Dispersion tests in generalised linear mixed-effects models - a methods comparison and practical guide

Melina de Souza Leite[1*], Daniel Rettelbach[1,2] & Florian Hartig[1]

1. Theoretical Ecology, University of Regensburg, Germany
2. coTrial Associates, Department of Surgery, University Hospital Regensburg, Germany (current address)

*corresponding author: melina.souza-leite@ur.de

Author contributions

MSL, FH, and DR conceived the ideas and designed the methodology. MSL wrote the simulation code and created the final version of the graphs and tables. MSL and FH led the writing of the manuscript. All authors contributed critically to the drafts and gave final approval for publication.

Data availability

All data were simulated. The code for the simulations, analyses, and figures is available on Zenodo (https://doi.org/10.5281/zenodo.17611061).

Acknowledgements

This study was funded by the Deutsche Forschungsgemeinschaft (DFG), project number 528747641.

Conflict of interest

The authors, MSL and FH, are developers of the R package DHARMa, which implements the dispersion tests used in this study.

Abstract

1. Underdispersion and overdispersion are common issues when analysing ecological data with generalised linear (mixed) models (GLMs/GLMMs). Overdispersion, the phenomenon where observations spread wider than expected by the fitted model, leads to anti-conservative p-values and, thus, to inflated type I error. In contrast, underdispersion, a narrower spread of the data than expected, causes overly conservative p-values and, therefore, a reduction in power. A range of tests has been suggested to detect such dispersion problems, but there are few comparative studies of their performance across a range of models and analysis situations.
2. The goal of this study is to identify a general dispersion test for GLMs/GLMMs that is applicable across all standard distributions and random-effects structures. After an initial assessment of available tests, we selected two classes of dispersion tests as candidates: (1) parametric and nonparametric tests based on Pearson residuals and (2) simulation-based tests that compare the expected to the observed variance in the response.
3. Comparing their performance by type I error, power, and dispersion estimate, across a range of GLMs and GLMMs, we find that a nonparametric Pearson residuals test performed best across all metrics, especially for data with low incidence or count rates and/or sample sizes; however, at the cost of high computational expenses. The parametric Pearson residuals test, which is recommended in many books and guidelines, is faster and performs excellently for GLMs, but can be seriously biased towards underdispersion for GLMMs. We show that the reason for this bias, which increases with the number of random effect clusters/groups, lies in the naïve computations of the degrees of freedom

for the random effects. The simulation-based response variance test is slightly less powerful than the nonparametric Pearson test, but it showed overall good calibration and is much faster to compute. It offers a compromise between the strengths and weaknesses of the two Pearson-based tests.

4. We conclude that for GLMs, the parametric Pearson residuals test offers the best combination of speed and accuracy. For GLMMs, we recommend either the computationally demanding non-parametric Pearson residuals test or the faster, although somewhat less powerful, simulation-based response variance test.

Keywords: overdispersion/underdispersion, multilevel/hierarchical models, hypothesis test, Pearson residuals, type I error, power, dispersion parameter

Introduction

Generalised linear models (GLMs) and generalised linear mixed models (GLMMs) are the most commonly used tools for the statistical analysis of ecological data (Bolker et al., 2009; Lai et al., 2019; Touchon & McCoy, 2016). By incorporating mixed and random effect structures with a wide array of distributional assumptions (e.g., binomial, Poisson), GLMMs allow researchers to model nonnormal response variables (e.g., counts, proportions, or presence-absence) while properly accounting for variation clustered in sampling units, sites, or study years (Bolker et al. 2009; McMahon & Diez 2007). However, as for all parametric statistics, these models rely on the fact that residuals scatter around the regression mean with the specified distribution, and their inferential results can be seriously biased if these distributional assumptions are violated.

A particularly common and dreaded violation of distributional assumptions in generalised linear (mixed) models is overdispersion. Overdispersion refers to a higher variation in the observed data (and particularly the model residuals) than the fitted model assumes (Campbell, 2021; McCullagh & Nelder, 1989). Strong overdispersion usually appears in GLM distributions that assume a fixed mean-variance relationship, such as the Poisson model for count data (Harrison, 2014; J. M. Hilbe, 2014) or the binomial model for discrete proportions (Dunn & Smyth, 2018; Harrison, 2015). For example, a Poisson process assumes that we count randomly distributed points in space, but when observations are subject to spatial/temporal clustering and/or imperfect detection (Rhodes, 2015), we typically find higher dispersion than expected from a Poisson distribution. Alternatively, overdispersion may also arise from misfit, for example, by failing to include important predictors and interactions or by specifying the incorrect link function (J. M. Hilbe, 2011).

Overdispersion is a major concern in practical data analyses because it can have substantial anti-conservative effects on p-values, confidence intervals, and all other goodness-of-fit and precision metrics (Fig. 1, see also Rhodes, 2015). Anti-conservative means that p-values and confidence intervals are too small, leading to associated inflated false positive results (type I errors). In practice, we have encountered analyses where an overdispersed model had very small and significant p-values (<0.001) that became nonsignificant after changing to a GLM with more appropriate dispersion (see also example in Fig. 1).

The counterpart to overdispersion is underdispersion, where the variation in the observed data (and, thus, model residuals) is lower than assumed by the fitted model. Reasons for underdispersion can again be that the data-generating process differs from what is assumed by the model (Lynch et al., 2014). However, in practice, it is often the result of model overfitting, i.e., having a too complex model that overfits the data. Underdispersion is somewhat less discussed in the literature, both because it is less frequent, but also because it leads to over-conservative model metrics (Fig. 1). This may seem less problematic as it does not lead to reporting "wrong" effects, but underdispersion reduce overall power and thus increase type II error. Therefore, accurate statistical inference demands that we identify and adequately deal with both underdispersion and overdispersion to minimise the risk of wrong inference.

Due to the central importance of dispersion for all statistical indicators, statisticians have pondered how to detect and address dispersion problems since the early days of modern statistics (see Quine & Seneta (1987) and Xekalaki (2014) for a historical perspective). The first attempts to describe the phenomenon date back at least to the end of the 19th century, likely with Lexis's ratio (Lexis 1879, apud Xekalaki, 2014) for binomial clustered data, where Q is the ratio of the between-clusters variance

to the total variance (Xekalaki, 2014). Bortkiewicz later (1898) coined the term "divergence coefficient" (Q^2), which is the variance divided by the mean of the sample, as a test statistic for the Poisson model (Quine & Seneta, 1987). William Gosset, the inventor of the t-test, also considered the problem of dispersion in the Poisson model (Student, 1919). Since then, a large variety of approaches have been proposed and discussed to deal with the "dispersion problem", ranging from (1) comparing models with or without free dispersion parameters through likelihood ratio test, such as Poisson and negative binomial (e.g. Yang et al., 2007), (2) designing specific hypothesis tests for the "extra" variation (e.g. Fisher, 1950), such as score tests (Dean, 1992; Dean & Lawless, 1989; Lawless, 1987), (3) using goodness-of-fit tests, such as tests on Pearson or Deviance residuals (Dunn & Smyth, 2018; McCullagh, 1985) (although the distinction between categories (2) and (3) can be blurry, see (Collings & Margolin, 1985; Dean, 1992; Dean & Lawless, 1989) or (4) using simulation-based non-parametric tests to compare observed and predicted variance of the response data (Hartig, 2024).

Somewhat confusing for the practical data analyst, however, many of these approaches have been designed and tested only in very specific scenarios (e.g. only for a Poisson GLM), and there is a surprising lack of systematic evaluation of these tests and strategies across a range of more complex GLMMs. Moreover, a quick review of current methods available in the R environment (R Core Team, 2024) revealed that existing dispersion tests are scattered across different packages (Table 1), and most of these only work for a restricted set of models. All this makes it challenging to decide which test should be used in an applied data analysis.

The goals of this study are: (1) to review and order the diversity of dispersion tests for GLMs and GLMMs, and (2) to identify tests that can reliably work across a

range of models with diverse distributions and complex hierarchical structures. Based on our literature review (next section), we identified two groups of tests that appeared to be generally applicable: parametric and non-parametric tests on Pearson residuals, as well as a new simulation-based non-parametric test that directly compares observed and predicted variance of the response data. We then used simulated data to compare the performance of these tests in terms of type I error, power, and the interpretability of the dispersion statistics. Based on this, we provide recommendations on the most suitable tests for detecting over- or underdispersion, depending on model complexity and software availability (i.e., currently available packages and functions in R).

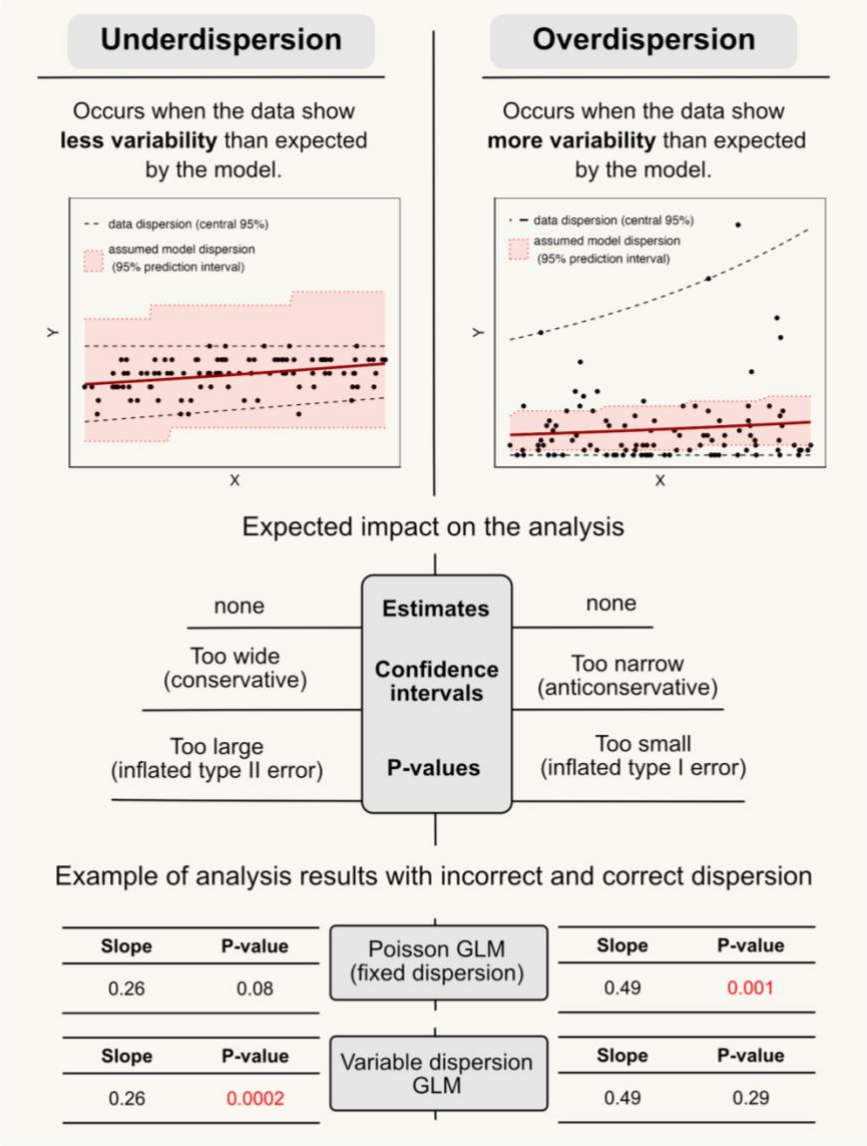


Figure 1. Definition, statistical consequences, and a practical analysis example of under-/overdispersion in generalised linear (mixed) models. The top row shows examples of a data analysis using a Poisson GLM with simulated under- and overdispersed count data. The data points in black are contrasted to the Poisson model's 95% prediction interval (in red). Black dashed lines illustrate the data dispersion (central 95% quantiles of the data). In the example data analysis, we present slope estimates and p-values for the GLM Poisson model fitted to the under- and overdispersed data above, as well as the results using more appropriate models with correct dispersion, here a Conway-Maxwell-Poisson GLM for underdispersed data and a negative binomial GLM for overdispersed data.

Table 1. Different types of dispersion evaluation and tests for GLMs and GLMMs with examples of available R packages and functions.

Test	Principle	Details/Limitations	R package:: function	Supported models	References
Likelihood Ratio Test (LRT)	Compare two models with and without free dispersion parameters, for example: - Poisson and negative binomial or generalized Poisson - binomial and beta-binomial	Requires fitting two models, requires defining an alternative model. Not a dispersion test.	`pscl::odTest()`	GLM Poisson -> negative binomial with `MASS::glm.nb()`	Jackman (2024)
			`DCluster::test.nb.pois()`	GLM Poisson -> negative binomial with `MASS::glm.nb()`	Lopez-Quílez (2005)
			`anova(…,test="LRT")`	Many GLM/GLMMs (Different packages have the S3 method for anova functions to perform LRT)	
			`lmtest::lrtest()`	Any GLM	(Zeileis & Hothorn, 2002)
Score test	Evaluate score of restricted dispersion parameter	Requires score calculation for specific models. R functions only for Poisson GLM.	`DCluster::DeanB()` `DCluster::DeanB2()`	GLM Poisson. Score tests based on Dean (1992)	Lopez-Quílez (2005)
			`Rfast2::overdispreg.test()`	GLM Poisson (own model implementation)	Papadakis et al. (2025)
	Regression-based test for overdispersion from Cameron & Trivedi (1990)	Distribution specific (Poisson-based only).	`overdisp::overdisp()`	GLM Poisson (own model implementation)	Cameron & Trivedi (2023)
			`AER::dispersiontest()`	GLM Poisson from `stats::glm()`	Kleiber & Zeileis (2008)
Standard. residuals dispersion	A goodness-of-fit test to evaluate residual dispersion, e.g. via sum of Pearson residuals.	**Parametric Pearson residuals test:** Assume Pearson residuals are Chi-squared distributed. For complex models, difficult to define parametric null distribution (unclear residual degrees of freedom).	`msme::P__disp()`	GLMs	Hilbe & Robinson (2025)
			`DHARMa::testDispersion(…, type="Pearson")`	GLMs/GLMMs (naïve residual *df*)	Hartig (2024)
			`performance::check_overdispersion()`	GLMs/GLMMs (naïve residual *df*)	Lüdecke (2021)
			`RVAideMemoire::overdisp.glmer()`	GLMMs (from `lme4` package, naïve residual *df*, calculates only dispersion statistic, no test)	Herve (2025)
		Nonparametric Pearson residuals test: Parametric bootstrapping of the model to generate a nonparametric estimate of the null distribution of the Pearson statistic. Computational costly.	`DHARMa::testDispersion(…, refit=T, type="Pearson")`	GLMs/GLMMs	Hartig (2024)
Response variance	Compares the expected to the observed variance in the response variable.	Expected variance of response variable calculated through simulations of fitted model. Fast nonparametrics but possibly less exact than working on the residual dispersion.	`DHARMa::testDispersion(…, type="DHARMa")`	GLMs/GLMMs	Hartig (2024)

A short review of existing approaches to dispersion tests

After reviewing the available literature, we divided the different strategies proposed for checking dispersion problems into four classes (Table 1). Here, we discuss these broad strategies in more detail and explain why we focused on two of these classes as the most suitable competitors for a general dispersion test for GLMs and GLMMs. We note that, in addition to the four approaches mentioned here, dispersion problems may also show up in general goodness-of-fit tests (e.g., Feng et al., 2020). However, as they are not specifically designed to react to dispersion, we did not consider them further.

Likelihood ratio tests

A first general strategy for detecting dispersion problems is to compare a model with fixed dispersion to its nearest "relative" with variable dispersion using a likelihood ratio test (LRT) or another model selection technique, such as AIC (Yang et al., 2007). For count data, a practical example would be to compare a Poisson GLM as a null hypothesis to a negative binomial or generalised Poisson GLM (J. M. Hilbe, 2014), or to compare a binomial GLM to a beta-binomial GLM (Dunn & Smyth, 2018). While relatively easy to implement, the downside of this approach, apart from the higher computational costs resulting from fitting two models, is that it doesn't provide any direct diagnostics of over- or underdispersion, but only compares a base model against an alternative. The alternative model, however, might also fit better or worse for reasons other than a dispersion problem. Moreover, using LRTs for detecting dispersion problems has also been discouraged as it may provide unreliable results (Dean, 1992) because it tends to underestimate the evidence against the base model (Lawless, 1987).

Therefore, we do not find this approach suitable as a general dispersion test and do not consider it further.

Score tests

A second traditional option for assessing overdispersion is the use of a score test (Dean, 1992; Dean & Lawless, 1989; Lawless, 1987). Score tests, also known as Lagrange Multiplier (LM) tests, evaluate the gradient of the log-likelihood (called the score or LM statistic) of a restricted parameter estimator (e.g., an overdispersion estimator restricted to zero). Under the null hypothesis that the overdispersion is indeed zero, the score will have an asymptotic chi-square distribution (Rao, 1948). In performance comparisons, score tests have been found to have good power (Ohara Hines, 1997), but their disadvantage is that they are usually model specific (in the sense that different tests are needed for Poisson or binomial GLMs); their implementation can be computationally demanding; and, as they require access to the score, they must usually be implemented with the model and cannot be calculated on top of a fitted model object. Perhaps because of these issues, we were unable to find any R function that computes score tests beyond the Poisson GLM (Table 1), although score tests have been developed for other models, such as the binomial GLM (Dean, 1992).

An equivalent test related to the score test under certain conditions is the regression-based overdispersion test proposed by Cameron & Trivedi (1990). Under a Poisson model, the squared deviation of the observations from their fitted mean, after subtracting the observation itself and scaling by the fitted mean, has expectation zero. In contrast, under the negative binomial, it increases systematically with the mean. This motivates an auxiliary regression of the transformed variable against the fitted mean, with a significant slope indicating extra-Poisson variation. The main advantage of this

test against other score tests is its ease of implementation: it can be carried out after fitting a standard Poisson GLM. However, similar to an LRT, the linear regression imposes a particular form of overdispersion as an alternative hypothesis, and therefore, seems less general than the test based on Pearson residuals described below.

We discarded score tests in general, and the Cameron & Trivedi (1990) test in particular, from our further analysis, as it seems impractical to implement them across a wide range of existing GLMM software.

Tests based on residual dispersion

A third class of testing approaches, arguably the most intuitive, directly calculates a test statistic or goodness-of-fit metric on standardised model residuals. The most widely used test of this kind is based on the sum of the model's Pearson residuals. As Pearson residuals divide the raw residuals by the expected residual standard deviation, a correctly specified model is expected to have a Pearson residual of around 1 for each observation. A dispersion statistic is then defined as the sum of squared Pearson residuals divided by the residual degrees of freedom. Models with a so-defined dispersion statistic > 1 are considered overdispersed, while dispersion statistics < 1 are underdispersed. Sometimes, a modification of this metric is often recommended by replacing the sum of squared Pearson residuals with the model deviance, which is typically more readily available. However, as Venables & Ripley (2002) discuss, this metric should be avoided, as it often deviates from 1, even for correctly specified GLMs.

Defining dispersion via the Pearson statistic has the added advantage that for a GLM, the expected distribution under the null hypothesis of a correctly specified model asymptotically follows a Chi-square distribution (McCullagh, 1985). This allows a

straightforward construction of a hypothesis test, where we compare the Pearson statistic to the chi-squared distribution with the respective residual degrees of freedom (*df*). This test is referred to with different terminologies, such as Pearson chi-squared dispersion test, Pearson residuals-based test for overdispersion, or simply Pearson dispersion test. Hereafter, we refer to this test as the **parametric Pearson residuals test**, to differentiate it from the nonparametric test based on Pearson residuals, discussed below.

An alternative approach to constructing a dispersion test based on the Pearson dispersion statistic involves generating a null distribution through parametric bootstrapping. A parametric bootstrap means that new data is simulated from the fitted model, and then the statistic of interest (in this case: the Pearson statistic of a fitted model) is calculated based on this data. The parametric bootstrap has been previously used for hypothesis tests in mixed-effects models where parametric null distributions were difficult to obtain (e.g., Barr et al., 2013; Luke, 2017), and thus it seems a logical alternative for more complicated models where the Chi-square distribution of the Pearson dispersion statistic cannot be taken for granted (see methods for GLMMs below). Nevertheless, implementing parametric bootstrapping in complex models can be less efficient for at least two reasons: it is time-consuming and prone to errors in model refits (Luke, 2017; Moral et al., 2017). A dispersion test based on this principle was implemented in R by Hartig (2024). Hereafter, we will refer to this test as the **nonparametric Pearson residuals test**.

Tests based on response variable variance

Simulation approaches can also be useful to generate null distributions for alternative metrics of dispersion. A last class of dispersion test approaches, which, to

our knowledge, was introduced in the DHARMa R package (Hartig, 2024) but has not been discussed in the literature so far, involves defining a test statistic based on the dispersion of the response variable, rather than the residuals. More specifically, the test compares the observed data variance with the simulated data variances (which can be created conditional or unconditional on the fitted random effects for GLMMs). The dispersion statistic is then defined as the ratio between the observed variance and the mean simulated variance. Similar to the Pearson statistic, a ratio > 1 indicates overdispersion, a ratio < 1 indicates underdispersion, and a significance test is constructed based on the distribution of simulated variances.

From a theoretical viewpoint, this approach seems less elegant compared to the idea of using Pearson residuals, because the latter, by "standardising" the residual dispersion with the expected dispersion, allows each data point to contribute similarly to the dispersion statistic. In contrast, the test on the unstandardized response variable will be more influenced by large data points. However, the primary advantage of this approach is computational, as it enables the creation of a nonparametric estimate of the test statistic without requiring a re-fit of the model (in contrast to the nonparametric Pearson residuals test). Hereafter, we will refer to this test as the **simulation-based response variance test** to differentiate it from the tests based on Pearson residuals.

Methods

Selected models and setup of the performance comparisons

After reviewing the available approaches, we identified three tests as potential candidates for a generally applicable dispersion test that could be implemented across a wide range of GLMs and GLMMs:

(1) The parametric Pearson residuals test

(2) The nonparametric Pearson residuals test

(3) The simulation-based response variance test

To compare the performance of these three tests, we simulated datasets based on the two main distributions that often present over- or underdispersion problems: the Poisson and the binomial (N/K) proportions. We varied the sample size (from 10 to 10,000, depending on the simulation) and intercept (from -3 to 3, at the link function scale) of the simulated data for both distributions. We simulated a gradient of overdispersed data by adding noise to the linear predictor with values from a Gaussian distribution with a mean of zero and ten standard deviation values varying from 0 to 1. We evaluated the performance of the tests by comparing type I error, power, and dispersion statistics for all combinations of parameters in the simulated datasets.

All models were fitted using the functions *glm* from the stats package or *glmer* from the lme4 package (Bates et al., 2015) in R (v4.4; R Core Team, 2024). All dispersion tests were performed with the DHARMa package (Hartig, 2024). For the simulation-based response variance test and the nonparametric Pearson residuals tests, we set the number of simulations fixed at 250 (the default parameter in DHARMa). All simulations and analysis codes are available at this repository (https://anonymous.4open.science/r/dispersion_test_GLMM/README.md). The supplementary material provides a script file with instructions and examples for applying dispersion tests using the DHARMa package.

Theoretical expectations

The classical (1) parametric Pearson residuals test assumes that the sample size (n-asymptotic) and the expected values are sufficiently large (phi-asymptotic) (Venables

296 & Ripley, 2002). This implies that, when the expected counts (or intercept) and/or the
297 number of observations are small, Pearson residuals may not provide reliable
298 information about model fit (see S1). Some corrections for Pearson residuals with small
299 sample sizes were suggested (e.g., Cordeiro, 2004; Cordeiro & Simas, 2009), but they
300 are not currently implemented in the most common packages in R. Therefore, we expect
301 the parametric Pearson residuals test to perform well for GLMs, except for very small
302 sample sizes and expected counts (hereafter "small-data" situations).

303 Moreover, it is unclear whether the parametric Pearson residuals test approach
304 can be extended to GLMMs or other hierarchical models, where counting the residual
305 degrees of freedom (*df)* is not straightforward (Bolker et al., 2009; Luke, 2017). In
306 mixed-effects models, the *df* used by a random effect are data-specific (adaptive
307 shrinkage) and expected to be somewhere between one and the number of grouping
308 factors (Baayen et al., 2008; Bolker et al., 2009; Luke, 2017). There exist approaches to
309 approximate *df* for random effects in LMMs (e.g. Schaalje et al., 2002), but their
310 generalisation to GLMMs is still an area of active research. Current R packages that
311 implement the parametric Pearson residuals test approximate the *df* by the so-called
312 naïve *df* (e.g., n = 1 per random effect) for testing LMMs/GLMMs (Table 1). We expect
313 that the error imposed by this approximation increases with the number of random
314 effect groups. To test this, we varied the number of groups in the random intercept (10,
315 50, and 100 groups) of our simulated data.

316 In contrast to the parametric Pearson residual test, we expect the (2)
317 nonparametric Pearson residuals test to be robust to small-data problems as well as the
318 presence of random effects, as it doesn't rely on a particular parametric distribution.
319 However, since the test uses parametric bootstrapping, we expected it to run much

slower than the other tests, especially for more complex GLMMs. For this purpose, we compared the runtime of the tests with a small set of simulated data (see S6).

For GLMMs tested with the (3) simulation-based response variance test, we compared the performance of the test under the two simulation approaches, conditional and unconditional to random effects. We expect to see lower power for the unconditional simulation results, as overdispersion is a phenomenon at the level of the model distribution (i.e., at a higher level). We evaluated the circumstances under which this test is reliable as a fast alternative to both dispersion tests based on Pearson residuals.

Results

Performance on Poisson and binomial GLMs

For Poisson GLMs, we found the expected distribution problems (Fig. 2): type I error rates for the parametric Pearson residuals test were substantially high for the smallest intercepts (-3), and they did not reach the nominal value of 0.05 even for very large sample sizes (n = 10,000). The type I error rates for the nonparametric Pearson residuals test were well calibrated, except for the smallest intercept (-3), with slightly conservative type I error rates (< 0.05). For the simulation-based response variance test, type I errors were independent of sample size, but exhibited an intercept-dependent conservative bias, ranging from almost 0 for the smallest intercept to 0.06 for the largest intercepts.

For binomial GLMs, the type I error rates for the parametric Pearson residuals test were generally conservatively calibrated around 0.04 (Fig. 2). Type I error rates for the nonparametric Pearson residuals test averaged around 0.05 and 0.06, except for the

very low and very high intercepts (-3 and 3). For the simulation-based response variance test, type I error rates were conservatively very low for all simulated parameters, bouncing below 0.01.

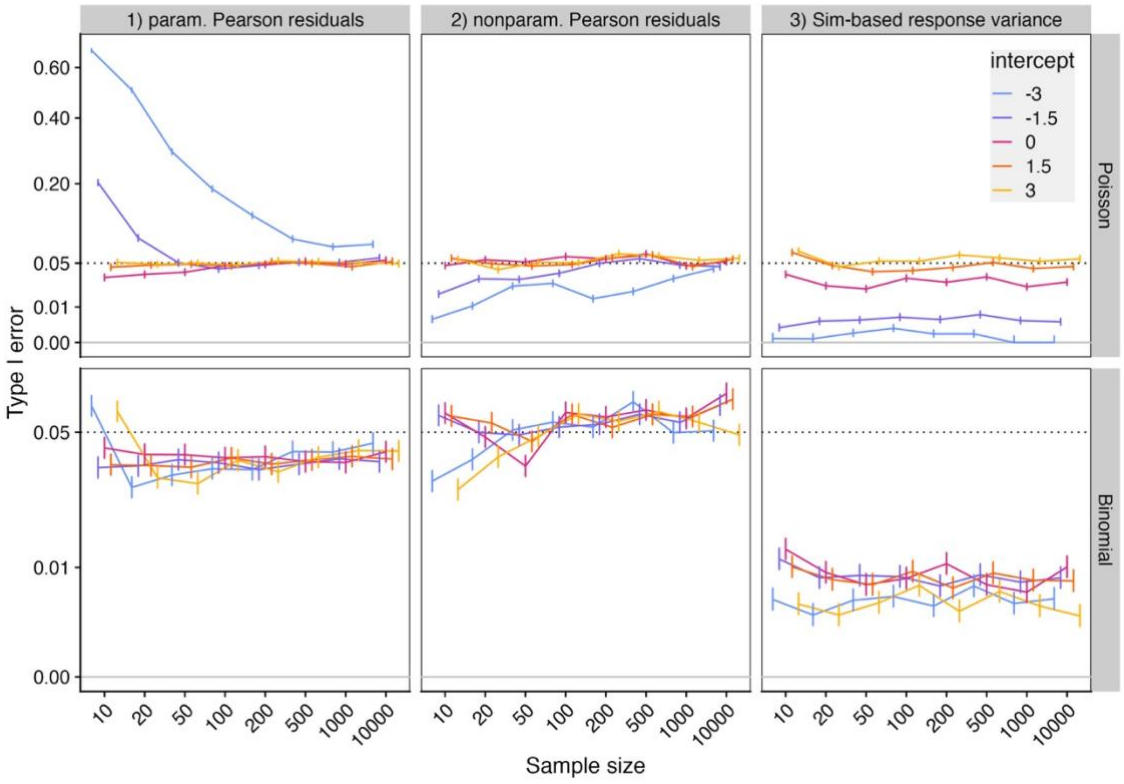


Figure 2. Simulation-based response variance tests have more conservative type I error rates than both Pearson residuals tests. The three dispersion tests were applied to Poisson (upper panels) and binomial proportion (lower panels) GLMs: 1a) parametric Pearson residuals test, 1b) nonparametric Pearson residuals test, and 2) simulation-based response variance test (see Table 1 for explanations). Simulations under different sample sizes (x-axis) and intercepts (colours, values at the link function scale). In B), the model is a binomial proportion with ten trials. All points include a 95% confidence interval calculated based on exact binomial tests for the 10,000 simulations. Note the square-root scale of the y-axis in plot A. The Dotted horizontal black line shows the 0.05 nominal value for type I error.

The statistical power of the simulation-based response variance test was lower than the parametric and nonparametric Pearson residuals tests for both binomial and Poisson GLMs, but tended to be similar with larger sample sizes (Fig. 3). We found that the reason for this is the very conservative type I error rates (Fig. 2). When power is

calibrated by using the p-value at the 5% quantile of its empirical distribution for each simulation (details in S5), the differences disapear (Fig. 3).

The dispersion statistics of the simulation-based response variance test were highly dependent on the intercept, slope, and number of trials for the binomial model (see S4, Fig. S4.2), and they tended to be smaller than those based on the Pearson residuals. In contrast, for Poisson models, the values tended to be larger than those of Pearson statistics (Fig. S3.5). This may also explain the lower uncorrected power for the simulation-based response variance test, especially for binomial models.

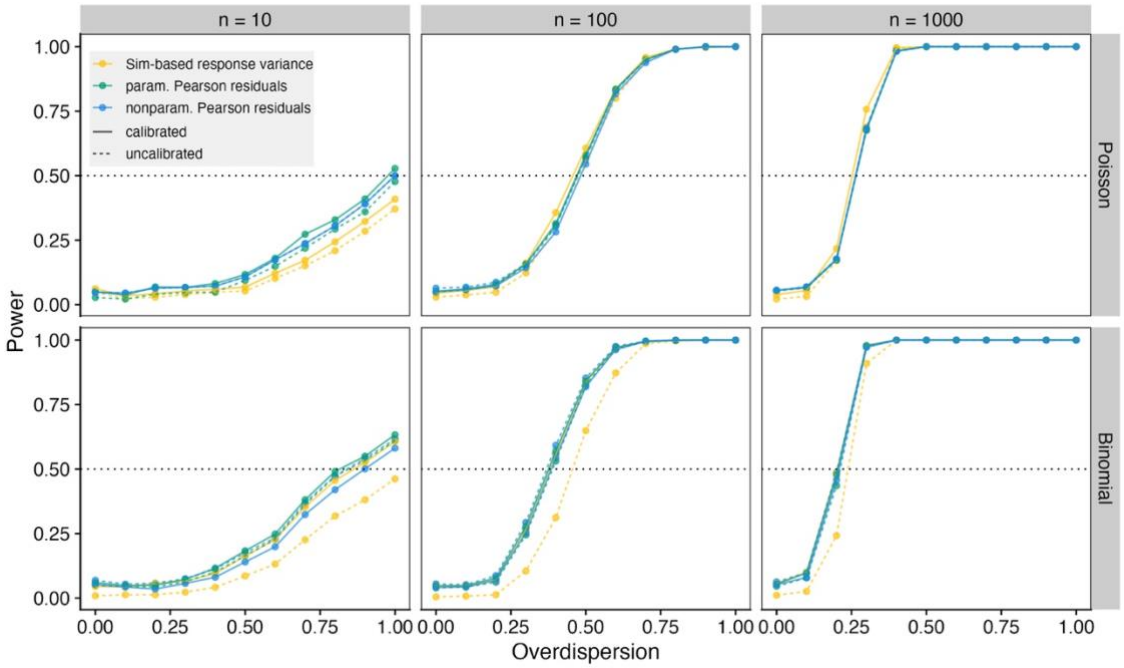


Figure 3. The simulation-based response variance test (in yellow) has lower power than both Pearson residuals tests (green and blue) for GLMs unless power is calibrated by type I error rates (dashed lines). Lower power is more evident for binomial models (upper panel) and smaller sample sizes (first two columns). Results based on 10,000 simulations per combination of parameters for an intercept = 0 and slope = 1. For all simulation results, see Fig. S5.1 and S5.2.

GLMM performance

For the GLMMs, we first compared the performance of the parametric Pearson residuals test (two-sided) for an increasing number of groups (m) in the random

intercepts. As expected, the performance of the test failed for a large number of groups in the random effects (Fig. 4A). The dispersion statistic was underestimated, and the type I error rates were too high because the test detected significant underdispersion. Testing only for overdispersion ("greater" test) when using the parametric Pearson residuals test appears to be the only reasonable approach for GLMMs (Fig. 4A). Still, it doesn't prevent the dispersion statistics from being biased to lower values.

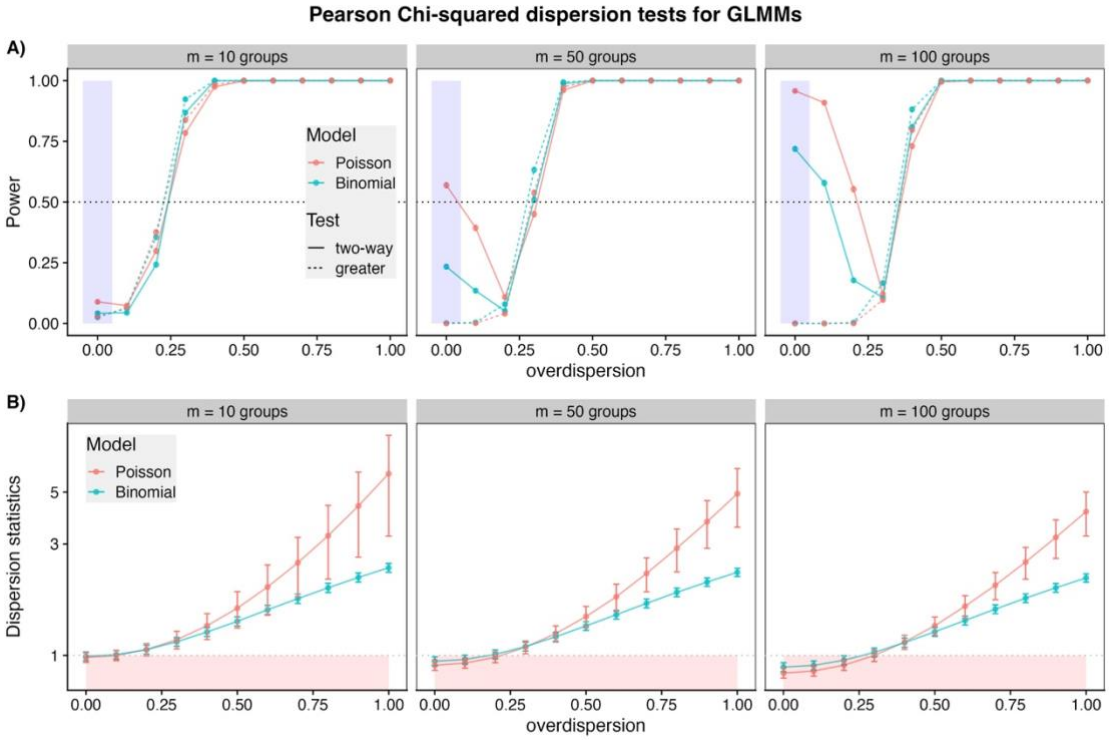
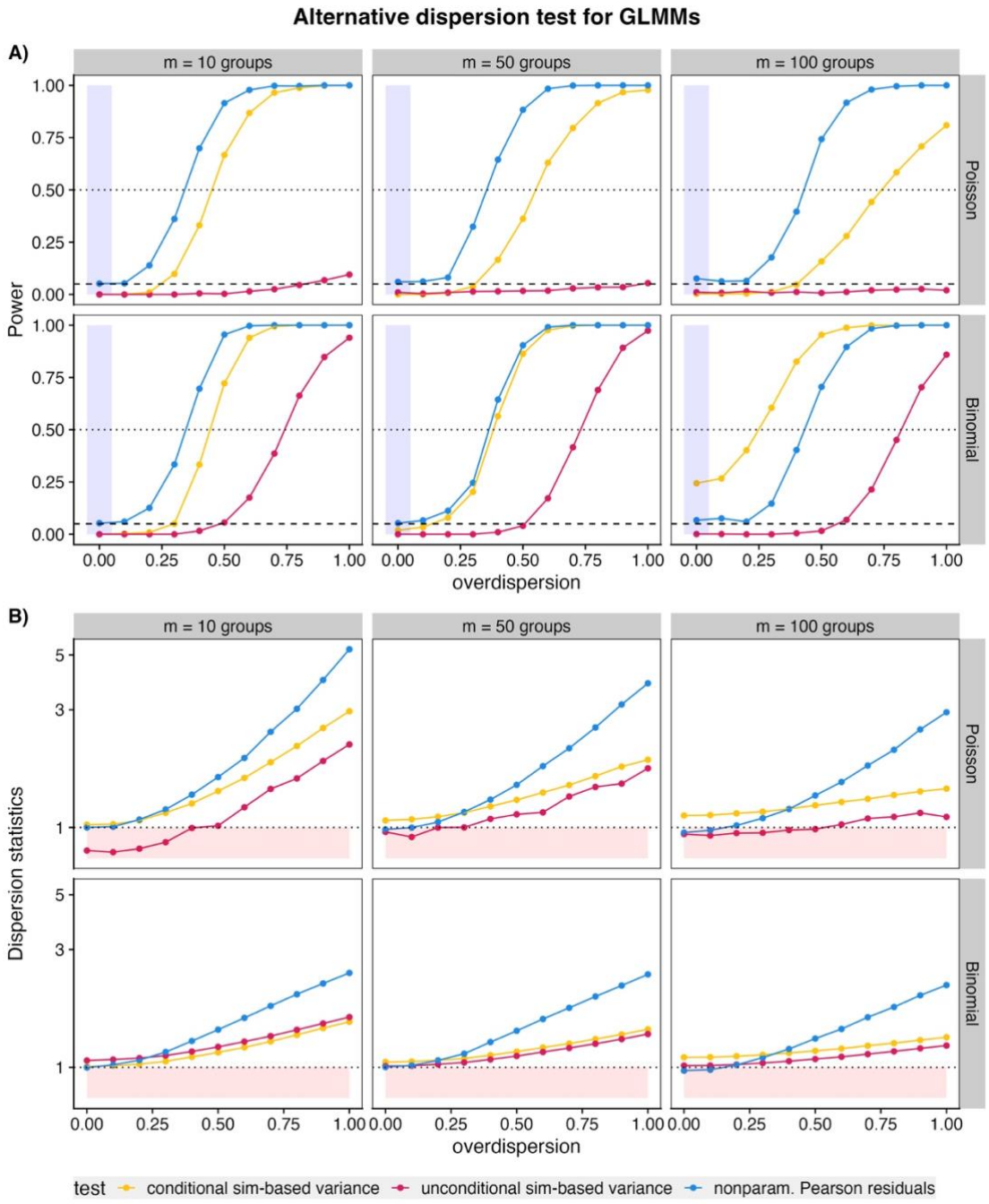


Figure 4. The parametric Pearson residuals test failed for GLMMs with many groups in the random intercepts (plot panels). A) Power and type I error rates (blue shaded area) for the "two-sided" (solid lines) and "greater" (dotted lines) Chi-squared tests for the Pearson statistic. B) Pearson dispersion statistics with the red shaded area indicating dispersion statistics estimated below 1 (underdispersion). Notice that the y-axis of plot B is on a logarithmic scale of 10. Results with 10,000 simulations for an intercept of 0 and a sample size (n) of 1,000 data points.

When comparing the alternative dispersion tests for GLMMs, the nonparametric Pearson residuals test presented very good results, with a type I error rate around 0.05 (Fig. S6.1 and S6.2) and higher power than the simulation-based response variance tests (Fig. 5). As expected, the unconditional simulation-based response variance test had the

worst performance: very low type I errors (Fig. S6.1 and S6.2), very low power, and dispersion statistics below 1 (Fig. 5B), especially for Poisson models. The conditional simulation-based response variance test also had very small type I errors (Fig. S6.1 and S6.2), but power increased with the simulated overdispersion. The performance of both simulation-based response variance tests (unconditional and conditional) didn't change much with the number of groups for the Poisson GLMMs, but it improved for the binomial GLMMs with the increasing number of groups in the random intercept.



404

405 **Figure 5**. The nonparametric Pearson residuals test showed correct Type 1 error, higher
406 power, and larger dispersion statistics than the simulated-based response variance tests
407 (conditional and unconditional to all random effects) for Poisson and binomial GLMMs.
408 Power (A), type I error (shaded blue area in A), and dispersion statistics (B) for the
409 alternative dispersion tests for Poisson and binomial GLMMs with different numbers of
410 groups in random intercepts. The dashed horizontal line in (A) indicates the nominal
411 value of 0.05 for type I error. The dotted horizontal line in (A) indicates the 50% power,
412 and the dotted horizontal line in (B) indicates the dispersion statistics of 1. The results
413 are based on 1,000 simulations per combination of parameters, with an intercept of 0
414 and a sample size (n) of 1,000.

415 ## Discussion

The goal of this study was to find a dispersion test that is widely applicable across different GLM and GLMM distributions and random-effects structures. Our conclusion is that the nonparametric Pearson residuals test is the most reliable general test currently available. For GLMs, this test exhibited similar power as the parametric Pearson residuals test but with more reliable type I error rates in small-data situations. The downside of this test is that it can be computationally expensive, with runtimes in the order of minutes for larger GLMMs.

The two alternative tests that we considered have advantages in particular situations. The simulation-based response variance test for GLMs is fast to compute, but has a dispersion statistic that is more difficult to interpret and often too conservative type I errors. This resulted in low power if not additionally calibrated by a simulated p-value distribution. The parametric Pearson residuals test is computationally efficient, but it is unreliable in small-data situations and in the presence of random effects. Below, we discuss these points in more detail and provide recommendations for general users who rely on already implemented R packages for model fit and diagnostics.

Why and when does the parametric Pearson residuals test fail?

We showed that the parametric Pearson residuals test, although popular, quick, and relatively easy to compute, has two main disadvantages: it does not perform well in (1) small-data situations (Fig. 2) and (2) in the presence of random effects (Fig. 4). The reason for the first problem can be attributed to the mismatch between the Pearson statistic distribution and the Chi-squared distribution under small-data conditions (Fig. S1.1 and S1.2). This phenomenon has already been studied (e.g., Fletcher, 2012; Kuss, 2002), with suggested corrections (Farrington, 1996; McCullagh, 1985). However, none of these corrections are implemented in the current R packages (Table 1), and we

440 believe that it will be difficult to devise corrections that work across a wide range of
441 distributions.

442 The reason for the second problem is that counting 1 degree of freedom (*df*) for
443 a random effect, as done in most implementations of this test, is typically an
444 underestimation of the true model *df,* which increases in magnitude with the increasing
445 number of levels of the random effect. The result is a bias in the dispersion statistic
446 towards underdispersion that increases with the number of random effect levels (Fig. 4).
447 Two-sided tests would therefore often wrongly detect significant underdispersion
448 problems in perfectly valid GLMMs, which is likely the reason why most R
449 implementations of this test only test for overdispersion. When applying this test for
450 GLMMs, we recommend following the same approach and ignoring dispersion statistics
451 smaller than 1. Nevertheless, it is an unsatisfactory solution since the biased dispersion
452 statistic will also cause a loss of power.

453 A possible solution for GLMMs could be using a better approximation of the
454 residual degrees of freedom (*df*). For LMMs, approximations for denominator *df* have
455 been successfully used for hypothesis testing (Luke, 2017), for example, the
456 Satterthwaite (1946) and the Kenward-Roger (2009). Although there is some evidence
457 that these approximations are also accurate for GLMMs (Stroup, 2015), the main R
458 packages implementing some of these approximation methods are currently limited to
459 LMMs (e.g., *pbkrtest* Halekoh & Højsgaard, 2014; *lmerTest* Kuznetsova et al., 2017).
460 However, the recently released package *glmmrBase* (Watson, 2024) allows these
461 methods to be applied to GLMMs. We performed some parametric Pearson residuals
462 tests for Poisson GLMMs using a modified residual *df* approximation (see S8).
463 Although the parametric Pearson residuals tests with the approximated residual *df*
464 performed much better than those with the naïve residual *df*, they still underperformed

compared to the nonparametric Pearson test when having a large number of groups in the random effects (Fig. S8.4), especially for small-data situations.

When are simulation-based response variance tests an alternative?

The simulation-based response variance test developed in the R package DHARMa (Hartig, 2024) is the main alternative to the family of Pearson residuals tests. Its principle is simple: when the model is correctly specified, the variance of the observed data should match the variance of the data simulated from the model. The main advantage of this approach is that it is a non-parametric test that can be applied to any model structure and it does not require refitting the model, which makes it both considerably faster and easier to implement in statistical software. We also note that for GLMMs, simulations should be performed conditionally to avoid a loss of power, presumably due to the increased variability created by re-simulating the random effects (unconditional simulations).

The disadvantages of this approach are that it is often overly conservative, resulting in lower power compared to the Pearson residuals tests. Additionally, the calculated dispersion statistic differs from the Pearson dispersion statistic, making it difficult to compare the two approaches. We conjectured that both problems could be related to the fact that the test statistic is based on the raw variance (and not a scaled variance, as for the Pearson statistics), and therefore observations with large values may be overrepresented in the statistics. We considered scaling each observation with expected variance, but this is not readily available for a wide class of models, and using simulations to approximate it fails for discrete-valued distributions (see S7).

Conclusions and recommendations

In conclusion, while neither of the considered options excelled in all dimensions (Fig. 6), our base recommendation is that for standard GLMs with sufficient data, the parametric Pearson Chi-square test, available in many packages, can be safely used. In complex situations, particularly for GLMMs, we recommend the nonparametric Pearson residuals test. It has very few weaknesses, other than being computationally costly to calculate. If the nonparametric Pearson residuals test cannot be calculated due to speed or convergence problems with refitting complex models, we recommend using the simulation-based response variance test with simulations performed conditionally on the fitted random effects. All three approaches are available via the *testDispersion* function in the DHARMa R package (Hartig, 2024). We provide a tutorial with instructions and an example for applying dispersion tests using the DHARMa package on the repository website (https://theoreticalecology.github.io/dispersionTest/).

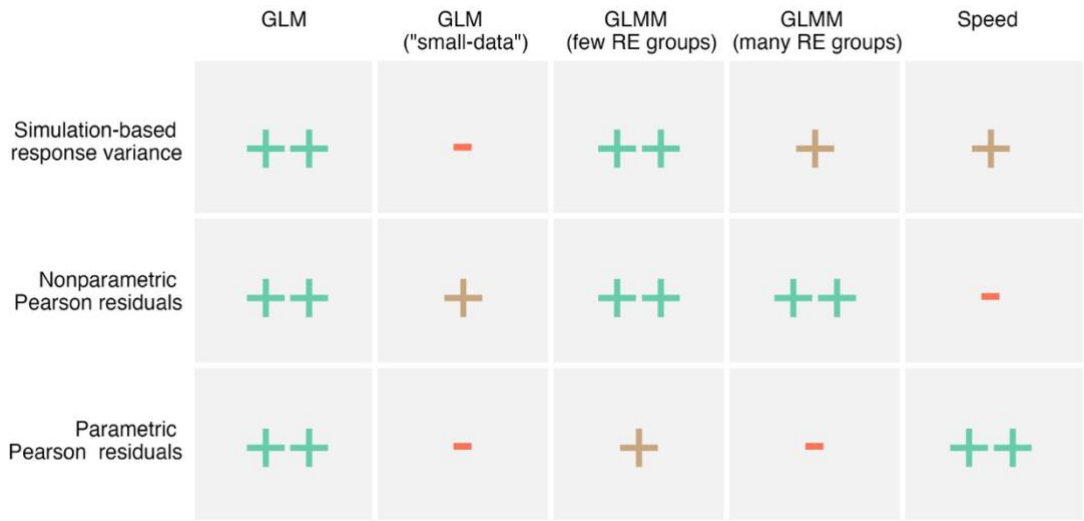


Figure 6. Performance comparisons of the dispersion tests evaluated for each "dimension" for Poisson and binomial models: GLMs in general, GLMs with small sample size or intercept ("small data"), GLMMs with one random effect with few groups/levels, GLMMs with many groups/levels in a random effect, and computational time for calculating the test (speed). The symbols mean: "-" bad performance, "+" good performance, "++" very good performance.

Although our simulation examples concentrated on overdispersion, the tests under consideration in our study can equally be used to detect underdispersion problems by testing the dispersion "two-sided" or "less than" against null statistics. The clear exception would be testing for underdispersion using the parametric Pearson residuals test for GLMMs, which would be anti-conservative due to the discussed bias towards underdispersion in the presence of random effects.

Recommendations for practical data analysis when using dispersion tests

For the interpretation and applied data analysis, we stress that getting a significant over-/underdispersion result does not necessarily indicate that the distribution must be changed. First, hypothesis tests famously evaluate statistical rather than practical significance. In other words, a significant test for overdispersion indicates that the overdispersion signal deviates from a null expectation, but the p-value does not measure the strength of the deviation. The first step in a dispersion test should thus be to examine how much the dispersion statistic deviates from the expected value of 1. For very large sample sizes, small departures from 1 may be statistically significant, but they may not necessarily warrant a change to the model. Second, after finding that a dispersion problem is both significant and meaningful, we suggest first checking for problems other than the distribution, such as heteroscedasticity, missing predictors, incorrect link function, excess of zeros, or overfitting. In our experience, these types of model misspecifications often cause over-/underdispersion, but can be distinguished from a "real" distributional problem through careful residual checks. Blindly changing the distribution only masks the problem, without offering a real remedy to the underlying problems.

Finally, after having convinced ourselves through these previous investigations that we are facing an 'intrinsic' under-/overdispersion problem, we should consider changing the GLM distribution. A traditional and flexible solution is using the 'quasi' distributions (Wedderburn, 1974), which essentially correct p-values, but have the disadvantage that they do not represent an explicit data-generating process with associated likelihood, which does not allow, for example, to simulate from the fitted model. A second alternative to add dispersion is using observation-level random effects (Bolker et al., 2009; Elston et al., 2001; Harrison, 2014; Ozgul et al., 2009). While often offering a reasonable solution, we feel that the excessive use of REs tends to create problems in the calculation of other statistical indicators (such as p-values) that we would rather avoid. For that reason, we feel the best solution to address 'intrinsic' under-/overdispersion is to switch to the corresponding variable-dispersion distributions, such as the negative binomial (Harrison, 2014) for overdispersed and the Conway-Maxwell-Poisson distribution (Lynch et al., 2014) for underdispersed Poisson models, or the beta-binomial distribution for overdispersed binomial models (Harrison, 2015). Regardless of the approach, an "over-/underdispersion-free" GLM/GLMM is essential for better interpreting model results and facilitating sound scientific discoveries.

References

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*(4), 390–412. https://doi.org/10.1016/j.jml.2007.12.005

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278. https://doi.org/10.1016/j.jml.2012.11.001

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Bolker, B. M., Brooks, M. E., Clark, C. J., Geange, S. W., Poulsen, J. R., Stevens, M. H. H., & White, J.-S. S. (2009). Generalized linear mixed models: A practical guide for ecology and evolution. *Trends in Ecology & Evolution*, *24*(3), 127–135. https://doi.org/10.1016/j.tree.2008.10.008

Cameron, A. C., & Trivedi, P. (2023). *overdisp: Overdispersion in count data multiple regression analysis* (Version 0.1.2) [Computer software]. https://CRAN.R-project.org/package=overdisp

Cameron, A. C., & Trivedi, P. K. (1990). Regression-based tests for overdispersion in the Poisson model. *Journal of Econometrics*, *46*(3), 347–364. https://doi.org/10.1016/0304-4076(90)90014-K

Campbell, H. (2021). The consequences of checking for zero-inflation and overdispersion in the analysis of count data. *Methods in Ecology and Evolution*, *12*(4), 665–680. https://doi.org/10.1111/2041-210X.13559

Collings, B. J., & Margolin, B. H. (1985). Testing Goodness of Fit for the Poisson Assumption When Observations Are Not Identically Distributed. *Journal of the American Statistical Association*, *80*(390), 411–418.

Cordeiro, G. M. (2004). On Pearson's residuals in generalized linear models. *Statistics & Probability Letters*, *66*(3), 213–219. https://doi.org/10.1016/j.spl.2003.09.004

Cordeiro, G. M., & Simas, A. B. (2009). The distribution of Pearson residuals in generalized linear models. *Computational Statistics & Data Analysis*, *53*(9), 3397–3411. https://doi.org/10.1016/j.csda.2009.02.025

Dean. (1992). Testing for Overdispersion in Poisson and Binomial Regression Models. *Journal of the American Statistical Association*, *87*(418), 451–457. https://doi.org/10.2307/2290276

Dean, C., & Lawless, J. F. (1989). Tests for Detecting Overdispersion in Poisson Regression Models. *Journal of the American Statistical Association*, *84*(406), 467–472. https://doi.org/10.1080/01621459.1989.10478792

Dunn, P. K., & Smyth, G. K. (2018). *Generalized Linear Models With Examples in R*. Springer New York. https://doi.org/10.1007/978-1-4419-0118-7

Elston, D. A., Moss, R., Boulinier, T., Arrowsmith, C., & Lambin, X. (2001). Analysis of aggregation, a worked example: Numbers of ticks on red grouse chicks. *Parasitology*, *122*(05), 563–569.

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.

Feng, C., Li, L., & Sadeghpour, A. (2020). A comparison of residual diagnosis tools for diagnosing regression models for count data. *BMC Medical Research Methodology*, *20*(1), 175. https://doi.org/10.1186/s12874-020-01055-2

Fisher, R. A. (1950). The Significance of Deviations from Expectation in a Poisson Series. *Biometrics*, *6*(1), 17–24. https://doi.org/10.2307/3001420

Fletcher, D. J. (2012). Estimating overdispersion when fitting a generalized linear model to sparse data. *Biometrika*, *99*(1), 230–237. https://doi.org/10.1093/biomet/asr083

Halekoh, U., & Højsgaard, S. (2014). A Kenward-Roger Approximation and Parametric Bootstrap Methods for Tests in Linear Mixed Models – The R Package pbkrtest. *Journal of Statistical Software*, *59*, 1–32. https://doi.org/10.18637/jss.v059.i09

Harrison, X. A. (2014). Using observation-level random effects to model overdispersion in count data in ecology and evolution. *PeerJ*, *2*, e616. https://doi.org/10.7717/peerj.616

Harrison, X. A. (2015). A comparison of observation-level random effect and Beta-Binomial models for modelling overdispersion in Binomial data in ecology & evolution. *PeerJ*, *3*, e1114. https://doi.org/10.7717/peerj.1114

Hartig, F. (2024). *DHARMa: Residual Diagnostics for Hierarchical (Multi-Level / Mixed) Regression Models* (Version 0.4.7) [Computer software]. https://CRAN.R-project.org/package=DHARMa

Herve, M. (2025). *RVAideMemoire: Testing and plotting procedures for biostatistics* (Version 0.9-83-11) [Computer software]. https://CRAN.R-project.org/package=RVAideMemoire

Hilbe, J. M. (2011). *Negative Binomial Regression*.

Hilbe, J. M. (2014). *Modeling count data*. Cambridge University Press.

Hilbe, J., & Robinson, A. (2025). *msme: Functions and datasets for "methods of statistical model estimation"* (Version 0.5.4) [Computer software]. https://CRAN.R-project.org/package=msme

Jackman, S. (2024). *pscl: Classes and methods for R developed in the political science computational laboratory* (Version 1.5.9) [Computer software]. University of Sydney. https://github.com/atahk/pscl/

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

Kleiber, C., & Zeileis, A. (2008). *Applied econometrics with R [R package AER version 1.2-14]*. Springer-Verlag. https://doi.org/10.1007/978-0-387-77318-6

Kuss, O. (2002). Global goodness-of-fit tests in logistic regression with sparse data. *Statistics in Medicine*, *21*(24), 3789–3801. https://doi.org/10.1002/sim.1421

Kuznetsova, A., Brockhoff, P., & Christensen, R. (2017). lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software, Articles*, *82*(13). https://doi.org/10.18637/JSS.V082.I13

Lai, J., Lortie, C. J., Muenchen, R. A., Yang, J., & Ma, K. (2019). Evaluating the popularity of R in ecology. *Ecosphere*, *10*(1), e02567. https://doi.org/10.1002/ecs2.2567

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

Lopez-Quílez, V. G.-R. J. F.-F. A. (2005). Detecting clusters of disease with R. *Journal of Geographical Systems*, *7*(2), 189–206. https://doi.org/10.1007/s10109-005-0156-5

Lüdecke, D., Ben-Shachar, M. S., Patil, I., Waggoner, P., & Makowski, D. (2021). performance: An R package for assessment, comparison and testing of statistical

models. *Journal of Open Source Software*, *6*(60), 3139. https://doi.org/10.21105/joss.03139

Luke, S. G. (2017). Evaluating significance in linear mixed-effects models in R. *Behavior Research Methods*, *49*(4), 1494–1502. https://doi.org/10.3758/s13428-016-0809-y

Lynch, H. J., Thorson, J. T., & Shelton, A. O. (2014). Dealing with under- and over-dispersed count data in life history, spatial, and community ecology. *Ecology*, *95*(11), 3173–3180. https://doi.org/10.1890/13-1912.1

McCullagh, P. (1985). On the Asymptotic Distribution of Pearson's Statistic in Linear Exponential-Family Models. *International Statistical Review / Revue Internationale de Statistique*, *53*(1), 61–67. https://doi.org/10.2307/1402880

McCullagh, P., & Nelder, J. (1989). Generalized linear models. *Journal of the Royal Statistical Society*, *135*(3), 370–384.

Moral, R. A., Hinde, J., & Demétrio, C. G. B. (2017). Half-Normal Plots and Overdispersed Models in R: The hnp Package. *Journal of Statistical Software*, *81*, 1–23. https://doi.org/10.18637/jss.v081.i10

Ohara Hines, R. J. (1997). A comparison of tests for overdispersion in generalized linear models. *Journal of Statistical Computation and Simulation*, *58*(4), 323–342. https://doi.org/10.1080/00949659708811838

Ozgul, A., Oli, M. K., Bolker, B. M., & Perez-Heydrich, C. (2009). Upper respiratory tract disease, force of infection, and effects on survival of gopher tortoises. *Ecological Applications*, *19*(3), 786–798.

Papadakis, M., Tsagris, M., Fafalios, S., Dimitriadis, M., & Lasithiotakis, M. (2025). *Rfast2: A collection of efficient and extremely fast R functions II* (Version 0.1.5.4) [Computer software]. https://CRAN.R-project.org/package=Rfast2

Quine, M. P., & Seneta, E. (1987). Bortkiewicz's Data and the Law of Small Numbers. *International Statistical Review / Revue Internationale de Statistique*, *55*(2), 173–181. https://doi.org/10.2307/1403193

R Core Team. (2024). *R: a language and environment for statistical computing* (Version v4.4.1) [Computer software]. R Foundation for Statistical Computing. https://www.R-project.org/

Rao, C. R. (1948). Large sample tests of statistical hypotheses concerning several parameters with applications to problems of estimation. *Mathematical Proceedings of the Cambridge Philosophical Society*, *44*(1), 50–57. https://doi.org/10.1017/S0305004100023987

Rhodes, J. R. (2015). Mixture models for overdispersed data. In G. A. Fox, V. J. Sosa, & S. M. Negrete-Yankelevich, *Ecological Statistics: Contemporary theory and applications*. Oxford University Press.

Satterthwaite, F. E. (1946). An Approximate Distribution of Estimates of Variance Components. *Biometrics Bulletin*, *2*(6), 110–114. https://doi.org/10.2307/3002019

Schaalje, G. B., McBride, J. B., & Fellingham, G. W. (2002). Adequacy of approximations to distributions of test statistics in complex mixed linear models.

687 *Journal of Agricultural, Biological, and Environmental Statistics*, *7*(4), 512–
688 524. https://doi.org/10.1198/108571102726

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

692 Student. (1919). An explanation of deviations from poisson's law in practice.
693 *Biometrika*, *12*, 211.

694 Touchon, J. C., & McCoy, M. W. (2016). The mismatch between current statistical
695 practice and doctoral training in ecology. *Ecosphere*, *7*(8), e01394.
696 https://doi.org/10.1002/ecs2.1394

697 Venables, W. N., & Ripley, B. D. (2002). *Modern Applied Statistics with S*. Springer
698 New York. https://doi.org/10.1007/978-0-387-21706-2

699 Watson, S. I. (2024). *Generalised Linear Mixed Model Specification, Analysis, Fitting,*
700 *and Optimal Design in R with the glmmr Packages* (No. arXiv:2303.12657).
701 arXiv. https://doi.org/10.48550/arXiv.2303.12657

702 Wedderburn, R. W. M. (1974). Quasi-likelihood functions, generalized linear models,
703 and the Gauss—Newton method. *Biometrika*, *61*(3), 439–447.
704 https://doi.org/10.1093/biomet/61.3.439

705 Xekalaki, E. (2014). On the distribution theory of over-dispersion. *Journal of Statistical*
706 *Distributions and Applications*, *1*(1), 19. https://doi.org/10.1186/s40488-014-
707 0019-z

708 Yang, Z., Hardin, J. W., Addy, C. L., & Vuong, Q. H. (2007). Testing Approaches for
709 Overdispersion in Poisson Regression versus the Generalized Poisson Model.
710 *Biometrical Journal*, *49*(4), 565–584. https://doi.org/10.1002/bimj.200610340

711 Zeileis, A., & Hothorn, T. (2002). Diagnostic checking in regression relationships. *R*
712 *News*, *2*(3), 7–10.

713

Supplementary Material of

Dispersion tests in generalised linear mixed-effects models - a methods comparison and practical guide

Melina de Souza Leite[1*], Daniel Rettelbach[1,2] & Florian Hartig[1]

1. Theoretical Ecology, University of Regensburg, Germany
2. coTrial Associates, Department of Surgery, University Hospital Regensburg, Germany (current address)

*corresponding author: melina.souza-leite@ur.de

S1. Pearson statistics and Chi-squared distribution

For GLMs, the parametric Pearson residuals test assumes that the sample size (n-asymptotic) and the expected values are sufficiently large (phi-asymptotic). Therefore, when the expected counts (or intercept) and/or the number of observations are small, Pearson residuals may not provide reliable information about model fit. To test boundaries where Pearson statistics fail, we simulated data with very different sample sizes (from 10 to 10,000, depending on the simulation) and intercepts (from -3 to 3, at the link function scale) for Poisson and binomial proportion GLMs. For each distribution and parameter combination, we used the Kolmogorov-Smirnov test (KS test) of adherence to compare the empirical distribution of 1000 simulations of the Pearson residuals with the Chi-squared distribution having the same residual degrees of freedom. We repeated this procedure 100 times and recorded the proportion of significant KS tests.

For the Poisson GLMs, the Pearson statistics distribution clearly departed from the Chi-square distribution for very small intercepts (-3, -1.5) and sample sizes (10, 20 and 50) (Figure S1.1 A). Even for very large sample sizes (10,000), the distribution did

not approximate the Chi-squared distribution for the smallest simulated intercept (-3). Consequently, the KS tests showed all significant results for all simulations with the intercept at -3, except for the largest sample size (10,000), where it decreased to 60%. As expected, the proportion of significant results decreased with sample size for intercepts at -1.5 and 0. For larger intercepts, it remained around 5% for all sample sizes (Figure S1.2A).

For the **binomial GLMs**, the Pearson statistics distribution clearly departed from the Chi-squared distribution for very small and large intercepts (-3, 3) and small sample sizes (10, 20, 50) (Figure S1.1B). The proportion of significant KS tests decreased with sample size, but did not reach the nominal value of 0.05, even for very large sample sizes and intermediate intercept values (-1.5, 0, 1.5).

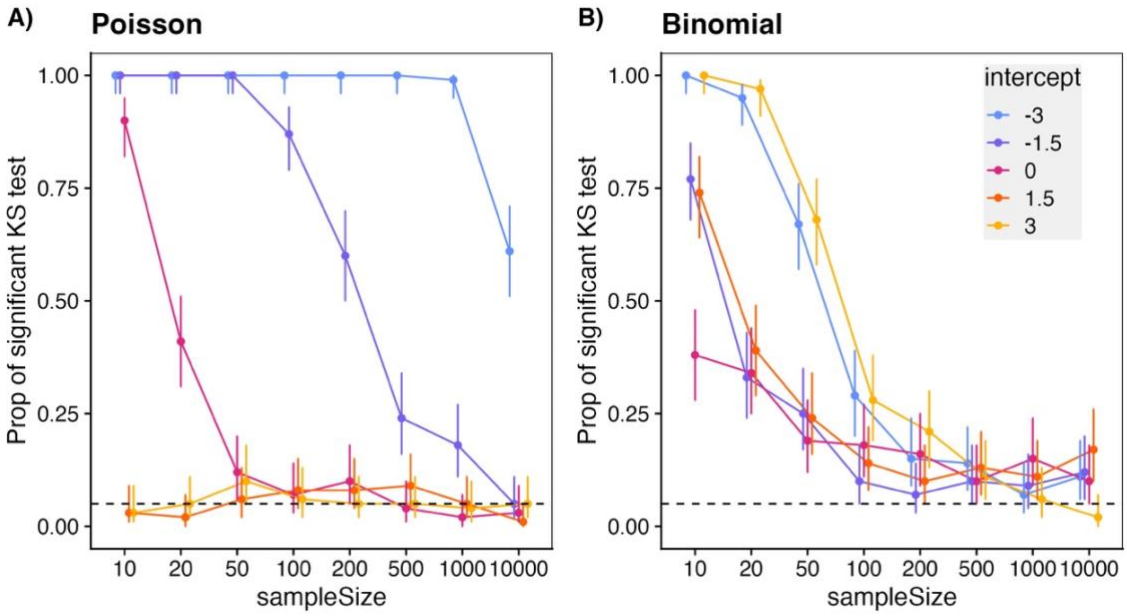


Figure S1.1. Proportion of significant Kolmogorov-Smirnov adherence tests between the empirical distribution of 1000 simulations of the Pearson statistics and a Chi-squared distribution with the same residual degrees of freedom for A) Poisson and B) binomial GLMs. Proportions were calculated from 100 simulations for each combination of the data parameters (sample size and intercept). For binomial data, the number of trials was fixed at 10. The 95% confidence intervals (vertical lines) were drawn from binomial exact tests for each result with p = 0.05.

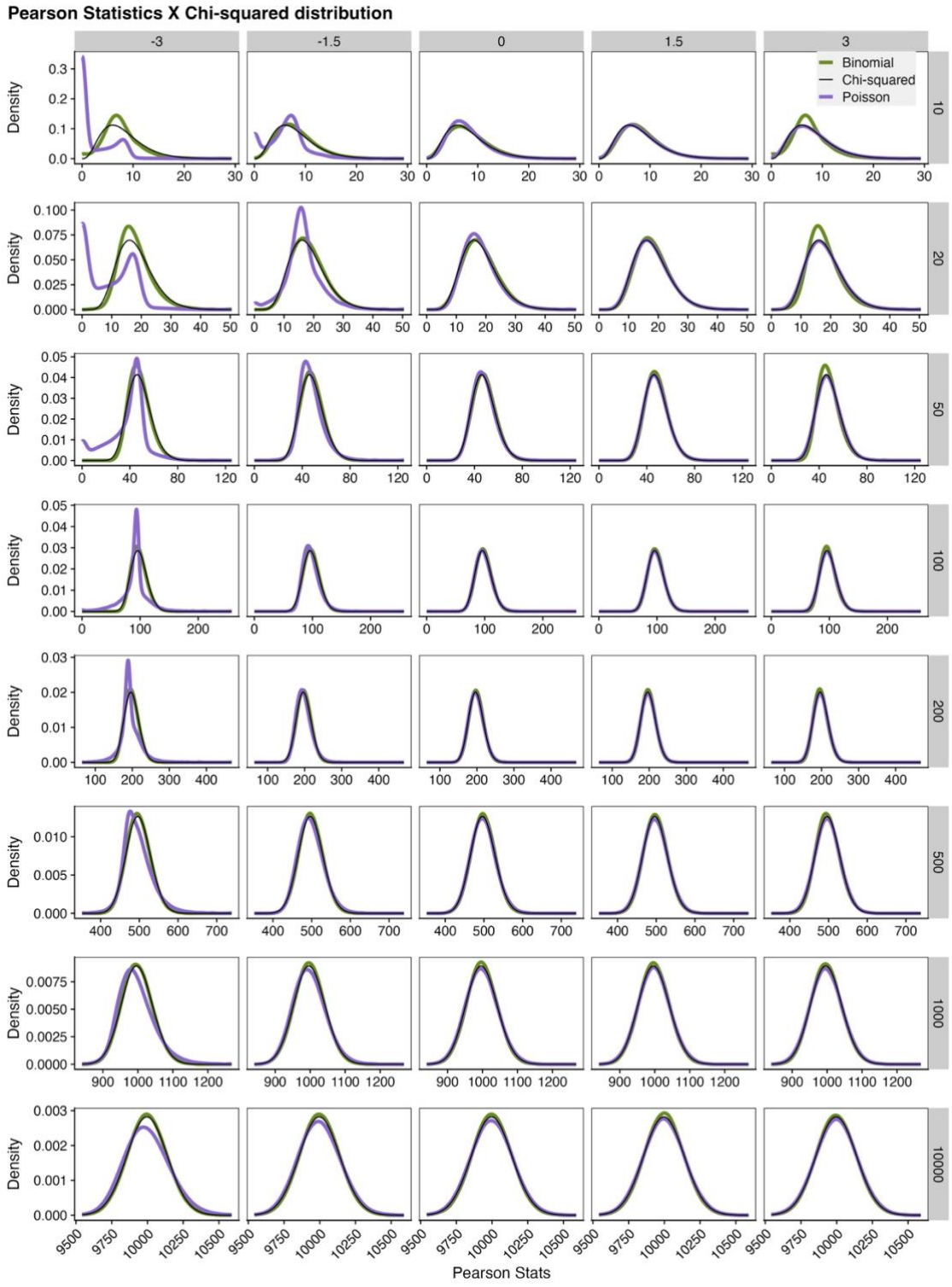


Figure S1.2. Mean Pearson statistics distribution (from 100 simulated curves) for the binomial (green) and Poisson (purple), and the Chi-square distribution in black.

S2. Type I error rates for the GLMs

Figures S2.1 and S2.2 show the distribution of the p-values for the dispersion tests applied to the Poisson and binomial GLMs, respectively, with 10,000 simulations for each combination of intercept and sample size. For the dispersion tests with correct type I error rates around the nominal value of 0.05, the distributions of p-values should present a uniform distribution with density 1.

For the Poisson GLMs (Figure S2.1), the simulation-based response variance test (in red) presented the largest departure of the expected distribution for the smallest intercepts (-3, -1.5) across all sample sizes. This explains why the type I error rates for the simulation-based residual tests were so low and varied according to the intercept but didn't change with the sample size (main text Figure 2A). The parametric Pearson test had the opposite pattern with very low p-values for the smallest intercept (-3), but it tended to approximate the uniform distribution (decreasing the peak for the low p-values) with sample size. The p-values for the nonparametric Pearson test also showed a departure from the uniform distribution for the smallest intercept (-3), but tended to approach the uniform distribution with larger sample sizes and intercepts.

For the binomial GLMs (Figures S2.2), the p-values distribution of the simulation-based response variance test also presented the largest departure from the uniform distribution, but for all intercepts and sample sizes. The p-values for both parametric Pearson and nonparametric Pearson tests were similar and tended towards the uniform distribution with larger sample sizes.

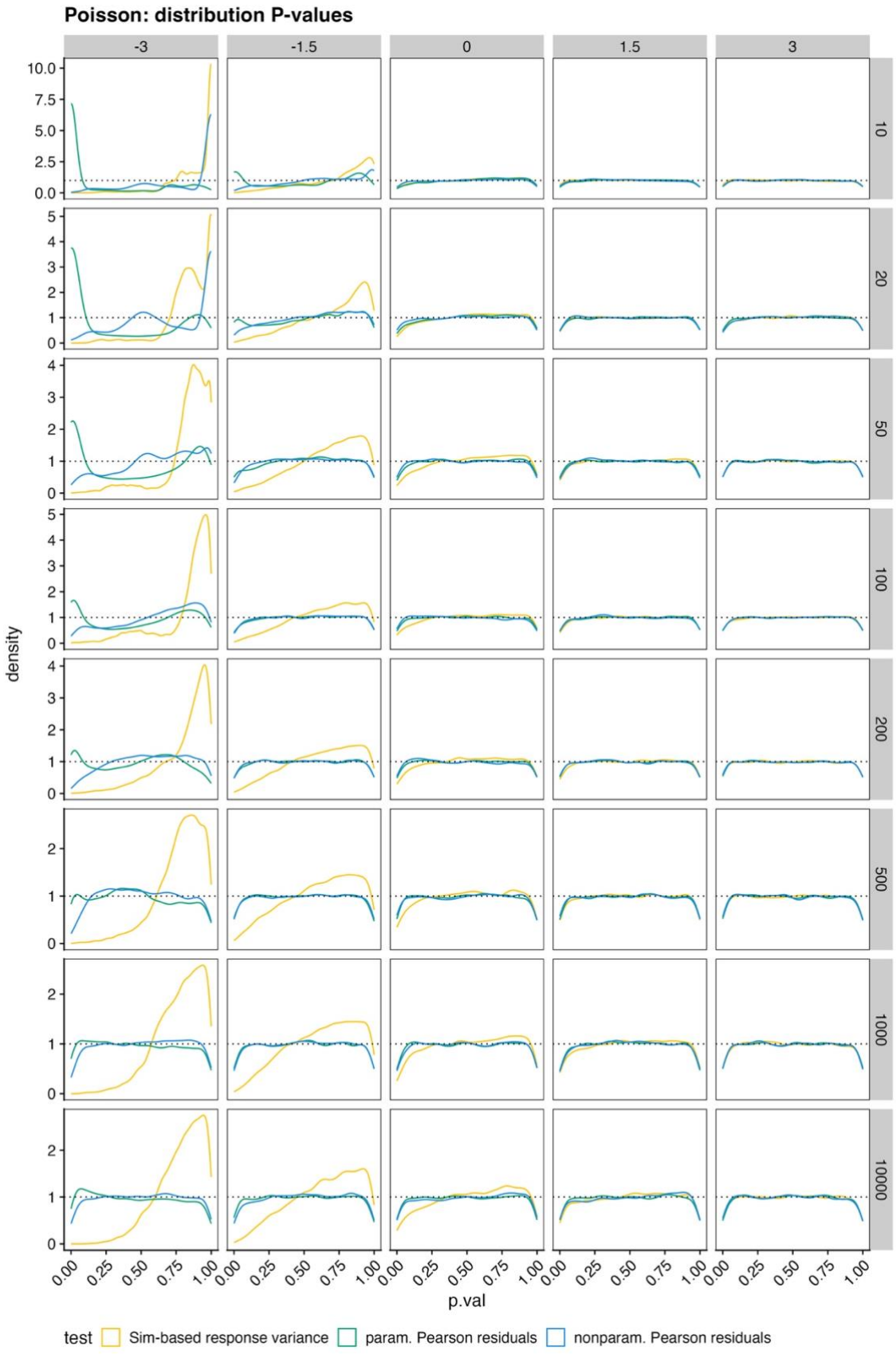


Figure S2.1. Distribution of p-values for the Poisson GLMs for each dispersion test. 10,000 simulations per simulation set (intercept x sample size).

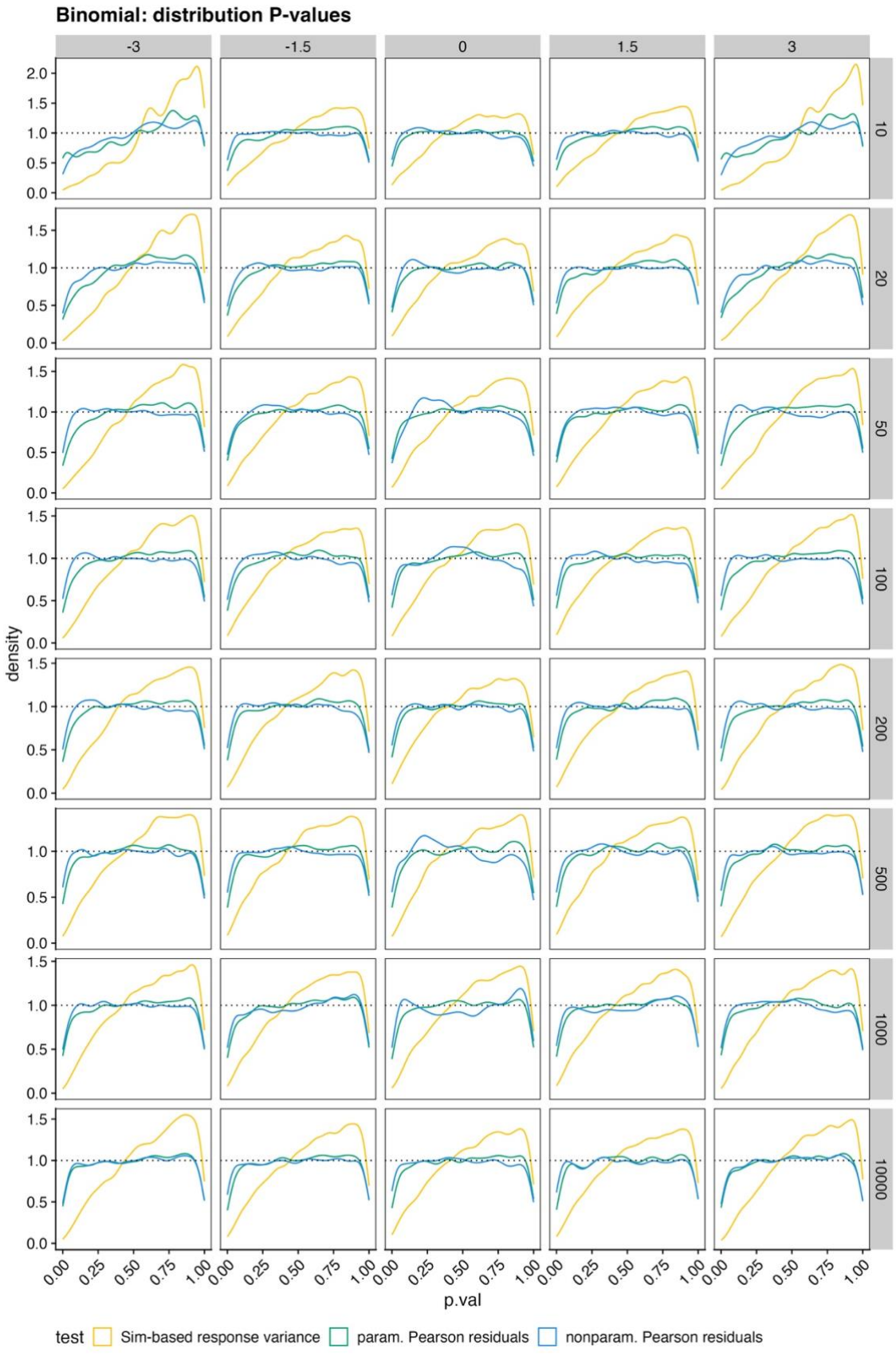
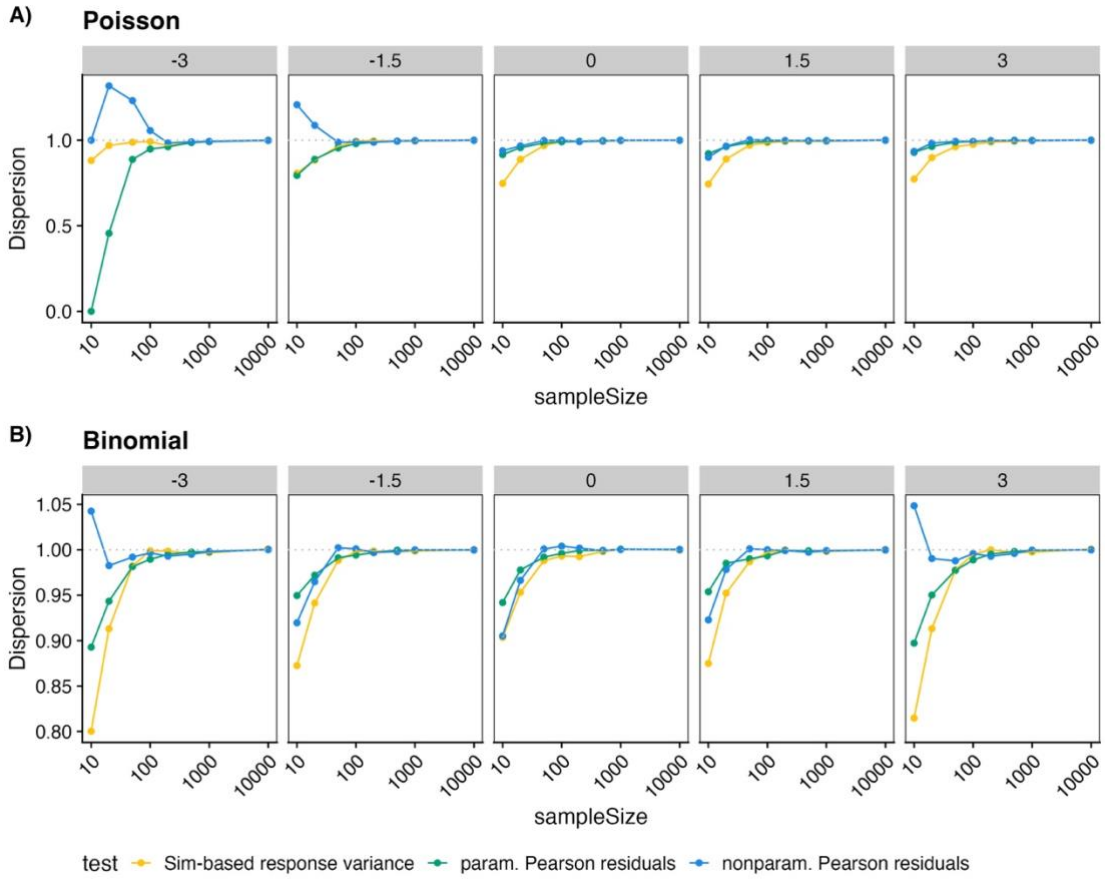


Figure S2.2. Distribution of p-values for the binomial GLMs for each dispersion test. 10,000 simulations per simulation set (intercept x sample size).

S3. Dispersion statistics for GLMs

The dispersion statistics of the tests for GLMs tended to be smaller than 1 (expected value) when there was no overdispersion simulated for very small sample sizes for both binomial and Poisson distributions (Figure S3.1). The exception was the nonparametric Pearson test that presented values larger than 1 for the very small intercepts (-3 in both distributions, 3 in binomial only). When comparing dispersion statistics for the simulated overdispersed data (Figures S3.2 and S3.3), we found that both Pearson-based dispersion statistics presented similar values. In contrast, the dispersion statistic of the simulation-based response variance presented lower values for small sample sizes. The differences in dispersion statistics between tests tended to increase with the increase of simulated overdispersion, but in opposite directions for binomial and Poisson GLMs (Figure S3.4 and S3.5). Moreover, we found out that the dispersion statistics of the simulation-based response variance test depend heavily on the slope parameter of the simulated data (Figure S3.6).



88

89 **Figure S3.1.** Median of the dispersion statistics of the tests for A) Poisson and B)
90 binomial GLMs, simulated without overdispersion for different intercepts (panels) and
91 sample sizes (x-axis) for the three dispersion tests: parametric Pearson test,
92 nonparametric Pearson test, and simulation-based response variance test. The dotted
93 horizontal line indicates the ratio of 1. Values below the line are considered
94 underdispersion, and above the line are overdispersion. For all simulations, the slope
95 was fixed at 1.

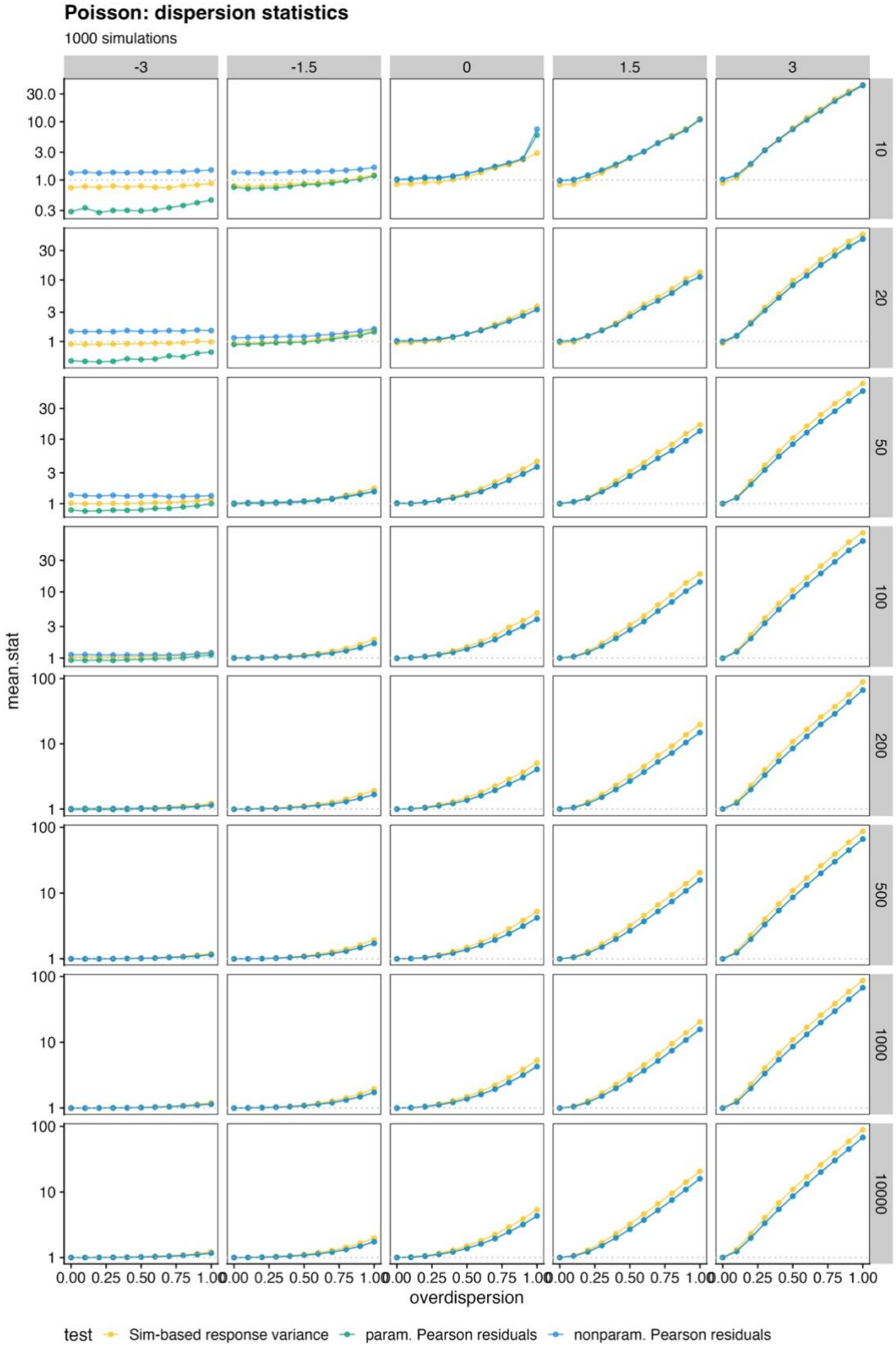


Figure S3.2. Dispersion statistics (median) for GLM Poisson. Notice the different y-axis scales across sample sizes.

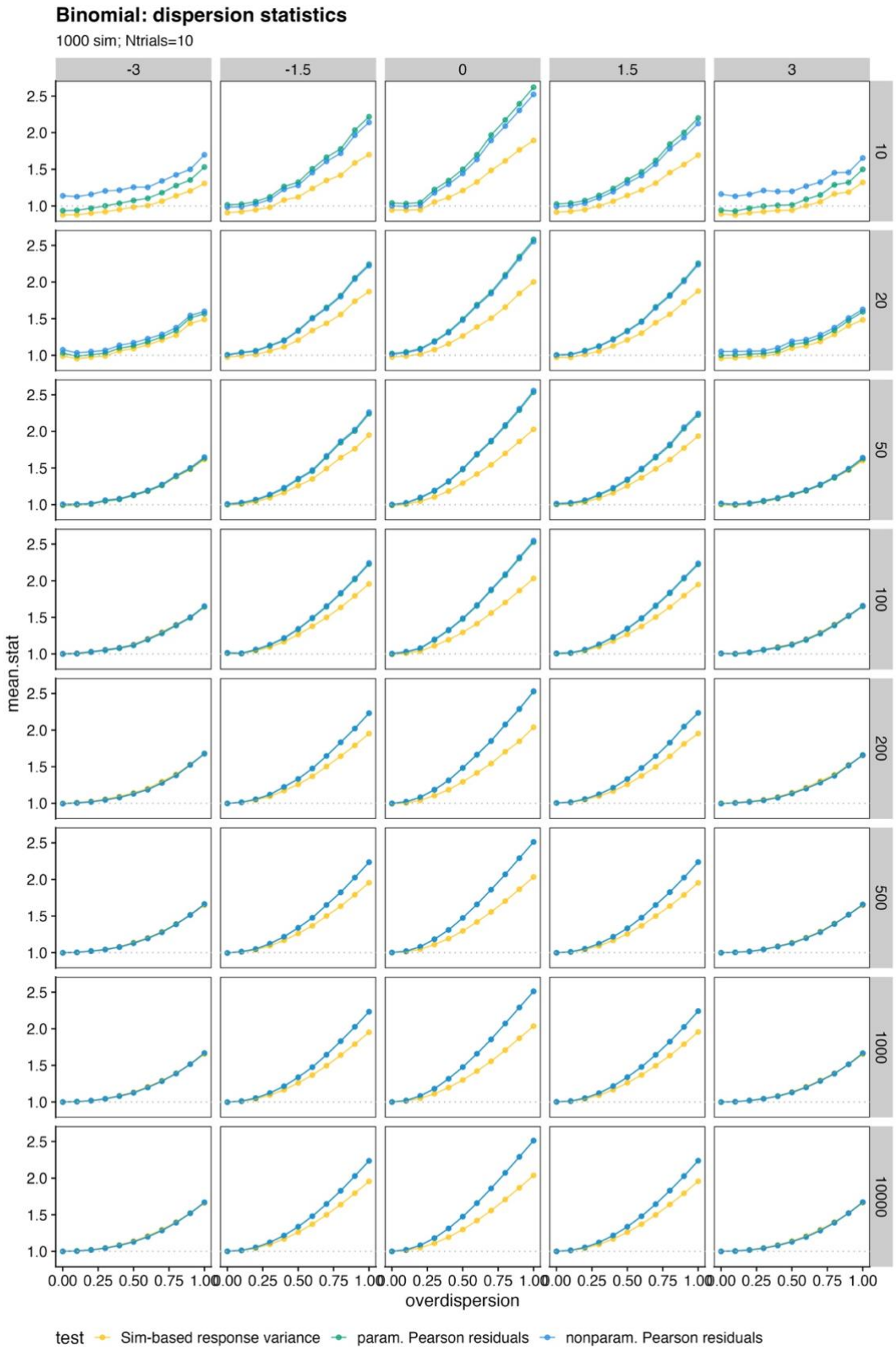


Figure S3.3. Dispersion statistics (median) for GLM binomial.

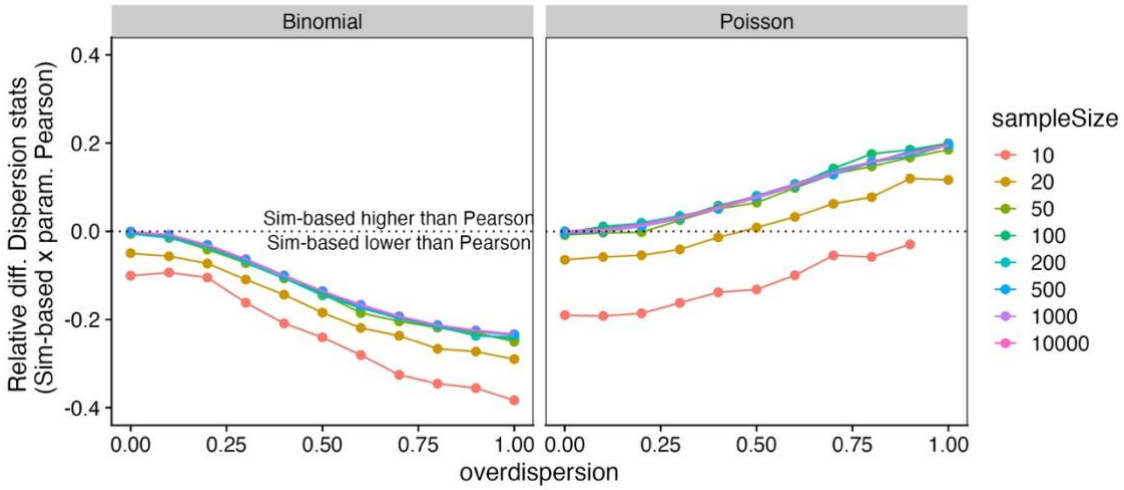


Figure S3.4. The dispersion statistics of the simulation-based response variance test are smaller than the parametric Pearson test statistics for all binomial models and for small sample sizes in Poisson models. The differences between the two dispersion statistics decrease with increasing sample size (coloured lines) and increase with simulated overdispersion in the data (x-axis). The relative differences (y-axis) were calculated by subtracting the simulation-based dispersion statistics from the parametric Pearson statistic, then dividing by the simulation-based statistic, and can be interpreted as the difference in the percentage of the simulation-based statistics. The results presented are based on 1,000 simulations with zero intercepts.

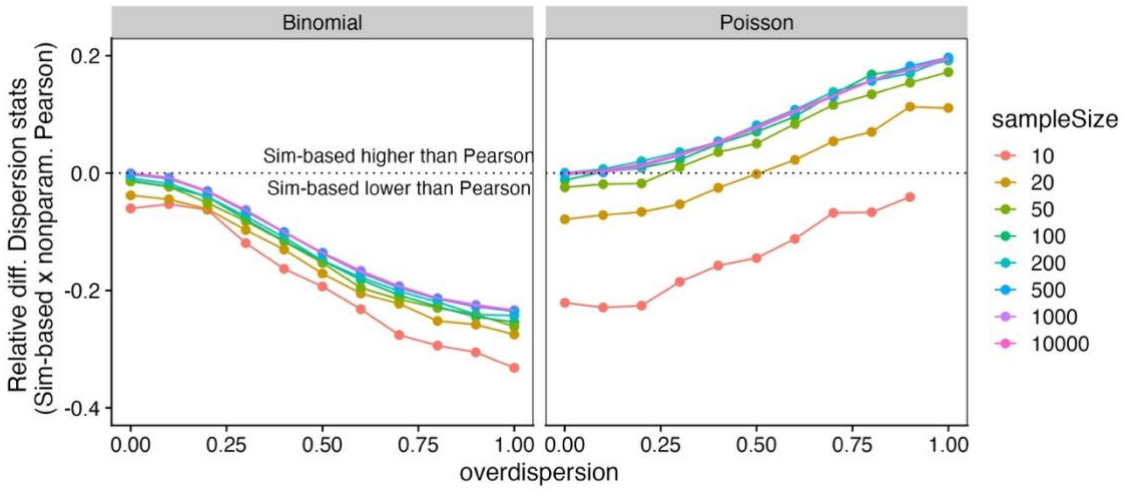


Figure S3.5. The dispersion statistics of the simulation-based response variance test are smaller than nonparametric Pearson dispersion statistics for all binomial models and for small sample sizes in Poisson models. The differences between the two dispersion statistics decrease with increasing sample size (coloured lines) and increase with simulated overdispersion in the data (x-axis). The relative differences (y-axis) were calculated by subtracting the Parametric Bootstrapping statistics from the simulation-based dispersion statistics, then dividing by the simulation-based statistics, and can be interpreted as the difference in the percentage of the simulation-based statistics. The results presented are based on 1,000 simulations with zero intercepts.

S4: Expanding simulation parameters for GLMs

Here, we investigated the possible influence of other parameters used to generate the datasets for binomial and Poisson GLMs. In Figure S4.1, we investigated the power and dispersion statistic for datasets simulated with different slopes (the default slope in all other simulations was 1). In Figure S4.2, we investigated the effect of varying the number of trials on the binomial GLMs in terms of power, type I error, and dispersion statistics.

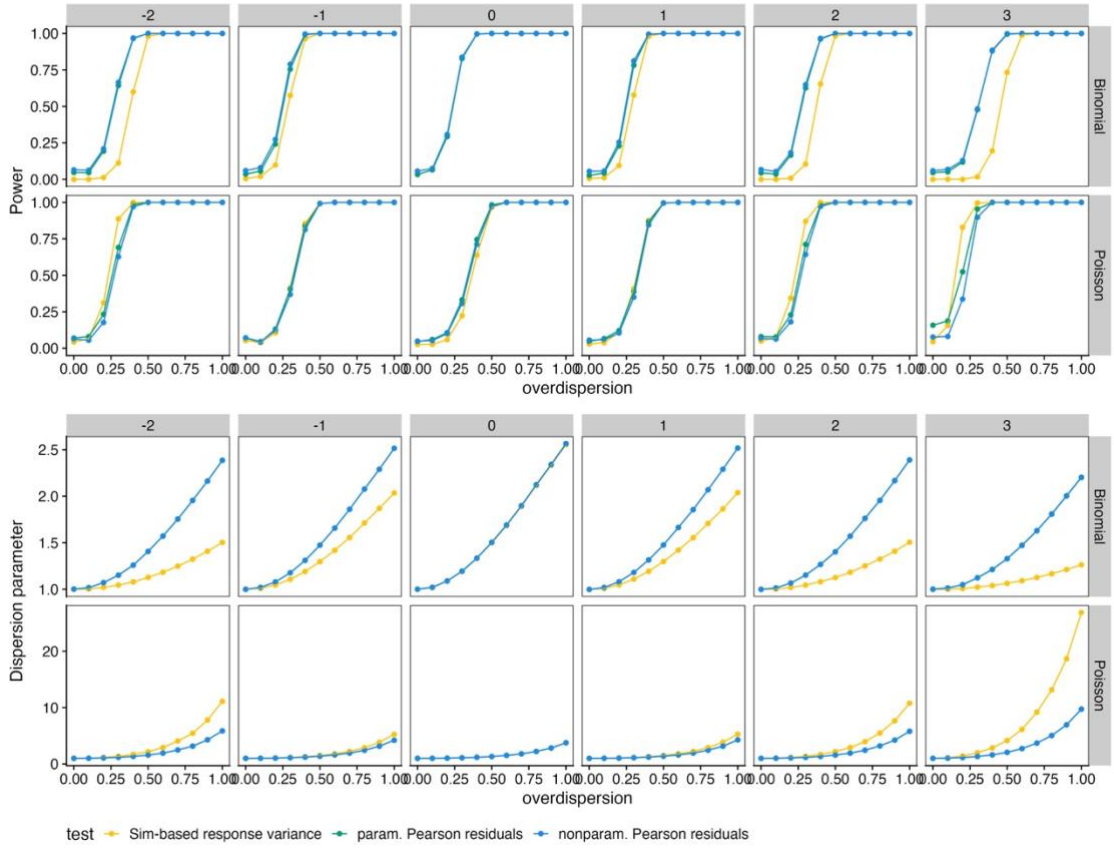


Figure S4.1. Power and dispersion statistics for simulations with different slopes (panel columns) for binomial and Poisson GLMs. Number of simulations = 500; intercept = 0, number of trials for the binomial = 10.

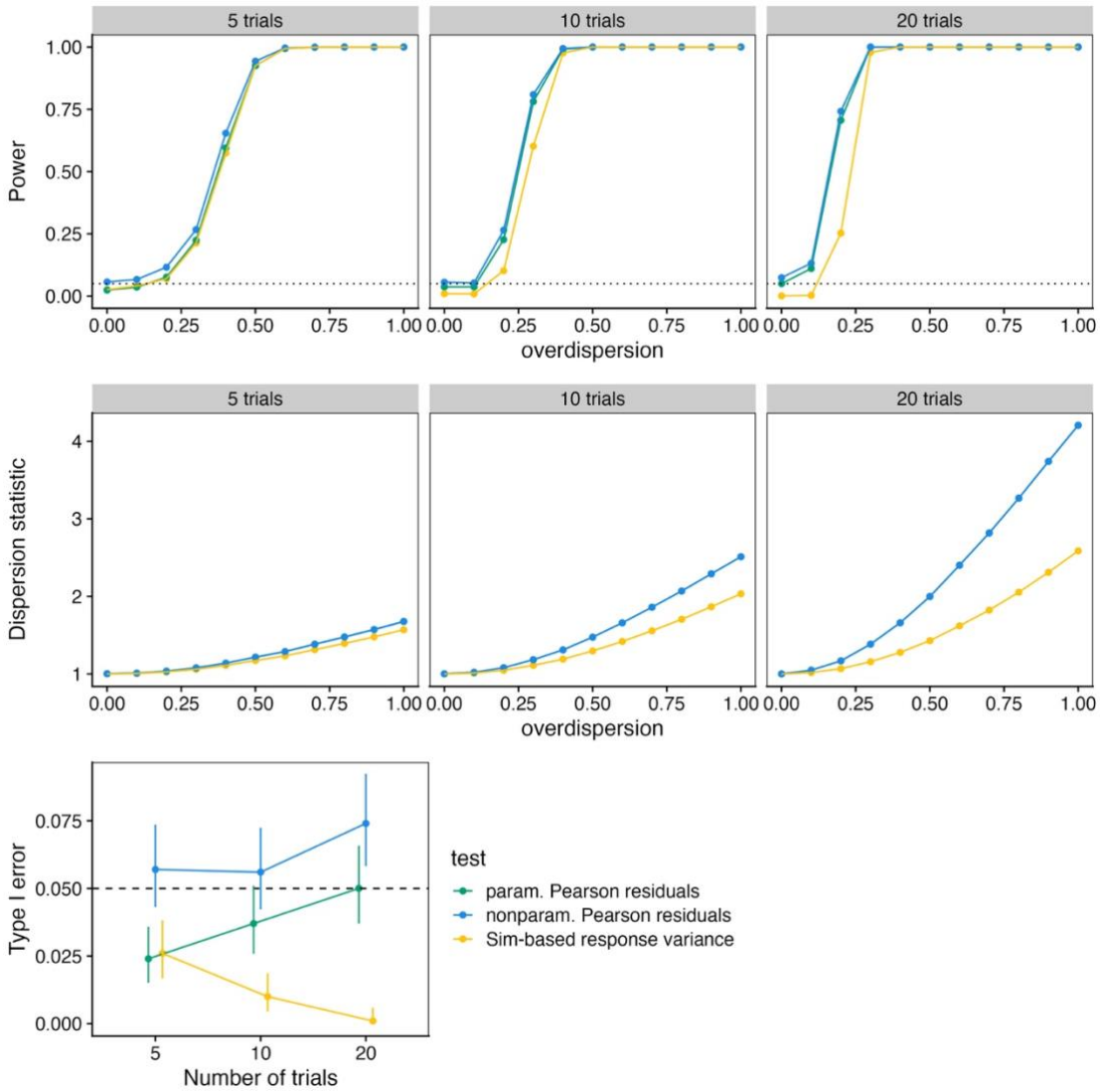
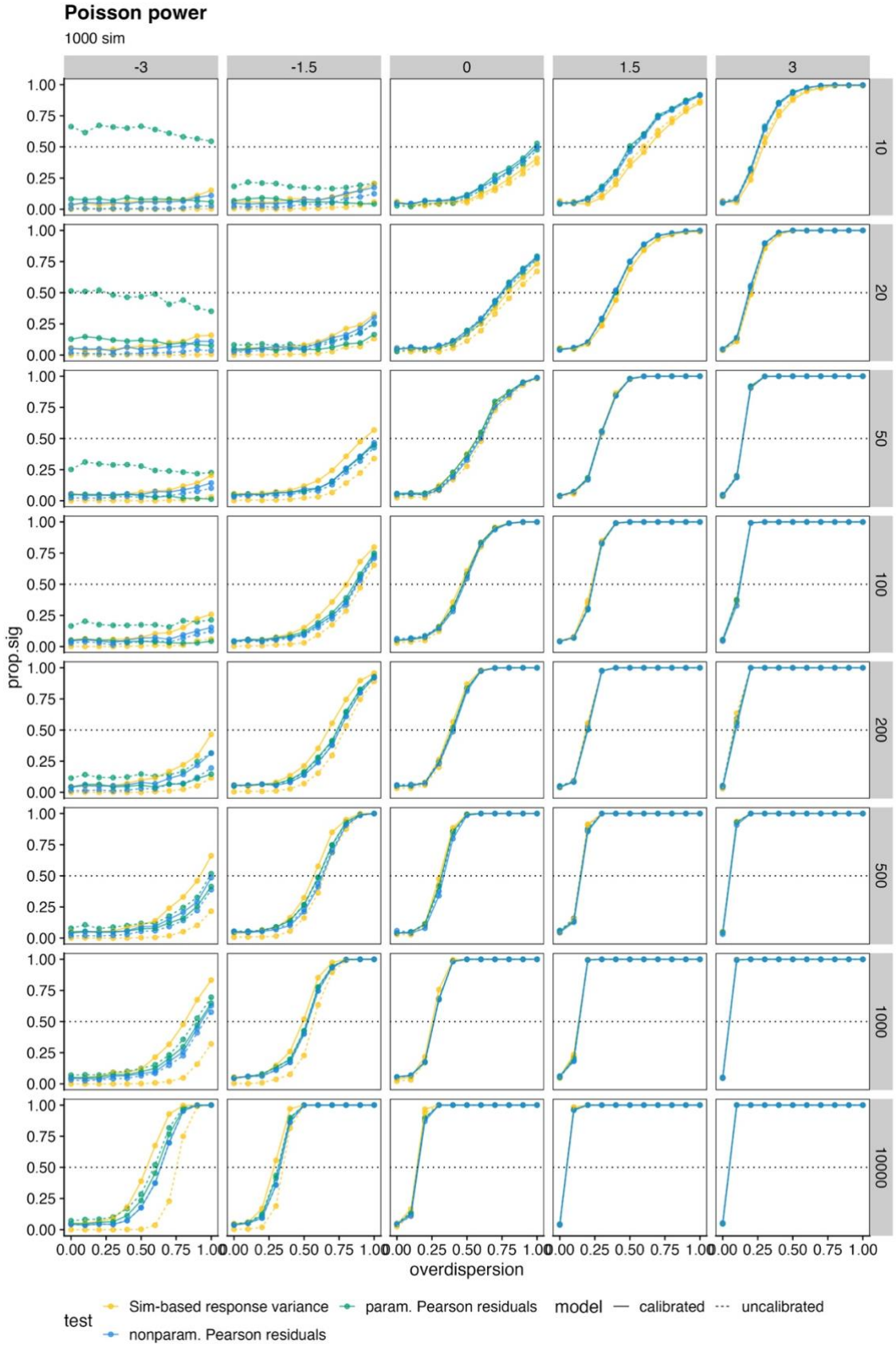


Figure S4.2. Power, dispersion statistics, and type I error of dispersion tests for binomial data simulations with different numbers of trials (panel columns). The fixed parameters are: intercept = 0, sample size = 500, slope = 1. Results for 1000 simulations.

S5. Power for the GLMs

Power calibration

To investigate if the lower power of the simulation-based response variance test is a consequence of the very conservative type I error rates, we calibrated the power using the p-value at the 5% quantile of the empirical distribution of p-values where the null hypothesis was true for each set of simulations (Figures S2.1 and S2.2). This method should provide an estimate of differences in power, controlling for type I error rate (Luke et al. 2017). Figures S5.1 and S5.2 show the power (calibrated and uncalibrated) of the dispersion tests for each simulation set (intercept, sample size and overdispersion) for Poisson and binomial GLMs, respectively.



147

148 **Figure S5.1.** Power for GLM Poisson.

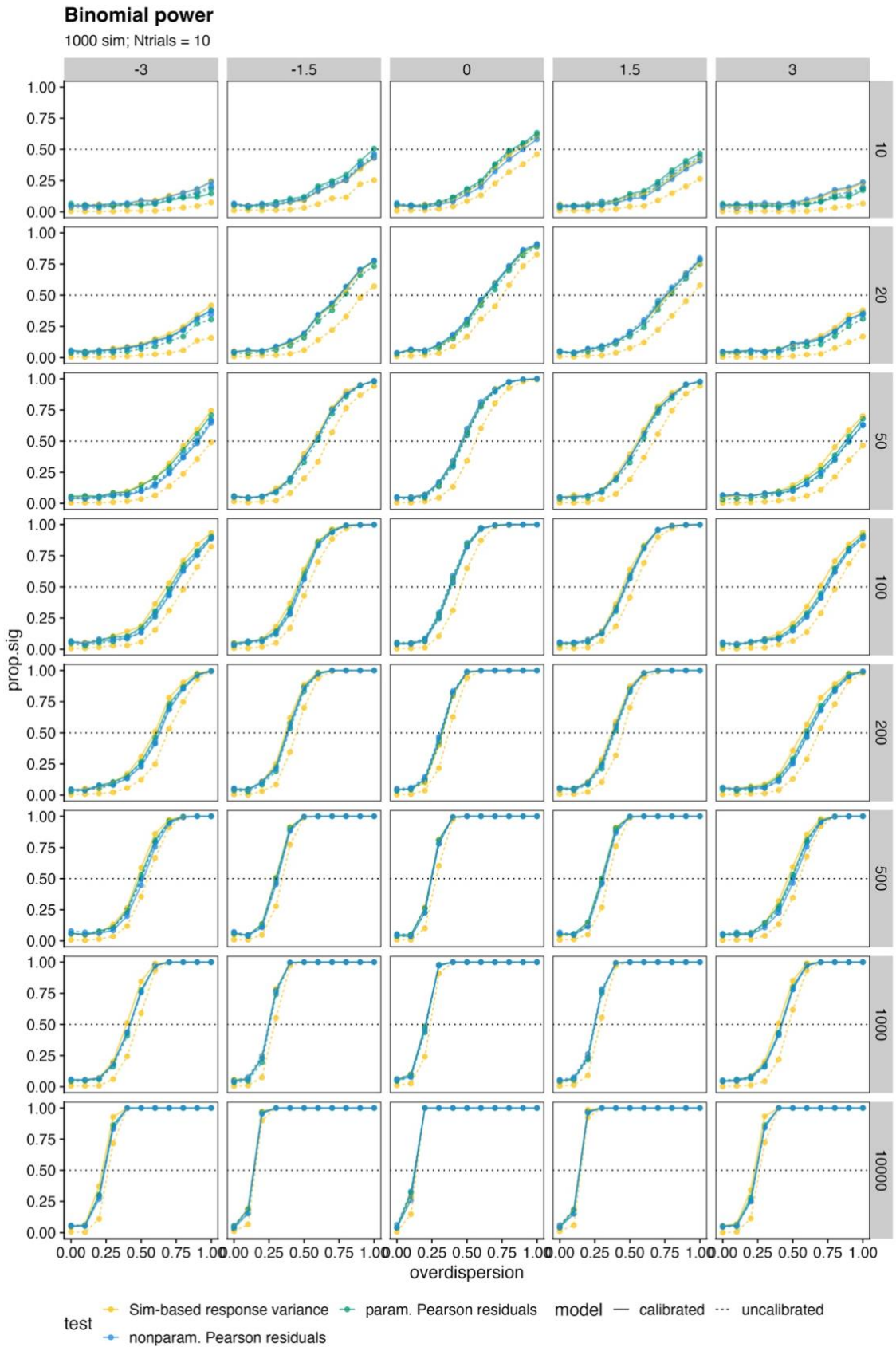


Figure S5.2. Power for GLM binomial.

S6. Additional GLMM results

Type I error rate of the alternative dispersion tests

In Figures S6.1 and S6.2, we present the type I error rates for the four alternative dispersion tests for the Poisson and binomial GLMMs, respectively, using simulated sets of parameters: number of observations, number of groups, and intercepts.

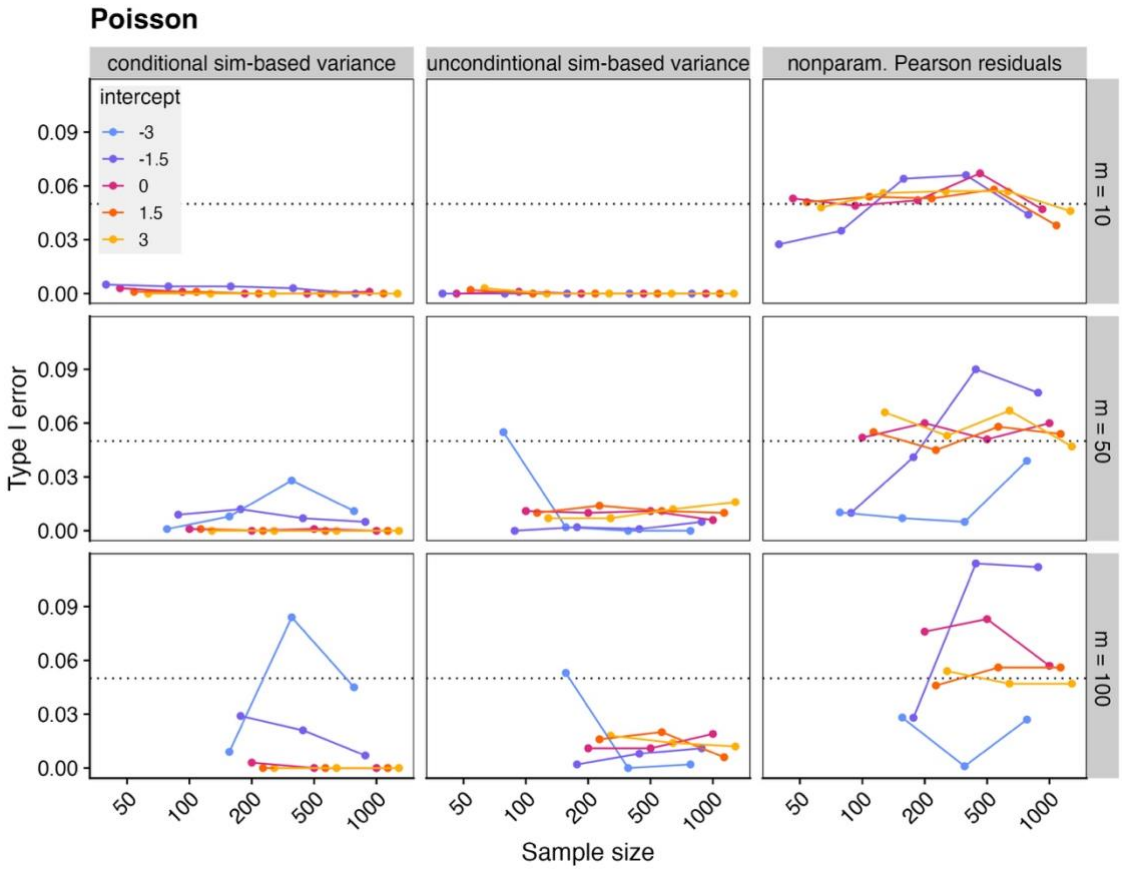


Figure S6.1. Type I error rate for the three alternative dispersion tests for the Poisson GLMMs. 1000 simulations for each parameter set. To improve visualising the different intercept lines, the values in the x-axis were slightly displaced around the sample size values.

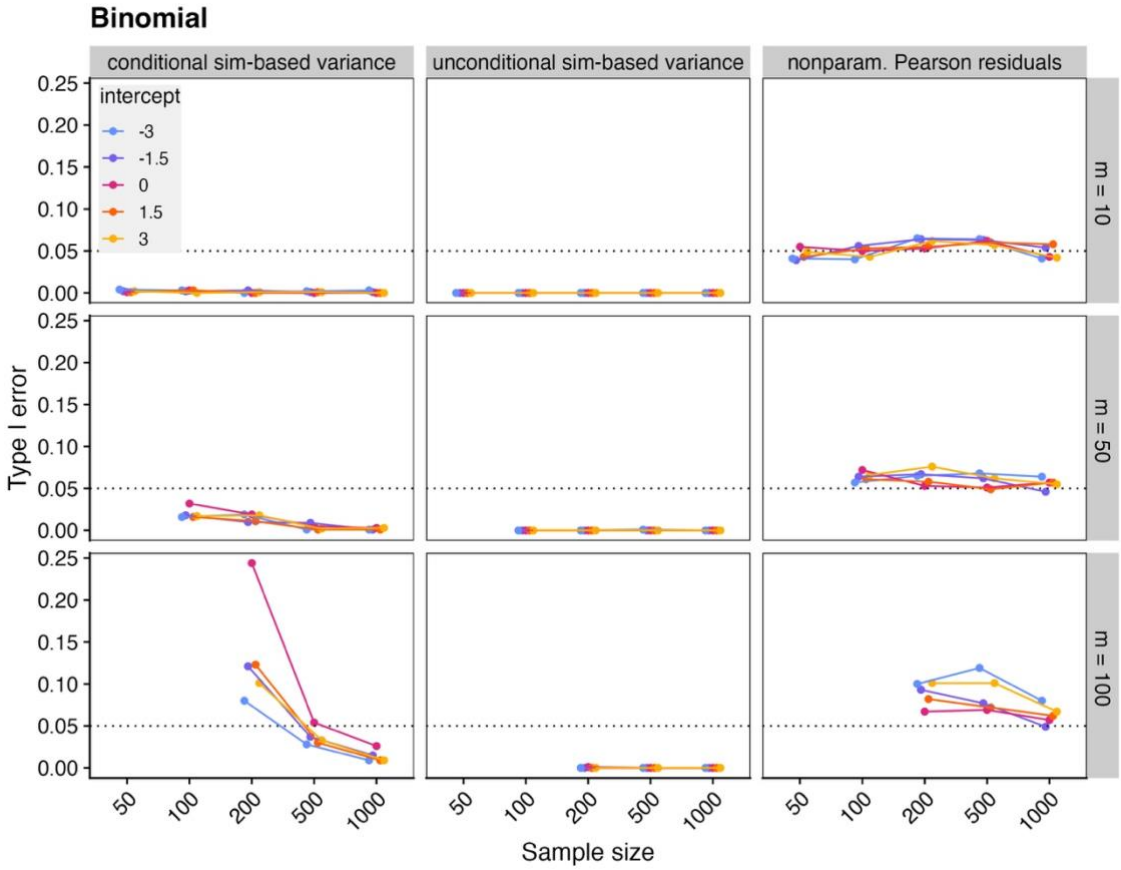


Figure S6.2. Type I error rate for the three alternative dispersion tests for binomial GLMMs. 1000 simulations for each parameter set. To improve visualising the different intercept lines, the values in the x-axis were slightly displaced around the sample size values.

Power of the alternative dispersion tests

In Figures S6.3 and S6.4, we show the Power for the three alternative dispersion tests for the Poisson and binomial GLMMs, respectively, for the simulated sets of parameters: number of observations, number of groups, and intercepts.

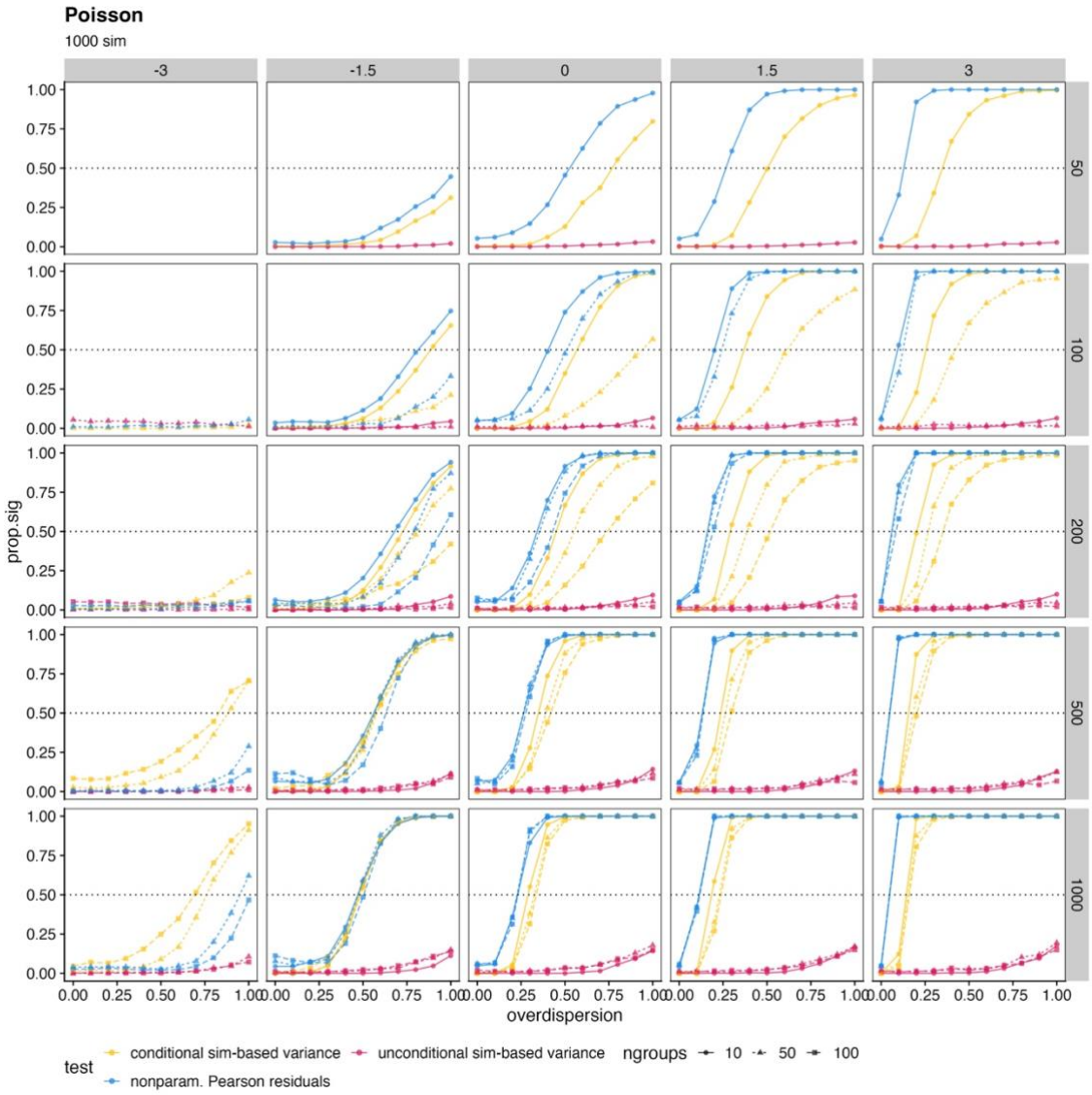
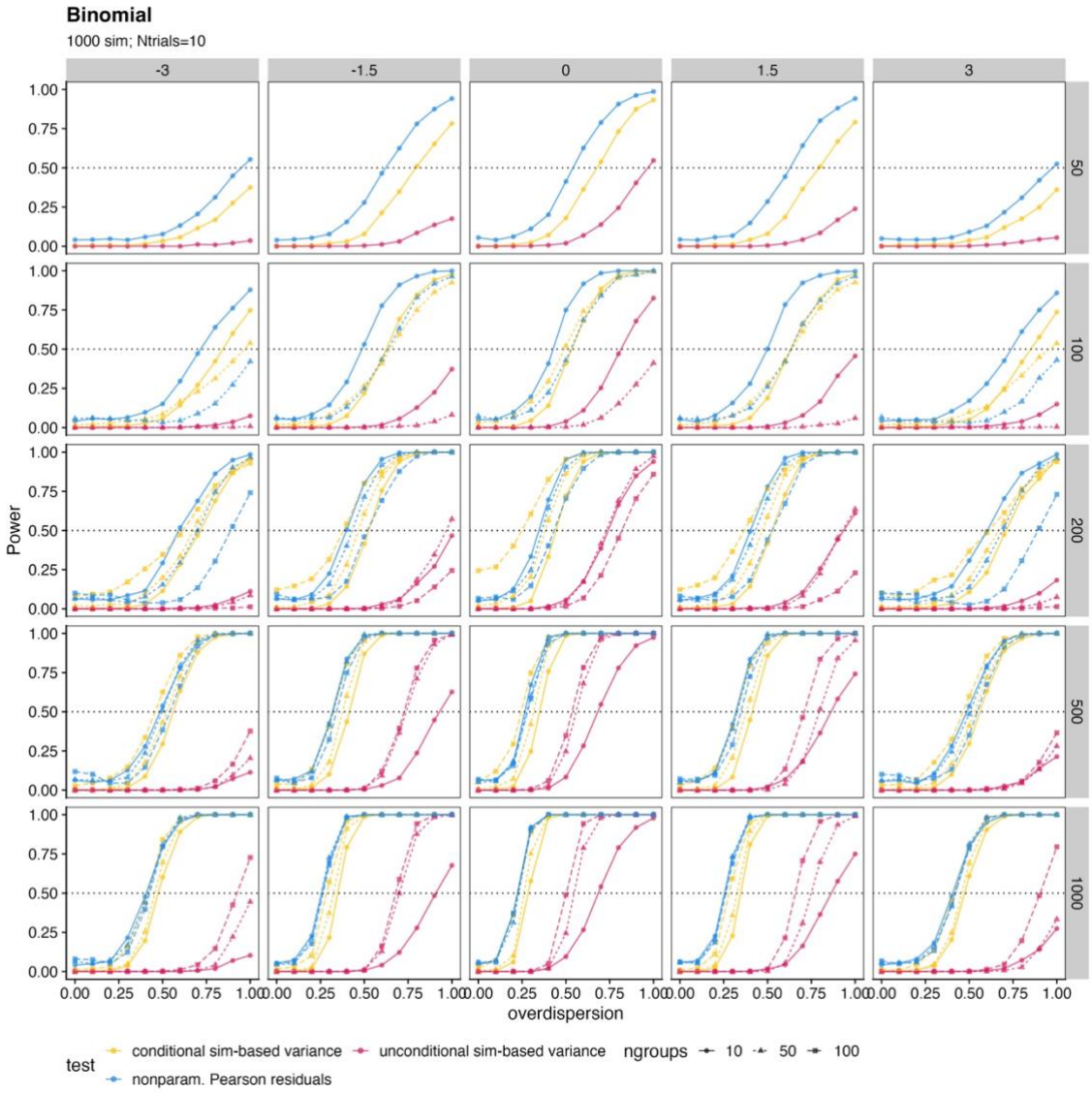


Figure S6.3. Power of the three alternative dispersion tests for the Poisson GLMMs, with different sample sizes (rows), intercepts (columns), and number of groups for the random intercept (line types). The missing lines for the first panel (intercept = -3 and sample size = 50 are due to simulation errors for some tests. For each parameter set, we ran 1000 simulations.

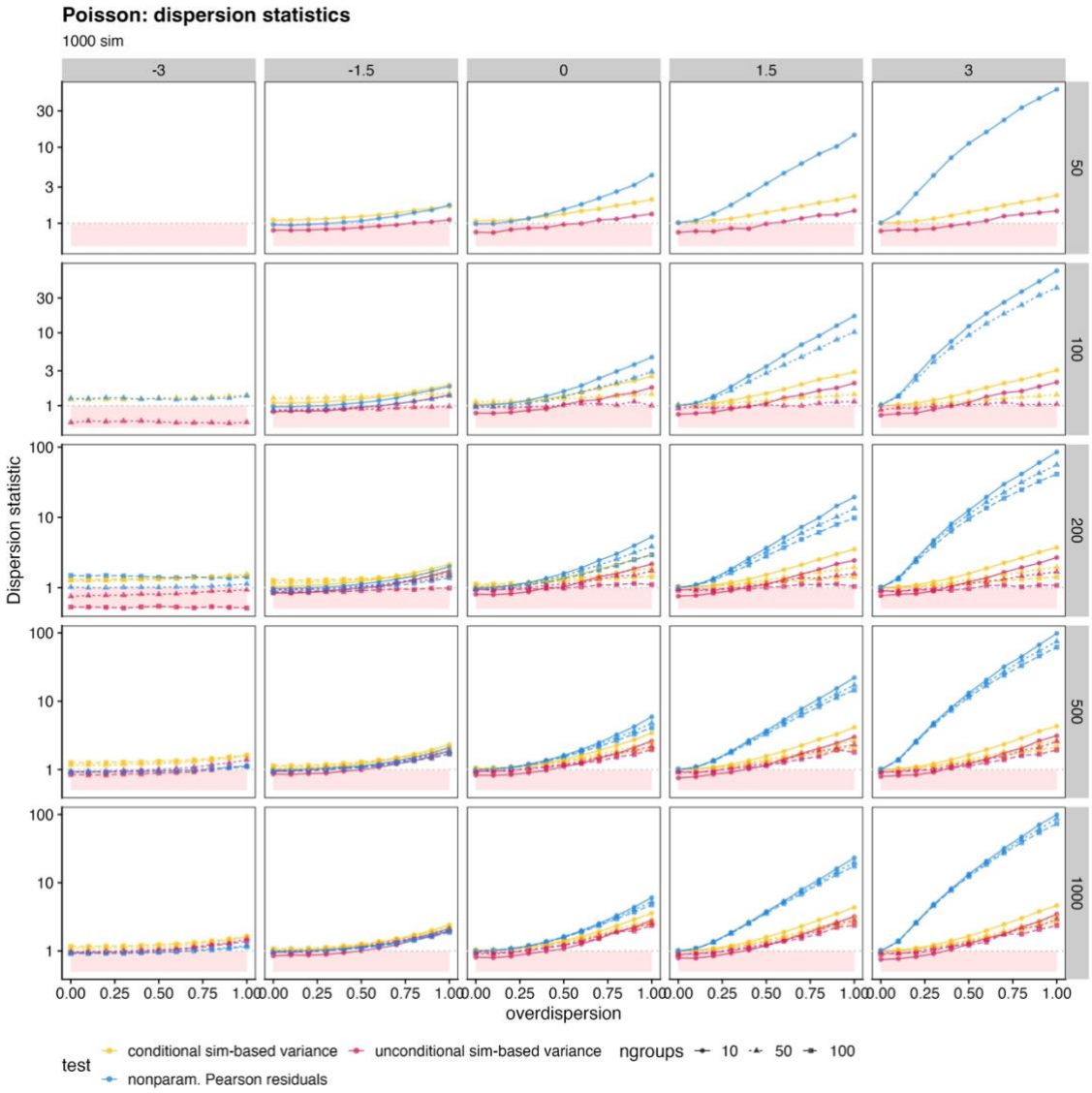


176

177 **Figure S6.4.** Power of the three alternative dispersion tests for binomial GLMMs, with
178 different numbers of observations (rows), intercepts (columns), and number of groups
179 for the random intercept (line types). 1000 simulations for each parameter set.

180 *Dispersion statistics of the alternative dispersion tests*

181 In Figures S6.5 and S6.6, we show the dispersion statistics for the three
182 alternative dispersion tests for the Poisson and binomial GLMMs, respectively, for the
183 simulated sets of parameters: number of observations, number of groups, and intercepts.



184

185 **Figure S6.5.** Dispersion statistics of the three alternative dispersion tests for the Poisson
186 GLMMs, with different numbers of observations (rows), intercepts (columns) and
187 number of groups for the random intercept (line types). The missing lines for the first
188 panel (intercept = -3 and sample size = 50 are due to simulation errors for some tests.
189 For each parameter set, we ran 1000 simulations.

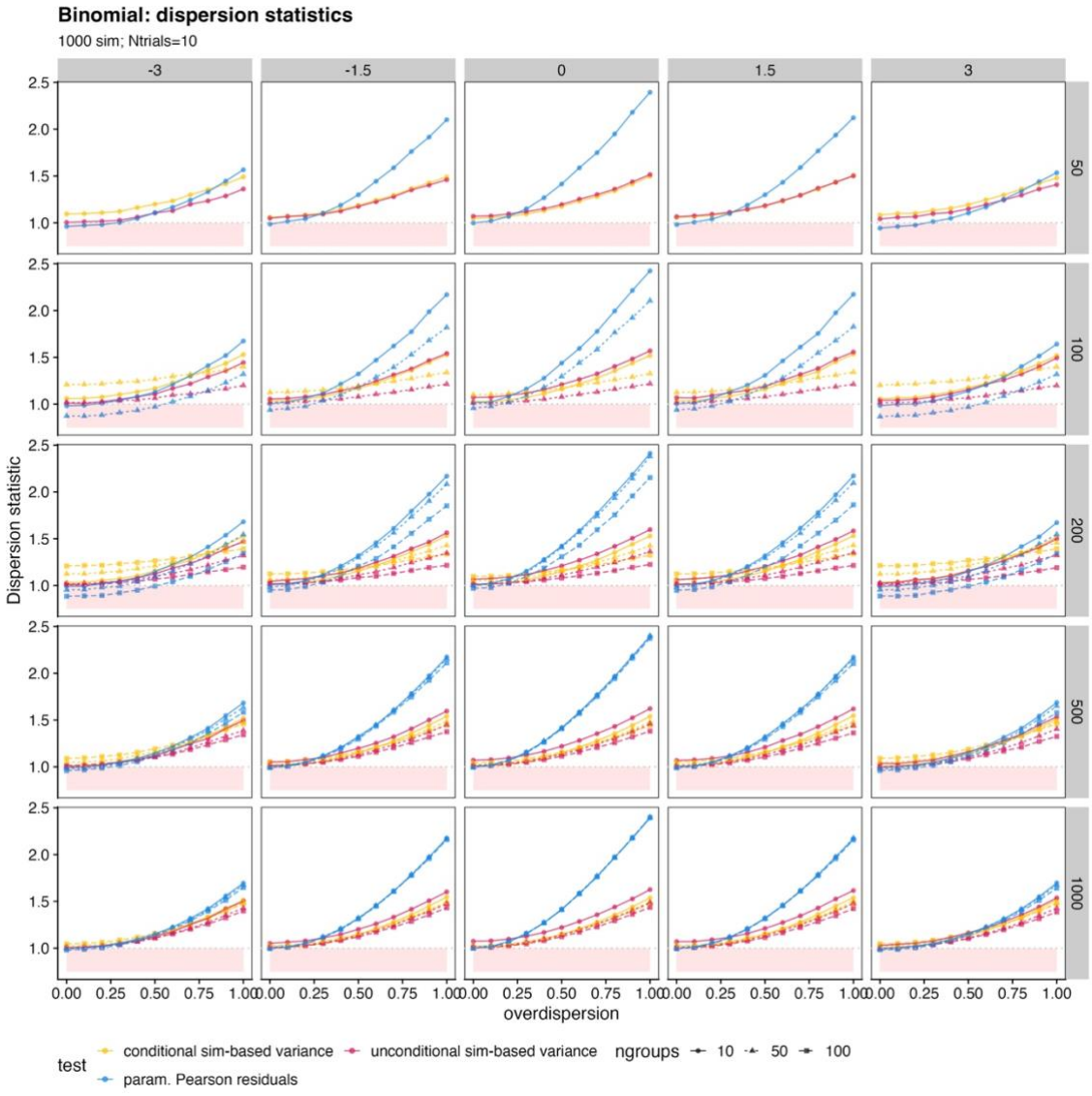
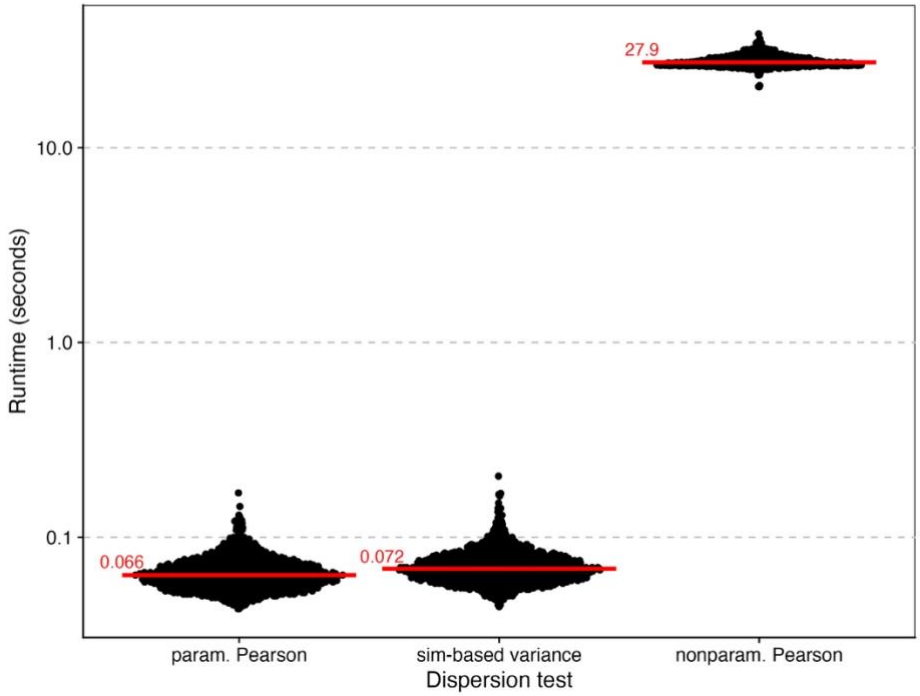


Figure S6.6. Dispersion statistics of the three alternative dispersion tests for binomial GLMMs, with different numbers of observations (rows), intercepts (columns), and number of groups for the random intercept (line types). 1000 simulations for each parameter set.

Computational runtime for tests with GLMMs

We computed the run time for the three tests used for GLMMs: the parametric Pearson test, the nonparametric Pearson test, and the simulation-based response variance test with conditional simulations (Figure S6.7). We used 1,000 simulations of the Poisson GLMM as an example, with an overdispersion parameter of 0.4, an intercept of 0, a sample size of 1,000, and 100 groups. There was almost no difference

in computational time between the parametric Pearson test (median at 0.066 seconds) and the simulation-based response variance test (median at 0.072 seconds). As expected, the nonparametric Pearson residuals presented the largest runtime, with a median of 27.9 seconds.



Figures S6.7. Runtime (in seconds) for each dispersion test for a Poisson GLMM simulated 1000 times with the following parameters: overdispersion parameter of 0.4, an intercept of 0, a sample size of 1,000, and a number of groups of 100. Note the y-axis at the log 10 scale.

S7: Alternative simulation-based residuals dispersion test

Another possibility for improving dispersion tests for GLMMs is to develop a simulation-based approach that shows better type I, power, and a dispersion statistic that could be interpreted similarly to the Pearson dispersion. To explore future possibilities, we briefly considered an alternative simulation-based test that attempts to approximate the Pearson residuals by dividing the observed raw residuals (observed – fitted values) by the variance of the simulated values for each observation (Equations S7.1 and S7.2). We evaluated and compared this test for Poisson and binomial GLMs and GLMMs (conditional simulations only), as we did for the other tests.

$$Approx.\,Pearson\ observed\ residuals:\ \ r_i = \frac{(y_i - \hat{\mu})}{var(y_{is})} \qquad \text{(Equation S7.1)}$$

$$Approx.\,Pearson\ simulated\ residuals:\ \ r_{is} = \frac{(y_{is} - \hat{\mu})}{var(y_{is})} \qquad \text{(Equation S7.2)}$$

One obstacle with calculating the denominator of the approximate Pearson residuals for each observation is that the variance depends on the number of simulations and the model parameters, such as the intercept or the number of trials in the binomial GLM/GLMMs. If there are too few simulations or the intercept is very small, the chance of resulting in zero variance (all simulated values are the same) is higher for data points with small variance. To overcome this, we first evaluated the minimum number of simulations for different intercepts and sample sizes, in which all observations have estimated variances that are different from zero. For all combinations of parameters, we found that 1,000 simulations were sufficient to ensure that all variances in the simulated observations were positive (Figures S7.1 and S7.2). However, 250 simulations (the default parameter of the DHARMa package) also presented reasonable results, with the only exception being the Poisson GLMs with 30 out of 1,000 simulations (sample size

of 100 and intercept of -1.5) with a very low percentage of zero variances in the simulated observations (mean of 0.01, maximum of 0.06). We are aware that the number of zero variances in the simulations depends heavily on the simulation set, e.g., the number of trials for the binomial GLM. To develop an effective dispersion test, one should consider alternatives to address this issue. For the subsequent analyses, we excluded the simulations with zero variance in any simulated observation to compare the alternative dispersion test with the simulation-based residuals test and the Pearson Chi-squared dispersion test.

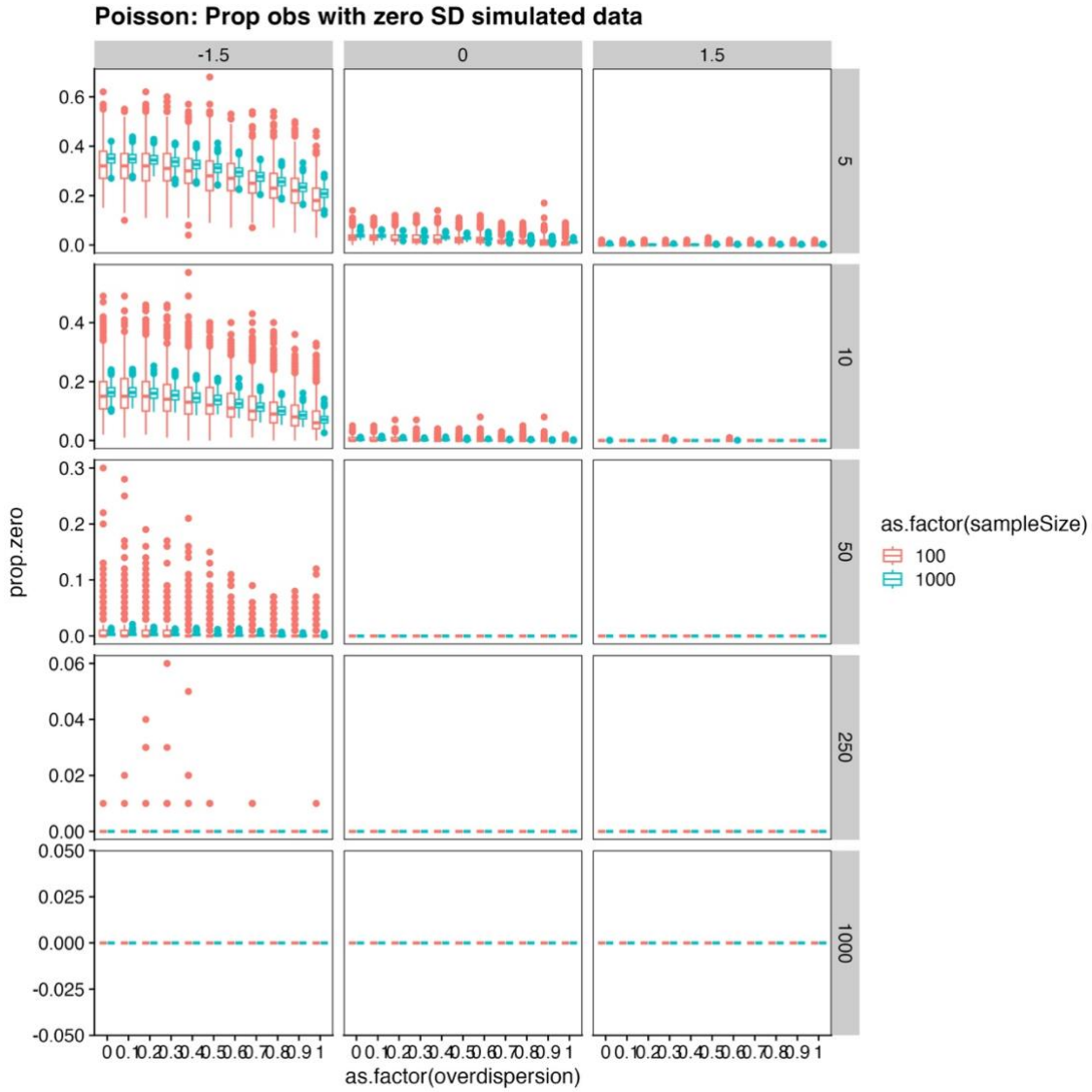


Figure S7.1. Poisson GLM: Proportion of observations with simulated zero variance in the dataset for different combinations of intercept (columns), number of simulations (rows) and sample sizes (colours).

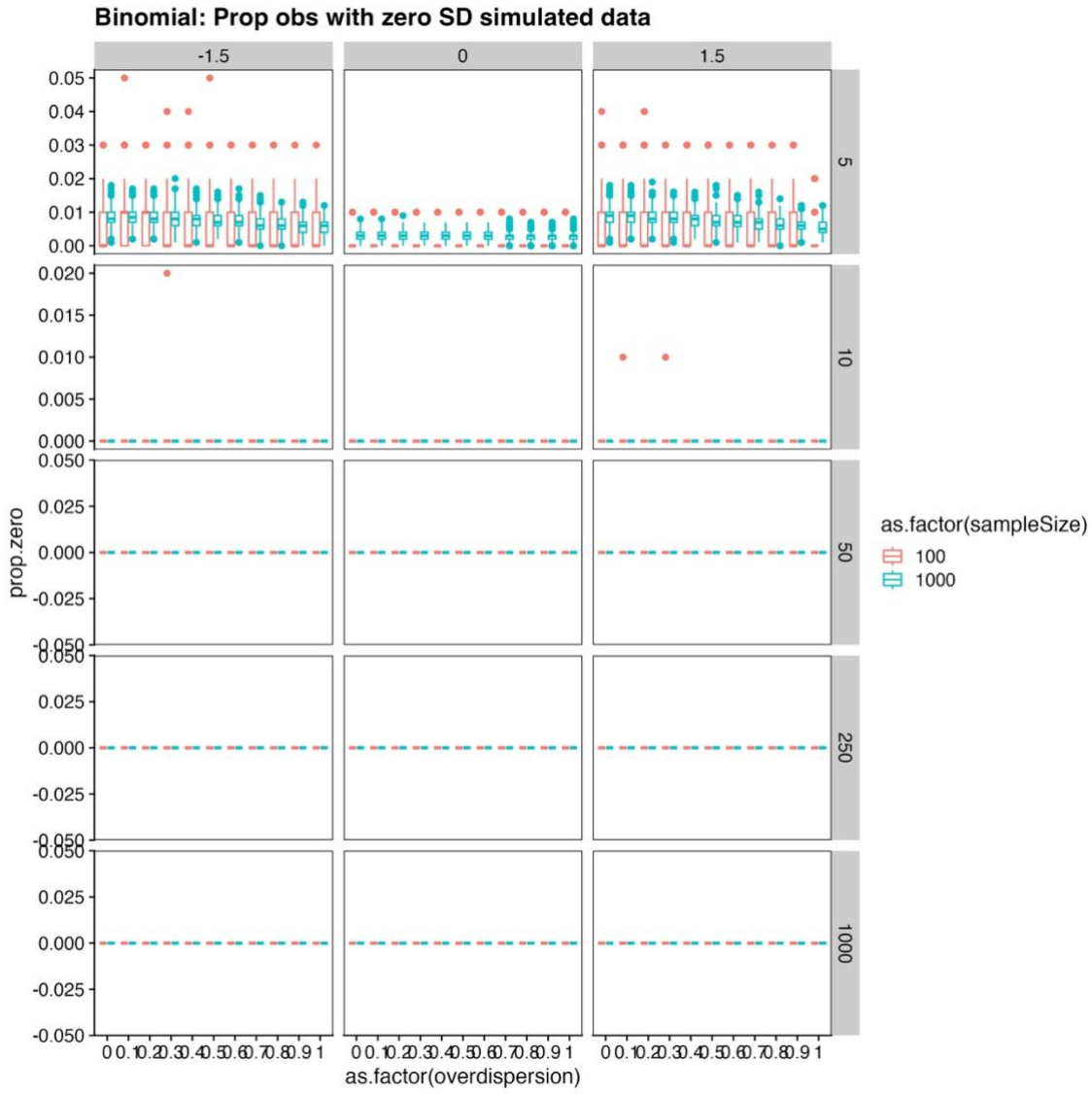


Figure S7.2. Binomial GLM: Proportion of observations with simulated zero variance in the data set for different combinations of intercept (columns), number of simulations (rows) and sample sizes (colours). The number of trials of the binomial was set to 10 in all simulations.

First, we compared the approximate Pearson residuals for GLMs with the Pearson residuals by regressing the difference between them as the response variable and the Pearson residuals as the predictor for the Poisson GLMs (Figure S7.3). The intercepts for all simulation sets were nearly zero. The slope of the regression was positive and very small for the larger number of simulations and intercepts. It means

that the approximate Pearson tends to be slightly larger than the Pearson for larger residuals. We did not ca

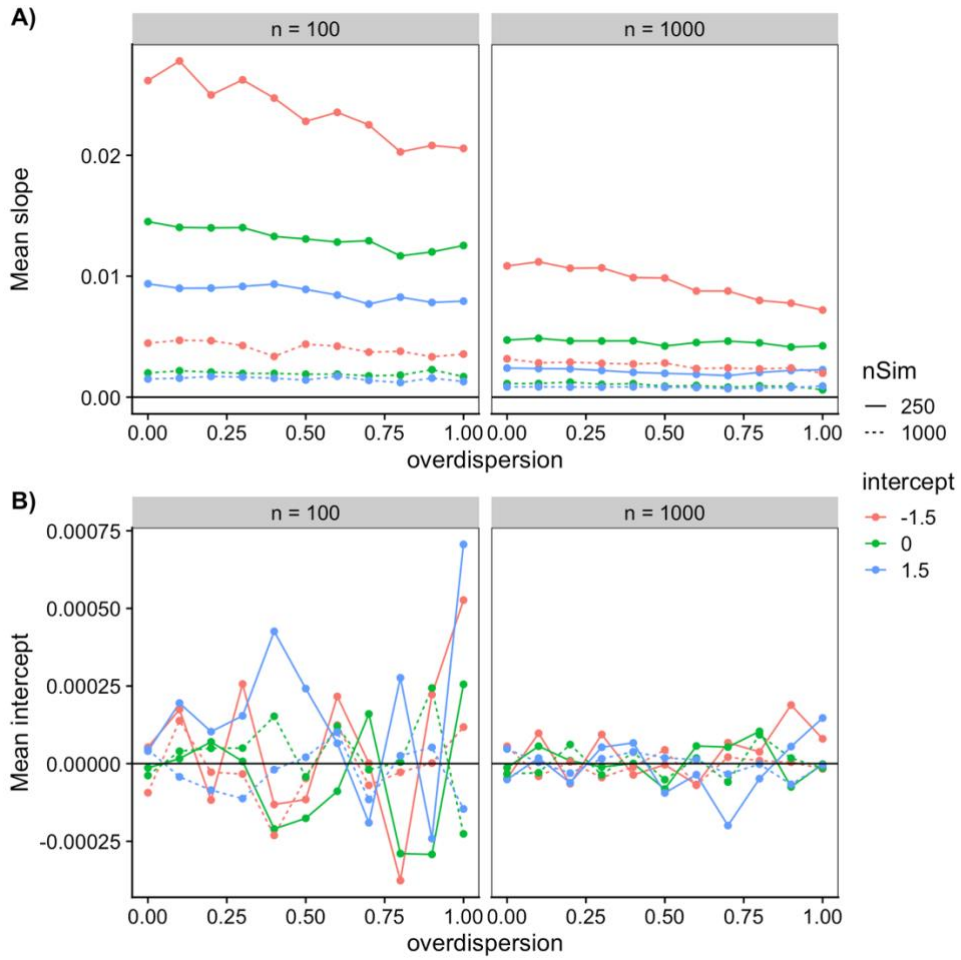


Figure S7.3. Mean slope (A) and intercept (B) of the regression of the difference between the Approximate Pearson residuals and Pearson residuals as response variable and the Pearson residuals as predictor for the Poisson GLMs.

Type I error rates for the alternative simulation-based test, based on the approximate Pearson residuals for GLMs, were similar to those for the simulation-based response variance test for the Poisson model. They tended to be conservative for small intercepts (Figure S7.4). However, for the binomial model, type I error rates were more similar to the parametric Pearson residuals test, with values closer to 0.05 (Figure S7.4).

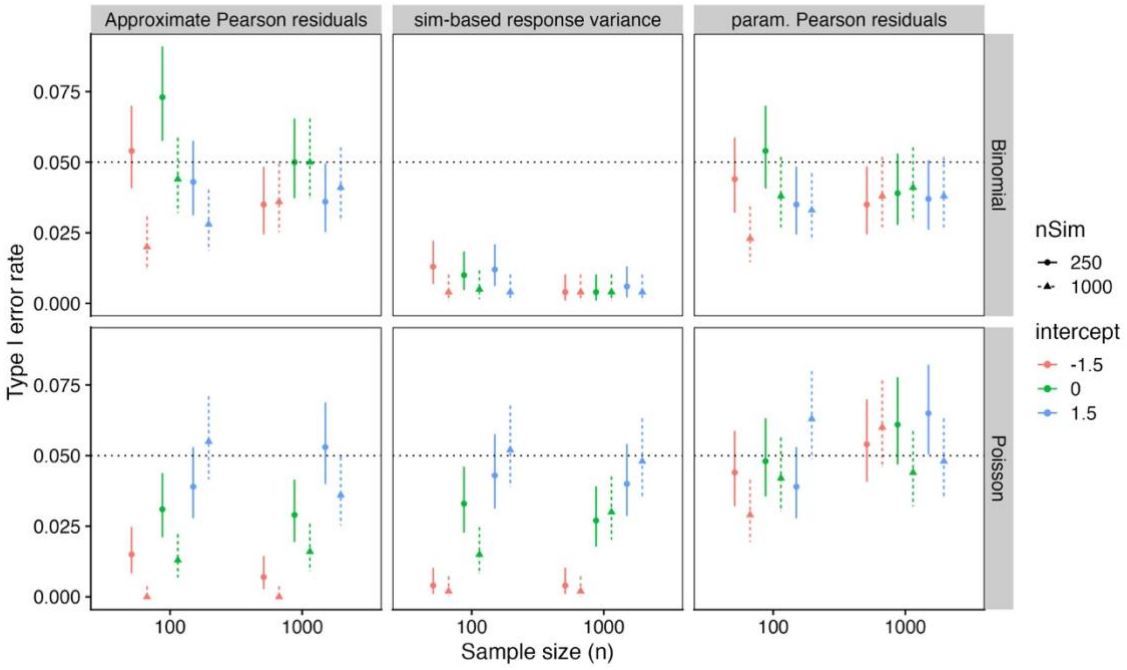
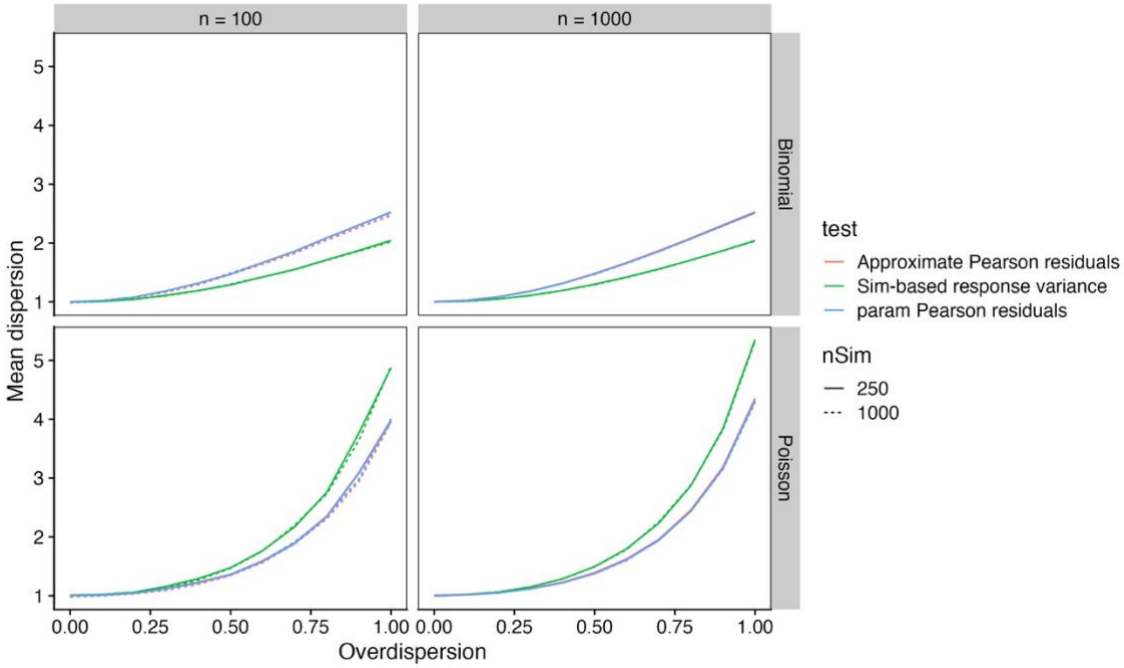


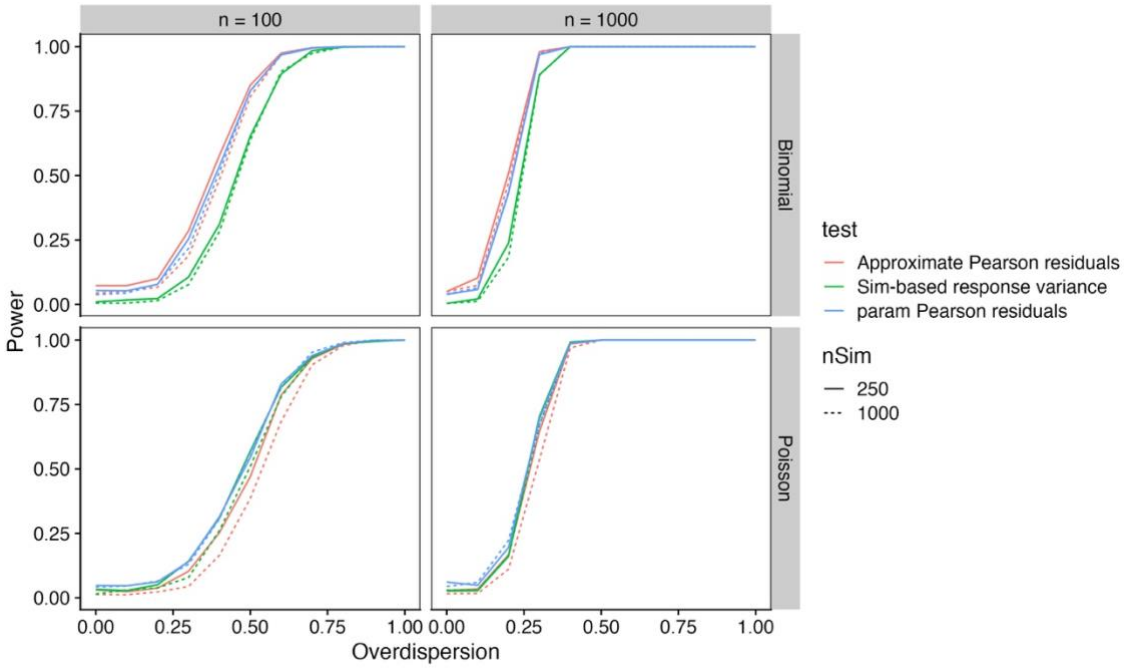
Figure S7.4. Type I error rates for GLMs comparing the parametric Pearson residuals tests, the simulation-based response variance test and the simulation-based approximate Pearson test.

The dispersion statistics for the alternative simulation-based response variance test didn't change depending on the number of simulations and were very similar to the parametric Pearson dispersion statistics for both GLMs (Figure S7.5). Power was very similar among the tests for the Poisson GLM (Figure S7.6). For binomial GLMs, the power of the alternative simulation-based residual test was high and similar to the parametric Pearson residuals test.



278

279 **Figure S7.5**. Dispersion statistics GLMs. Simulation set with intercept = 0.



280

281 **Figure S7.6**. Power GLMs. Simulation set with intercept = 0.

282 For the GLMM simulations, we fixed the number of groups at 100 and the
283 number of simulations at 250 to compare with the cases where the Pearson Chi-squared
284 test fails. We compared sample sizes of 200, 500, and 1000 observations and intercepts

of -1.5, 0, and 1.5. We excluded simulations with zero variance in the simulated observations (specifically, for Poisson GLMMs, which accounted for less than 0.1% of the simulations). For GLMMs, we used only the conditional simulations, which have been proven to yield better dispersion test results.

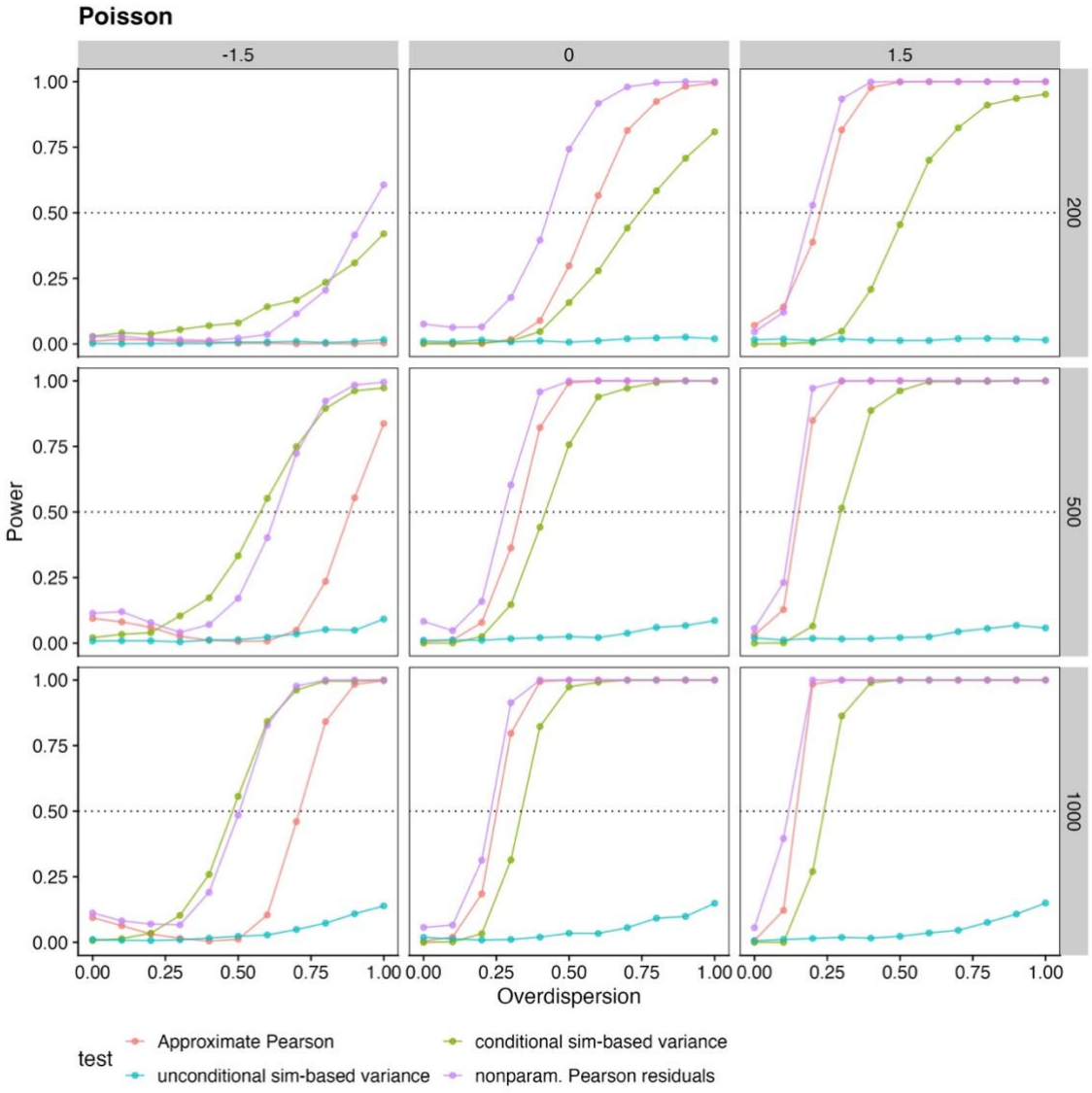
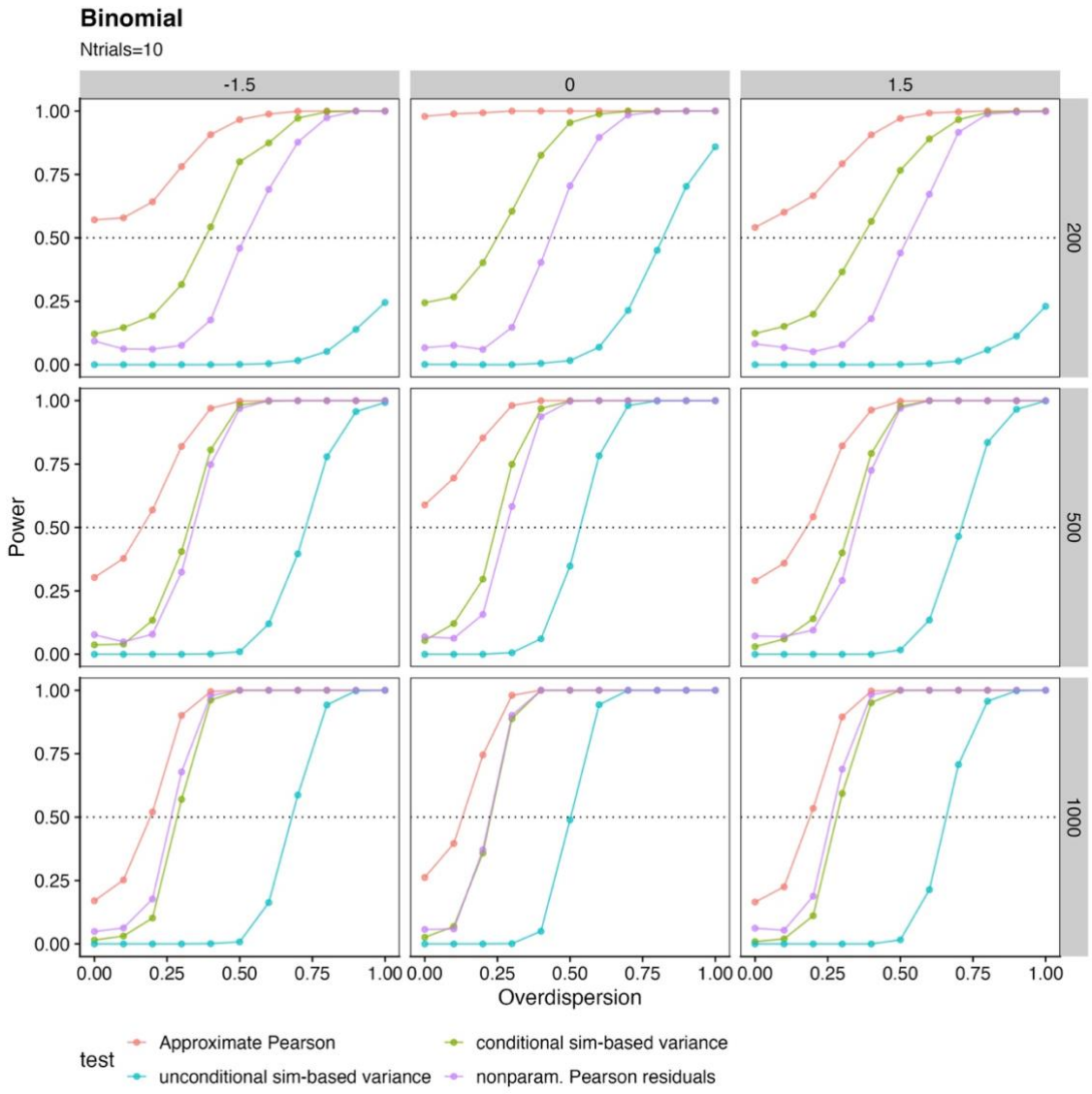


Fig S7.7. Power for Poisson GLMMs for the alternative simulation-based test using an approximation for Pearson residuals compared with the other tests assessed in the study. 1000 simulations for each parameter set: intercept (panel columns) and sample size (panel rows). The fixed parameters are slope = 1, number of groups = 100, and random effects variance = 1.



295

296 **Fig S7.8.** Power for binomial GLMMs for the alternative simulation-based test using an
297 approximation for Pearson residuals compared with the other tests assessed in the study.
298 1000 simulations for each parameter set: intercept (panel columns) and sample size
299 (panel rows). The fixed parameters are slope = 1, number of groups = 100, random
300 effects variance = 1, number of trials = 10.

301

302

S8. Parametric Pearson test with approximated residual degrees of freedom for GLMMs

Degrees of freedom (*df*) are not always known for GLMMs with complex hierarchical structures and limit the use of the parametric Pearson test because it depends on it for evaluating overdispersion with the Chi-squared distribution. Moreover, our results show that using the naïve *df* is problematic for testing dispersion when you have a large number of groups in the random intercept. The two most suggested methods to approximate *df* of mixed-effect models, the Satterthwaite (1946) and the Kenward-Roger (Kenward & Roger 2009), were developed for LMMs to account for the effect of the covariance structure on *df* and standard errors. Stroup et al. (2013) suggested that the adjustment is also accurate for GLMMs. However, none of the most used R packages use any correction for the degrees of freedom for GLMMs. The few R packages that provide those approximations, e.g. *lmerTest* (Kuznetsova et al., 2017; Kuznetsova et al., 2020) that relies on *pbkrtest* (Halekoh & Højsgaard 2014), are only implemented for LMMs.

Recently, we found that the R package *glmmrBase* (Watson 2024) provides those approximation methods for GLMMs. Thus, we compared the parametric Pearson test with the three corrections for degrees of freedom available in the package for the Poisson GLMMs. The corrections are:

- The Kenward-Roger (KR) bias-corrected variance-covariance matrix for the fixed effect parameters and degrees of freedom from Kenward & Roger (1997).
- The improved correction of the Kenward-Roger (KR2) returns an improved correction given in Kenward & Roger (2009).
- The Satterthwaite correction (Sat) from Satterthwaite (1946).

Our test results show that all three correction methods presented very similar residual *df* for all simulation settings (Figure S8.1), which resulted also in very similar test results (e.g., Figure S8.2 for type I error). Given the high similarity among tests for the different residual *df* corrections, we show and discuss the results for the KR2 test in comparison with the parametric Pearson "naïve" test and the alternative GLMM tests (nonparametric Pearson and simulation-based response variance test with conditional simulations). In Figure S8.3, we observe that the correction for the residual *df* corrected the dispersion statistics towards 1 for simulations without overdispersion, except for the very small intercept (-1.5). This results in the two-sided dispersion test being less prone to being significant, given the very low dispersion parameter (detecting underdispersion instead of overdispersion).

Although the parametric Pearson tests with the approximated residual degrees of freedom performed much better than those with the "naïve" residual *df*, they still underperformed compared to the nonparametric version when having a large number of groups in the random effects (Figure S8.4), especially for very small intercepts and sample sizes.

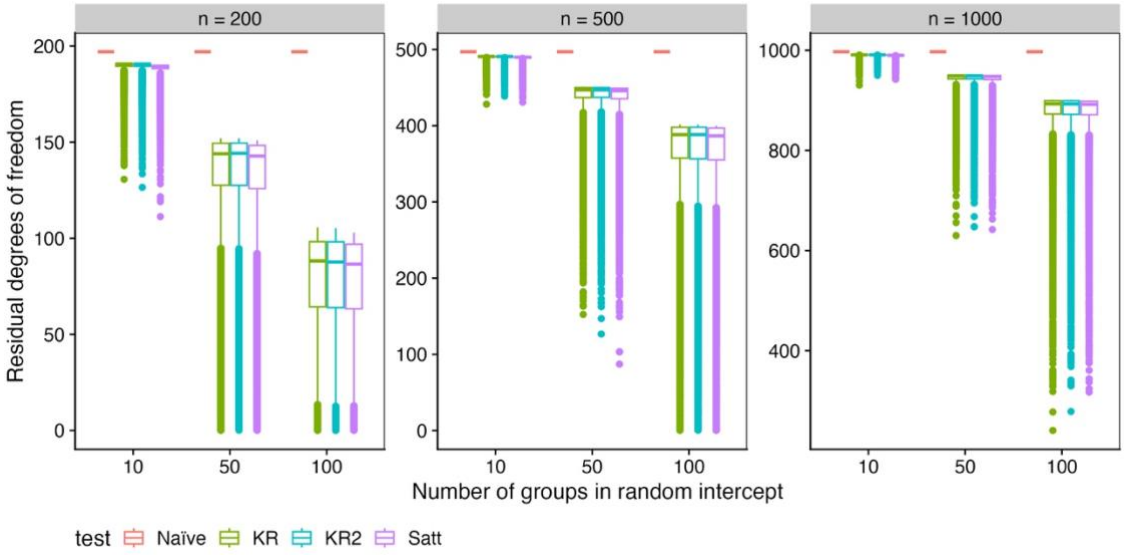
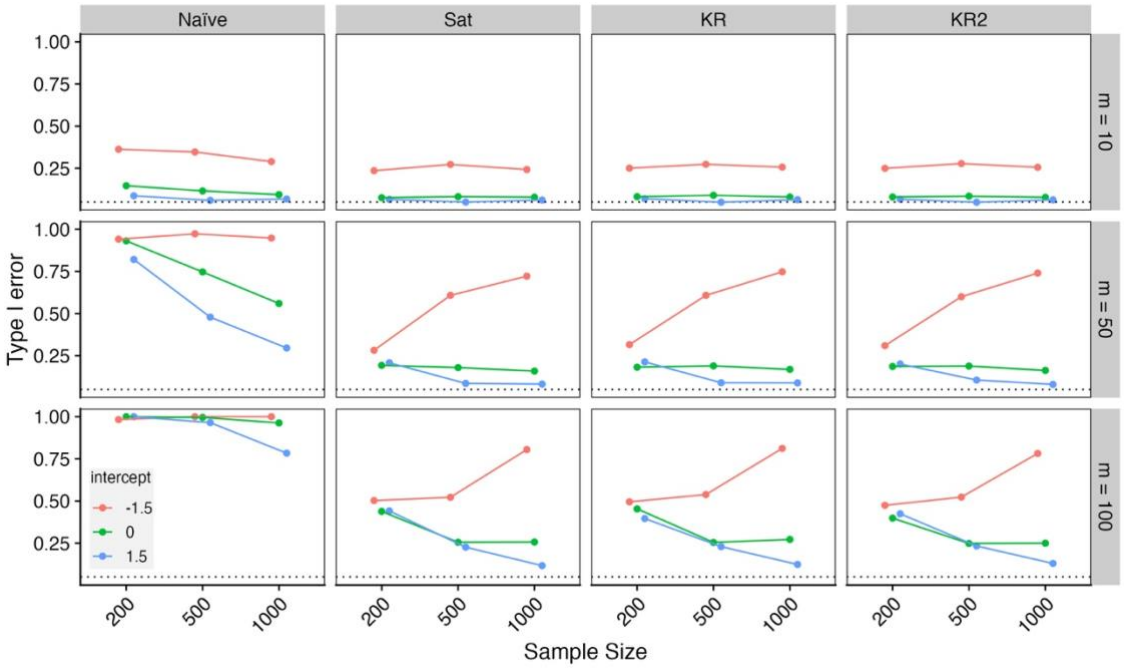


Figure S8.1. Residual degrees of freedom for the different correction methods for Poisson GLMMs with different numbers of groups in the random intercept (x-axis) and sample sizes (panel columns). Please refer to the main text above to relate to each applied correction. 1,000 simulations for each parameter setting, slope = 1, random intercept variance = 1.



Figures S8.2. Type I error for the parametric Pearson test for Poisson GLMMs performed with different corrections for the residual degrees of freedom (panel columns), number of groups in the random intercept (panel rows) and sample size (x-axis). Data were simulated from a Poisson GLMM with different intercepts (colours). Please refer to the main text above to relate to each applied correction. 1000 simulations for each parameter setting, slope = 1, random intercept variance = 1.

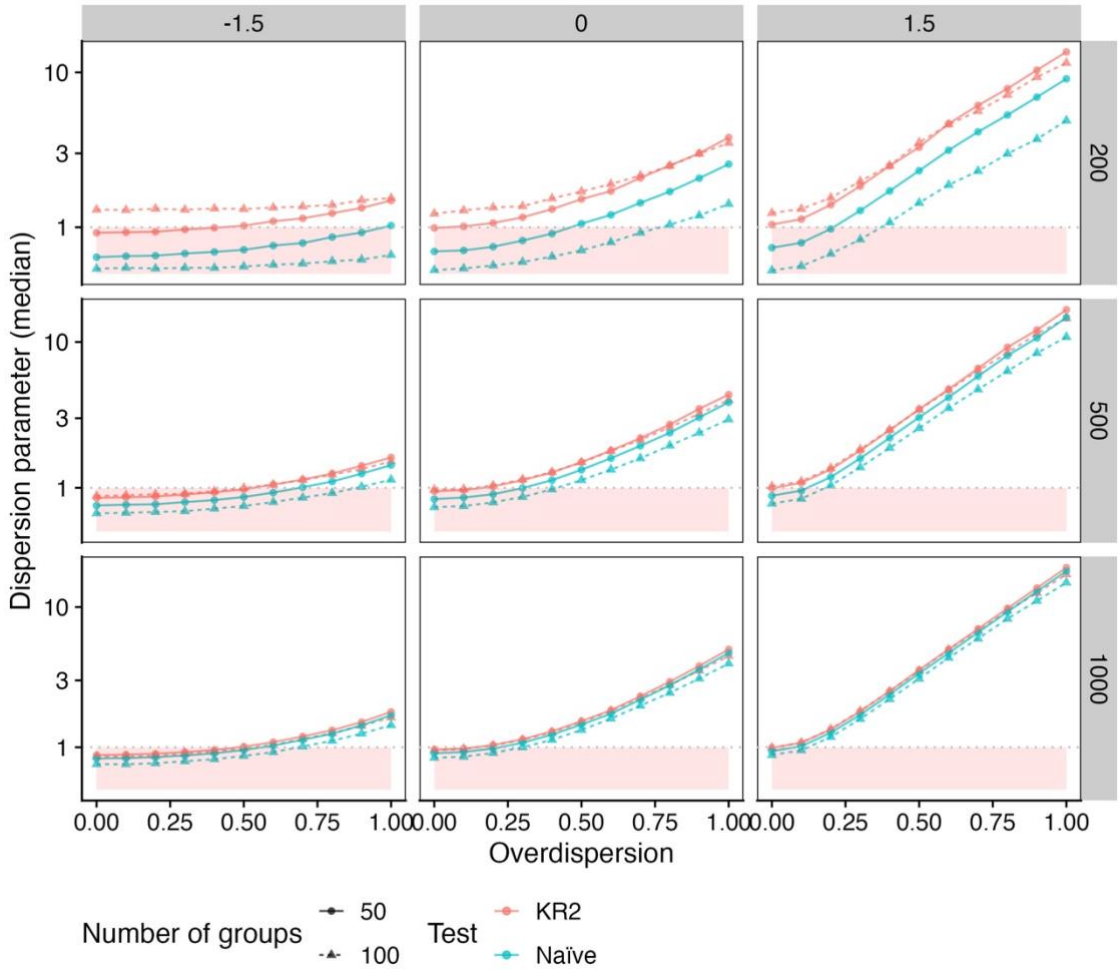


Figure S8.3. Dispersion parameters for the parametric Pearson test for Poisson GLMMs performed with different corrections for the residual degrees of freedom (colours), number of groups in the random intercept (linetype and shape), sample size (panel rows), and intercept (panel columns). Please refer to the main text above to relate to each applied correction. To improve clarity, we omitted the other corrections because they are too similar to each other. 1000 simulations for each parameter setting, slope = 1, random intercept variance = 1.

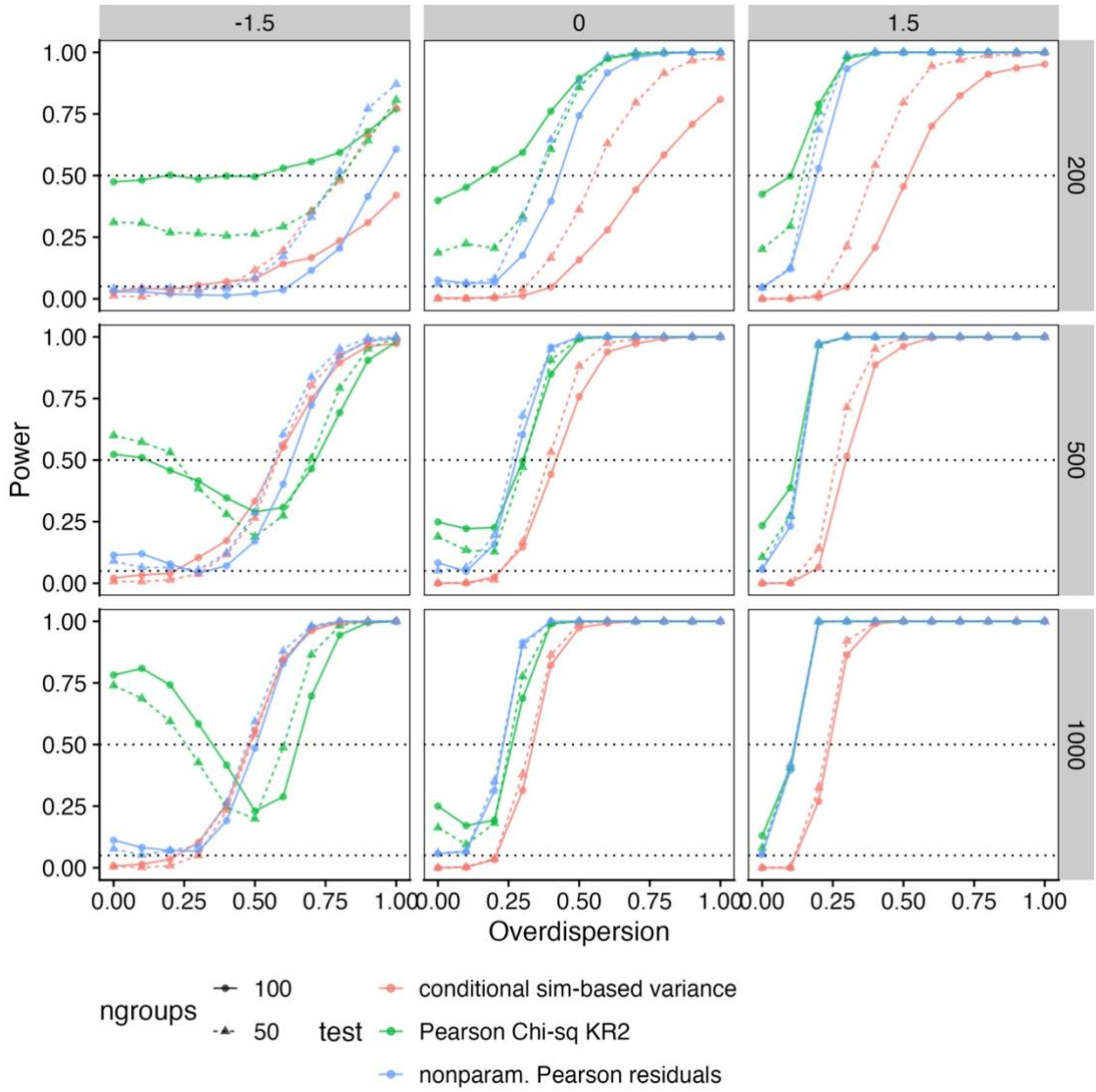


Figure S8.4. Power of dispersion tests for Poisson GLMMs (colours) performed with different numbers of groups in the random intercept (linetype and shape), sample size (panel rows), and intercept (panel columns). Please refer to the main text above to relate to the applied correction for residual degrees of freedom. To improve clarity, we omitted other corrections for residual degrees of freedom because they are too similar to each other. 1000 simulations for each parameter setting, slope = 1, random intercept variance = 1.