 2016). Yet, reliable and applicable diagnostic metrics to evaluate GLMM performance in terms of goodness of fit to data, identification of sources of lack of fit, and appropriate remedial measures, when needed, are lacking. For the statistical practitioner working with real data, this poses a very tangible problem of potentially serious inferential consequences (Stroup, 2015).

Overdispersion is among the most common issues encountered with GLMMs and occurs when there are more sources of variation in the data than are accounted for in a given model (Stroup, 2012). Particularly susceptible to overdispersion are models assuming distributions characterized by a single parameter, for instance the Poisson distribution, often used for count data (Lawless, 1987; Stroup, 2012). Indeed, the Poisson distribution imposes a strict structural assumption between mean and variance while count data frequently exhibit variances greater than the Poisson assumptions allow (Lawless, 1987). Another non-normal distribution also used for count data is the Negative Binomial (Lawless, 1987), which includes an additional scale parameter. Even in this case, the variance is often still constrained by a function of the mean (Lawless, 1987). These constraints are often violated in real data applications.

Fortunately, overdispersion is typically diagnosed in real data by way of the Pearson chi-sq/df statistic, an umbrella-type statistic that compares the dispersion described by a model

against the dispersion present in the data (Farrington, 1996; Payne et al., 2018; Stroup, 2012). Values of this statistic substantially greater than 1 are considered indicative of symptoms of overdispersion in the data. Yet, the statistic does not necessarily inform the practitioner about the cause of overdispersion and thus, what remedial measures to pursue. Possible contributing causes for overdispersion may include extremely large counts, zero inflation, a misspecified linear predictor or incorrect distributional assumptions (Stroup, 2012), amongst others, each of them calling for a different remedial approach.

Specification of distributional assumptions is of interest when multiple options are amenable to the nature of the data. For example, for count data, either Poisson or Negative Binomial distributions may be considered acceptable, as well as variations on the Poisson distribution; yet, these specifications differ substantially in the assumed relationship between means and variances (Stroup, 2012, 2015; Ver Hoef & Boveng, 2007). These differences are non-trivial, as a poorly chosen GLMM may perform worse inferentially than normal approximations based on linear mixed models (LMMs) (Stroup, 2015).

Recently proposed fit diagnostics for GLMM include extensions of the coefficient of determination (Piepho, 2019), though its scope seems limited to an umbrella-type assessment with applicability constrained to specific cases of (co)variance structures. Additionally, graphical representations of a frequentist take on posterior predictive checks, known as centipede plots, have been proposed to assess model fit to individual observations (Kramer, 2018).

In this study, we developed and evaluated statistics derived from model predicted values $\hat{y}$ as a way of assessing model fit to data y. Our underlying rationale is that predicted values $\hat{y}$, and functions thereof, calculated from competing GLMM specifications ought to be able to guide selection of the models better aligned with the underlying data generation process. Thus, models

yielding predicted values, and functions thereof, that match closely with those from observed data may be considered good fitting. In contrast, areas of discrepancies between predicted and observed values might indicate specific instances where a model performs poorly and thus motivate remedial measures.

In the context of a simulation study, we implemented major axis regression (MAR) to evaluate discrepancy between functions of model predictions and correspondingly of simulated data (Mesplé et al., 1996). Estimates of the MAR intercept close to 0 and MAR slope close to 1 indicate alignment between the fitted model and the observed data (Mesplé et al., 1996). On the other hand, MAR intercept estimates substantially greater (less) than 0 indicates over(under) prediction of low data counts (Mesplé et al., 1996). An analogous interpretation can be made of the MAR slope estimates relative to high data counts.

Also of interest was the evaluation of the ability of a competing model to accommodate and regenerate specific features of the data, such as extreme values, (a)symmetry and relative variation. These features were assessed with statistics defined as the ratio of predicted values over observed values for the 97.5^{th} percentile, the skewness coefficient, and the CV, respectively. These ratios were expected to yield values close to 1 if the specific data feature was properly reproduced by a model.

We acknowledge that predicted values $\hat{y}$, as computed and utilized in this study, are of little predictive use beyond in-sample prediction (Burman, 1989; Hardle & Marron, J.S., 1985). Indeed, thus computed, predicted values do little to inform out-of-sample predictions in mixed models, as the expectations of random effects is assumed to be 0 (Hardle & Marron, J.S., 1985; Stroup, 2012). Here, we do not intend the proposed approach for forecasting or prediction of

future observations. Instead, we consider assessments of model fit to specific data features, such as extreme values, (a)symmetry and relative variation.

The objective of this study was to develop and explore potential diagnostic metrics to assess GLMM fit and performance to ultimately inform modeling choice, specifically for discrete count responses. This objective was addressed by way of a simulation study with multiple scenarios, followed by a proof-of-concept data application in the entomological sciences. These will be defined in more detail in coming sections. The simulation study utilized three realistic data generation processes (DGPs), namely Poisson-Unit, Poisson-Gamma, and Additive Means processes, to simulate data under a 2x2 factorial treatment structure in a randomized block design, specified with either 10 or 50 blocks. These sample sizes are consistent with studies conducted in the animal sciences (Bello & Renter, 2018; Gonçalves et al., 2018). Competing models fitted to count data included alternative specifications of traditional LMMs assuming normality with a constant or non-constant variance, as well as variance-stabilizing transformations. Competing GLMMs assumed Poisson or Negative Binomial distributions of count data, both recognizing its positive discrete nature. Also, a GLMM assuming a Gamma distribution was fitted as an alternative continuous approximation to count data limited to the positive line. For two of the DGP evaluated in the simulation study, one of the competing models was the true DGP model, thus enabling an assessment of inferential robustness of the competing models.

Methodology

2.1 Simulation Scenarios

Data were simulated under a treatment structure consisting of a 2x2 factorial arrangement given by the combination of levels of treatment A and treatment B within a completely randomized block design. Data were generated as discrete counts under three distinct data generation processes (DGP) adapted from Stroup (2013) and designed to represent plausible realistic mechanisms, namely (i) Poisson-Unit, (ii) Poisson-Gamma, and (iii) Additive Means, to be described below.

For each DGP, four simulation scenarios were designed, consisting of the combination of number of blocks (i.e., 10 vs. 50) and equal or unequal treatment means. Simulation scenarios of equal or unequal treatment means were intended to characterize Type I Error and statistical power (i.e., 1 – Type II error). The choice of number of blocks, namely 10 and 50, was intended to mirror data collection conditions commonly aligned with experimental studies and observational studies respectively.

For each simulation scenario under a DGP, 100 simulated datasets were generated, as follows:

2.1.1 Data Generation Process 1: Poisson-Unit

1. Sample differential block effects $b_k \sim \text{NIID}(\mu = 0, \sigma_b^2 = 0.75^2)$, where either $k \in [1,2,\dots,10]$ or $k \in [1,2,\dots,50]$, depending on the simulation scenario.
2. Define an intercept $\eta = 1$ and define differential treatment effects τ_{ij} where $i = \{1,2\}$ corresponds to levels of treatment factor A and $j = \{1,2\}$ corresponds to levels of treatment factor B, such that:

a. $\tau_{ij} = 1.5$ for all ij^{th} combinations under the simulation scenario of equal treatment means, or

b. $\tau_{11} = \tau_{12} = 1.5,\ \tau_{21} = 3, \text{and } \tau_{22} = 0$ under the simulation scenario of unequal treatment means.

3. Sample differential unit-level effects identified by the ij^{th} treatment combination within the k^{th} block, such that $u_{ijk} \sim \text{NIID}(\mu = 0, \sigma_u^2 = 0.75^2)$

4. Define the linear predictor using a canonical log link function such that $\eta_{ijk} = \log(\lambda_{ijk}) = \eta + \tau_{ij} + b_k + u_{ijk}$, where λ_{ijk} is the rate parameter of a Poisson distribution.

5. Generate a random sample $y_{ijk}|b_k, u_{ijk} \sim \text{Poisson}(\lambda_{ijk})$

In this setting, the expected value of the conditional treatment means $\lambda_{11}, \lambda_{12}, \lambda_{21}, \text{and } \lambda_{22}$ in the scale of the observed data are equal to 12.2, 12.2, 54.6, and 2.7, respectively.

2.1.2 Data Generation Process 2: Poisson-Gamma

1. Steps 1 and 2 are as described under the Poisson-Unit DGP explained in section 2.1.1.

2. Sample differential unit-level effects $u_{ijk} \sim \text{Gamma}\left(\alpha^* = \frac{1}{\phi_{DGP}}, \beta^* = \phi_{DGP}\right)$, where α^* is a shape parameter and β^* is a scale parameter. Also, $\phi_{DGP} = 0.5$ such that $E(u_{ijk}) = 1$.

3. Define the linear predictor using a canonical log link function such that $\log(\lambda_{ijk}) = \eta_{ijk} = \eta + \tau_{ij} + b_k$, where λ_{ijk} is the rate parameter for a Poisson distribution.

4. Generate a random sample $y_{ijk}|b_k, u_{ijk} \sim \text{Poisson}(\lambda_{ijk} \cdot u_{ijk})$.

In this DGP setting, the expected value of the conditional treatment means in the scale of the observed data were equal to those in the Poisson-Unit DGP described in section 2.1.1, where $\lambda_{11} = \lambda_{12} = 12.2,\ \lambda_{21} = 54.6,$ and $\lambda_{22} = 2.7.$

In this DGP, the observed values y_{ijk} , conditional on block and unit-level effects, were generated from a Poisson distribution with expectation given by the rate parameter multiplied by a gamma-distributed unit-level effect, as described in steps 2 and 4, thus the Poisson-Gamma label for this DGP. In fact, thus generated, $y_{ijk}|b_k$ can be shown to follow a negative binomial distribution with mean λ_{ijk} and scale parameter ϕ_{DGP}, whereby $y_{ijk}|b_k \sim NB(\lambda_{ijk}, \phi_{DGP})$ (Stroup, 2013).

2.1.3 Data Generation Process 3: Additive Means

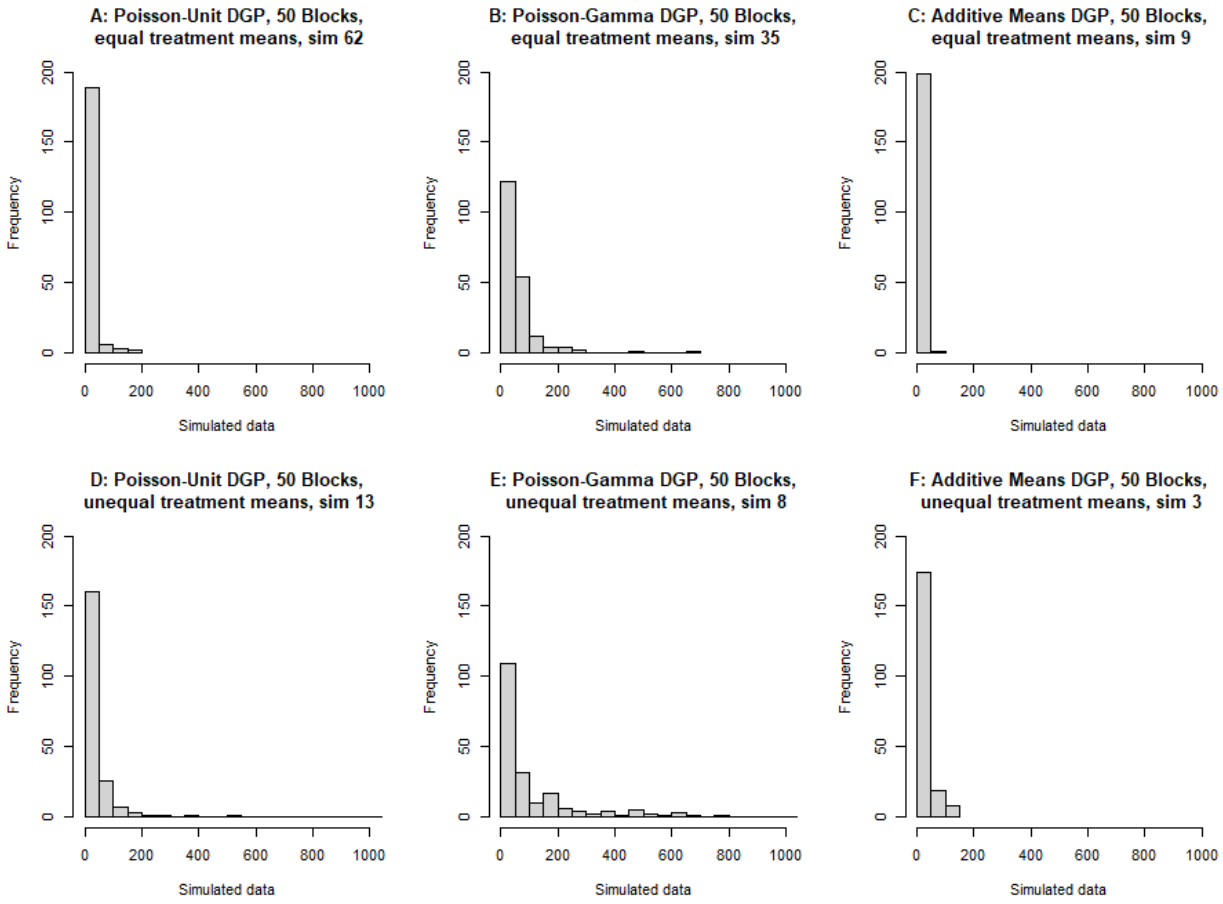
1. Define an intercept term η equal to 1.
2. To define combined differential treatment and block effects,
    a. For simulation scenarios of equal treatment means, sample combined differential treatment effect and block effects as $(\tau_{ij} + b_k)$~Gamma$(\alpha^* = 11, \beta^* = 1)$, or,
    b. For simulation scenarios of unequal treatment means, sample the combined differential treatment effect and block effects as $(\tau_{ij} + b_k)$~Gamma$(\alpha^*_{ij}, \beta^* = 1)$, where $\alpha^*_{11} = \alpha^*_{12} = 11,\ \alpha^*_{21} = 54,$ and $\alpha^*_{22} = 2.$
3. Sample unit level effects u_{ijk}~Gamma $\left(\alpha^* = \frac{1}{\phi_{DGP}}, \beta^* = \phi_{DGP}\right)$, where $\phi_{DGP} = 0.5.$
4. Define the linear predictor using an identity link function, namely $\eta_{ijk} = \lambda_{ijk} = \eta + (\tau_{ij} + b_k)$.

5. Generate a random sample $y_{ijk}|b_k, u_{ijk} \sim \text{Poisson}(\lambda_{ijk} \cdot u_{ijk})$.

In this setting, the expectations of the conditional treatment means were $\lambda_{11} = \lambda_{12} = 12$, $\lambda_{21} = 55, \text{and } \lambda_{22} = 3$ and thus, of similar order of magnitude relative to those used for the Poisson-Unit DGP and Poisson-Gamma DGP. The key feature of this Additive Means DGP was that the Poisson rate parameter λ_{ijk} was a direct function of the fixed and random effects in the linear predictor (i.e., the "link" function was just the identity function).

For illustration purposes, Figure 1 depicts six simulated datasets corresponding to the 50-block scenario and corresponding to either equal or unequal treatment means under each of the three DGPs described above.

Figure 1: Histograms for illustration of simulated datasets with equal and unequal treatment means generated under the Poisson-Unit DGP (panel A and D), Poisson-Gamma DGP (panel B and E) and Additive Means DGP (panel C and F). Simulation scenarios with 50 blocks are shown.



2.2 Specification of Competing Models

Each simulated dataset was fitted with competing model specifications labeled 1 to 16 and explained next. In general terms, competing models 1 through 4 represented alternative specifications of LMMs that used a normal distribution to approximate the behavior of the response variable, expressed either in the original scale (models 1 and 2) or following a variance-stabilizing transformation (models 3 and 4). Models 5 to 16 consisted of GLMM specifications

that explicitly recognized the non-normal nature of the response variable. Further, consistent with GLMM methods of estimation, models 5 to 16 were fitted using either likelihood approximations (i.e. Laplace approximation and Adaptive Quadrature, Pinheiro & Chao, 2006) or linearization (i.e. Pseudolikelihood, Wolfinger & O'Connell, 1993). Each of the competing Models 5 to 16 are thus identified by the suffix "-L", "-Q" or "-P", respectively, to indicate method of estimation.

A detailed description of the competing models considered in this study follows.

2.2.1 Linear Mixed Models

LMMs were used as normal approximations to the count response, as is common practice in many disciplines (Stroup, 2015). The linear predictor included fixed effects for the intercept η, the differential effects α_i and β_j for treatment A and B respectively, and their differential combined effect $\alpha\beta_{ij}$, as well as a random differential block effect b_k assumed $NIID(0, \sigma_b^2)$. Specifically, for the LMMs, μ_{ijk} was defined as the expectation for the ij^{th} treatment combination conditional on the k^{th} block such that $\mu_{ijk} = \eta_{ijk} = \eta + \alpha_i + \beta_j + \alpha\beta_{ij} + b_k$.

Model 1 was defined as a homoskedastic LMM (LMM-hm) whereby $y_{ijk}|b_k \sim NIID(\mu_{ijk},\ \sigma_e^2)$. Model 2 extended Model 1 to accommodate heteroskedasticity (LMM-ht), whereby $y_{ijk}|b_k \sim NIID(\mu_{ijk}, \sigma_{e_{ij}}^2)$, such that the residual variance was specific to the ij^{th} treatment combination.

Models 3 and 4 involved variance-stabilizing transformations of y_{ijk} and thus consisted of LMMs analogous to Model 1 (LMM-hm) fitted to $y_{ijk}^* = \log(y_{ijk})$ or $\tilde{y}_{ijk} = \sqrt{y_{ijk} + 3/8}$, such that $y_{ijk}^*|b_k \sim NIID(\mu_{ijk}^*, \sigma_e^2)$ or $\tilde{y}_{ijk}|b_k \sim NIID(\tilde{\mu}_{ijk}, \sigma_e^2)$, respectively.

2.2.2 Generalized Linear Mixed Models

Competing Models 5, 6 and 7, labeled GLMM-Ps-L, GLMM-Ps-Q and GLMM-Ps-P, respectively, assumed a conditional Poisson distribution of the response variable such that $y_{ijk}|b_k \sim Poisson(\lambda_{ijk})$, $\log(\lambda_{ijk}) = \eta_{ijk} = \eta + \alpha_i + \beta_j + \alpha\beta_{ij} + b_k$. Recall that suffixes -L, -Q, and -P indicate methods of GLMM estimation, as previously indicated.

Competing models 8, 9 and 10, labeled GLMM-PsU-L, GLMM-PsU -Q and GLMM-PsU-P, respectively, also assumed a conditional Poisson distribution of the response variable. In this case, the linear predictor further included random unit-level effects $u_{ijk} \sim NIID(0, \sigma_u^2)$, such that $y_{ijk}|b_k, u_{ijk} \sim Poisson(\lambda_{ijk})$, $\log(\lambda_{ijk}) = \eta_{ijk} = \eta + \alpha_i + \beta_j + \alpha\beta_{ij} + b_k + u_{ijk}$ (Stroup, 2013).

Competing models 11, 12 and 13, labeled GLMM-NB-L, GLMM-NB-Q, and GLMM-NB-P, respectively, assumed a conditional Negative Binomial distribution of the response variable such that $y_{ijk}|b_k \sim NB(\lambda_{ijk}, \phi_{NB})$, whereby $\log(\lambda_{ijk}) = \eta_{ijk} = \eta + \alpha_i + \beta_j + \alpha\beta_{ij} + b_k$ and ϕ_{NB} is the Negative Binomial scale parameter that accommodates variation at the level of observation.

Finally, competing models 14, 15, and 16, labeled GLMM-Gm-L, GLMM-Gm-Q, and GLMM-Gm-P, respectively, assumed a Gamma distribution to approximate the behavior of the discrete count response variable, whereby $y_{ijk}|b_k \sim Gamma(\alpha^*_{ijk}, \beta^*_{ijk})$ with shape parameter α^* and scale parameter β^*, such that $E(y_{ijk}|b_k) = \mu_{ijk} = \alpha^*_{ijk}\beta^*_{ijk}$ and $\log(\mu_{ijk}) = \eta_{ijk} = \eta + \alpha_i + \beta_j + \alpha\beta_{ij} + b_k$.

For models 5 to 16, the differential random effects, b_k and u_{ijk}, fitted by the GLMMs were assumed $b_k \sim NIID(0, \sigma_b^2)$ and $u_{ijk} \sim NIID(0, \sigma_u^2)$, where applicable.

Competing models 14 to 16, namely GLMM-Gm, were considered as an alternative to the normal approximation of Models 1 to 4. Our rationale for exploring a Gamma approximation instead was based on the inherent skewness and positive-line support of the Gamma distribution, which may better align with the asymmetric and zero-bounded nature of count data.

To note: the specification of competing models 8 through 10, namely, GLMM-PsU was directly aligned with the Poisson-Unit DGP used for data simulation. Similarly, the specification of models 11,12, and 13, namely GLMM-NB were directly aligned with the Poisson-Gamma DGP, respectively. Thus, competing models GLMM-PsU and GLMM-NB were, by design, the true data generation models in their respective case. In turn, none of the competing models were aligned with the Additive Means DGP used for data simulation. As a result, data generated under the Additive Means DGP did not have a true model amongst the specifications evaluated.

2.3 Model Comparison

2.3.1 Existing Performance Statistics

Inferential performance of competing models was assessed using Type I Error and statistical power of Type III F-tests to assess treatment differences where H_0: $\mu_{ij} = \mu_{i'j} = \mu_{ij'} = \mu_{i'j'}$ and H_A: at least one treatment mean is different. For each DGP under equal treatment means, Type I Error was computed as the proportion of simulated datasets for which the null hypothesis was incorrectly rejected. In turn, statistical power was computed for scenarios of unequal treatment means under each DGP as the proportion of simulated datasets that correctly rejected the null hypothesis of equal treatment means.

Fit statistics Akaike's Information Criterion (AIC, Akaike, 1973) and Bayesian Information Criterion (BIC, Schwarz, 1978) were utilized to assess the fit of the competing

models. Although neither are proper scoring rules, AIC is asymptotically equivalent to the mean logarithmic score, which is a proper scoring rule which serves to summarize discrepancies between predictions and observations (Correndo et al., 2021; Czado et al., 2009; Gneiting, 2011). These information criteria were computed for the subset of competing models fitted using a true likelihood-based method of estimation, as follows:

$$AIC = -2\log\left(L\left(\widehat{\boldsymbol{\theta}}\middle|\boldsymbol{y}\right)\right) + 2d, \text{ and}$$

$$BIC = -2\log\left(L\left(\widehat{\boldsymbol{\theta}}\middle|\boldsymbol{y}\right)\right) + d\log(n),$$

whereby $L\left(\widehat{\boldsymbol{\theta}}\middle|\boldsymbol{y}\right)$ is the likelihood function of the data, $\widehat{\boldsymbol{\theta}}$ is a vector of parameter estimates, $\boldsymbol{y}$ is a vector of observed response variables, *d* is the dimension of the model equivalent to the number of effective covariance parameters, and *n* is the size of the data equivalent to the number of effective subjects. Fit criteria calculated on this subset of models are directly comparable across LMMs (all fitted using true data likelihoods) and those GLMMs fitted based on integral approximations to the corresponding likelihood (Stroup, 2012). In contrast, for GLMMs estimated through linearization (i.e. pseudolikelihood, Schabenberger, 2005; Wolfinger & O'Connell, 1993), specifically competing models 7, 10, 13, and 16, AIC and BIC were considered non-interpretable (Schabenberger, 2005) and thus were not calculated in this study. For a given dataset, smaller values of AIC and BIC are considered as indications of better fitting models (Hastie et al., 2009).

Overdispersion was assessed using the Pearson Chi-square statistic over degrees of freedom (McCullagh & Nelder, 1989a). In general terms, computation of this statistic involves squaring the difference between observations and group means, then dividing by the variance weighted by the overall degrees of freedom (McCullagh & Nelder, 1989). In the context of GLMM, the Pearson Chi-square/DF overdispersion statistic is computed for conditional

distributions of the data (given random effects) using estimates approximated by Laplace approximation or Adaptive Quadrature (Stroup, 2012). For GLMMs estimated through linearization (i.e., pseudolikelihood, Stroup 2012), namely models 7, 10, 13, and 16, only the generalized Chi-square/DF statistic can be computed; this statistic is not considered appropriate for assessing overdispersion (Stroup 2012) and is thus not reported in this study. Values of the Pearson Chi-sq/DF statistic substantially greater than 1 are considered indicative of overdispersion (Stroup, 2012).

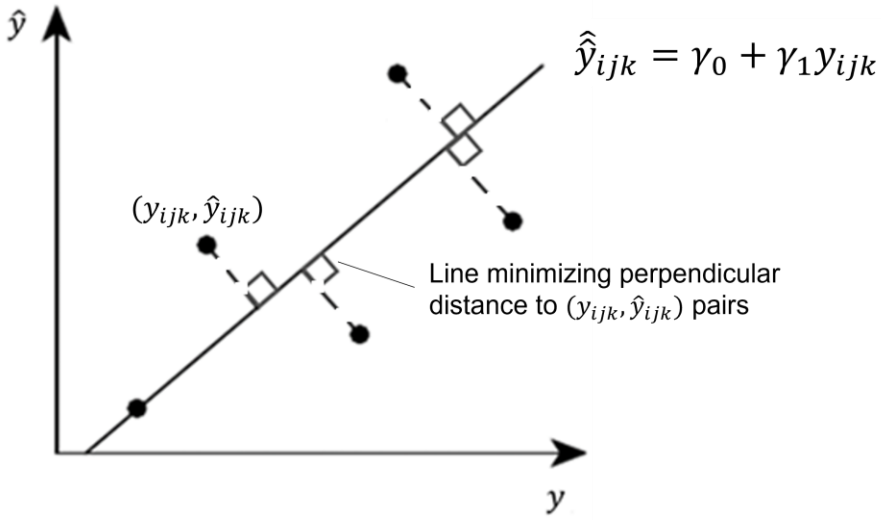
2.3.2 Additional Performance Statistics Considered

To further assess model fit and performance, we developed and evaluated an array of statistics intended to characterize fit to selected data features based on discrepancies between observed values y_{ijk} and predicted values $\hat{y}_{ijk}$ obtained from fitting competing models. Specifically, the predicted value $\hat{y}_{ijk}$ corresponding to the ijk^{th} observation was obtained from point estimates of model parameters included in the linear predictor of the corresponding competing model, such that $\hat{y}_{ijk} = g^{-1}(\hat{\eta}_{ijk})$, whereby $g(.)$ indicates the link function of the corresponding model. For competing models 1 through 4, labeled LMM-hm, LMM-ht, LMM-l, and LMM-s, respectively, $g(.)$ indicates the identity function, such that $\hat{y}_{ijk} = \hat{\eta}_{ijk}$. For all GLMMs indicated by models 5 through 16, $g(.) = \log(.)$.

First, for each competing model fitted to a simulated dataset, predicted values $\hat{y}_{ijk}$ were regressed over observed values y_{ijk}, namely $\hat{y}_{ijk} = \gamma_0 + \gamma_1 * y_{ijk} + e_{ijk}$ using MAR, also known as Model II regression analyses (Mesplé et al., 1996). The MAR approach minimizes the sum of squares of the perpendicular distances between each point (y_{ijk}, $\hat{y}_{ijk}$) and the fitted MAR line $\hat{\hat{y}}_{ijk}$ (Correndo et al., 2021; Legendre & Legendre, 2012; Mesplé et al., 1996) (Figure 2).

Specifically, the criterion minimized is $D(\boldsymbol{y}, \hat{\boldsymbol{y}})^2 = \sum_{ijk} \sqrt{\left(y_{\hat{y}\,ijk} - y_{ijk}\right)^2 + \left(\hat{\hat{y}}_{ijk} - \hat{y}_{ijk}\right)^2}$, where $\boldsymbol{y} = \{\boldsymbol{y_{ijk}}\}$ and $\hat{\boldsymbol{y}} = \{\hat{y}_{ijk}\}$ represent the vector of *ijk*th observed values and corresponding model-predicted values, respectively. Similarly, $\boldsymbol{y}_{\hat{\boldsymbol{y}}} = \{y_{\hat{y}ijk}\}$ and $\hat{\hat{y}}_{ijk}$ denote the vectors of *ijk*th coordinate pairs on the MAR fitted line (Legendre & Legendre, 2012). Recall that MAR differs from ordinary least squares regression in that the latter minimizes the sum of the squared deviations along the vertical axis (i.e., residuals). MAR is usually recommended when comparing model predictions to observations in a simulation study (Legendre & Legendre, 2012; Mesplé et al., 1996).

Figure 2: Simple visualization of major axis regression



In this study, we estimated and summarized MAR coefficients, both intercept γ_0 and slope γ_1, for each competing model fitted to a simulation dataset within a DGP. Point estimates of 0 and 1 for the MAR intercept γ_0 and MAR slope γ_1, respectively, suggest good data fit of the competing model used to generate predicted values (Mesplé et al., 1996). On the other hand, a

MAR intercept substantially greater than 0 indicates that a model is over estimating low counts (Mesplé et al., 1996). Equivalently, MAR slopes smaller than 1 indicate underestimation of high counts. Note, these coefficients are dependent on each other and are strongly negatively correlated at approximately -0.73. Moreover, plotting observed and predicted values alongside a MAR line can also provide a visual aide for identifying specific areas in which a model struggles to describe features of the data.

Next, we characterized fit to selected data features, specifically to dispersion and asymmetry, as follows. For each competing model, we computed:

i) The ratio of the 97.5th-percentile of $\hat{\boldsymbol{y}}$ over the 97.5th-percentile of $\boldsymbol{y}$, defined as $P_{97.5}^{(\hat{y})}/\ P_{97.5}^{(y)}$, where $P_{97.5}^{(.)}$ is the value below which 97.5% of the frequency distribution of $\hat{\boldsymbol{y}}$ (or $\boldsymbol{y}$, as appropriate) lies;

ii) The ratio of the skewness coefficient of $\hat{\boldsymbol{y}}$ over the skewness coefficient of $\boldsymbol{y}$, defined as $m_3^{(\hat{y})}/m_3^{(y)}$, where $m_3^{(y)} = \frac{\sum_{i=1}^{n}(y_{ijk}-\bar{y})^3}{(n-1)s_y^3}$ with sample size n, sample mean $\bar{y}$, and sample standard deviation s_y, and similarly for $m_3^{(\hat{y})}$;

iii) The ratio of coefficients of variation (CV) of $\hat{\boldsymbol{y}}$ over the coefficients of variation of $\boldsymbol{y}$, defined as $CV^{(\hat{y})}/CV^{(y)}$ where $CV^{(y)} = s_y/\bar{y}$ and similarly for $CV^{\hat{y}}$.

Values of the proposed ratios i), ii) or iii) closer to 1 are indicative of better model fit to the corresponding data feature. For as $P_{97.5}^{(\hat{y})}/\ P_{97.5}^{(y)}$ and $CV^{(\hat{y})}/CV^{(y)}$, values less than 1 may be indicative of a competing model suffering from overdispersion.

2.4 Data Application

For a proof-of-concept data application, we used entomological data presented by Li, Cloyd, and Bello (2019). In that study, the objective was to determine the effectiveness of using the entomopathogenic fungus *Beauveria bassiana* in conjunction with the rove beetle *Dalotia coriaria* to suppress western flower thrips *Frankliniella occidentalis* populations under greenhouse conditions. Here, we worked with a data subset collected during Summer 2016 and consisting of observations on thrip counts collected on yellow chrysanthemum plants *Dendranthema x grandiflorum* (Ramat.) Kitam. at four weeks after treatment application. Further experimental details can be accessed at Li, Cloyd, and Bello (2019). Briefly, a total of 35 individually potted plants, each contained inside an individual plastic cage, were arranged in 6 blocks of 5 plants each and two additional blocks of 2 and 3 plants each, respectively, as deemed necessary by logistical constraints in the greenhouse. Blocks were defined by location within the greenhouse to control for gradients in light, temperature, and humidity. Within each block, plants were randomly assigned to one of five treatments consisting of (1) spinosad, pyridalyl, chlorfenapyr, and abamectin insecticides; (2) entomopathogenic fungus *B. bassiana*; (3) rove beetle *D. coriaria*; (4) a combination of *B. bassiana* and *D. coriaria*; and (5) a water control. Twenty thrips were applied to each of the plants and allowed to establish populations two weeks prior to the application of treatments. Each week after treatment application, a yellow sticky card was placed adjacent to each plant and the number of western flower thrips captured on the sticky card was recorded. Observations on thrip counts collected four weeks after treatment application are considered in this data application. Competing models 1 through 16 were fitted to the entomological dataset and performance statistics were computed to determine which of the competing models best characterized the data.

2.5 Software implementation

Simulated data were generated and results were summarized and tabulated using R software (R Core Team, 2021, version 4.1.0). All competing models were fitted using the GLIMMIX procedure of SAS software (Version 9.4, SAS Institute Inc, 2016). For fitting competing models 11 through 16, namely GLMM-NB and GLMM-Gm, the maximum number of iterations of the nonlinear optimization process was increased to 10,000 iterations (i.e., MAXITER option in the NLOPTIONS statement of proc GLIMMIX) to facilitate model convergence.

For competing model LMM-ht, the PARMS statement in proc GLIMMIX was used to provide starting values for estimation of variance components due to failure of the default Quasi-Newton (QUANEW) optimization algorithm to yield starting values amenable with a valid objective function.

Results

Table 1 shows the number of convergence failures in fitting competing models to the 100 datasets generated under each DGP and simulation scenario; only models for which convergence failure was detected are listed. Overall, model convergence rates were acceptable, except for competing model 2, namely LMM-ht, which showed the greatest frequency of failure to converge. This convergence misbehavior was particularly problematic under the Poisson-Unit DGP, despite technical fine-tuning of the estimation algorithm to inform starting values for variance components, as explained in section 2.5. Results reported here are based on models that properly converged to simulated data.

3.1 Simulation Study: Poisson-Unit DGP

Table 2 presents empirical Type I error and Statistical power for simulation scenarios with equal and unequal treatment means, respectively, following a Poisson-Unit DGP with 10 or 50 blocks. In all cases, linear mixed modeling approaches for normal approximations to the data, either expressed as discrete counts (model 1, LMM-hm) or their transformations (models 3 and 4, LMM-l and 4-LMM-s) seemed to control Type I Error at or close to the nominal value. Of note is the relative stringency of model 2, namely LMM-ht, which yielded the lowest observed empirical Type I error in both the 10- and 50-block simulation scenarios. Most GLMMs (models 8 to 16) showed some level of Type I Error inflation in the 10-block scenario, though false positives were better controlled with larger datasets consisting of 50 blocks. This was particularly apparent for competing models 8 through 10, namely GLMM-PsU, which was the true model for this DGP. Meanwhile, competing models 5 through 7, namely GLMM-Ps yielded large Type I error inflation, above 0.60 in all cases, regardless of the number of blocks considered. This inflated probability of false positives was to be expected given strong evidence

for overdispersion of GLMM-Ps, as apparent from Pearson chi-sq /df statistics ranging from 2.5 to 43.4 for this model (Table 3). In contrast, the remaining GLMMs (models 8 through 16) showed Pearson chi-sq /df statistics close to, or below, 1 (Table 3). This suggests that, for models 8 through 16, any additional variability in the data had been accounted for, either by the inclusion of a unit-level effect in the linear predictor (i.e., GLMM-PsU), or by the incorporation of a scale parameter (i.e., GLMM-NB or GLMM-Gm). Notably, the true model aligned with the DGP, namely GLMM-PsU, yielded Pearson chi-sq/df statistics between 0.1 and 0.4, well below the established threshold of 1.

Regarding statistical power, most competing models yielded values close to or above 0.7, regardless of sample size (Table 2). The exception was model 2, namely LMM-ht which showed only 0.22 probability of correctly detecting treatment differences when data from only 10 blocks was available. For some of the competing models, specifically models 1, 2, 4, and 16, power increased further with more blocks, though log transformations seemed to lag behind in power gain relative to other models. Empirical power for competing models 5 through 7, namely GLMM-Ps, is not presented due to the reported failure to control Type I Error.

Table 4 contains a summary of estimates of expectations of treatments and 95% confidence interval coverage of these expectations for competing models fitted to data generated under the Poisson-Unit DGP. Overall, competing models seemed to overestimate treatment means, particularly for larger treatment means (i.e., ij = 2,1) with small datasets (i.e., 10 blocks), as indicated by point estimates above the true value and confidence interval (CI) coverage below the nominal value. Normal approximations by models 1 and 2, namely LMM-hm and LMM-ht, overestimated treatment means by the largest margin and with the broadest CI width. For treatment means of smaller magnitude or when larger datasets were available (i.e., 50 blocks),

accuracy of estimated treatment expectations was improved, particularly for GLMM-PsU. Indeed, point estimates were closer to the truth and coverage was closer to nominal values without compromising CI width. Overestimation of treatment means by competing models 11 through 16, namely GLMM-NB and GLMM-Gm was substantial, as indicated by both point estimates and lack of CI coverage; larger sample sizes partially mitigated the former but worsened the latter problem. Regardless of sample size, the logarithmic transformation of model 3, namely LMM-l yielded better CI coverage rates for higher than for lower treatment means. Meanwhile, the reverse was true for the square-root-plus-3/8 transformation of model 4 whereby CI coverage often improved for higher treatment expectations compared to that of lower treatment expectations.

Table 5 shows AIC and BIC fit statistics for simulation scenarios of unequal treatment means with 10 or 50 blocks. In all cases, models 14-16, namely GLMM-Gm, yielded the lowest values of AIC and BIC, both on average and for most individual datasets in these simulation scenarios, thus indicating best fit of all competing models considered, even the true model GLMM-PsU. Specifically, GLMM-Gm returned the lowest AIC and BIC values in 66%, 94% and 100% of the scenarios of equal treatment means with 10 blocks, unequal treatment means with 10 blocks, and all remaining scenarios, respectively. Meanwhile, values of AIC and of BIC for GLMM-PsU or GLMM-NB were very close in magnitude, many times within the 2-point rule of thumb considered discriminatory of model fit (Burnham & Anderson, 2004). Overall, neither AIC nor BIC were able to reliably discriminate between the true data generation model and other alternative GLMM specifications in these simulations with 10 or 50 blocks.

Next, we consider the performance of the proposed diagnostic metrics based on predicted values. Ideally, if a model were able to perfectly replicate the observed data, the MAR intercept

(γ_0) and slope parameter (γ_1) would be expected, on average, to take values 0 and 1, respectively. Table 6 shows point estimates for MAR coefficients obtained from regressing predicted values obtained from competing models over observed values generated under a Poisson-Unit DGP. Overall, when competing models were fitted to data, the point estimates of the corresponding MAR coefficients showed substantial variability, though in most cases the range of estimates overlapped the expected values (Table 6). Of particular interest were models 8 to 10, namely GLMM-PsU, which the reader may recall was the true model under the Poisson-Unit DGP. Point estimates of MAR coefficients under the GLMM-PsU models were slightly biased upward for intercept (ranging from 0.04 to 0.08 across scenarios) and downward for slope (ranging from 0.96 to 1 across scenarios), though with the least uncertainty of the remaining competing models (Table 6). In fact, despite the slight bias, MAR coefficient estimates from the GLMM-PsU were closest (i.e., in terms of least absolute distance) to their respective target values compared to those of other competing models, both on average and for most individual datasets. Specifically, in 80% of the datasets of the 10-block unequal-treatment means scenario, and in at least 99% of all other simulation scenarios under this Poisson-Unit DGP, the GLMM-PsU yielded MAR estimates closest to the corresponding target values relative to any other model evaluated. The MAR coefficients appeared useful in identifying the Poisson-Unit GLMM as the best fitting model.

Table 7 shows descriptive statistics for the ratio of 97.5th percentiles on predicted values $\hat{y}_{ijk}$ vs. observed values y_{ijk} for the simulation scenarios of unequal treatment means under a Poisson-Unit DGP. Regardless of competing GLMM, values for this ratio overlapped with the target value 1. However, compared to all other competing models, the GLMM-PsU yielded estimated ratios closest to the target, both on average and for most simulated datasets;

specifically, for 93% of the datasets under the Poisson-Unit DGP scenarios. Meanwhile, models 1 through 4, namely LMMs, seemed to consistently underestimate this ratio, indicating a limited ability to reproduce extreme values in the data.

Table 8 contains the ratio of the skewness statistic computed from predicted values $\hat{y}_{ijk}$ vs. that from observed values y_{ijk}, under the unequal treatment means scenarios in the Poisson-Unit DGP. Here too, competing GLMMs returned skewness ratios with ranges that overlapped the target value 1. However, the estimated ratios obtained from the Poisson-Unit GLMMs were closest to the target, both on average and in at least 93% of the simulated datasets under the Poisson-Unit DGP. A similar pattern was also observed for the CV ratio computed on predicted values $\hat{y}_{ijk}$ over that on observed values y_{ijk}, with the GLMM-PsU yielding ratios closest to the target for at least 97% of the datasets (Table 9).

3.2 Simulation Study: Poisson-Gamma DGP

Recall, that in simulation scenarios under the Poisson-Gamma DGP, the true model is represented by models 11 through 13, namely GLMM-NB.

For both Type 1 error and statistical power, model comparisons were analogous to those reported for the Poisson-Unit DGP (Table 10) in that LMMs seemed to control Type I error. Models 1 through 4, namely GLMM-NB, showed slightly inflated Type I error that was not corrected with larger sample size. Meanwhile, models 8 through 10, namely GLMM-PsU, were able to control Type I error at the nominal level.

Table 11 contains a summary of estimates of treatment expectations and 95% confidence interval coverage of these expectations for competing models fitted to data generated under the Poisson-Gamma DGP. Overall, model 4, namely LMM-s and models 11 through 16, namely

GLMM-NB, GLMM-Gm, returned estimates of treatment expectations closest to the true values, whereas models 3, namely LMM-l and models 5 through 10, namely GLMM-Ps, GLMM-PsU, consistently underestimated treatment expectations. Remaining LMMs substantially overestimated the expectations of treatments and yielded the broadest CI width. For models 11 through 13, namely GLMM-NB, CI coverage was closest to target, particularly with larger sample sizes. For models 14 to 16, namely GLMM-Gm, CI coverage typically performed as well as that of GLMM-NB, except for small treatment expectations ($ij = 22$).

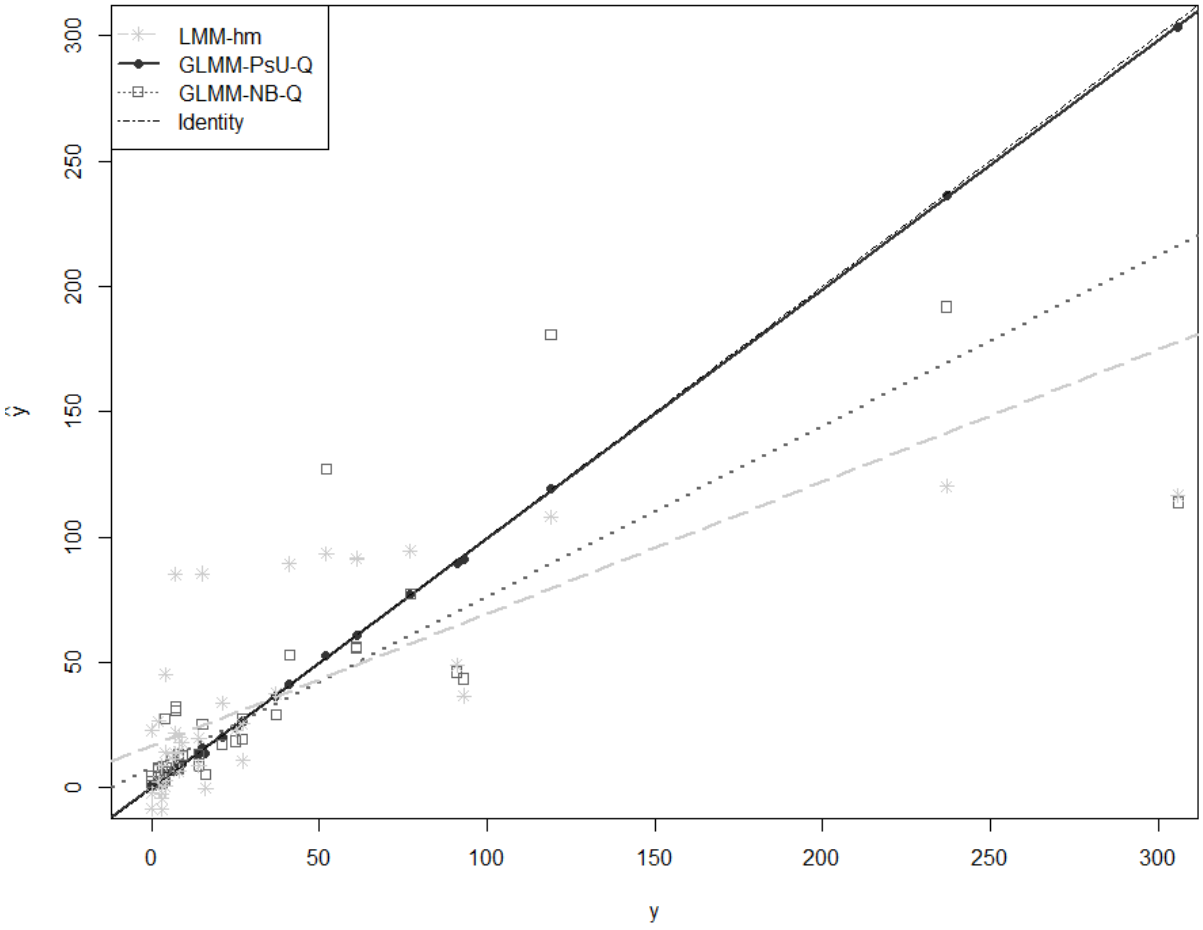
Overdispersion based on the Pearson chi-sq/df statistic seemed to be well controlled for most GLMM, except models 5 through 7, namely GLMM-Ps (Table A 1 in Appendix A).

Estimates of fit statistics AIC and BIC showed similar behavior under the Poisson-Gamma DGP as they had for the Poisson-Unit DGP (Table 12). In particular, models 14 through 16, namely, GLMM-Gm models, yielded the smallest values of AIC and BIC, both on average and in at least 74% of all simulated datasets under the Poisson-Gamma DGP. Meanwhile, models 8 through 10, namely GLMM-PsU and models 11 through 13, namely GLMM-NB showed small numerical differences in AIC or BIC, thus indicating little discriminatory power to detect the true data generation distribution.

Table 13 shows estimated MAR coefficients for the unequal treatment means scenarios under the Poisson-Gamma DGP. Here too, models 8 through 10, namely GLMM-PsU, showed MAR estimates closest to the target values for both intercept and slope, both on average and in at least 74 and 94% of the simulated datasets, respectively. Meanwhile, the true model, namely GLMM-NB yielded substantial variability the estimates of MAR coefficients. Below, figure 3 shows lines described by MAR coefficients from competing models LMM-hm, GLMM-PsU-Q, and GLMM-NB-Q fitted to a random simulation with 10 blocks and unequal treatment means

under the Poisson-Gamma DGP. A similar pattern was observed for the ratio statistics obtained from predicted values of competing models. In all cases, models 8 through 10, namely GLMM-PsU, yielded estimates closest to the target value 1 compared to the remaining competing models (Tables A 2, A 3, and A 4). Specifically, for at least 91%, 81% and 93% of the simulated datasets, the estimated ratios of percentiles, skewness, and coefficients of variation obtained from the fitted GLMM-PsU were closest to the target value 1 of all models considered, including the true model GLMM-NB.

Figure 3: Fitted major axis regression (MAR) lines of predicted values $\hat{y}_{ijk}$ over observed values y_{ijk} under selected competing models, namely LMM-hm (-- - -*- - --), GLMM-PsU-Q (--●--), and GLMM-NB-Q (square). Observed values y_{ijk} were generated under a Poisson-Gamma DGP with 10 blocks and unequal treatment means. Identity line represented by -·-·-.



3.3 Simulation Study: Additive Means DGP

The reader may recall that, unlike the Poisson-Unit DGP and the Poisson-Gamma DGP examined earlier in this study, none of the competing models align with the Additive Means

DGP. That is, for this simulation scenario, there is no true model amongst the competing specifications.

Table 14 presents the Type I Error rates and statistical power for models 1 through 16 fitted to data generated under the Additive Means simulation scenarios.

Models 1 through 4, or the LMMs, controlled Type I error within nominal expectations. LMM-l yielded power near 0.70 for both the 10- and 50-block scenarios whereas the LMM-ht increased in power from 0.45 with 10 blocks to 0.83 with 50 blocks. Remaining LMMs returned statistical power greater than 0.80.

Of all the GLMMs, only model 10, namely GLMM-PsU-P, controlled Type I Error rate on target while maintaining power at approximately 0.70. For many of these Additive Means simulations, specifically, at least 60% of GLMM-PsU-L and at least 80% of GLMM-PsU-Q, estimates of variance components converge to 0 which, in turn, made inference based on F-tests impossible to compute. Of the subset of simulations that avoided the problem of variance component estimates converging to 0, competing models 8 and 9, namely GLMM-PsU-L and GLMM-PsU-Q, showed slight inflation of Type I error that was somewhat mitigated by the larger sample size. On the other hand, GLMM-PsU-P returned Type I error rates at the nominal level. Remaining models 11 through 16, namely GLMM-NB and GLMM-Gm, showed slight inflation of Type I error that did not seem to abate with larger sample sizes.

Table 15 shows a summary of estimates of treatment expectations and corresponding 95% CI coverage for competing models fitted to data generated under the Additive Means DGP. Competing models generally tended to underestimate the treatment means by varying degrees, though GLMM-Gm would overestimate the highest and lowest means (i.e., ij = 21 and 22, respectively) at 10 blocks and the lowest mean (i.e., ij = 22) with 50 blocks. Model LMM-hm

returned 100% CI coverage for lower treatment means (ij = 11 and 22) due to a large CI width that precluded estimation precision. GLMM-NB and GLMM-Gm CI coverage of treatment expectations was between 0.90 and 0.95 across equal and unequal expectations of treatments, except for model 13, namely GLMM-NB-P. GLMM-Gm CI coverage also fell to about 0.7 and overestimated the lowest treatment expectation (i.e., ij = 22) with 50 blocks.

As with previously considered DGP, the Pearson Chi-Squared/DF statistic consistently detected overdispersion in models 5 to 7, namely GLMM-Ps, when fitted to data generated under the Additive Means DGP (Table 16). Meanwhile, remaining GLMMs under consideration (i.e., models 8, 9, 11, 12, 14, 15) showed no evidence for overdispersion, with values of the Pearson Chi-Squared/DF statistic close to or below 1 (Table 16).

Table 17 shows AIC and BIC values for competing models fitted to data generated under Additive Means DGP scenario with unequal treatment means. Results indicate a similar pattern of relative model fit as reported for previously considered DGPs, whereby models 14 to 16, namely GLMM-Gm, returned the smallest values of AIC and BIC, both on average and in about 78% of the cases. Notably, model 1, namely LMM-hm, yielded smallest AIC and BIC values in about 8% of the cases. Taken together, under an Additive Means DGP, both AIC and BIC criteria seemed to indicate as best fitting, models that assumed continuous approximations to count data, namely GLMM-Gm or LMM-hm, as opposed to GLMMs that recognized the discrete distributional nature of the response.

Table 18 shows point estimates of MAR coefficients obtained from regressing predicted values ($\hat{y}_{ijk}$) of each competing model over observed (simulated) values (y_{ijk}) for the unequal treatment means simulation scenarios in this DGP. As seen in previous DGPs, models 8 through 10, namely GLMM-PsU, showed MAR estimates closest to the respective targets (Table 18) in at

least 99% of the datasets. Additionally, models 8 through 10, namely GLMM-PsU, yielded the least variability in estimates of MAR of all the models considered (Table 18). All competing models considered for this study consistently overestimated the major axis intercept parameter, as all estimates were greater than 0, except for the minimum estimate returned by LMM-l of -0.63. Similarly, all competing models underestimated major axis slope parameters as all estimates were less than 1.

Table 19 shows descriptive statistics for the ratio of 97.5th percentiles on predicted ($\hat{y}_{ijk}$) vs. observed values (y_{ijk}) for the simulation scenarios of unequal treatment means under this DGP. GLMM-PsU returned ratios of the predicted vs. observed 97.5th percentile that were consistently closest to the target value 1, both on average and in at least 94% of all simulated datasets under the Additive Means DGP. GLMM-PsU appears to yield predicted values that best describe extreme values in these simulated data, specifically the 97.5th percentile.

Table 20 shows presents descriptive statistics for the ratio of skewness statistics on predicted ($\hat{y}_{ijk}$) vs. observed (y_{ijk}) values for the simulation scenarios of unequal treatment means under this DGP. The GLMM-PsU returned ratios for skewness within the narrowest range and closest to the target value 1, both on average and in at least 84% of the datasets under the Additive Means DGP. All remaining competing models yielded ratios consistently below 1, thus indicating failure to capture skewness in the data.

Table 21 contains descriptive statistics for the CV ratios on predicted ($\hat{y}_{ijk}$) vs. observed values (y_{ijk}) for the simulation scenarios of unequal treatment means under the Additive Means simulation scenarios. Models 8 through 10, namely GLMM-PsU, consistently yielded estimates of this ratio closest to the target of 1, both on average and across all datasets. For remaining

models, CV ratios were consistently below 1, indicating failure to describe the relative variation within the data.

3.4 Data application

For a proof-of-concept application, competing models 1 through 16 were fit to a selected data subset from the entomological study by Li et al (2019), as explained previously in section 2.4. The data subset was selected to match the design structure of the simulation study under consideration, namely that of randomized blocks.

Table 22 shows the Pearson Chi-Sq/DF statistics, as well as AIC and BIC fit statistics for the models fitted to the data. Models 5 and 6, namely GLMM-Ps, showed clear signs of overdispersion, with a Pearson Chi-Squared/DF statistic of 4.28. Therefore, no additional results are presented for these models. Remaining GLMMs fitted to data yielded values of the Pearson Chi-Squared/DF statistic ranging between 0.26 and 0.61, thus indicating no evidence of overdispersion.

Models 14 through 16, namely GLMM-Gm, yielded smallest values of AIC and BIC, thus indicating best fit amongst the models considered. In turn, competing models 11 and 12, namely GLMM-NB ranked second in goodness of fit based on AIC and BIC, followed closely by models 8 and 9, namely GLMM-PsU, and model 2, namely LMM-ht.

Table 23 shows the estimated MAR coefficients, and corresponding 95% CI, obtained from regressing predicted values $\hat{y}_{ijk}$ over observed values y_{ijk}. For predictions generated under all competing models, estimates of MAR coefficients overestimated the intercept parameter and underestimated the slope parameter, as indicated by corresponding 95% confidence intervals that were above and below, respectively, of the target values for these parameters. Models 8 through

10, namely GLMM-PsU, returned coefficient estimates closest to the respective target values than any of the other competing models.

Table 24 shows the ratios of 97.5th percentile, skewness coefficient, and CV for predicted values ($\hat{y}_{ijk}$) over observed values (y_{ijk}) for the data application. Models 8 through 10, namely GLMM-PsU, returned estimated ratios of the 97.5 percentile and CV closest to the target value 1 compared to any of the competing models considered. For the ratio of skewness coefficients, it was model 3, namely LMM-l, which yielded an estimate marginally closer to 1 – about 0.03 units close – than GLMM-PsU. Results suggest that models 8 through 10, namely GLMM-PsU, recovered the data features of interest at least as well, if not better, relative to any of the competing model alternatives.

Table 25 shows F-test statistics assessing the null hypothesis of equal treatment means under competing models fitted to the entomological data subset. Of all the competing models, only GLMM-Ps reports a significance at the 5% level, though this is likely explained by the overdispersion previously observed. Models 1 and 4, namely LMM-hm and LMM-s show marginally significant p-values for treatment difference (i.e. 0.061 and 0.078, respectively). Models 8 through 16, namely GLMM-PsU, GLMM-NB and GLMM-Gm show no evidence for treatment difference, though the magnitude of P-values reflects the method of estimation used for variance components.

Table 26 shows estimated treatment means and corresponding 95% CI from LMM-hm, LMM-ht, LMM-l, GLMM-PsU, GLMM-NB, and GLMM-Gm. Models 8 through 10, namely GLMM-PsU, consistently yielded estimated treatment means smaller in magnitude and narrower in CI than models 11 through 13, namely GLMM-NB. This is consistent with results from our simulation study and with previous studies (Stroup 2012; 2015).

Discussion

The objective of this study was to explore potential diagnostic metrics to assess GLMM fit and performance, to ultimately inform model choice for non-normal response variables, specifically discrete counts. To this end, we conducted a simulation study under various data generation processes and a proof-of-concept application using entomological data. In each case, we fitted an array of competing models, including GLMMs that explicitly specified count-amenable distributional assumptions and models intended to work by approximation. We evaluated the performance of an array of statistics developed to assess GLMM fit.

Since the early 20th century, linear models have been the go-to methods in the applied statistician's toolbox to deal with count data, mostly by way of variance-stabilizing transformations of the data, the behavior of which was to be approximated by a normal distribution (Miller, 1997). This type of approach has inevitably led to problems, mostly related to loss of power and estimates outside of the parameter space; for example, treatment mean estimates below 0 (Stroup, 2015). It was not until 1972 that generalized linear model theory opened up a formal framework to incorporate distributional assumptions beyond normality into the model process (Lee & Nelder, 1996; Nelder & Wedderburn, 1972). These developments were soon followed by estimation algorithms (Laird & Ware, 1982; Wolfinger & O'Connell, 1993) that enabled implementation of GLMMs into software tools. For example, the advent of the GLIMMIX macro from SAS in 2005 introduced the use of non-normal distributional assumptions to fit discrete data in the context of correlated data, and laid the foundation for the GLIMMIX procedure (Stroup, 2015).

Yet, the practical implementation of GLMM for the modeling of discrete data is not without shortcomings due to challenges unique to discrete data (e.g., overdispersion) as well as

alternative, albeit conceptually appropriate, modeling specifications (e.g., distributional assumptions). Further complicating the development of GLMM diagnostics are the many components that must be simultaneously gauged, such as the accuracy of distributional assumptions, the specification of effects in the linear predictor, both fixed and random, and the recognition of the underlying covariance structure (Kramer, 2018).

In this study, we considered models spanning the history of count data modeling, starting with various LMM specifications, and ranging to GLMMs that accommodate Poisson or Negative Binomial distributions on the (conditional) response. In addition, we explored GLMM-Gm; while not commonly used in practice, the Gamma distribution offers support along the positive line with consequential asymmetry, thereby supporting its exploration as a potential candidate for approximating the behavior of count data. Furthermore, in recognizing that the true process by which count data are generated is often unknown, our simulation study evaluated model behavior across three plausible data generation processes.

Results from the simulation study indicated that regardless of DGP, GLMM-PsU and GLMM-NB showed slight inflation of Type I Error when sample size was limited, though this problem was mitigated as sample size increased. This was also the case when modeling by Gamma approximation using GLMM-Gm. Meanwhile, modeling by normal approximation, namely through LMMs, seemed to control Type I error at or below the nominal level. LMMs maintained high power, except for LMM-ht. By contrast, GLMMs, specifically GLMM-PsU and GLMM-NB, showed a clear advantage over LMM in terms of decreasing bias in the estimation of treatment means. Indeed, GLMMs yielded point estimates closer to the truth and narrower CI for comparable coverage relative to LMMs. The notable exception to GLMM control over Type I error inflation is GLMM-PsU-L and GLMM-PsU-Q in the Additive Means scenario, which

experienced severe Type I Error inflation. As expected, GLMM-Ps consistently showed symptoms of overdispersion with concomitant inflation of Type I Error, which rendered GLMM-Ps ineligible for modeling purposes. Regarding GLMM-Gm, there does not seem to be any practical benefit in terms of estimation or inference for using a Gamma approximation for the modeling of count data, compared to fitting properly discrete distributional assumptions, namely GLMM-PsU or GLMM-NB.

A common issue encountered when fitting GLMM to count data is that of overdispersion, whereby variation present in the data is unaccounted for by the model (Stroup, 2012). Single-parameter probability distributions, such as the Poisson distribution, are especially vulnerable (Stroup, 2012). The Pearson Chi-Sq/DF overdispersion statistic is a well-described umbrella-type statistic that can reliably detect problems of overdispersion (Farrington, 1996; Payne et al., 2018; Stroup, 2012). Indeed, in this study, the Pearson Chi-Sq/DF statistic consistently returned values substantially greater than 2 for GLMM-Ps, thus supporting the diagnostic value of this statistic to detect overdispersion. However, the Pearson Chi-Sq/DF statistic lacks specificity to identify the source of overdispersion and thus, provides no guidance on implementation of remedial measures to adjust model fit, when needed.

As a side note, the reverse problem, that of model underdispersion, can be of interest in some applications. Underdispersion is characterized by less variability present in the data than assumed by a model (Winkelmann & Zimmermann, 1995). It has been proposed that values of the Pearson Chi-Sq/DF below the accepted threshold of 1 may be indicative of underdispersion (Payne et al., 2018). In our study, this was not the case. Specifically, when data generated under a Poisson-Unit DGP were fitted with the true model, namely GLMM-PsU, the Pearson Chi-Sq/DF statistic was considerably smaller than 1, in some cases closer to 0. The same result was

observed when data generated under a Poisson-Gamma DGP were fitted with their respective true model, namely GLMM-NB. In either case, by definition and construction, the model specification directly reflected the DGP and thus, could not possibly be underdispersed; yet values of the Pearson Chi-square statistic were found to be as low as 0.07. Taken together, our results indicated that Pearson Chi-Sq/DF values between 0 and 1 are not necessarily indicative of underdispersion and should not be interpreted as such.

Fit statistics Akaike's and Bayesian Information Criterion are statistics often used to compare fit of alternative models to a given dataset (Burnham & Anderson, 2004). Our simulation study showed that neither AIC nor BIC was able to reliably identify the true data generation models, even when true models were available amongst competing models. Rather surprisingly, GLMM-Gm – intended as an approximation to count data – was found to be better fitting based on both AIC and BIC, regardless of the DGP involved. Moreover, true models yielded AIC and BIC values that, at best, differed from those of other competing models by very small magnitudes. Specifically, both fit statistics struggled to discriminate fit between the distributional assumptions embedded in GLMM-PsU and GLMM-NB. Although model dimension differs slightly across competing models, the magnitude of the differences in information criteria are attributed to the differences in likelihood and likelihood approximations calculations under each distributional assumption. As a result, the value of AIC or BIC to aid the practicing statistician in assessing GLMM fit given these sample sizes seems questionable, at best. However, it is known that BIC will generally tend to select the true model if it is included in the set of competing models as the sample size increases to infinity (Burnham & Anderson, 2004). Thus, it is possible that the failure of BIC in this study was due to an insufficiently large sample size.

Recent proposals for umbrella coefficients of determination statistics for GLMMs attempt to accommodate diversity in (co)variance structure with varying degrees of success (Jaeger et al., 2019; Nakagawa & Schielzeth, 2013; Piepho, 2019). The measure proposed by Nakagawa, Johnson, and Schielzeth is R^2_{GLMM}, a ratio of fixed-effect variation over the sum of fixed-effect, individual-effect, and observational-level variances (Nakagawa et al., 2017). This method has an advantage in its simplicity of calculation and interpretation, but has been found to be limited to simple variance-covariance structures (Jaeger et al., 2019; Piepho, 2019). Extensions to more complex (co)variance structures include R^2_Σ (Jaeger et al., 2019), which leverages standardized generalized variances, Kenwood-Rogers estimates of degrees of freedom, and adjusted Wald F-statistics, and is interpreted as the proportion of generalized variance accounted for by the fixed effects in the GLMM (Jaeger et al., 2019). An alternative calculation of a coefficient of determination for fixed effects (Ω_β), random effects (Ω_u), and both fixed and random effects ($\Omega_{\beta u}$), introduces average semivariances to measure the total variance in terms of a mean variance of a difference in observations across any variance-covariance structure, and can respectively be interpreted as the proportion of the variation accounted for by fixed effects, random effects, and both fixed and random effects (Piepho, 2019). Although these coefficients of determination appear to have desirable properties, it is presently unclear if these options are useful for model selection when competing models differ in their distributional assumptions of the response. Additionally, the various coefficients of determination continue to be omnibus, umbrella-type statistics. Diagnostics that inform on specific features of model lack-of-fit to data would be most helpful to inform remedial measures, when needed.

This study attempted to develop useful methods for assessment of GLMM fit and performance through the utilization of predicted values $\hat{y}_{ijk}$(as defined in section 2.3.2) as a

method of investigating specific data features in model assessment. It should be noted that predicted values $\hat{y}$ are direct functions of a fitted linear predictor, that is, the combination of fixed and random effects in a linear model, and thus are influenced by the number of parameters included in the linear predictor. Of the competing models that were evaluated in this study, one such model included one additional effect in its linear predictor relative to the others being evaluated, that model being GLMM-PsU. Thus, the predicted values from this model were calculated slightly differently. It is unclear if comparison of statistics calculated on predicted values from models with different specifications of linear predictors is appropriate.

MAR coefficients were used to evaluate discrepancies between simulated data and model prediction, as were ratios of 97.5^{th} percentiles, skewness coefficients, and coefficients of variation calculated on both predicted and simulated values. In our study, the MAR coefficients consistently identified GLMM-PsU as a best-fitting model regardless of the DGP. Indeed, predicted values obtained from GLMM-PsU most closely mapped the data. Likewise, the 97.5^{th} percentile, skewness, and CV ratio statistics yielded by GLMM-PsU consistently returned values closest to the target of 1 with very little variation compared to any other competing model. We interpret these results to indicate that the GLMM-PsU best reflected the data features of interest. In this study, we also considered analogous ratios of other percentiles including the 0^{th} (minimum), 2.5^{th}, 16^{th}, 25^{th}, 75^{th}, 84^{th}, and 100^{th} (maximum). None of these percentiles yielded patterns useful for fit assessment due to extreme variability of ratios (for specifically the 2.5^{th}, 16^{th} and 25^{th} percentiles) or poor separation of values (for higher percentiles). Therefore, results were not shown here and are not discussed further.

These findings were not necessarily surprising because MAR coefficients and the selected ratios of descriptive statistics were based on predicted values generated from point

estimates of each model's linear predictor. Of all the competing GLMM specifications considered in this study, GLMM-PsU was the only one that incorporated a unit-level effect term, namely u_{ijk}, within the linear predictor. This unit-level term is intended to accommodate any left-over variability in the data that might stem from the unit of observation, which is otherwise left out of consideration in a Poisson GLMM (Stroup, 2012). To note, this unit-level term can be considered to play an equivalent role to that of residuals in the LMM framework assuming normality. By contrast, GLMMs that assume Negative Binomial or Gamma distributions specify scale parameters to accommodate unit-level variability. However, these scale parameters are not necessarily included in the linear predictor per se and thus, cannot be reflect unit-level variability in predicted values. Thus, it may not be appropriate to only utilize statistics on predicted values alone to compare assessments of model fit among competing models with different specifications of linear predictors. However, these statistics may be useful in comparing model fit between GLMMs fitted with different methods of estimations, e.g., GLMM-NB-L vs. GLMM-NB-Q vs. GLMM-NB-P. Other diagnostic statistics that similarly and exclusively focus on the estimation of centrality parameters and related functions may also tend to select GLMMs with more parameters present in the linear predictor than alternative specifications. Further work on GLMM diagnostics may be well served by incorporating estimates of dispersion into the assessment of fit to data.

Taken together, GLMM-PsU proved to be a comparatively robust model for capturing data features of dispersion and skewness and for producing predicted values that best reflected such features, regardless of DGP. The close alignment of predicted values returned by GLMM-PsU to simulated data may indicate the usefulness of GLMM-PsU for predictive purposes. Predictive modeling is concerned with accuracy of the prediction of future or new observations,

and will occasionally sacrifice theoretical accuracy to benefit empirical prediction (Shmueli, 2010). Meanwhile, inference and explanatory modeling is focused on explaining the nature of the relationships between variables in a mechanistic way to accurately describe an underlying data generation process (Shmueli, 2010). When inference is prioritized, the main interest is in minimizing bias between true values of parameters and their estimators. In this study, it was not possible to reliably diagnose the true DGP with these metrics when different from the Poisson-Unit DGP. Of the other two simulated DGP, the Poisson-Gamma DGP had true models in GLMM-NB, while the Additive Means DGP was deliberately designed to be an inherently arbitrary mechanism. Our development of potential diagnostic statistics for GLMM using predicted values, and functions thereof, may be considered better aligned with predictive purposes.

Regarding the data application, our findings were mostly consistent with the results from the simulation studies. The data application also reinforced the relative behaviors of GLMM-PsU and GLMM-NB observed by Stroup (2015) for estimation of treatment means, whereby the former yields lower estimates and narrower CI than the latter. As it stands, the appropriate characterization of the dispersion of the data and the most accurate estimates of the treatment means are unclear to the practicing statistician, and cloud determinations on which model specification to use to ensure proper inference for the data at hand. Further work on diagnostics to address the appropriateness of distributional assumptions made by GLMM is warranted.

A recent practical proposal for GLMM diagnostics drew from the Bayesian literature and explored discrepancies between observed data and pseudo-data generated from candidate models using so-called “centipede plots” (Kramer, 2018). This approach compared data against multiple sets of pseudo-data generated by plugging parameter estimates into the data likelihood, then

using the likelihood as a stochastic data generation process. The alignment between the original datapoints and the generated pseudo-data can be explored visually after ranking the data (Kramer, 2018), with the resulting plots showing a remarkable similarity to centipedes. These plots can be useful in identifying specific ways in which a candidate model fails to align with the data for goodness-of-fit purposes (Kramer, 2018). By inserting parameter estimates into a candidate model as a stochastic process to generate pseudo-data, information on dispersion parameters is naturally incorporated in the evaluation of competing models, thereby avoiding the inherent limitations of relying solely on predicted values, as previously discussed.

Moving forward, the Bayesian statistical framework may provide additional insight to advance GLMM diagnostics, specifically through posterior predictive checks (Gelman et al., 1996; Guttman, 1967; Rubin, 1984). Recognizing that, in a Bayesian setting, both location and dispersion parameters themselves have distributions, a distribution of model predictions can be generated a posteriori by numerically integrating across the joint posterior distribution of the parameters of interest (Gelman et al., 2004). This approach, in turn, propagates uncertainty from the estimation process onto predictions to be used for GLMM diagnostics and model fit assessments. So generated, the posterior predictive distributions (or functions thereof) can then be contrasted for alignment against those of the data.

Finally, in this study we focused on diagnostics computed on predicted values, which, by definition, are expressed in the observable scale, also known as the data scale or the inverse link scale. By contrast, GLMMs are, by definition, fitted on a non-observable scale defined by a link function (Stroup, 2012, 2015). Thus, as defined, a link function imposes a non-linear relationship between the linear predictor in the model and the expectation of the response variable. As such, it is plausible that assessments of model fit in the observable scale be distorted. It may be worth

considering GLMM diagnostics that assess goodness-of-fit in the link scale in which the model is fitted. Further research is warranted to better understand the complexities of GLMM diagnostics.

Conclusion

Potential diagnostic statistics to assess fit of generalized linear mixed models were developed and examined in a simulation study involving realistic data generation processes, followed by a proof-of-concept data application in the entomological sciences.

Diagnostics were developed from functions of predicted values obtained from fitting competing models to simulated data. Specifically, we considered MAR coefficients of predicted values regressed over observed data, as well as ratios of selected features of count data, namely the 97.5^{th} percentile, skewness, and CV. Overall, none of the diagnostics considered here allowed for recovery of the true data generation process; this may be a limitation if research focus is on identifying the underlying mechanisms. In contrast, a Poisson GLMM that included unit-level effects was comparatively robust for inference and produced predicted values that best reflected features of dispersion and skewness commonly encountered in count data.

Further research on diagnostics to assess fit of GLMMs to non-normal data is warranted. Future developments may be well served by incorporating uncertainty of estimation into diagnostic metrics.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. *Second International Symposium on Information Theory*.

Bello, N. M., & Renter, D. G. (2018). Invited review: Reproducible research from noisy data: Revisiting key statistical principles for the animal sciences. *Journal of Dairy Science, 101*(7), 5679–5701. https://doi.org/10.3168/jds.2017-13978

Burman, P. (1989). A Comparative Study of Ordinary Cross-Validation, v-Fold Cross-Validation and the Repeated Learning-Testing Methods. *Biometrika, 76*, 503–514.

Burnham, K. P., & Anderson, D. R. (2004). Multimodel Inference: Understanding AIC and BIC in Model Selection. *Sociological Methods & Research, 33*(2), 261–304. https://doi.org/10.1177/0049124104268644

Correndo, A. A., Hefley, T. J., Holzworth, D. P., & Ciampitti, I. A. (2021). Revisiting linear regression to test agreement in continuous predicted-observed datasets. *Agricultural Systems, 192,* 103194. https://doi.org/10.1016/j.agsy.2021.103194

Czado, C., Gneiting, T., & Held, L. (2009). Predictive Model Assessment for Count Data. *Biometrics, 65*(4), 1254–1261. https://doi.org/10.1111/j.1541-0420.2009.01191.x

Farrington, C. P. (1996). On Assessing Goodness of Fit of Generalized Linear Models to Sparse Data. *Journal of the Royal Statistical Society: Series B (Methodological), 58*(2), 349–360. https://doi.org/10.1111/j.2517-6161.1996.tb02086.x

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2004). *Bayesian Data Analysis, Third Edition, 3rd Edition* (Third). CRC Press.

https://learning.oreilly.com/library/view/bayesian-data-analysis/9781439898222/OEBPS/ref.htm

Gelman, A., Meng, X.-L., & Stern, H. (1996). Posterior Predictive Assessment of Model Fitness via Realized Discrepancies. *Statistica Sinica*, *6*, 733–807.

Gneiting, T. (2011). Making and Evaluating Point Forecasts. *Journal of the American Statistical Association*, *106*(494), 746–762. https://doi.org/10.1198/jasa.2011.r10138

Gonçalves, M. A. D., Tokach, M. D., Dritz, S. S., Bello, N. M., Touchette, K. J., Goodband, R. D., DeRouchey, J. M., & Woodworth, J. C. (2018). Standardized ileal digestible valine:lysine dose response effects in 25- to 45-kg pigs under commercial conditions. *Journal of Animal Science*. https://doi.org/10.1093/jas/skx059

Guttman, I. (1967). The Use of the Concept of a Future Observation in Goodness-Of-Fit Problems. *Journal of the Royal Statistical Society: Series B (Methodological)*, *29*(1), 83–100. https://doi.org/10.1111/j.2517-6161.1967.tb00676.x

Hardle, W., & Marron, J.S. (1985). Asymptotic Nonequivalence of Some Bandwidth Selectors in Nonparametric Regression. *Biometrika*, *72*, 5.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning* (2nd ed.). Springer New York. https://doi.org/10.1007/978-0-387-84858-7

Jaeger, B. C., Edwards, L. J., & Gurka, M. J. (2019). An R^2 statistic for covariance model selection in the linear mixed model. *Journal of Applied Statistics*, *46*(1), 164–184. https://doi.org/10.1080/02664763.2018.1466869

Kramer, M. (2018). *Using the Posterior Predictive Distribution as a Diagnostic Tool for Mixed Models*. 19.

Laird, N. M., & Ware, J. H. (1982). Random-Effects Models for Longitudinal Data. *Biometrics*, *38*(4), 963–974. JSTOR. https://doi.org/10.2307/2529876

Larrabee, B., Scott, H. M., & Bello, N. M. (2014). Ordinary Least Squares Regression of Ordered Categorical Data: Inferential Implications for Practice. *Journal of Agricultural, Biological, and Environmental Statistics*, *19*(3), 373–386. https://doi.org/10.1007/s13253-014-0176-z

Lawless, J. F. (1987). Negative Binomial and Mixed Poisson Regression. *Canadian Journal of Statistics*, *15*(3), 209–225. https://doi.org/10.2307/3314912

Lee, Y., & Nelder, J. A. (1996). Hierarchical Generalized Linear Models. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(4), 619–656. https://doi.org/10.1111/j.2517-6161.1996.tb02105.x

Legendre, P., & Legendre, L. (2012). *Numerical Ecology*. Elsevier. http://ebookcentral.proquest.com/lib/ksu/detail.action?docID=982554

Li, Y., Cloyd, R. A., & Bello, N. M. (2019). Effect of Integrating the Entomopathogenic Fungus (Hypocreales: Cordycipitaceae) and the Rove Beetle (Coleoptera: Staphylinidae) in Suppressing Western Flower Thrips (Thysanoptera: Thripidae) Populations Under Greenhouse Conditions. *Journal of Economic Entomology*, *112*(5), 2085–2093. https://doi.org/10.1093/jee/toz132

McCullagh, P., & Nelder, J. A. (1989a). *Generalized Linear Models* (2nd ed.). Chapman and Hall.

McCullagh, P., & Nelder, J. A. (1989b). *Generalized Linear Models*. CRC Press LLC. http://ebookcentral.proquest.com/lib/ksu/detail.action?docID=5631551

Mesplé, F., Troussellier, M., Casellas, C., & Legendre, P. (1996). Evaluation of simple statistical criteria to qualify a simulation. *Ecological Modelling*, *88*(1), 9–18. https://doi.org/10.1016/0304-3800(95)00033-X

Miller, R. G., Jr. (1997). *Beyond ANOVA: Basics of Applied Statistics*. CRC Press.

Nakagawa, S., Johnson, P. C. D., & Schielzeth, H. (2017). The coefficient of determination R2 and intra-class correlation coefficient from generalized linear mixed-effects models revisited and expanded. *J.R. Soc. Interface*, *14*(134), 11.

Nakagawa, S., & Schielzeth, H. (2013). A general and simple method for obtaining R^2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution*, *4*(2), 133–142. https://doi.org/10.1111/j.2041-210x.2012.00261.x

Nelder, J. A., & Wedderburn, R. W. M. (1972). *Generalized Linear Models*. 16.

Payne, E. H., Gebregziabher, M., Hardin, J. W., Ramakrishnan, V., & Egede, L. E. (2018). An empirical approach to determine a threshold for assessing overdispersion in Poisson and negative binomial models for count data. *Communications in Statistics - Simulation and Computation*, *47*(6), 1722–1738. https://doi.org/10.1080/03610918.2017.1323223

Piepho, H. (2019). A coefficient of determination (R^2) for generalized linear mixed models. *Biometrical Journal*, bimj.201800270. https://doi.org/10.1002/bimj.201800270

Pinheiro, J. C., & Chao, E. C. (2006). Efficient Laplacian and Adaptive Gaussian Quadrature Algorithms for Multilevel Generalized Linear Mixed Models. *Journal of Computational and Graphical Statistics*, *15*(1), 58–81. https://doi.org/10.1198/106186006X96962

R Core Team. (2021). *R: A Language and Environment for Statistical Computing* (4.1.0) [Computer software]. R Foundation for Statistical Computing. https://www.R-project.org/

Rubin, D. B. (1984). Bayesianly Justifiable and Relevant Frequency Calculations for the Applied Statistician. *The Annals of Statistics*, *12*(4), 1151–1172. https://doi.org/10.1214/aos/1176346785

SAS Institute Inc. (2016). *SAS* (9.4) [Computer software]. SAS Institute, Inc.

Schabenberger, O. (2005). Paper 196-30: Introducing the GLIMMIX Procedure for Generalized Linear Mixed Models. *SAS Global Forum Proceedings*, 20.

Schwarz, G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics*, *6*(2). https://doi.org/10.1214/aos/1176344136

Shmueli, G. (2010). To Explain or to Predict? *Statistical Science*, *25*(3). https://doi.org/10.1214/10-STS330

Stroup, W. W. (2012). *Generalized Linear Mixed Models: Modern Concepts, Methods and Applications*. Taylor & Francis Group. http://ebookcentral.proquest.com/lib/ksu/detail.action?docID=1570060

Stroup, W. W. (2013). Non-Normal Data in Agricultural Experiments. *Conference on Applied Statistics in Agriculture*. https://doi.org/10.4148/2475-7772.1018

Stroup, W. W. (2015). Rethinking the Analysis of Non-Normal Data in Plant and Soil Science. *Agronomy Journal*, *107*(2), 811–827. https://doi.org/10.2134/agronj2013.0342

Ver Hoef, J. M., & Boveng, P. L. (2007). Quasi-Poisson vs. Negative Binomial Regression: How Should We Model Overdispersed Count Data? *Ecology*, *88*(11), 2766–2772. https://doi.org/10.1890/07-0043.1

Winkelmann, R., & Zimmermann, K. F. (1995). Recent Developments in Count Data Modelling: Theory and Application. *Journal of Economic Surveys*, *9*(1), 1. Business Source Premier.

Wolfinger, R., & O'Connell, M. (1993). Generalized linear mixed models a pseudo-likelihood approach. *Journal of Statistical Computation and Simulation*, *48*(3–4), 233–243. https://doi.org/10.1080/00949659308811554

Tables

Table 1: Frequencies of model failures to converge, by simulation scenario

Simulation scenarios	N	LMM-ht	GLMM-PsU-L	GLMM-PsU-Q	GLMM-NB-P	GLMM-Gm-P
Poisson-Unit DGP 10 Blocks, Equal Treatment Means	100	34	1	1	2	1
Poisson-Unit DGP 10 Blocks, Unequal Treatment Means	100	51	0	0	4	0
Poisson-Unit DGP 50 Blocks, Equal Treatment Means	100	13	0	0	1	0
Poisson-Unit DGP 50 Blocks, Unequal Treatment Means	100	50	0	1	1	0
Poisson-Gamma DGP 10 Blocks, Equal Treatment Means	100	35	0	0	0	0
Poisson-Gamma 10 Blocks, Unequal Treatment Means	100	27	0	0	0	0
Poisson-Gamma 50 Blocks, Equal Treatment Means	100	1	0	0	0	0
Poisson-Gamma 50 Blocks, Unequal Treatment Means	100	0	0	0	0	0
Additive Means DGP 10 Blocks, Equal Treatment Means	100	1	0	0	0	0
Additive Means DGP 10 Blocks, Unequal Treatment Means	100	10	0	0	0	0
Additive Means DGP 50 Blocks, Equal Treatment Means	100	0	0	0	1	0
Additive Means DGP 50 Blocks, Unequal Treatment Means	100	0	0	0	1	0

Table 2: Empirical Type I Error & Statistical Power, Poisson-Unit DGP

Results for the true GLMMs used for data generation are **bolded**

	10 Blocks		50 Blocks	
	Type I Error	**Power**	**Type I Error**	**Power**
1-LMM-hm	0.05	0.72	0.05	0.99
2-LMM-ht	0.02	0.22	0.00	0.81
3-LMM-l	0.06	0.67	0.06	0.69
4-LMM-s	0.06	0.79	0.04	0.98
5-GLMM Ps-L	0.61	.	0.68	.
6-GLMM Ps-Q	0.61	.	0.68	.
7-GLMM Ps-P	0.61	.	0.68	.
8-GLMM PsU-L	**0.12**	**0.69**	**0.05**	**0.70**
9-GLMM PsU-Q	**0.14**	**0.70**	**0.05**	**0.70**
10-GLMM PsU-P	**0.07**	**0.68**	**0.04**	**0.70**
11-GLMM NB-L	0.09	0.69	0.07	0.70
12-GLMM NB-Q	0.09	0.69	0.07	0.70
13-GLMM NB-P	0.09	0.69	0.07	0.70
14-GLMM Gm-L	0.09	0.71	0.06	0.71
15-GLMM Gm-Q	0.09	0.71	0.06	0.71
16-GLMM Gm-P	0.09	0.71	0.07	0.99

Table 3: Average, minimum and maximum values of the Pearson chi-sq/df statistic under simulation scenario of unequal treatment means with 10 or 50 blocks and a Poisson-Unit DGP

Results for the true GLMMs used for data generation are **bolded**

	10 Blocks, Unequal Treatment Means			50 Blocks, Unequal Treatment Means		
	Mean	**Min**	**Max**	**Mean**	**Min**	**Max**
5-GLMM Ps-L	9.73	2.53	43.37	10.93	6.60	27.38
6-GLMM Ps-Q	9.73	2.53	43.37	10.93	6.60	27.38
8-GLMM PsU-L	**0.18**	**0.08**	**0.39**	**0.16**	**0.12**	**0.21**
9-GLMM PsU-Q	**0.18**	**0.07**	**0.39**	**0.16**	**0.11**	**0.21**
11-GLMM NB-L	0.77	0.58	1.17	0.78	0.69	0.92
12-GLMM NB-Q	0.77	0.58	1.17	0.78	0.69	0.92
14-GLMM Gm-L	0.38	0.14	0.90	0.43	0.29	0.56
15-GLMM Gm-Q	0.38	0.14	0.90	0.43	0.29	0.56

Table 4: Summary of estimates of expectations of treatments and 95% confidence interval coverage for competing models fitted to data generated under the Poisson-Unit DGP

Results for true GLMMs used for data generation are **bolded**, and results for estimates of treatment mean λ_{12} are excluded as $\lambda_{12} = \lambda_{11}$

Model	Blocks	Treatment levels (*ij*)	True Mean	Avg Treatment Mean Estimate	Avg Diff from True Mean	Avg width of 95% CI for Means	95% CI Coverage
1-LMM-hm	10	1,1; 1,2	12.2	23.00	10.80	83.19	0.96
2-LMM-ht	10	1,1; 1,2	12.2	29.92	17.72	382.46	0.88
3-LMM-l	10	1,1; 1,2	12.2	12.47	0.27	23.98	0.94
4-LMM-s	10	1,1; 1,2	12.2	17.26	5.06	33.25	0.94
5-GLMM-Ps-L	10	1,1; 1,2	12.2	15.47	3.27	19.13	0.79
6-GLMM-Ps-Q	10	1,1; 1,2	12.2	15.47	3.27	19.13	0.79
7-GLMM-Ps-P	10	1,1; 1,2	12.2	15.56	3.36	20.24	0.80
8-GLMM-PsU-L	**10**	**1,1; 1,2**	**12.2**	**13.29**	**1.09**	**19.76**	**0.91**
9-GLMM-PsU-Q	**10**	**1,1; 1,2**	**12.2**	**13.29**	**1.09**	**19.30**	**0.87**
10-GLMM-PsU-P	**10**	**1,1; 1,2**	**12.2**	**13.76**	**1.56**	**21.55**	**0.93**
11-GLMM-NB-L	10	1,1; 1,2	12.2	16.91	4.71	25.57	0.86
12-GLMM-NB-Q	10	1,1; 1,2	12.2	16.90	4.70	25.56	0.86
13-GLMM-NB-P	10	1,1; 1,2	12.2	16.20	4.00	25.18	0.85
14-GLMM-Gm-L	10	1,1; 1,2	12.2	17.62	5.42	25.24	0.83
15-GLMM-Gm-Q	10	1,1; 1,2	12.2	17.61	5.41	25.28	0.83
16-GLMM-Gm-P	10	1,1; 1,2	12.2	16.93	4.73	25.09	0.83
1-LMM-hm	10	2,1	54.6	102.08	47.48	83.19	0.43
2-LMM-ht	10	2,1	54.6	122.16	67.56	369.18	0.80
3-LMM-l	10	2,1	54.6	59.42	4.82	114.68	0.97
4-LMM-s	10	2,1	54.6	78.43	23.83	71.28	0.66
5-GLMM-Ps-L	10	2,1	54.6	67.38	12.78	81.97	0.83
6-GLMM-Ps-Q	10	2,1	54.6	67.38	12.78	81.97	0.83
7-GLMM-Ps-P	10	2,1	54.6	67.76	13.16	86.94	0.86
8-GLMM-PsU-L	**10**	**2,1**	**54.6**	**60.11**	**5.51**	**85.80**	**0.93**
9-GLMM-PsU-Q	**10**	**2,1**	**54.6**	**60.10**	**5.50**	**85.22**	**0.91**
10-GLMM-PsU-P	**10**	**2,1**	**54.6**	**60.75**	**6.15**	**91.26**	**0.94**
11-GLMM-NB-L	10	2,1	54.6	77.17	22.57	112.20	0.79
12-GLMM-NB-Q	10	2,1	54.6	77.16	22.56	112.26	0.79
13-GLMM-NB-P	10	2,1	54.6	73.50	18.90	110.01	0.85
14-GLMM-Gm-L	10	2,1	54.6	78.26	23.66	110.60	0.78
15-GLMM-Gm-Q	10	2,1	54.6	78.23	23.63	110.76	0.79
16-GLMM-Gm-P	10	2,1	54.6	74.49	19.89	108.58	0.82
1-LMM-hm	10	2,2	2.7	4.99	2.29	83.19	1.00
2-LMM-ht	10	2,2	2.7	5.60	2.90	318.64	0.98
3-LMM-l	10	2,2	2.7	2.31	-0.39	4.32	0.88
4-LMM-s	10	2,2	2.7	3.68	0.98	15.84	0.98
5-GLMM-Ps-L	10	2,2	2.7	3.40	0.70	4.60	0.84
6-GLMM-Ps-Q	10	2,2	2.7	3.40	0.70	4.60	0.84
7-GLMM-Ps-P	10	2,2	2.7	3.42	0.72	4.83	0.84

8-GLMM-PsU-L	**10**	**2,2**	**2.7**	**2.93**	**0.23**	**4.96**	**0.96**
9-GLMM-PsU-Q	**10**	**2,2**	**2.7**	**2.93**	**0.23**	**4.96**	**0.95**
10-GLMM-PsU-P	**10**	**2,2**	**2.7**	**3.21**	**0.51**	**5.55**	**0.96**
11-GLMM-NB-L	10	2,2	2.7	3.73	1.03	6.15	0.90
12-GLMM-NB-Q	10	2,2	2.7	3.73	1.03	6.15	0.90
13-GLMM-NB-P	10	2,2	2.7	3.65	0.95	6.23	0.91
14-GLMM-Gm-L	10	2,2	2.7	4.52	1.82	6.87	0.75
15-GLMM-Gm-Q	10	2,2	2.7	4.52	1.82	6.88	0.75
16-GLMM-Gm-P	10	2,2	2.7	4.31	1.61	6.74	0.81
1-LMM-hm	50	1,1; 1,2	12.2	20.89	8.69	38.04	0.96
2-LMM-ht	50	1,1; 1,2	12.2	22.56	10.36	87.29	0.38
3-LMM-l	50	1,1; 1,2	12.2	11.14	-1.06	8.26	0.87
4-LMM-s	50	1,1; 1,2	12.2	15.65	3.45	13.34	0.88
5-GLMM-Ps-L	50	1,1; 1,2	12.2	13.84	1.64	7.25	0.78
6-GLMM-Ps-Q	50	1,1; 1,2	12.2	13.84	1.64	7.25	0.78
7-GLMM-Ps-P	50	1,1; 1,2	12.2	13.94	1.74	7.33	0.76
8-GLMM-PsU-L	**50**	**1,1; 1,2**	**12.2**	**12.02**	**-0.18**	**7.52**	**0.91**
9-GLMM-PsU-Q	**50**	**1,1; 1,2**	**12.2**	**12.06**	**-0.14**	**7.48**	**0.91**
10-GLMM-PsU-P	**50**	**1,1; 1,2**	**12.2**	**12.59**	**0.39**	**7.71**	**0.94**
11-GLMM-NB-L	50	1,1; 1,2	12.2	15.58	3.38	9.73	0.66
12-GLMM-NB-Q	50	1,1; 1,2	12.2	15.57	3.37	9.73	0.66
13-GLMM-NB-P	50	1,1; 1,2	12.2	14.96	2.76	9.22	0.75
14-GLMM-Gm-L	50	1,1; 1,2	12.2	15.99	3.79	9.56	0.54
15-GLMM-Gm-Q	50	1,1; 1,2	12.2	15.98	3.78	9.57	0.54
16-GLMM-Gm-P	50	1,1; 1,2	12.2	15.16	2.96	8.96	0.69
1-LMM-hm	50	2,1	54.6	96.53	41.93	38.04	0.05
2-LMM-ht	50	2,1	54.6	105.63	51.03	113.43	0.24
3-LMM-l	50	2,1	54.6	54.49	-0.11	40.59	0.96
4-LMM-s	50	2,1	54.6	72.74	18.14	28.58	0.31
5-GLMM-Ps-L	50	2,1	54.6	63.55	8.95	32.68	0.74
6-GLMM-Ps-Q	50	2,1	54.6	63.55	8.95	32.68	0.74
7-GLMM-Ps-P	50	2,1	54.6	64.01	9.41	33.03	0.75
8-GLMM-PsU-L	**50**	**2,1**	**54.6**	**55.18**	**0.58**	**33.20**	**0.93**
9-GLMM-PsU-Q	**50**	**2,1**	**54.6**	**55.35**	**0.75**	**33.25**	**0.94**
10-GLMM-PsU-P	**50**	**2,1**	**54.6**	**55.97**	**1.37**	**33.05**	**0.92**
11-GLMM-NB-L	50	2,1	54.6	71.85	17.25	43.80	0.58
12-GLMM-NB-Q	50	2,1	54.6	71.83	17.23	43.82	0.58
13-GLMM-NB-P	50	2,1	54.6	68.39	13.79	41.03	0.69
14-GLMM-Gm-L	50	2,1	54.6	72.84	18.24	43.20	0.53
15-GLMM-Gm-Q	50	2,1	54.6	72.79	18.19	43.29	0.53
16-GLMM-Gm-P	50	2,1	54.6	69.03	14.43	40.53	0.63
1-LMM-hm	50	2,2	2.7	4.76	2.06	38.04	1.00
2-LMM-ht	50	2,2	2.7	4.86	2.16	87.06	1.00
3-LMM-l	50	2,2	2.7	2.10	-0.60	1.56	0.66
4-LMM-s	50	2,2	2.7	3.46	0.76	6.51	1.00
5-GLMM-Ps-L	50	2,2	2.7	3.16	0.46	1.80	0.68
6-GLMM-Ps-Q	50	2,2	2.7	3.16	0.46	1.80	0.68
7-GLMM-Ps-P	50	2,2	2.7	3.18	0.48	1.82	0.67

8-GLMM-PsU-L	**50**	**2,2**	**2.7**	**2.73**	**0.03**	**1.91**	**0.93**
9-GLMM-PsU-Q	**50**	**2,2**	**2.7**	**2.73**	**0.03**	**1.91**	**0.93**
10-GLMM-PsU-P	**50**	**2,2**	**2.7**	**3.06**	**0.36**	**2.07**	**0.83**
11-GLMM-NB-L	50	2,2	2.7	3.54	0.84	2.41	0.65
12-GLMM-NB-Q	50	2,2	2.7	3.54	0.84	2.41	0.65
13-GLMM-NB-P	50	2,2	2.7	3.49	0.79	2.36	0.67
14-GLMM-Gm-L	50	2,2	2.7	4.27	1.57	2.69	0.20
15-GLMM-Gm-Q	50	2,2	2.7	4.26	1.56	2.69	0.20
16-GLMM-Gm-P	50	2,2	2.7	4.05	1.35	2.52	0.29

Table 5: Average, minimum, and maximum values of Akaike's and Bayesian Fit Criteria under simulation scenarios of unequal treatment means with 10 or 50 blocks and a Poisson-Unit DGP

Results for the true GLMMs used for data generation are **bolded**

	Akaike's Information Criterion						Bayesian Information Criterion					
	10 Blocks, Unequal Treatment Means			50 Blocks, Unequal Treatment Means			10 Blocks, Unequal Treatment Means			50 Blocks, Unequal Treatment Means		
	Mean	Min	Max	Mean	Min	Max	Mean	Min	Max	Mean	Min	Max
1-LMM-hm	396.2	287.9	539.1	2196.5	1840.3	2796.5	396.8	288.5	539.7	2200.3	1844.1	2800.3
2-LMM-ht	420.8	288.7	602.4	2269.8	1872.2	2932.4	422.3	290.2	603.9	2279.3	1881.8	2942.0
5-GLMM Ps-L	594.9	306.5	2035.9	3131.7	2320.9	5698.1	596.4	308.1	2037.4	3141.2	2330.4	5707.6
6-GLMM Ps-Q	594.9	306.5	2035.9	3131.7	2320.9	5698.1	596.4	308.1	2037.4	3141.2	2330.4	5707.6
8-GLMM PsU-L	**321.9**	**249.6**	**395.8**	**1576.5**	**1460.1**	**1684.9**	**323.7**	**251.2**	**397.6**	**1588.0**	**1471.5**	**1696.4**
9-GLMM PsU-Q	**321.7**	**249.6**	**395.8**	**1576.5**	**1458.6**	**1726.1**	**323.5**	**251.2**	**397.6**	**1588.0**	**1470.1**	**1735.7**
11-GLMM NB-L	322.0	250.1	399.0	1579.0	1462.6	1691.6	323.8	251.6	400.8	1590.5	1474.1	1703.1
12-GLMM NB-Q	322.0	250.1	399.0	1579.0	1462.5	1691.5	323.8	251.6	400.8	1590.4	1474.0	1703.0
14-GLMM Gm-L	310.6	231.3	397.9	1527.1	1391.1	1642.3	312.4	233.1	399.7	1538.6	1402.6	1653.7
15-GLMM Gm-Q	310.6	231.3	397.8	1526.8	1391.0	1641.9	312.4	233.1	399.6	1538.3	1402.5	1653.4

Table 6: Average, minimum, and maximum values for point estimates of MAR coefficients, both intercept and slope, in simulation scenarios of unequal treatment means with 10 or 50 blocks under a Poisson-Unit DGP

Results for the true GLMMs used for data generation are **bolded**

	Major Axis Regression (Data Scale) Intercept Coefficients						Major Axis Regression (Data Scale) Slope Coefficients					
	10 Blocks, Unequal Treatment Means			50 Blocks, Unequal Treatment Means			10 Blocks, Unequal Treatment Means			50 Blocks, Unequal Treatment Means		
	Mean	Min	Max	Mean	Min	Max	Mean	Min	Max	Mean	Min	Max
1-LMM-hm	19.03	3.04	68.27	21.77	8.41	58.91	0.53	0.20	0.83	0.40	0.10	0.65
2-LMM-ht	25.01	4.71	80.28	25.11	10.35	58.32	0.50	0.15	0.77	0.36	0.11	0.67
3-LMM-l	6.81	-10.02	26.50	10.00	3.42	25.67	0.56	0.15	1.18	0.42	0.11	0.68
4-LMM-s	14.20	1.96	38.43	16.12	7.51	34.26	0.47	0.18	0.81	0.36	0.09	0.56
5-GLMM Ps-L	4.18	-2.30	23.36	4.97	-0.13	10.82	0.89	0.63	1.03	0.86	0.75	1.01
6-GLMM Ps-Q	4.18	-2.30	23.36	4.97	-0.13	10.82	0.89	0.63	1.03	0.86	0.75	1.01
7-GLMM Ps-P	4.17	-2.31	23.34	4.97	-0.13	10.82	0.89	0.63	1.03	0.86	0.75	1.01
8-GLMM PsU-L	**0.29**	**0.08**	**0.80**	**0.21**	**0.07**	**0.44**	**0.99**	**0.96**	**1.00**	**0.99**	**0.98**	**1.00**
9-GLMM PsU-Q	**0.28**	**0.08**	**0.80**	**0.21**	**0.04**	**0.43**	**0.99**	**0.96**	**1.00**	**0.99**	**0.98**	**1.00**
10-GLMM PsU-P	**0.30**	**0.08**	**0.80**	**0.25**	**0.09**	**0.49**	**0.99**	**0.96**	**1.00**	**0.99**	**0.98**	**1.00**
11-GLMM NB-L	7.83	-13.28	25.49	11.13	2.00	32.27	0.73	0.37	1.50	0.61	0.25	0.92
12-GLMM NB-Q	7.80	-13.28	25.43	11.11	1.98	32.27	0.73	0.37	1.50	0.61	0.25	0.92
13-GLMM NB-P	7.75	-12.82	24.41	11.24	2.28	31.78	0.72	0.35	1.48	0.59	0.24	0.89
14-GLMM Gm-L	9.13	-12.98	29.44	12.22	3.63	32.27	0.69	0.34	1.35	0.57	0.23	0.86
15-GLMM Gm-Q	9.10	-12.98	29.29	12.18	3.58	32.24	0.69	0.34	1.35	0.57	0.23	0.86
16-GLMM Gm-P	9.12	-13.17	36.05	12.15	3.07	32.11	0.68	0.26	1.38	0.56	0.22	0.87

Table 7: Average, minimum, and maximum values for 97.5th Percentile Ratios from simulation scenarios with 10 or 50 blocks and unequal treatment means under a Poisson-Unit DGP

Results for the true GLMMs used for data generation are **bolded**

	10 Blocks, Unequal Treatment Means			50 Blocks, Unequal Treatment Means		
	Mean	**Min**	**Max**	**Mean**	**Min**	**Max**
1-LMM-hm	0.60	0.40	1.06	0.54	0.37	0.88
2-LMM-ht	0.60	0.33	1.10	0.59	0.31	1.16
3-LMM-l	0.03	0.01	0.09	0.02	0.01	0.04
4-LMM-s	0.06	0.03	0.14	0.05	0.03	0.07
5-GLMM Ps-L	0.95	0.70	1.53	0.94	0.65	1.33
6-GLMM Ps-Q	0.95	0.70	1.53	0.94	0.65	1.33
7-GLMM Ps-P	0.95	0.70	1.53	0.94	0.65	1.33
8-GLMM PsU-L	**0.99**	**0.97**	**1.01**	**0.99**	**0.98**	**1.01**
9-GLMM PsU-Q	**0.99**	**0.97**	**1.01**	**0.99**	**0.98**	**1.01**
10-GLMM PsU-P	**0.99**	**0.97**	**1.01**	**0.99**	**0.98**	**1.01**
11-GLMM NB-L	0.80	0.39	2.00	0.78	0.46	1.14
12-GLMM NB-Q	0.80	0.39	2.00	0.79	0.46	1.14
13-GLMM NB-P	0.80	0.37	1.88	0.76	0.44	1.11
14-GLMM Gm-L	0.76	0.36	1.95	0.75	0.45	1.13
15-GLMM Gm-Q	0.76	0.36	1.96	0.75	0.45	1.13
16-GLMM Gm-P	0.76	0.35	1.86	0.74	0.44	1.13

Table 8: Average, minimum, and maximum values for Skewness Ratios from simulation scenarios with 10 or 50 blocks and unequal treatment means under a Poisson-Unit DGP

Results for the true GLMMs used for data generation are bolded

	10 Blocks, Unequal Treatment Means			50 Blocks, Unequal Treatment Means		
	Mean	**Min**	**Max**	**Mean**	**Min**	**Max**
1-LMM-hm	0.30	-0.26	0.71	0.20	0.10	0.47
2-LMM-ht	0.29	0.02	0.50	0.21	0.09	0.49
3-LMM-l	-0.03	-0.49	0.27	-0.02	-0.15	0.06
4-LMM-s	0.20	-0.35	0.52	0.13	0.05	0.24
5-GLMM Ps-L	0.95	0.57	1.81	0.89	0.71	1.59
6-GLMM Ps-Q	0.95	0.57	1.81	0.89	0.71	1.59
7-GLMM Ps-P	0.95	0.57	1.81	0.89	0.71	1.59
8-GLMM PsU-L	**1.01**	**0.99**	**1.05**	**1.01**	**1.00**	**1.02**
9-GLMM PsU-Q	**1.01**	**0.99**	**1.05**	**1.01**	**1.00**	**1.02**
10-GLMM PsU-P	**1.01**	**0.99**	**1.05**	**1.01**	**1.00**	**1.02**
11-GLMM NB-L	0.85	0.19	2.42	0.70	0.30	1.62
12-GLMM NB-Q	0.85	0.19	2.42	0.71	0.30	1.62
13-GLMM NB-P	0.86	0.28	2.42	0.70	0.30	1.58
14-GLMM Gm-L	0.81	-0.18	2.37	0.67	0.28	1.47
15-GLMM Gm-Q	0.81	-0.18	2.37	0.67	0.28	1.47
16-GLMM Gm-P	0.81	-0.07	2.40	0.67	0.27	1.47

Table 9: Average, minimum, and maximum values for Coefficient of Variation Ratios from simulation scenarios with 10 or 50 blocks and unequal treatment means from a Poisson-Unit DGP

Results for the true GLMMs used for data generation are **bolded**

	10 Blocks, Unequal Treatment Means			**50 Blocks, Unequal Treatment Means**		
	Mean	**Min**	**Max**	**Mean**	**Min**	**Max**
1-LMM-hm	0.62	0.38	0.85	0.53	0.28	0.70
2-LMM-ht	0.61	0.32	0.82	0.52	0.28	0.76
3-LMM-l	0.35	0.15	0.70	0.28	0.13	0.42
4-LMM-s	0.35	0.22	0.50	0.29	0.14	0.38
5-GLMM Ps-L	0.90	0.73	1.03	0.87	0.78	1.01
6-GLMM Ps-Q	0.90	0.73	1.03	0.87	0.78	1.01
7-GLMM Ps-P	0.90	0.73	1.03	0.87	0.78	1.01
8-GLMM PsU-L	**0.99**	**0.96**	**1.00**	**0.99**	**0.98**	**1.00**
9-GLMM PsU-Q	**0.99**	**0.96**	**1.00**	**0.99**	**0.98**	**1.00**
10-GLMM PsU-P	**0.99**	**0.96**	**1.00**	**0.99**	**0.98**	**1.00**
11-GLMM NB-L	0.81	0.50	1.32	0.72	0.41	0.95
12-GLMM NB-Q	0.81	0.50	1.32	0.72	0.41	0.95
13-GLMM NB-P	0.81	0.50	1.32	0.71	0.40	0.95
14-GLMM Gm-L	0.77	0.48	1.25	0.69	0.39	0.91
15-GLMM Gm-Q	0.77	0.48	1.25	0.69	0.39	0.91
16-GLMM Gm-P	0.77	0.48	1.27	0.69	0.38	0.93

Table 10: Empirical Type I Error & Statistical Power, Poisson-Gamma DGP

Results for the true GLMMs used for data generation are **bolded**

	10 Blocks		**50 Blocks**	
	Type I Error	**Power**	**Type I Error**	**Power**
1-LMM-hm	0.03	0.77	0.07	0.99
2-LMM-ht	0.01	0.13	0.01	0.27
3-LMM-l	0.04	0.68	0.05	0.70
4-LMM-s	0.04	0.80	0.06	0.99
5-GLMM Ps-L	0.55	.	0.80	.
6-GLMM Ps-Q	0.55	.	0.80	.
7-GLMM Ps-P	0.55	.	0.80	.
8-GLMM PsU-L	0.05	0.70	0.05	0.69
9-GLMM PsU-Q	0.07	0.71	0.05	0.69
10-GLMM PsU-P	0.04	0.68	0.05	0.69
11-GLMM NB-L	**0.06**	**0.69**	**0.07**	**0.69**
12-GLMM NB-Q	**0.06**	**0.69**	**0.07**	**0.69**
13-GLMM NB-P	**0.07**	**0.69**	**0.08**	**0.71**
14-GLMM Gm-L	0.08	0.69	0.06	0.70
15-GLMM Gm-Q	0.08	0.69	0.06	0.70
16-GLMM Gm-P	0.10	0.70	0.09	0.71

Table 11: Summary of estimates of expectations of treatments and 95% confidence interval coverage for competing models fitted to data generated under the Poisson-Gamma DGP

Model	Blocks	Treatment levels (*ij* =)	True Mean	Avg Treatment Mean Estimate	Avg Diff from True Mean	Avg width of 95% CI for Means	95% CI Coverage
1-LMM-hm	10	1,1; 1,2	12.2	16.56	4.36	53.62	1.00
2-LMM-ht	10	1,1; 1,2	12.2	19.89	7.69	33.55	0.78
3-LMM-l	10	1,1; 1,2	12.2	8.95	-3.25	17.90	0.90
4-LMM-s	10	1,1; 1,2	12.2	12.72	0.52	23.53	1.00
5-GLMM-Ps-L	10	1,1; 1,2	12.2	11.54	-0.66	14.06	0.88
6-GLMM-Ps-Q	10	1,1; 1,2	12.2	11.54	-0.66	14.06	0.88
7-GLMM-Ps-P	10	1,1; 1,2	12.2	11.63	-0.57	14.86	0.91
8-GLMM-PsU-L	10	1,1; 1,2	12.2	10.03	-2.17	14.71	0.88
9-GLMM-PsU-Q	10	1,1; 1,2	12.2	10.03	-2.17	13.99	0.84
10-GLMM-PsU-P	10	1,1; 1,2	12.2	10.46	-1.74	16.16	0.93
11-GLMM-NB-L	**10**	**1,1; 1,2**	**12.2**	**12.60**	**0.40**	**18.44**	**0.96**
12-GLMM-NB-Q	**10**	**1,1; 1,2**	**12.2**	**12.60**	**0.40**	**18.44**	**0.96**
13-GLMM-NB-P	**10**	**1,1; 1,2**	**12.2**	**12.11**	**-0.09**	**17.98**	**0.95**
14-GLMM-Gm-L	10	1,1; 1,2	12.2	13.29	1.09	18.30	0.96
15-GLMM-Gm-Q	10	1,1; 1,2	12.2	13.28	1.08	18.32	0.96
16-GLMM-Gm-P	10	1,1; 1,2	12.2	12.64	0.44	17.59	0.95
1-LMM-hm	10	2,1	54.6	73.81	19.21	53.62	0.61
2-LMM-ht	10	2,1	54.6	28.26	-26.34	48.76	0.22
3-LMM-l	10	2,1	54.6	44.72	-9.88	90.95	0.96
4-LMM-s	10	2,1	54.6	58.50	3.90	50.12	0.76
5-GLMM-Ps-L	10	2,1	54.6	50.96	-3.64	60.92	0.86
6-GLMM-Ps-Q	10	2,1	54.6	50.96	-3.64	60.92	0.86
7-GLMM-Ps-P	10	2,1	54.6	51.34	-3.26	64.57	0.88
8-GLMM-PsU-L	10	2,1	54.6	45.68	-8.92	64.00	0.84
9-GLMM-PsU-Q	10	2,1	54.6	45.67	-8.93	64.60	0.84
10-GLMM-PsU-P	10	2,1	54.6	46.30	-8.30	68.34	0.89
11-GLMM-NB-L	**10**	**2,1**	**54.6**	**56.84**	**2.24**	**80.27**	**0.90**
12-GLMM-NB-Q	**10**	**2,1**	**54.6**	**56.83**	**2.23**	**80.32**	**0.90**
13-GLMM-NB-P	**10**	**2,1**	**54.6**	**53.37**	**-1.23**	**76.58**	**0.89**
14-GLMM-Gm-L	10	2,1	54.6	58.12	3.52	78.87	0.90
15-GLMM-Gm-Q	10	2,1	54.6	58.09	3.49	78.96	0.90
16-GLMM-Gm-P	10	2,1	54.6	55.30	0.70	76.21	0.89
1-LMM-hm	10	2,2	2.7	3.79	1.09	53.62	1.00
2-LMM-ht	10	2,2	2.7	4.09	1.39	43.62	0.92
3-LMM-l	10	2,2	2.7	1.97	-0.73	3.89	0.80
4-LMM-s	10	2,2	2.7	2.96	0.26	11.75	1.00
5-GLMM-Ps-L	10	2,2	2.7	2.68	-0.02	3.64	0.87
6-GLMM-Ps-Q	10	2,2	2.7	2.68	-0.02	3.64	0.87
7-GLMM-Ps-P	10	2,2	2.7	2.70	0.00	3.81	0.87
8-GLMM-PsU-L	10	2,2	2.7	2.45	-0.25	4.11	0.91
9-GLMM-PsU-Q	10	2,2	2.7	2.45	-0.25	4.15	0.90

10-GLMM-PsU-P	10	2,2	2.7	2.70	0.00	4.66	0.93
11-GLMM-NB-L	**10**	**2,2**	**2.7**	**3.02**	**0.32**	**4.94**	**0.91**
12-GLMM-NB-Q	**10**	**2,2**	**2.7**	**3.02**	**0.32**	**4.94**	**0.91**
13-GLMM-NB-P	**10**	**2,2**	**2.7**	**2.98**	**0.28**	**4.93**	**0.92**
14-GLMM-Gm-L	10	2,2	2.7	3.78	1.08	5.58	0.85
15-GLMM-Gm-Q	10	2,2	2.7	3.77	1.07	5.58	0.85
16-GLMM-Gm-P	10	2,2	2.7	3.60	0.90	5.36	0.87
1-LMM-hm	50	1,1; 1,2	12.2	16.14	3.9	28.5	1
2-LMM-ht	50	1,1; 1,2	12.2	61.32	49.1	725.4	0.86
3-LMM-l	50	1,1; 1,2	12.2	8.64	-3.6	6.8	0.58
4-LMM-s	50	1,1; 1,2	12.2	12.36	0.2	10.4	0.98
5-GLMM-Ps-L	50	1,1; 1,2	12.2	10.77	-1.4	5.7	0.73
6-GLMM-Ps-Q	50	1,1; 1,2	12.2	10.77	-1.4	5.7	0.73
7-GLMM-Ps-P	50	1,1; 1,2	12.2	10.87	-1.3	5.7	0.74
8-GLMM-PsU-L	50	1,1; 1,2	12.2	9.66	-2.5	6.1	0.68
9-GLMM-PsU-Q	50	1,1; 1,2	12.2	9.66	-2.5	6.1	0.68
10-GLMM-PsU-P	50	1,1; 1,2	12.2	10.20	-2.0	6.3	0.81
11-GLMM-NB-L	**50**	**1,1; 1,2**	**12.2**	**12.34**	**0.1**	**7.6**	**0.95**
12-GLMM-NB-Q	**50**	**1,1; 1,2**	**12.2**	**12.33**	**0.1**	**7.6**	**0.95**
13-GLMM-NB-P	**50**	**1,1; 1,2**	**12.2**	**11.84**	**-0.4**	**7.1**	**0.94**
14-GLMM-Gm-L	50	1,1; 1,2	12.2	12.95	0.7	7.5	0.91
15-GLMM-Gm-Q	50	1,1; 1,2	12.2	12.94	0.7	7.5	0.91
16-GLMM-Gm-P	50	1,1; 1,2	12.2	12.57	0.4	7.1	0.92
1-LMM-hm	50	2,1	54.6	75.23	20.6	28.5	0.23
2-LMM-ht	50	2,1	54.6	277.46	222.9	555.7	0.82
3-LMM-l	50	2,1	54.6	41.63	-13.0	32.6	0.78
4-LMM-s	50	2,1	54.6	56.85	2.3	22.1	0.83
5-GLMM-Ps-L	50	2,1	54.6	49.95	-4.6	25.6	0.88
6-GLMM-Ps-Q	50	2,1	54.6	49.95	-4.65	25.63	0.88
7-GLMM-Ps-P	50	2,1	54.6	50.41	-4.19	25.91	0.88
8-GLMM-PsU-L	50	2,1	54.6	43.02	-11.58	25.94	0.62
9-GLMM-PsU-Q	50	2,1	54.6	43.01	-11.59	26.02	0.62
10-GLMM-PsU-P	50	2,1	54.6	43.83	-10.77	25.83	0.68
11-GLMM-NB-L	**50**	**2,1**	**54.6**	**55.96**	**1.36**	**33.59**	**0.95**
12-GLMM-NB-Q	**50**	**2,1**	**54.6**	**55.95**	**1.35**	**33.62**	**0.95**
13-GLMM-NB-P	**50**	**2,1**	**54.6**	**52.87**	**-1.73**	**30.55**	**0.96**
14-GLMM-Gm-L	50	2,1	54.6	57.04	2.44	32.79	0.95
15-GLMM-Gm-Q	50	2,1	54.6	56.99	2.39	32.84	0.95
16-GLMM-Gm-P	50	2,1	54.6	55.98	1.38	31.18	0.95
1-LMM-hm	50	2,2	2.7	3.66	0.96	28.51	1
2-LMM-ht	50	2,2	2.7	13.33	10.63	540.85	1
3-LMM-l	50	2,2	2.7	1.63	-1.07	1.28	0.28
4-LMM-s	50	2,2	2.7	2.73	0.03	5.13	1
5-GLMM-Ps-L	50	2,2	2.7	2.44	-0.26	1.44	0.84
6-GLMM-Ps-Q	50	2,2	2.7	2.44	-0.26	1.44	0.84
7-GLMM-Ps-P	50	2,2	2.7	2.47	-0.23	1.46	0.86
8-GLMM-PsU-L	50	2,2	2.7	2.19	-0.51	1.57	0.78
9-GLMM-PsU-Q	50	2,2	2.7	2.20	-0.50	1.58	0.78

10-GLMM-PsU-P	50	2,2	2.7	2.49	-0.21	1.73	0.95
11-GLMM-NB-L	**50**	**2,2**	**2.7**	**2.80**	**0.10**	**1.93**	**0.97**
12-GLMM-NB-Q	**50**	**2,2**	**2.7**	**2.80**	**0.10**	**1.93**	**0.97**
13-GLMM-NB-P	**50**	**2,2**	**2.7**	**2.76**	**0.06**	**1.86**	**0.97**
14-GLMM-Gm-L	50	2,2	2.7	3.59	0.89	2.24	0.57
15-GLMM-Gm-Q	50	2,2	2.7	3.59	0.89	2.24	0.57
16-GLMM-Gm-P	50	2,2	2.7	3.47	0.77	2.09	0.74

Table 12: Average, minimum, and maximum values of Akaike's and Bayesian Fit Criteria under simulation scenarios of unequal treatment means with 10 or 50 blocks and a Poisson-Gamma DGP

Results for the true GLMMs used for data generation are **bolded**

	Akaike's Information Criterion						Bayesian Information Criterion					
	10 Blocks, Unequal Treatment Means			50 Blocks, Unequal Treatment Means			10 Blocks, Unequal Treatment Means			50 Blocks, Unequal Treatment Means		
	Mean	**Min**	**Max**	**Mean**	**Min**	**Max**	**Mean**	**Min**	**Max**	**Mean**	**Min**	**Max**
1-LMM-hm	372.5	305.4	446.3	2603.6	2282.4	3375.0	373.1	305.7	446.9	2607.5	2286.2	3378.8
2-LMM-ht	333.2	198.2	445.1	2872.9	2278.7	3195.3	334.7	199.7	446.7	2882.5	2288.2	3204.9
5-GLMM Ps-L	448.5	291.2	699.8	6390.1	4324.6	9997.6	450.0	292.7	701.3	6399.7	4334.1	10007.2
6-GLMM Ps-Q	448.5	291.2	699.8	6390.1	4324.6	9997.6	450.0	292.7	701.3	6399.7	4334.1	10007.2
8-GLMM PsU-L	303.3	248.6	348.5	2018.9	1888.4	2162.8	305.2	250.4	350.3	2030.3	1899.8	2174.3
9-GLMM PsU-Q	303.1	248.4	348.2	2018.9	1888.4	2162.8	304.9	250.2	350.1	2030.3	1899.8	2174.3
11-GLMM NB-L	**302.4**	**249.8**	**345.8**	**2013.3**	**1889.6**	**2154.3**	**304.2**	**251.6**	**347.6**	**2024.8**	**1901.1**	**2165.7**
12-GLMM NB-Q	**302.3**	**249.8**	**345.8**	**2013.4**	**1889.6**	**2154.4**	**304.1**	**251.6**	**347.6**	**2024.9**	**1901.0**	**2165.9**
14-GLMM Gm-L	288.7	234.1	336.5	1995.7	1855.2	2140.5	290.5	236.0	338.3	2007.2	1866.7	2152.0
15-GLMM Gm-Q	288.7	234.1	336.5	1995.5	1854.9	2140.3	290.5	235.9	338.3	2007.0	1866.4	2151.8

Table 13: Average, minimum, and maximum values for point estimates of MAR coefficients, both intercept and slope, in simulation scenarios of unequal treatment means with 10 or 50 blocks under a Poisson-Gamma DGP

Results for the true GLMMs used for data generation are **bolded**

	Major Axis Regression (Data Scale) Intercept Coefficients						Major Axis Regression (Data Scale) Slope Coefficients					
	10 Blocks, Unequal Treatment Means			50 Blocks, Unequal Treatment Means			10 Blocks, Unequal Treatment Means			50 Blocks, Unequal Treatment Means		
	Mean	**Min**	**Max**	**Mean**	**Min**	**Max**	**Mean**	**Min**	**Max**	**Mean**	**Min**	**Max**
1-LMM-hm	12.32	4.00	26.47	60.59	25.69	168.96	0.55	0.25	0.86	0.46	0.12	0.70
2-LMM-ht	13.27	4.39	27.58	64.15	25.77	179.98	0.60	0.29	0.88	0.43	0.06	0.70
3-LMM-l	4.43	-8.09	12.48	27.44	0.72	60.71	0.56	0.17	1.13	0.47	0.13	0.83
4-LMM-s	9.59	1.56	18.62	46.26	20.52	93.58	0.50	0.22	0.87	0.41	0.12	0.65
5-GLMM Ps-L	2.45	-2.36	6.56	12.06	-1.71	24.97	0.90	0.74	1.06	0.89	0.78	1.02
6-GLMM Ps-Q	2.45	-2.36	6.56	12.06	-1.71	24.97	0.90	0.74	1.06	0.89	0.78	1.02
7-GLMM Ps-P	2.43	-2.37	6.54	12.06	-1.71	24.97	0.90	0.74	1.06	0.89	0.78	1.02
8-GLMM PsU-L	0.27	0.03	0.56	0.23	0.05	0.42	0.99	0.96	1.00	1.00	1.00	1.00
9-GLMM PsU-Q	0.27	0.03	0.56	0.23	0.05	0.42	0.99	0.96	1.00	1.00	1.00	1.00
10-GLMM PsU-P	0.29	0.05	0.57	0.25	0.05	0.44	0.99	0.96	1.00	1.00	1.00	1.00
11-GLMM NB-L	**5.67**	**-7.38**	**17.51**	**30.08**	**-4.26**	**69.54**	**0.72**	**0.33**	**1.33**	**0.64**	**0.26**	**1.03**
12-GLMM NB-Q	**5.66**	**-7.38**	**17.51**	**30.03**	**-4.36**	**69.52**	**0.72**	**0.33**	**1.33**	**0.65**	**0.27**	**1.03**
13-GLMM NB-P	**5.44**	**-6.89**	**16.85**	**29.06**	**-6.14**	**68.78**	**0.72**	**0.34**	**1.28**	**0.65**	**0.26**	**1.05**
14-GLMM Gm-L	6.95	-3.42	17.51	31.07	-1.32	70.09	0.67	0.33	1.19	0.63	0.26	0.98
15-GLMM Gm-Q	6.93	-3.42	17.51	30.97	-1.32	70.04	0.67	0.33	1.20	0.63	0.26	0.98
16-GLMM Gm-P	6.59	-3.79	16.42	29.69	-4.64	69.27	0.68	0.33	1.17	0.64	0.26	1.03

Table 14: Empirical Type I Error & Statistical Power, Additive Means DGP

	10 Blocks			50 Blocks		
	Convergence Failures	**Type I Error**	**Power**	**Convergence Failures**	**Type I Error**	**Power**
1-LMM-hm	0	0.06	0.91	0	0.05	1.00
2-LMM-ht	13	0.03	0.45	0	0.05	0.83
3-LMM-l	0	0.04	0.66	0	0.03	0.69
4-LMM-s	0	0.05	0.84	0	0.05	1.00
5-GLMM Ps-L	0	0.45	.	0	0.50	.
6-GLMM Ps-Q	0	0.45	.	0	0.50	.
7-GLMM Ps-P	0	0.45	.	0	0.49	.
8-GLMM PsU-L	61	0.08	.	62	0.06	.
9-GLMM PsU-Q	80	0.08	.	81	0.07	.
10-GLMM PsU-P	0	0.05	0.69	1	0.05	0.70
11-GLMM NB-L	0	0.07	0.69	0	0.07	0.70
12-GLMM NB-Q	0	0.07	0.69	0	0.07	0.70
13-GLMM NB-P	6	0.07	0.66	22	0.06	0.71
14-GLMM Gm-L	0	0.07	0.71	0	0.08	0.74
15-GLMM Gm-Q	0	0.07	0.71	0	0.08	0.74
16-GLMM Gm-P	2	0.07	0.71	1	0.07	0.73

Table 15: Summary of estimates of expectations of treatments and 95% confidence interval coverage for competing models fitted to data generated under the Additive Means DGP

Model	Blocks	Treatment levels (*ij* =)	True Mean	Avg Treatment Mean Estimate	Avg Diff from True Mean	Avg width of 95% CI for Means	95% CI Coverage
1-LMM-hm	10	1,1; 1,2	12.0	11.92	-0.08	29.65	1.00
2-LMM-ht	10	1,1; 1,2	12.0	13.29	1.29	13.91	0.89
3-LMM-l	10	1,1; 1,2	12.0	7.99	-4.01	13.32	0.85
4-LMM-s	10	1,1; 1,2	12.0	10.12	-1.88	14.62	0.95
5-GLMM-Ps-L	10	1,1; 1,2	12.0	10.50	-1.50	8.08	0.76
6-GLMM-Ps-Q	10	1,1; 1,2	12.0	10.50	-1.50	8.08	0.76
7-GLMM-Ps-P	10	1,1; 1,2	12.0	10.56	-1.44	8.45	0.78
8-GLMM-PsU-L	10	1,1; 1,2	12.0	8.97	-3.03	5.15	0.40
9-GLMM-PsU-Q	10	1,1; 1,2	12.0	8.96	-3.04	2.68	0.22
10-GLMM-PsU-P	10	1,1; 1,2	12.0	9.35	-2.65	11.20	0.87
11-GLMM-NB-L	10	1,1; 1,2	12.0	11.66	-0.34	12.90	0.92
12-GLMM-NB-Q	10	1,1; 1,2	12.0	11.66	-0.34	12.90	0.92
13-GLMM-NB-P	10	1,1; 1,2	12.0	12.00	0.00	13.32	0.91
14-GLMM-Gm-L	10	1,1; 1,2	12.0	12.05	0.05	12.70	0.94
15-GLMM-Gm-Q	10	1,1; 1,2	12.0	12.05	0.05	12.70	0.94
16-GLMM-Gm-P	10	1,1; 1,2	12.0	11.96	-0.04	13.00	0.94
1-LMM-hm	10	2,1	55.0	57.35	2.35	29.65	0.77
2-LMM-ht	10	2,1	55.0	39.78	-15.22	40.87	0.60
3-LMM-l	10	2,1	55.0	43.07	-11.93	72.89	0.95
4-LMM-s	10	2,1	55.0	50.27	-4.73	32.11	0.78
5-GLMM-Ps-L	10	2,1	55.0	50.31	-4.69	34.99	0.84
6-GLMM-Ps-Q	10	2,1	55.0	50.31	-4.69	34.99	0.84
7-GLMM-Ps-P	10	2,1	55.0	50.60	-4.40	36.92	0.88
8-GLMM-PsU-L	10	2,1	55.0	43.98	-11.02	27.59	0.52
9-GLMM-PsU-Q	10	2,1	55.0	43.96	-11.04	14.36	0.26
10-GLMM-PsU-P	10	2,1	55.0	44.49	-10.51	49.04	0.89
11-GLMM-NB-L	10	2,1	55.0	56.63	1.63	58.44	0.93
12-GLMM-NB-Q	10	2,1	55.0	56.63	1.63	58.45	0.93
13-GLMM-NB-P	10	2,1	55.0	51.36	-3.64	53.16	0.84
14-GLMM-Gm-L	10	2,1	55.0	56.72	1.72	58.53	0.96
15-GLMM-Gm-Q	10	2,1	55.0	56.72	1.72	58.53	0.96
16-GLMM-Gm-P	10	2,1	55.0	56.12	1.12	59.41	0.96
1-LMM-hm	10	2,2	3.0	2.96	-0.04	29.65	1.00
2-LMM-ht	10	2,2	3.0	6.66	3.66	7.36	0.60
3-LMM-l	10	2,2	3.0	1.69	-1.31	2.76	0.58
4-LMM-s	10	2,2	3.0	2.41	-0.59	7.46	1.00
5-GLMM-Ps-L	10	2,2	3.0	2.61	-0.39	2.70	0.82
6-GLMM-Ps-Q	10	2,2	3.0	2.61	-0.39	2.70	0.82
7-GLMM-Ps-P	10	2,2	3.0	2.62	-0.38	2.78	0.82
8-GLMM-PsU-L	10	2,2	3.0	2.24	-0.76	2.34	0.61

9-GLMM-PsU-Q	10	2,2	3.0	2.24	-0.76	1.27	0.31
10-GLMM-PsU-P	10	2,2	3.0	2.48	-0.52	3.55	0.93
11-GLMM-NB-L	10	2,2	3.0	2.90	-0.10	3.83	0.93
12-GLMM-NB-Q	10	2,2	3.0	2.90	-0.10	3.83	0.93
13-GLMM-NB-P	10	2,2	3.0	3.78	0.78	4.72	0.82
14-GLMM-Gm-L	10	2,2	3.0	3.55	0.55	4.11	0.92
15-GLMM-Gm-Q	10	2,2	3.0	3.55	0.55	4.11	0.92
16-GLMM-Gm-P	10	2,2	3.0	3.51	0.51	4.17	0.93
1-LMM-hm	50	1,1; 1,2	12.0	12.08	0.08	11.78	1.00
2-LMM-ht	50	1,1; 1,2	12.0	12.68	0.68	5.72	0.83
3-LMM-l	50	1,1; 1,2	12.0	8.21	-3.79	5.36	0.32
4-LMM-s	50	1,1; 1,2	12.0	10.30	-1.70	6.05	0.84
5-GLMM-Ps-L	50	1,1; 1,2	12.0	10.67	-1.33	3.50	0.57
6-GLMM-Ps-Q	50	1,1; 1,2	12.0	10.67	-1.33	3.50	0.57
7-GLMM-Ps-P	50	1,1; 1,2	12.0	10.75	-1.25	3.53	0.60
8-GLMM-PsU-L	50	1,1; 1,2	12.0	9.14	-2.86	2.16	0.21
9-GLMM-PsU-Q	50	1,1; 1,2	12.0	9.14	-2.86	0.98	0.10
10-GLMM-PsU-P	50	1,1; 1,2	12.0	9.61	-2.39	4.50	0.52
11-GLMM-NB-L	50	1,1; 1,2	12.0	11.98	-0.02	5.45	0.92
12-GLMM-NB-Q	50	1,1; 1,2	12.0	11.98	-0.02	5.45	0.92
13-GLMM-NB-P	50	1,1; 1,2	12.0	13.03	1.03	6.03	0.83
14-GLMM-Gm-L	50	1,1; 1,2	12.0	12.24	0.24	5.31	0.91
15-GLMM-Gm-Q	50	1,1; 1,2	12.0	12.24	0.24	5.31	0.91
16-GLMM-Gm-P	50	1,1; 1,2	12.0	12.20	0.20	5.39	0.91
1-LMM-hm	50	2,1	55.0	54.51	-0.49	11.78	0.70
2-LMM-ht	50	2,1	55.0	47.78	-7.22	19.74	0.80
3-LMM-l	50	2,1	55.0	40.60	-14.40	26.50	0.56
4-LMM-s	50	2,1	55.0	47.77	-7.23	12.87	0.41
5-GLMM-Ps-L	50	2,1	55.0	48.17	-6.83	14.21	0.53
6-GLMM-Ps-Q	50	2,1	55.0	48.17	-6.83	14.21	0.53
7-GLMM-Ps-P	50	2,1	55.0	48.51	-6.49	14.35	0.56
8-GLMM-PsU-L	50	2,1	55.0	41.68	-13.32	9.95	0.07
9-GLMM-PsU-Q	50	2,1	55.0	41.66	-13.34	4.66	0.04
10-GLMM-PsU-P	50	2,1	55.0	42.32	-12.68	18.46	0.30
11-GLMM-NB-L	50	2,1	55.0	54.16	-0.84	23.25	0.96
12-GLMM-NB-Q	50	2,1	55.0	54.16	-0.84	23.25	0.96
13-GLMM-NB-P	50	2,1	55.0	42.87	-12.13	18.91	0.70
14-GLMM-Gm-L	50	2,1	55.0	54.14	-0.86	23.10	0.95
15-GLMM-Gm-Q	50	2,1	55.0	54.14	-0.86	23.10	0.95
16-GLMM-Gm-P	50	2,1	55.0	53.96	-1.04	23.61	0.94
1-LMM-hm	50	2,2	3.0	2.96	-0.04	11.78	1.00
2-LMM-ht	50	2,2	3.0	4.54	1.54	2.32	0.80
3-LMM-l	50	2,2	3.0	1.53	-1.47	1.00	0.04
4-LMM-s	50	2,2	3.0	2.36	-0.64	3.06	0.99
5-GLMM-Ps-L	50	2,2	3.0	2.62	-0.38	1.14	0.65
6-GLMM-Ps-Q	50	2,2	3.0	2.62	-0.38	1.14	0.65
7-GLMM-Ps-P	50	2,2	3.0	2.64	-0.36	1.15	0.67
8-GLMM-PsU-L	50	2,2	3.0	2.20	-0.80	0.80	0.24

9-GLMM-PsU-Q	50	2,2	3.0	2.20	-0.80	0.41	0.15
10-GLMM-PsU-P	50	2,2	3.0	2.48	-0.52	1.39	0.69
11-GLMM-NB-L	50	2,2	3.0	2.93	-0.07	1.58	0.93
12-GLMM-NB-Q	50	2,2	3.0	2.93	-0.07	1.58	0.93
13-GLMM-NB-P	50	2,2	3.0	5.59	2.59	2.79	0.69
14-GLMM-Gm-L	50	2,2	3.0	3.63	0.63	1.73	0.68
15-GLMM-Gm-Q	50	2,2	3.0	3.63	0.63	1.73	0.68
16-GLMM-Gm-P	50	2,2	3.0	3.62	0.62	1.77	0.70

Table 16: Average, minimum and maximum values of the Pearson chi-sq/df statistic under simulation scenario of unequal treatment means with 10 or 50 blocks and an Additive Means DGP

	10 Blocks, Unequal Treatment Means			50 Blocks, Unequal Treatment Means		
	Mean	**Min**	**Max**	**Mean**	**Min**	**Max**
5-GLMM Ps-L	6.06	2.27	11.33	6.21	4.52	9.88
6-GLMM Ps-Q	6.06	2.27	11.33	6.21	4.52	9.88
8-GLMM PsU-L	0.23	0.14	0.38	0.22	0.17	0.27
9-GLMM PsU-Q	0.23	0.14	0.38	0.22	0.17	0.26
11-GLMM NB-L	0.91	0.61	1.38	0.99	0.78	1.21
12-GLMM NB-Q	0.91	0.61	1.38	0.99	0.78	1.21
14-GLMM Gm-L	0.49	0.17	0.92	0.54	0.35	0.78
15-GLMM Gm-Q	0.49	0.17	0.92	0.54	0.35	0.78

Table 17: Average, minimum, and maximum values of Akaike's and Bayesian Fit Criteria under simulation scenarios with 10 or 50 blocks and an Additive Means DGP

	Akaike's Information Criterion						Bayesian Information Criterion					
	10 Blocks, Unequal Treatment Means			50 Blocks, Unequal Treatment Means			10 Blocks, Unequal Treatment Means			50 Blocks, Unequal Treatment Means		
	Mean	Min	Max	Mean	Min	Max	Mean	Min	Max	Mean	Min	Max
1-LMM-hm	275.4	219.1	321.4	1478.7	1374.7	1580.4	275.9	219.4	322.0	1482.4	1378.5	1584.3
2-LMM-ht	271.8	242.6	336.9	1455.4	1362.9	1718.0	273.2	244.1	338.1	1464.8	1372.5	1727.6
5-GLMM Ps-L	350.3	238.4	478.9	1862.2	1608.0	2170.1	351.8	239.9	480.5	1871.7	1617.6	2179.7
6-GLMM Ps-Q	350.3	238.4	478.9	1862.2	1608.0	2170.1	351.8	239.9	480.5	1871.7	1617.6	2179.7
8-GLMM PsU-L	273.9	227.6	318.3	1362.4	1298.4	1439.2	275.7	229.2	320.1	1373.8	1309.9	1450.7
9-GLMM PsU-Q	273.5	227.7	318.0	1361.0	1297.3	1437.8	275.2	229.2	319.8	1372.2	1308.8	1449.3
11-GLMM NB-L	272.8	227.8	315.2	1356.9	1293.8	1427.5	274.5	229.3	317.1	1368.3	1305.2	1439.0
12-GLMM NB-Q	272.7	227.8	315.3	1356.9	1293.7	1427.5	274.5	229.3	317.1	1368.3	1305.2	1439.0
14-GLMM Gm-L	260.2	222.8	307.3	1293.0	1220.1	1378.1	261.9	224.3	309.1	1304.4	1231.6	1389.6
15-GLMM Gm-Q	260.2	222.8	307.3	1293.1	1220.3	1378.1	261.9	224.3	309.1	1304.5	1231.8	1389.6
	10 Blocks, Equal Treatment Means			50 Blocks, Equal Treatment Means			10 Blocks, Equal Treatment Means			50 Blocks, Equal Treatment Means		
1-LMM-hm	274.7	235.2	315.0	1472.9	1401.1	1556.2	275.1	235.5	315.3	1475.5	1403.0	1560.0
2-LMM-ht	276.2	215.5	328.5	1475.7	1403.3	1721.5	277.5	217.0	329.7	1484.1	1410.9	1729.2
5-GLMM Ps-L	382.6	264.1	562.0	2011.9	1798.6	2256.5	384.1	265.7	563.5	2021.5	1808.2	2266.1
6-GLMM Ps-Q	382.6	264.1	562.0	2011.9	1798.6	2256.5	384.1	265.7	563.5	2021.5	1808.2	2266.1
8-GLMM PsU-L	280.2	234.1	308.4	1395.7	1336.0	1451.5	281.8	235.9	309.9	1406.0	1345.6	1461.1
9-GLMM PsU-Q	280.7	246.6	308.0	1394.2	1334.9	1450.3	282.3	248.1	309.5	1404.2	1344.4	1459.8
11-GLMM NB-L	280.1	244.7	307.0	1389.8	1332.6	1439.8	281.7	246.2	308.5	1400.3	1342.2	1450.7
12-GLMM NB-Q	280.1	244.7	307.0	1389.8	1332.6	1439.8	281.7	246.2	308.5	1400.3	1342.2	1450.7
14-GLMM Gm-L	271.5	236.2	305.0	1344.6	1276.7	1404.0	273.1	237.7	306.5	1355.1	1286.2	1415.4
15-GLMM Gm-Q	271.5	236.2	305.0	1344.6	1276.7	1404.0	273.1	237.7	306.5	1355.1	1286.2	1415.4

Table 18: Average, minimum, and maximum values for point estimates of MAR coefficients, both intercept and slope, in simulation scenarios of unequal treatment means with 10 or 50 blocks under an Additive Means DGP

	Major Axis Regression (Data Scale) Intercept Coefficients						Major Axis Regression (Data Scale) Slope Coefficients					
	10 Blocks, Unequal Treatment Means			50 Blocks, Unequal Treatment Means			10 Blocks, Unequal Treatment Means			50 Blocks, Unequal Treatment Means		
	Mean	Min	Max	Mean	Min	Max	Mean	Min	Max	Mean	Min	Max
1-LMM-hm	8.02	2.81	15.24	8.13	4.93	11.42	0.61	0.26	0.89	0.60	0.44	0.74
2-LMM-ht	7.88	2.03	15.19	8.13	4.91	11.42	0.60	0.00	0.89	0.60	0.44	0.74
3-LMM-l	5.38	-0.63	10.77	5.92	3.20	8.26	0.47	0.12	0.86	0.43	0.26	0.63
4-LMM-s	6.94	2.07	13.19	7.23	4.25	9.99	0.53	0.18	0.86	0.51	0.34	0.68
5-GLMM Ps-L	3.31	1.01	6.33	3.36	2.34	4.89	0.84	0.67	0.95	0.83	0.76	0.89
6-GLMM Ps-Q	3.31	1.01	6.33	3.36	2.34	4.89	0.84	0.67	0.95	0.83	0.76	0.89
7-GLMM Ps-P	3.28	0.98	6.30	3.36	2.34	4.89	0.84	0.67	0.95	0.83	0.76	0.89
8-GLMM PsU-L	0.33	0.17	0.60	0.33	0.24	0.47	0.98	0.96	0.99	0.98	0.97	0.99
9-GLMM PsU-Q	0.33	0.17	0.60	0.33	0.24	0.47	0.98	0.96	0.99	0.98	0.97	0.99
10-GLMM PsU-P	0.38	0.25	0.64	0.43	0.32	0.56	0.98	0.96	0.99	0.98	0.97	0.99
11-GLMM NB-L	7.58	1.83	14.43	7.96	4.82	11.42	0.62	0.26	0.90	0.60	0.44	0.74
12-GLMM NB-Q	7.58	1.83	14.43	7.96	4.83	11.42	0.62	0.26	0.90	0.60	0.44	0.74
13-GLMM NB-P	7.34	1.93	14.43	8.18	5.65	11.42	0.63	0.26	0.89	0.59	0.44	0.73
14-GLMM Gm-L	8.08	2.83	14.83	8.37	5.11	11.77	0.62	0.27	0.89	0.60	0.43	0.75
15-GLMM Gm-Q	8.09	2.83	14.83	8.37	5.11	11.77	0.62	0.27	0.89	0.60	0.43	0.75
16-GLMM Gm-P	8.05	2.38	15.62	8.44	5.26	11.77	0.61	0.26	0.89	0.59	0.43	0.75

Table 19: Average, minimum, and maximum values for 97.5th Percentile Ratios from simulation scenarios with 10 or 50 blocks and unequal treatment means under an Additive Means DGP

	10 Blocks, Unequal Treatment Means			50 Blocks, Unequal Treatment Means		
	Mean	**Min**	**Max**	**Mean**	**Min**	**Max**
1-LMM-hm	0.59	0.34	0.86	0.52	0.43	0.64
2-LMM-ht	0.58	0.05	0.83	0.52	0.43	0.63
3-LMM-l	0.04	0.02	0.07	0.04	0.03	0.05
4-LMM-s	0.07	0.04	0.12	0.07	0.05	0.09
5-GLMM Ps-L	0.86	0.65	1.18	0.85	0.66	1.11
6-GLMM Ps-Q	0.86	0.65	1.18	0.85	0.66	1.11
7-GLMM Ps-P	0.86	0.65	1.18	0.85	0.66	1.11
8-GLMM PsU-L	0.99	0.96	1.00	0.99	0.98	0.99
9-GLMM PsU-Q	0.99	0.96	1.00	0.99	0.98	0.99
10-GLMM PsU-P	0.99	0.96	1.00	0.98	0.98	0.99
11-GLMM NB-L	0.62	0.34	0.97	0.54	0.43	0.65
12-GLMM NB-Q	0.62	0.34	0.97	0.54	0.43	0.65
13-GLMM NB-P	0.61	0.34	0.95	0.53	0.43	0.65
14-GLMM Gm-L	0.62	0.34	0.91	0.54	0.43	0.67
15-GLMM Gm-Q	0.62	0.34	0.91	0.54	0.43	0.67
16-GLMM Gm-P	0.62	0.34	0.94	0.53	0.43	0.65

Table 20: Average, minimum, and maximum values for Skewness Ratios from simulation scenarios with 10 or 50 blocks and unequal treatment means under an Additive Means DGP

	10 Blocks, Unequal Treatment Means			50 Blocks, Unequal Treatment Means		
	Mean	**Min**	**Max**	**Mean**	**Min**	**Max**
1-LMM-hm	0.43	0.20	0.88	0.39	0.22	0.54
2-LMM-ht	0.43	0.20	0.88	0.39	0.24	0.54
3-LMM-l	0.02	-0.41	0.49	-0.02	-0.20	0.20
4-LMM-s	0.31	0.04	0.70	0.28	0.14	0.44
5-GLMM Ps-L	0.83	0.60	1.19	0.80	0.66	0.96
6-GLMM Ps-Q	0.83	0.60	1.19	0.80	0.66	0.96
7-GLMM Ps-P	0.83	0.60	1.19	0.80	0.66	0.96
8-GLMM PsU-L	1.01	0.98	1.04	1.01	1.00	1.02
9-GLMM PsU-Q	1.01	0.98	1.04	1.01	1.00	1.02
10-GLMM PsU-P	1.01	0.99	1.04	1.01	1.00	1.02
11-GLMM NB-L	0.47	0.20	0.99	0.40	0.22	0.58
12-GLMM NB-Q	0.47	0.20	0.99	0.40	0.22	0.58
13-GLMM NB-P	0.48	0.20	0.99	0.40	0.22	0.63
14-GLMM Gm-L	0.47	0.21	0.92	0.41	0.23	0.57
15-GLMM Gm-Q	0.47	0.21	0.92	0.41	0.23	0.57
16-GLMM Gm-P	0.47	0.20	0.97	0.40	0.23	0.58

Table 21: Average, minimum, and maximum values for Coefficient of Variation Ratios from simulation scenarios with 10 or 50 blocks and unequal treatment means from an Additive Means DGP

	10 Blocks, Unequal Treatment Means			50 Blocks, Unequal Treatment Means		
	Mean	**Min**	**Max**	**Mean**	**Min**	**Max**
1-LMM-hm	0.70	0.47	0.90	0.69	0.59	0.78
2-LMM-ht	0.70	0.47	0.90	0.69	0.59	0.78
3-LMM-l	0.41	0.23	0.64	0.40	0.32	0.51
4-LMM-s	0.37	0.23	0.50	0.36	0.31	0.42
5-GLMM Ps-L	0.85	0.72	0.95	0.85	0.78	0.89
6-GLMM Ps-Q	0.85	0.72	0.95	0.85	0.78	0.89
7-GLMM Ps-P	0.85	0.72	0.96	0.85	0.78	0.89
8-GLMM PsU-L	0.98	0.96	0.99	0.98	0.97	0.99
9-GLMM PsU-Q	0.98	0.96	0.99	0.98	0.97	0.99
10-GLMM PsU-P	0.98	0.96	0.99	0.98	0.97	0.99
11-GLMM NB-L	0.71	0.47	0.93	0.70	0.59	0.79
12-GLMM NB-Q	0.71	0.47	0.93	0.70	0.59	0.79
13-GLMM NB-P	0.72	0.47	0.92	0.69	0.59	0.78
14-GLMM Gm-L	0.70	0.44	0.89	0.68	0.58	0.78
15-GLMM Gm-Q	0.70	0.44	0.89	0.68	0.58	0.78
16-GLMM Gm-P	0.70	0.46	0.90	0.68	0.58	0.78

Table 22: Pearson chi-sq/df statistic, AIC, and BIC, proof-of-concept data application

	Pearson chi-sq/df	AIC	BIC
1-LMM-hm	---	242.2	242.4
2-LMM-ht	---	233.6	234.1
5-GLMM Ps-L	4.28	297.9	298.4
6-GLMM Ps-Q	4.28	297.9	298.4
8-GLMM PsU-L	0.26	232.8	233.4
9-GLMM PsU-Q	0.26	232.4	233.0
11-GLMM NB-L	0.61	229.3	229.8
12-GLMM NB-Q	0.61	229.3	229.8
14-GLMM Gm-L	0.31	190.1	190.6
15-GLMM Gm-Q	0.31	190.1	190.6

Table 23: Estimates and 95% confidence interval bounds for MAR coefficients from the proof-of-concept data application

	Intercept			Slope		
	Est	95% Lower Bound	95% Upper Bound	Est	95% Lower Bound	95% Upper Bound
1-LMM-hm	3.87	2.21	5.30	0.58	0.42	0.76
2-LMM-ht	5.38	3.97	6.65	0.41	0.28	0.57
3-LMM-l	1.51	0.56	2.41	0.33	0.23	0.43
4-LMM-s	2.93	1.84	3.93	0.49	0.38	0.61
8-GLMM-PsU-L	0.29	0.12	0.45	0.97	0.95	0.98
9-GLMM-PsU-Q	0.28	0.12	0.44	0.97	0.95	0.98
10-GLMM-PsU-P	0.30	0.15	0.46	0.97	0.95	0.98
11-GLMM-NB-L	4.09	3.05	5.04	0.45	0.34	0.56
12-GLMM-NB-Q	4.07	3.02	5.02	0.45	0.35	0.56
13-GLMM-NB-P	3.50	2.40	4.50	0.51	0.40	0.63
14-GLMM-Gm-L	6.10	4.99	7.12	0.45	0.34	0.57
15-GLMM-Gm-Q	6.09	4.97	7.12	0.45	0.34	0.57
16-GLMM-Gm-P	5.93	4.79	6.97	0.46	0.34	0.58

Table 24: 97.5th Percentile, Skewness, and Coefficient of Variation Ratios from the proof-of-concept data application

	97.5th Percentile Ratio	Skewness Ratio	Coefficient of Variation Ratio
1-LMM-hm	0.728	0.373	0.648
2-LMM-ht	0.584	0.229	0.515
3-LMM-l	0.475	0.966	0.822
4-LMM-s	0.690	0.651	0.670
8-GLMM PsU-L	0.991	1.065	0.969
9-GLMM PsU-Q	0.991	1.064	0.969
10-GLMM PsU-P	0.991	1.061	0.967
11-GLMM NB-L	0.693	0.755	0.565
12-GLMM NB-Q	0.695	0.759	0.568
13-GLMM NB-P	0.755	0.847	0.627
14-GLMM Gm-L	0.642	0.650	0.459
15-GLMM Gm-Q	0.643	0.650	0.461
16-GLMM Gm-P	0.645	0.654	0.472

Table 25: F-test results for treatment effects of competing models from the proof-of-concept data application at α=0.05

	Effect	numDF	denDF	F-Value	P-value
1-LMM-hm	Trt	4	23	2.62	0.061
2-LMM-ht	Trt	4	23	1.60	0.208
3-LMM-l	Trt	4	23	1.61	0.205
4-LMM-s	Trt	4	23	2.42	0.078
8-GLMM-PsU-L	Trt	4	23	1.91	0.143
9-GLMM-PsU-Q	Trt	4	23	1.88	0.147
10-GLMM-PsU-P	Trt	4	23	1.66	0.194
11-GLMM-NB-L	Trt	4	23	1.83	0.157
12-GLMM-NB-Q	Trt	4	23	1.84	0.156
13-GLMM-NB-P	Trt	4	23	2.16	0.106
14-GLMM-Gm-L	Trt	4	15	1.80	0.181
15-GLMM-Gm-Q	Trt	4	15	1.80	0.181
16-GLMM-Gm-P	Trt	4	15	1.64	0.215

Table 26: Treatment mean estimates of LMM-hm, LMM-ht, LMM-l, GLMM-PsU, GLMM-NB, and GLMM-Gm from the proof-of-concept data application

	Trt	Estimate	95% Lower Bound	95% Upper Bound
1-LMM-hm	A	4.52	-4.78	13.83
2-LMM-ht	A	5.10	0.48	9.73
3-LMM-l	A	1.18	0.25	5.54
8-GLMM-PsU-L	A	2.28	0.72	7.20
9-GLMM-PsU-Q	A	2.27	0.71	7.25
10-GLMM-PsU-P	A	2.60	0.82	8.20
11-GLMM-NB-L	A	3.89	1.35	11.22
12-GLMM-NB-Q	A	3.88	1.34	11.19
13-GLMM-NB-P	A	3.51	1.33	9.27
14-GLMM-Gm-L	A	6.42	2.60	15.87
15-GLMM-Gm-Q	A	6.41	2.59	15.84
16-GLMM-Gm-P	A	6.04	2.40	15.22
1-LMM-hm	B	4.24	-5.07	13.54
2-LMM-ht	B	4.82	0.12	9.52
3-LMM-l	B	2.52	0.54	11.85
8-GLMM-PsU-L	B	3.06	1.04	9.01
9-GLMM-PsU-Q	B	3.06	1.03	9.06
10-GLMM-PsU-P	B	3.44	1.15	10.32
11-GLMM-NB-L	B	4.59	1.66	12.68
12-GLMM-NB-Q	B	4.58	1.66	12.65
13-GLMM-NB-P	B	4.29	1.65	11.16
14-GLMM-Gm-L	B	5.11	2.40	10.86
15-GLMM-Gm-Q	B	5.10	2.40	10.87
16-GLMM-Gm-P	B	4.84	2.20	10.62
1-LMM-hm	C	16.14	6.85	25.42
2-LMM-ht	C	15.92	2.05	29.79
3-LMM-l	C	5.65	1.20	26.52
8-GLMM-PsU-L	C	7.56	2.68	21.32
9-GLMM-PsU-Q	C	7.55	2.66	21.44
10-GLMM-PsU-P	C	8.10	2.81	23.37
11-GLMM-NB-L	C	11.88	4.43	31.84
12-GLMM-NB-Q	C	11.85	4.42	31.77
13-GLMM-NB-P	C	10.67	4.29	26.53
14-GLMM-Gm-L	C	13.73	6.43	29.32
15-GLMM-Gm-Q	C	13.71	6.42	29.31
16-GLMM-Gm-P	C	12.96	5.89	28.55
1-LMM-hm	D	13.99	4.71	23.28
2-LMM-ht	D	13.78	4.06	23.50
3-LMM-l	D	7.67	1.63	36.01
8-GLMM-PsU-L	D	7.95	2.85	22.19
9-GLMM-PsU-Q	D	7.93	2.82	22.33
10-GLMM-PsU-P	D	8.49	2.96	24.32

11-GLMM-NB-L	D	12.42	4.65	33.17
12-GLMM-NB-Q	D	12.40	4.64	33.12
13-GLMM-NB-P	D	11.41	4.60	28.30
14-GLMM-Gm-L	D	10.46	4.97	22.03
15-GLMM-Gm-Q	D	10.44	4.95	22.00
16-GLMM-Gm-P	D	9.83	4.67	20.69
1-LMM-hm	E	3.95	-5.35	13.26
2-LMM-ht	E	4.53	-1.56	10.62
3-LMM-l	E	1.15	0.24	5.39
8-GLMM-PsU-L	E	2.25	0.72	7.10
9-GLMM-PsU-Q	E	2.25	0.71	7.15
10-GLMM-PsU-P	E	2.56	0.82	8.06
11-GLMM-NB-L	E	4.25	1.49	12.14
12-GLMM-NB-Q	E	4.23	1.48	12.10
13-GLMM-NB-P	E	3.93	1.50	10.30
14-GLMM-Gm-L	E	8.81	3.53	22.00
15-GLMM-Gm-Q	E	8.81	3.53	22.02
16-GLMM-Gm-P	E	8.39	3.32	21.16

Appendix A - Supplemental Tables

A 1: Average, minimum and maximum values of the Pearson chi-sq/df statistic under simulation scenario of unequal treatment means with 10 or 50 blocks and a Poisson-Gamma DGP.

Results for the true GLMMs used for data generation are **bolded**

	10 Blocks, Unequal Treatment Means			50 Blocks, Unequal Treatment Means		
	Mean	**Min**	**Max**	**Mean**	**Min**	**Max**
5-GLMM Ps-L	5.88	2.09	11.86	26.51	14.95	51.46
6-GLMM Ps-Q	5.88	2.09	11.86	26.51	14.95	51.46
8-GLMM PsU-L	0.22	0.12	0.39	0.10	0.07	0.14
9-GLMM PsU-Q	0.22	0.12	0.39	0.10	0.07	0.14
11-GLMM NB-L	**0.72**	**0.56**	**1.27**	**0.73**	**0.63**	**0.85**
12-GLMM NB-Q	**0.72**	**0.56**	**1.27**	**0.73**	**0.63**	**0.85**
14-GLMM Gm-L	0.35	0.14	0.75	0.36	0.27	0.48
15-GLMM Gm-Q	0.35	0.14	0.75	0.36	0.27	0.48

A 2: Average, minimum, and maximum values for 97.5th Percentile Ratios from simulation scenarios with 10 or 50 blocks and unequal treatment means under a Poisson-Gamma DGP

Results for the true GLMMs used for data generation are **bolded**

	10 Blocks, Unequal Treatment Means			50 Blocks, Unequal Treatment Means		
	Mean	**Min**	**Max**	**Mean**	**Min**	**Max**
1-LMM-hm	0.61	0.40	0.95	0.55	0.42	0.75
2-LMM-ht	0.70	0.51	1.03	0.53	0.40	0.71
3-LMM-l	0.04	0.02	0.08	0.01	0.01	0.01
4-LMM-s	0.07	0.04	0.13	0.03	0.02	0.04
5-GLMM Ps-L	0.93	0.66	1.35	0.93	0.75	1.25
6-GLMM Ps-Q	0.93	0.66	1.35	0.93	0.75	1.25
7-GLMM Ps-P	0.93	0.66	1.35	0.93	0.75	1.25
8-GLMM PsU-L	0.99	0.96	1.01	1.00	1.00	1.00
9-GLMM PsU-Q	0.99	0.96	1.01	1.00	1.00	1.00
10-GLMM PsU-P	0.99	0.96	1.01	1.00	1.00	1.00
11-GLMM NB-L	**0.80**	**0.41**	**1.46**	**0.76**	**0.50**	**1.12**
12-GLMM NB-Q	**0.80**	**0.41**	**1.46**	**0.76**	**0.50**	**1.12**
13-GLMM NB-P	**0.80**	**0.43**	**1.47**	**0.76**	**0.51**	**1.12**
14-GLMM Gm-L	0.75	0.41	1.27	0.75	0.48	1.11
15-GLMM Gm-Q	0.75	0.41	1.27	0.75	0.48	1.11
16-GLMM Gm-P	0.76	0.41	1.24	0.76	0.50	1.13

A 3: Average, minimum, and maximum values for Skewness Ratios from simulation scenarios with 10 or 50 blocks and unequal treatment means under a Poisson-Gamma DGP

Results for the true GLMMs used for data generation are **bolded**

	10 Blocks, Unequal Treatment Means			**50 Blocks, Unequal Treatment Means**		
	Mean	**Min**	**Max**	**Mean**	**Min**	**Max**
1-LMM-hm	0.32	0.12	0.82	0.22	0.11	0.43
2-LMM-ht	0.37	0.19	0.58	0.22	0.08	0.44
3-LMM-l	0.00	-0.37	0.79	-0.02	-0.13	0.07
4-LMM-s	0.21	0.04	0.68	0.14	0.08	0.30
5-GLMM Ps-L	1.00	0.65	1.65	0.93	0.69	1.42
6-GLMM Ps-Q	1.00	0.65	1.65	0.93	0.69	1.42
7-GLMM Ps-P	1.00	0.65	1.65	0.93	0.69	1.42
8-GLMM PsU-L	1.01	1.00	1.05	1.00	1.00	1.00
9-GLMM PsU-Q	1.01	1.00	1.05	1.00	1.00	1.00
10-GLMM PsU-P	1.01	1.00	1.05	1.00	1.00	1.00
11-GLMM NB-L	**0.84**	**0.27**	**1.67**	**0.73**	**0.35**	**1.65**
12-GLMM NB-Q	**0.85**	**0.27**	**1.67**	**0.73**	**0.35**	**1.65**
13-GLMM NB-P	**0.87**	**0.29**	**1.86**	**0.75**	**0.35**	**1.69**
14-GLMM Gm-L	0.80	0.26	1.60	0.72	0.35	1.64
15-GLMM Gm-Q	0.80	0.26	1.60	0.72	0.35	1.64
16-GLMM Gm-P	0.82	0.27	1.64	0.74	0.35	1.67

A 4: Average, minimum, and maximum values for Coefficient of Variation Ratios from simulation scenarios with 10 or 50 blocks and unequal treatment means from a Poisson-Gamma DGP

Results for the true GLMMs used for data generation are **bolded**

	10 Blocks, Unequal Treatment Means			**50 Blocks, Unequal Treatment Means**		
	Mean	**Min**	**Max**	**Mean**	**Min**	**Max**
1-LMM-hm	0.64	0.42	0.87	0.57	0.24	0.74
2-LMM-ht	0.69	0.46	0.90	0.55	0.20	0.75
3-LMM-l	0.40	0.22	0.87	0.19	0.05	0.26
4-LMM-s	0.35	0.21	0.51	0.32	0.11	0.42
5-GLMM Ps-L	0.91	0.78	1.05	0.90	0.80	1.02
6-GLMM Ps-Q	0.91	0.78	1.05	0.90	0.80	1.02
7-GLMM Ps-P	0.91	0.78	1.06	0.90	0.80	1.02
8-GLMM PsU-L	0.99	0.96	1.00	1.00	1.00	1.00
9-GLMM PsU-Q	0.99	0.96	1.00	1.00	1.00	1.00
10-GLMM PsU-P	0.99	0.96	1.00	1.00	1.00	1.00
11-GLMM NB-L	**0.80**	**0.52**	**1.24**	**0.75**	**0.44**	**1.04**
12-GLMM NB-Q	**0.80**	**0.52**	**1.24**	**0.75**	**0.44**	**1.05**
13-GLMM NB-P	**0.81**	**0.52**	**1.22**	**0.75**	**0.44**	**1.06**
14-GLMM Gm-L	0.75	0.49	1.15	0.74	0.43	1.02
15-GLMM Gm-Q	0.75	0.49	1.15	0.74	0.43	1.02
16-GLMM Gm-P	0.76	0.49	1.15	0.75	0.43	1.05