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

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

Headline: Dispersion tests for GLMMs

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, usually 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, reduced power. A range of tests has been proposed to detect such dispersion problems, but there are few comparative studies of their performance across models and analysis settings.
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. Following 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 and observed residual variance.
3. Comparing their performance by type I error, power, and dispersion estimate, across a range of GLMs and GLMMs, we found that the nonparametric Pearson residuals test performed best across all metrics, especially for data with low incidence or count rates and/or small samples; however, at the cost of high computational expense. The parametric Pearson residuals test, recommended in many books and guidelines, was fast and effective for GLMs, but biased towards

underdispersion in GLMMs due to the naïve computation of the random-effect degrees of freedom. The simulation-based residual variance test was slightly less powerful, but showed overall good calibration and was much faster to compute. The latter 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 balance of speed and accuracy. For GLMMs, we recommend either the computationally demanding nonparametric Pearson residuals test or the faster, although somewhat less powerful, simulation-based residual variance test. We also provide additional recommendations for ecological data analysis to address dispersion issues using the most commonly used R packages, avoiding pitfalls and improving model fit and the interpretation of ecological datasets.

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

Code for Peer review:

https://anonymous.4open.science/r/dispersion_test_GLMM/README.md

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 GLMs/GLMMs is overdispersion. Overdispersion refers to greater 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 distributions that assume a fixed mean-variance relationship, such as the Poisson model for count data (Harrison, 2014; 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. However, when individuals are subject to spatial/temporal clustering due to different ecological mechanisms (e.g., patchy resource distribution, social behaviour, dispersal limitation) and/or imperfect detection (Rhodes, 2015), we typically find higher dispersion than expected from a Poisson distribution (Box 1). Alternatively, overdispersion may also arise from modelling misfit, for example, by failing to include important predictors and interactions or by specifying the incorrect link function (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-conservatism means that p-values and confidence intervals are too small, leading to 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 switching to a GLM with more appropriate dispersion (see 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 (e.g., a uniform distribution of individuals in space, Box 1) 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 ecological 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 reduces overall power and thus increases type II error. Therefore, accurate statistical inference demands that we identify and adequately address both underdispersion and overdispersion to minimise the risk of wrong inference.

Given the central importance of dispersion to all statistical indicators, statisticians have pondered how to detect and address dispersion problems since the early days of modern statistics in the 19^{th} century (see Quine & Seneta (1987) and

Xekalaki (2014) for a historical perspective). 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 residuals (Hartig, 2024).

Somewhat confusing for the ecological 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 in the R environment (R Core Team, 2024) revealed that existing dispersion tests are scattered across different packages (Table 1), and most of them work only for a restricted set of models. In the ecological literature, although awareness of dispersion problems has increased over the last 20 years (Box 2), there is still a clear lack of guidance on how dispersion problems are assessed or tested (Box 2). All this makes it challenging to decide which test to use in applied ecological 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, common situations for ecological data analyses. 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 raw residuals. 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 our results, we provide recommendations for the most suitable tests for detecting over- or underdispersion, depending on model complexity and software availability, i.e., currently available R packages and functions.

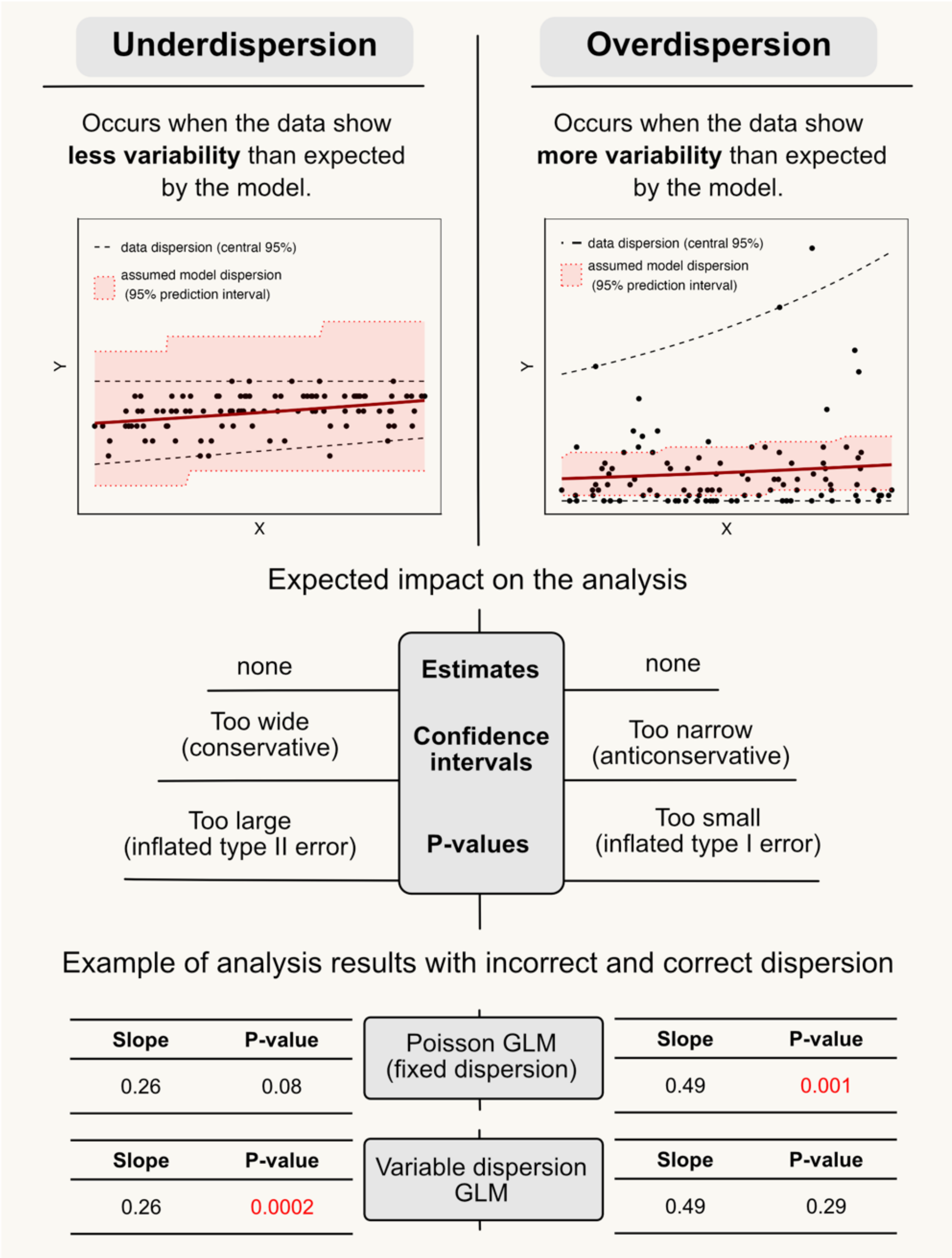


Figure 1. Definition, statistical consequences, and a practical analysis example of under-/overdispersion in GLM/GLMMs. The top row shows examples of a data analysis using a Poisson GLM with simulated under- and overdispersed count data. Data points in black are contrasted to the Poisson model's 95%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

BOX 1: Ecological causes of under- and overdispersion

Under/overdispersion in ecological data may not be only a statistical problem that we need to control for, but may also reflects key ecological processes of interest (Nakagawa et al., 2026; Rhodes, 2015). For example, ecological field data often consists of individual counts in space or time. If individuals are distributed completely at random, the sampling variability will follow a Poisson distribution. However, a range of ecological, observational and modelling processes can lead to deviations from this distribution, resulting in over- or underdispersion. For example, spatial aggregation (clustering) of individuals due to patchy resources distribution, social behaviour, or dispersal limitation increases sampling variability and thus creates overdispersion compared to the Poisson (Fig. B1, see also Lindén & Mäntyniemi, 2011). On the contrary, a uniform spatiotemporal distribution of individuals, for example due to territoriality, may create underdispersion (Lynch et al., 2014). Note that in these examples, but also in general, over- and underdispersion are always defined with respect to an expectation, in this case, the Poisson distribution.

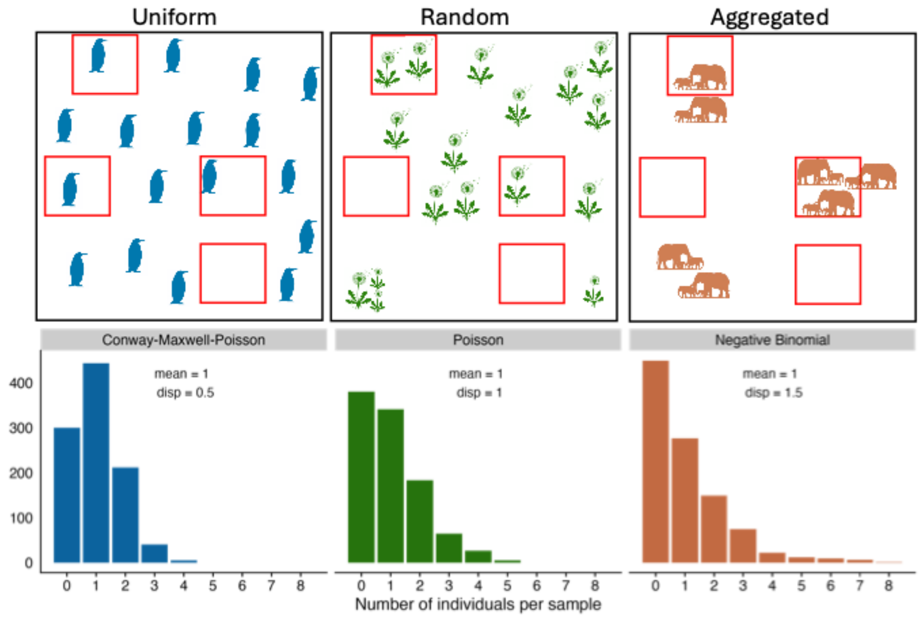


Figure B1. Examples of spatial distribution patterns of individuals under the same sampling design (quadrats in red) and the corresponding data-generating distributions for each pattern: Poisson (random pattern, green), negative binomial (aggregated pattern, light brown), and Conway-Maxwell-Poisson (uniform pattern, blue). Histograms are drawn from 1000 samples from the different distributions with the same mean (1) but varying dispersion (0.5, 1, 1.5). The figure silhouettes correspond to classical text-book examples (e.g., Campbell & Reece, 2005): most penguin species are territorials, tending to be uniformly spaced; dandelions have wind-dispersed seeds and tend to have a random distribution; elephants live in groups, and therefore exhibit an aggregated distribution.

More ecological causes for underdispersion appear, for example, in individual reproductive metrics (e.g, such as clutch size or seeds per fruit) with discrete counts upper limited by behavioural or physiological constrains, such as ovule number, parental care or resources availability (Brooks et al., 2019; Lynch et al., 2014; Puig et al., 2024).

It is worth noting that it doesn't mean, however, that dispersion problems always hint to an interesting ecological process. As discussed in the main text, dispersion problems may also arise from observational errors, for example imperfect detection (Rhodes, 2015), or from model misfit. When detecting dispersion problems, a careful consideration of their reasons is therefore paramount for an adequate ecological interpretation.

BOX 2: Current practices for dispersion issues in ecological studies

To understand the current practice for addressing dispersion problems in GLMs/GLMMs for count and discrete proportions data, we performed a text mining analysis of the ecological literature from the last 20 years (see S1 for details). Our results show that, over recent years, the percentage of all ecological studies using GLM/GLMMs for such data ahs remained around 8% (Fig. S1.1). Within these studies, we observed a steady increase in awareness about dispersion issues, with more than 28% of studies published in 2025 explicitly mentioning them (Fig. B2A).

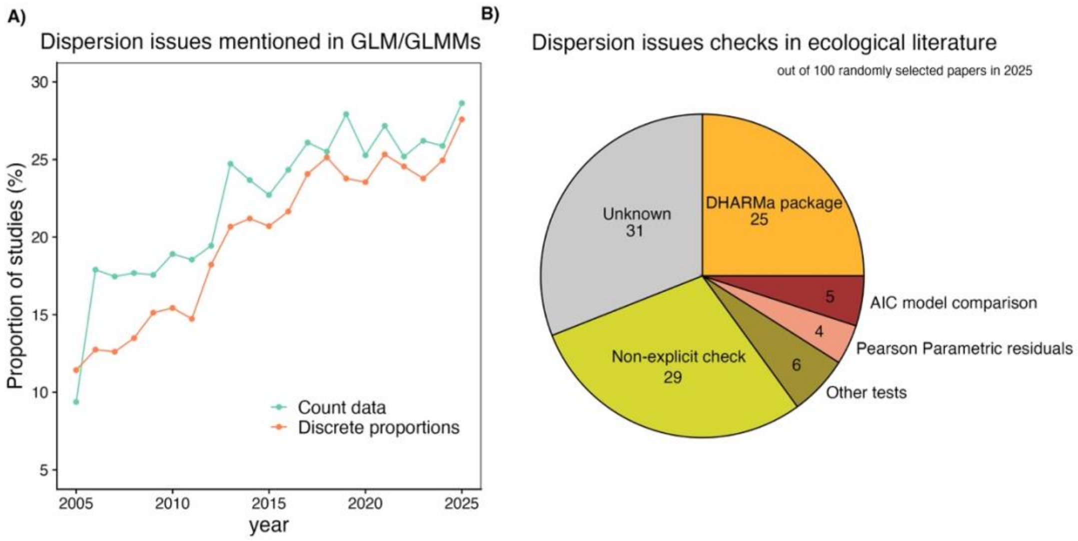


Figure B2. A) Annual trends for the proportion of ecological studies using GLMs/GLMMs for count and/or discrete proportion data **that mention dispersion terms** in the text**. B)** The type of checks and tools used for dispersion problems found in the papers that mentioned dispersion terms in 100 randomly selected papers from 2025. For details of the text mining analysis see S1.

We further analysed a subset of 100 randomly selected ecological papers in 2025 that used GLMs/GLMMs and mentioned dispersion issues in more detail. 81 mentioned overdispersion, 4 underdispersion and 4 tested for both issues (see S1). Among them, only 40 papers explicitly reported testing for dispersion problems (Fig. B2B): 25 using the DHARMa package (but not mentioning which test), 5 papers comparing models fit through AIC (Akaike Information Criterion) and 4 papers mentioning the parametric Pearson residuals test. We conclude that, although the awareness of dispersion problems are steadily increasing in ecology, there is still the need for proper and more standardized tools for checking and testing them.

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. Not a dispersion test.	Requires fitting two models, requires defining an alternative model. For example: - Poisson and negative binomial or generalized Poisson - binomial and beta-binomial	`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*	R Core Team (2024)
			`lmtest::lrtest()`	GLMs	Zeiles & Hothorn (2002)
Score-like test	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)
Standardized 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)
			`aods3::gof()`	GLMs	Lesnoff et al. (2024)
			`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)
		Deviance residuals: assumes the residual deviance are chi-squared distributed.	`aods3::gof()`	GLMs	Lesnoff et al. (2024)
		Dispersion metric: the square root of the penalized residual sum of squares divided by the number of observations.	`blmeco::dispersion_glmer()`	GLMMS from `lmer::glmer()` (computes dispersion parameter only, no test)	Korner-Nievergelt et al. (2019)
Raw residual variance	Compares the expected to the observed variance in raw residuals.	Expected variance of raw residuals calculated through simulations of fitted model. Fast nonparametrics but possibly less exact than working on the standardized residual dispersion.	`DHARMa::testDispersion(…, type="DHARMa")`	GLMs/GLMMs	Hartig (2024)

* Different packages have the S3 method for the *anova* function to perform LRT.

A short review of existing approaches to dispersion tests

After reviewing the available literature, we divided the proposed strategies for addressing 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, this could involve comparing a Poisson GLM to a negative binomial or generalised Poisson GLM (Hilbe, 2014), or comparing 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 cost of fitting two models, is that it doesn't provide any direct diagnostics of over- or underdispersion. The alternative model, however, might also fit better or worse for reasons other than a dispersion problem. Moreover, using LRTs to detect dispersion problems has also been discouraged, as they may yield unreliable results (Dean, 1992) and tend 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 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 constrained to zero). Under the null hypothesis that the overdispersion is indeed zero, the score will have an asymptotic chi-squared 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, this metric is modified 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-squared 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 the Pearson chi-squared dispersion test, the Pearson residuals-based test for overdispersion, or simply the

Pearson dispersion test. Hereafter, we refer to this test as the **parametric Pearson residuals test** to differentiate it from the nonparametric version 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 are 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 these 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-squared 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 simulated residual variances

Simulation approaches can also be useful for generating null distributions for alternative dispersion metrics. 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 raw residuals. The test compares the observed raw residual variance (differences between the observed data and the model predictions) with the simulated raw residual variances (differences between the simulated data and the model predictions). Both variances are scaled to the variance of all simulated observations to account for differences in the number of simulations across the fitted model. For GLMMs, data simulations can be generated conditionally or unconditionally on the fitted random effects. The dispersion statistic is then defined as the ratio of the observed residual variance to the mean of the simulated residual variances. 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 residual variances.

From a theoretical perspective , this approach seems less elegant than the use of Pearson residuals, because the latter, by "standardising" the residual dispersion relative to the expected dispersion, allows each data point to contribute similarly to the dispersion statistic. In contrast, the test on the unstandardised residuals will be more influenced by large data points. However, the primary advantage of this approach is computational, as it enables 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 residual variance test** to differentiate it from 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 residual variance test

To compare the performance of these three tests, we simulated datasets based on the two main distributions that often exhibit over- or underdispersion: the Poisson and the binomial (N/K) proportions. We varied the sample size (from 10 to 10,000) and the intercept (from -3 to 3, at the link function scale) for the simulated data from 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 test performance by comparing type I error, power, and dispersion statistics across all parameter combinations 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 residual variance test and the nonparametric Pearson residuals test, we set the number of simulations to 250 (the default 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 & Ripley, 2002). This implies that when the expected counts (or intercept) and/or the number of observations are small, Pearson residuals may not provide reliable information about model fit (see S2). Some corrections for Pearson residuals in small samples have been suggested (e.g., Cordeiro, 2004; Cordeiro & Simas, 2009), but they are not currently implemented in the most common R packages. Therefore, we expect the parametric Pearson residuals test to perform well for GLMs, except in very small sample sizes and expected counts (hereafter "small-data" situations).

It is unclear whether the parametric Pearson residuals test can be extended to GLMMs or other hierarchical models, where counting residual degrees of freedom (*df*) is not straightforward (Bolker et al., 2009; Luke, 2017). In mixed-effects models, the *df* associated with a random effect are data-specific (adaptive shrinkage) and expected to lie between one and the number of grouping factors (Baayen et al., 2008; Bolker et al., 2009; Luke, 2017). Approaches exist to approximate *df* for random effects in LMMs (e.g., Schaalje et al., 2002), but their generalisation to GLMMs remains an active area of research. Current R packages that implement the parametric Pearson residuals test approximate the *df* using the so-called naïve *df* (e.g., n = 1 per random effect) for testing LMMs/GLMMs (Table 1). We expect that the error introduced by this approximation increases with the number of random-effect groups. To test this, we varied the number of groups in the random intercept (10, 50, and 100) in our simulated data.

In contrast to the parametric Pearson residuals test, we expect the (2) nonparametric Pearson residuals test to be robust to small-data problems and to the presence of random effects, as it doesn't rely on a specific parametric distribution. However, because the test uses parametric bootstrapping, we expected it to run much more slowly than the other tests, especially for more complex GLMMs. For this purpose, we compared the runtimes of the tests using a small set of simulated data (see S7).

For GLMMs tested with the (3) simulation-based residual variance test, we compared the test's performance under the two simulation approaches, conditional and unconditional on random effects. We expect lower power in the unconditional simulation results, as overdispersion is a phenomenon at the model distribution level (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 residual 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 intercept.

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 residual variance test, type I error rates were conservatively very low for all simulated parameters, bouncing below 0.01.

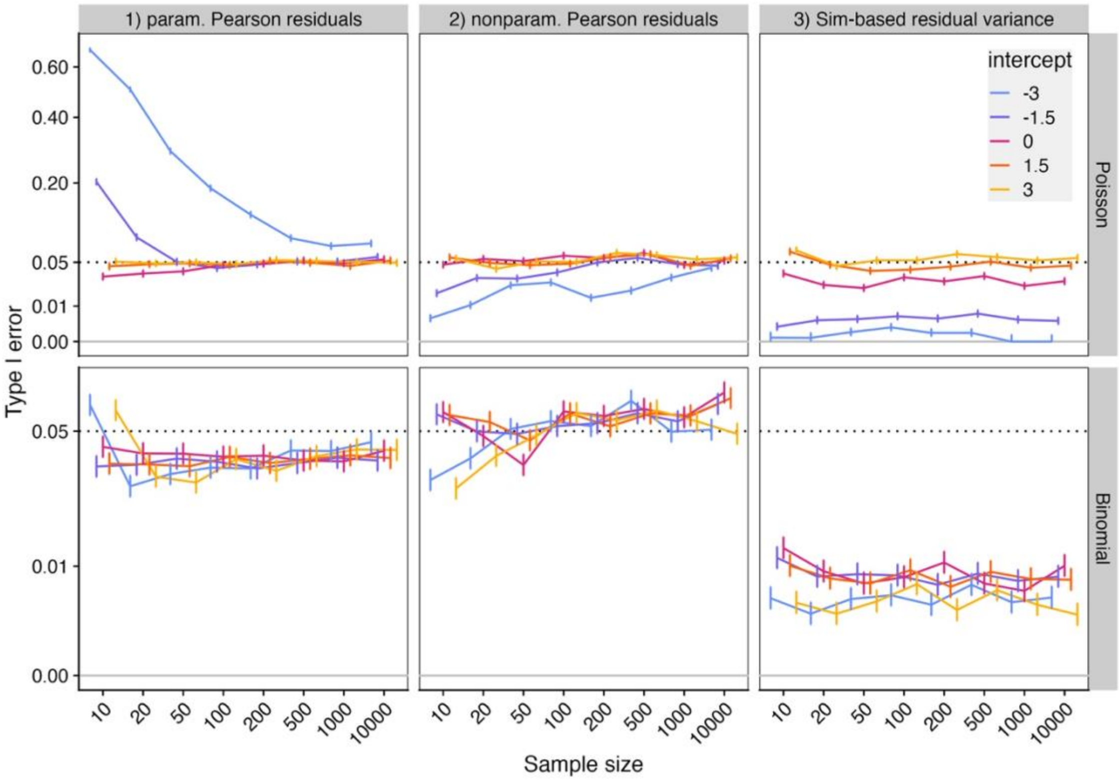


Figure 2. The simulation-based residual variance test has a more conservative type I error rate than both Pearson residual tests. The three dispersion tests were applied to Poisson (upper panels) and binomial proportion (lower panels) GLMs: 1a) parametric Pearson residuals test, 1b) nonparametric Pearson residuals test, and 2) simulation-based residual variance test (see Table 1 for explanations). Simulations were run across different sample sizes (x-axis) and intercepts (colours, values on the link function scale). In B), the model is a binomial proportion with ten trials. All points include a 95% confidence interval calculated from exact binomial tests across 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 type I error value.

The statistical power of the simulation-based residual 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 S6), the differences disapear (Fig. 3).

The dispersion statistics of the simulation-based residual variance test were highly dependent on the intercept, slope, and number of trials in the binomial model (see S5, Fig. S5.2), and tended to be smaller than those based on Pearson residuals. In contrast, for Poisson models, the values tended to be larger than those of Pearson statistics (Fig. S4.5). This may also explain the lower uncorrected power for the simulation-based residual variance test, especially for binomial models.

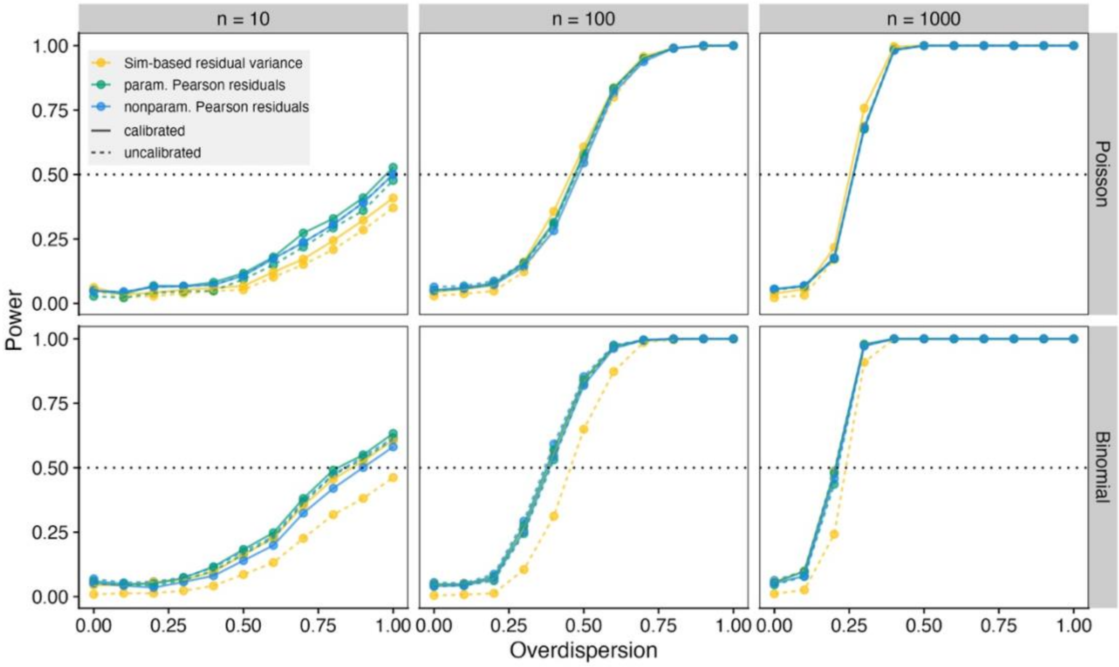


Figure 3. The simulation-based residual variance test (in yellow) has lower power than both Pearson residuals tests (green and blue) for GLMs unless power is calibrated by type I error rates (dashed lines). Lower power is more evident for binomial models (upper panel) and smaller sample sizes (first 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. S6.1 and S6.2.

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.

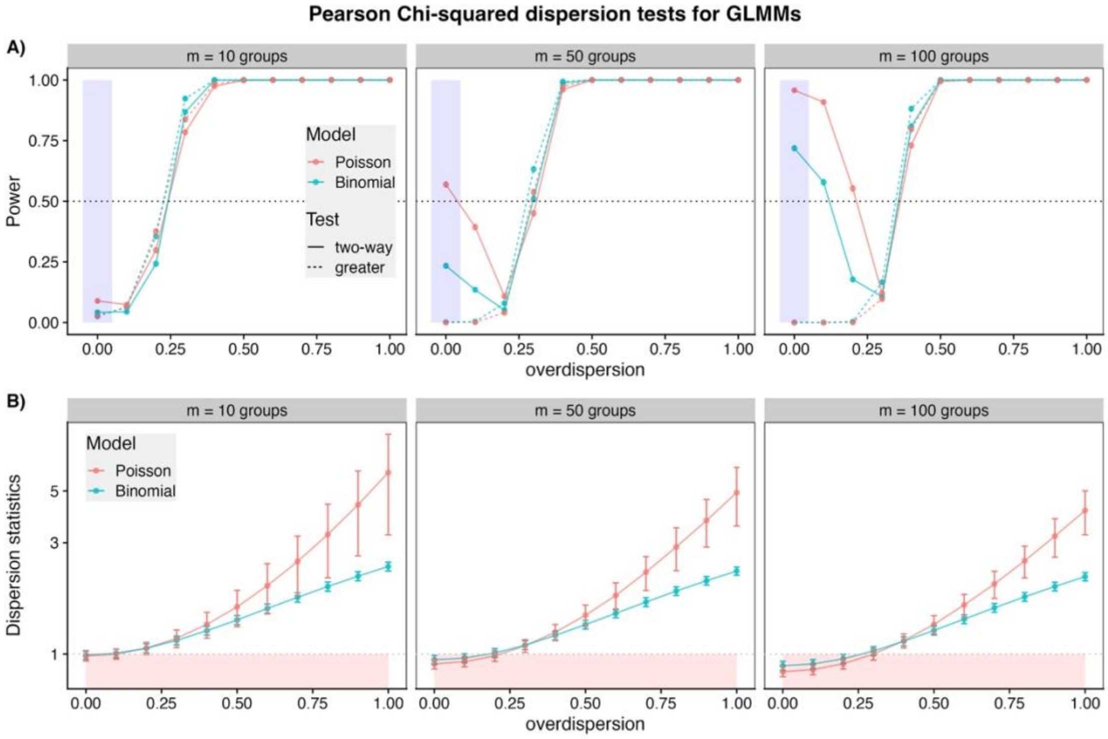


Figure 4. The parametric Pearson residuals test failed for GLMMs with many groups in the random intercepts (plot panels). A) Power and type I error rates (blue shaded area) for the "two-sided" (solid lines) and "greater" (dotted lines) chi-squared tests for the Pearson statistic. B) Pearson dispersion 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. S7.1 and S7.2) and higher power than the simulation-based residual variance tests (Fig. 5). As expected, the unconditional simulation-based residual variance test had the worst performance: very low type I errors (Fig. S7.1 and S7.2), very low power, and dispersion statistics below 1 (Fig. 5B), especially for Poisson models. The conditional simulation-based residual variance test also had very small type I errors (Fig. S7.1 and S7.2), but power increased with the simulated overdispersion. The performance of both simulation-based residual 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.

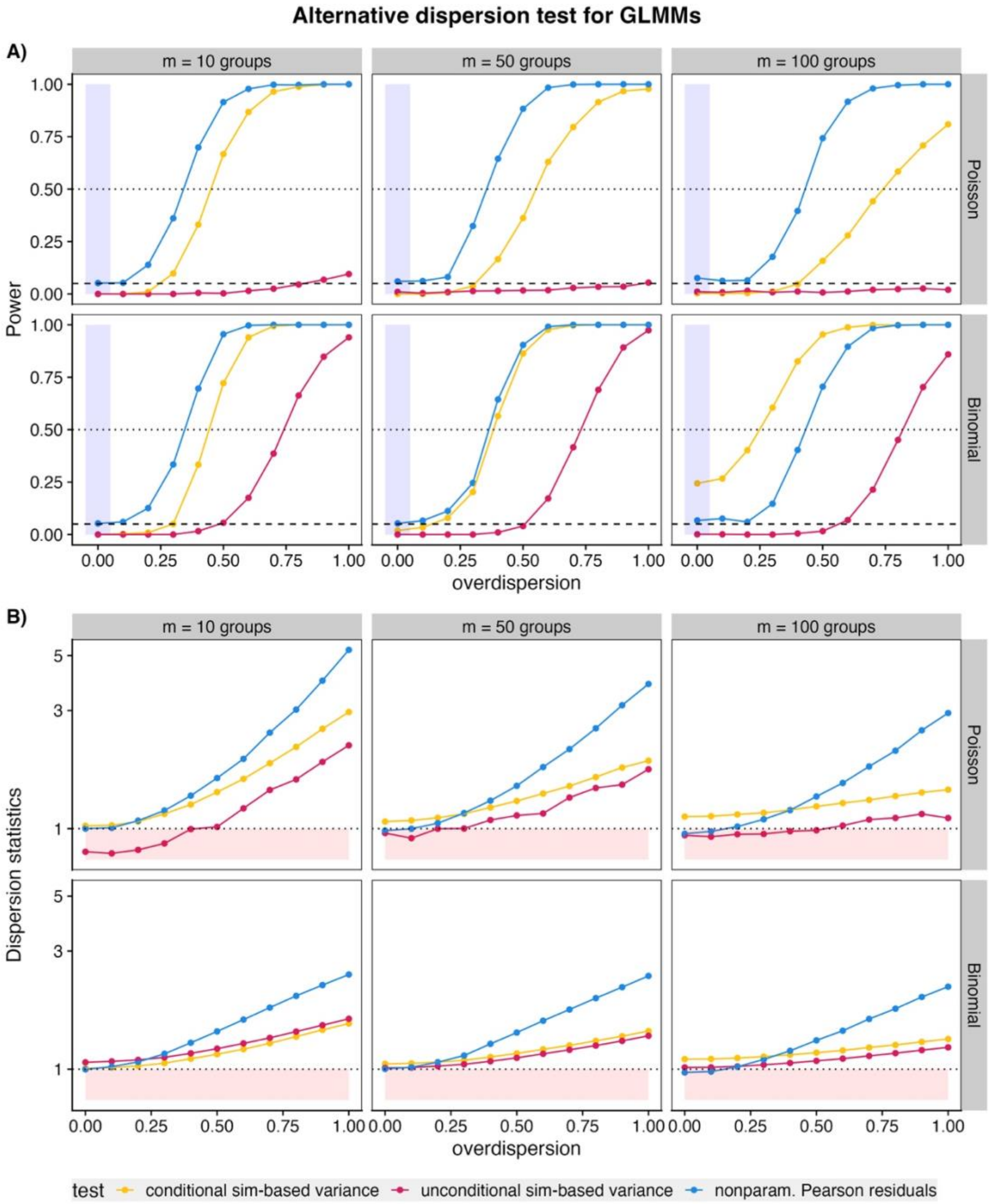


Figure 5. The nonparametric Pearson residuals test showed correct Type 1 error, higher power, and larger dispersion statistics than the simulated-based residual variance tests (conditional and unconditional to all random effects) for Poisson and binomial GLMMs. Power (A), type I error (shaded blue area in A), and dispersion statistics (B) for the alternative dispersion tests for Poisson and binomial GLMMs with different numbers of groups in random intercepts. The dashed horizontal line in (A) indicates the nominal

value of 0.05 for type I error. The dotted horizontal line in (A) indicates the 50% power, and the dotted horizontal line in (B) indicates the dispersion statistics of 1. The results are based on 1,000 simulations per parameter combination, with an intercept of 0 and a sample size (n) of 1,000.

Discussion

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

The simulation-based residual variance test for GLMs is fast to compute, but its dispersion statistic is more difficult to interpret and often leads to overly conservative type I errors. This results in low power unless it is additionally calibrated using 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 performs poorly in (1) small-data situations (Fig. 2) and (2) in the presence of random effects (Fig. 4). The first problem arises from a mismatch between the distribution of the Pearson statistic and the chi-squared distribution under small-data conditions (Fig. S2.1 and S2.2). 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 believe it will be difficult to devise corrections that work across a wide range of distributions.

The second problem arises because counting 1 degree of freedom (*df*) for a random effect, as done in most implementations of this test, typically underestimates the true model *df,* and this underestimation increases with the random effect's number of levels. The result is a bias in the dispersion statistic towards underdispersion that increases with the number of random-effect levels (Fig. 4). Two-sided tests would therefore often wrongly detect significant underdispersion in perfectly valid GLMMs, which is likely why most R implementations of this test only test for overdispersion. When applying this test to GLMMs, we recommend following the same approach and ignoring dispersion statistics smaller than 1. Nevertheless, this is an unsatisfactory solution, as the biased dispersion statistic also causes a loss of power.

A possible solution for GLMMs could be to use a better approximation of the residual degrees of freedom (*df*). For LMMs, approximations for denominator *df* have been successfully used for hypothesis testing (Luke, 2017), for example, the Satterthwaite (1946) and the Kenward-Roger (2009). Although there is some evidence that these approximations are also accurate for GLMMs (Stroup, 2015), the main R packages implementing these methods are currently limited to LMMs (e.g., *pbkrtest* Halekoh & Højsgaard, 2014; *lmerTest* Kuznetsova et al., 2017). However, the recently released package *glmmrBase* (Watson, 2024) allows these methods to be applied to GLMMs. We performed some parametric Pearson residuals tests for Poisson GLMMs using a modified residual *df* approximation (see S9). Although the parametric Pearson residuals tests with the approximated residual *df* performed much better than those with the naïve residual *df*, they still underperformed compared to the nonparametric Pearson residuals test when there were a large number of groups in the random effects (Fig. S9.4), especially in small-data situations.

When are simulation-based residual variance tests an alternative?

The simulation-based residual 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 that of data simulated from the model. The main advantage of this approach is that it is a non-parametric test applicable to any model structure and does not require refitting the model, making it 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 than 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 test statistic being based on the raw variance (rather than a scaled variance, as with the Pearson statistics), which may overrepresent observations with large values. We considered scaling each observation by the 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 S8).

Conclusions and recommendations

Although neither of the considered options for testing dispersion excelled in all dimensions (Fig. 6), our primary recommendation is that for standard GLMs with sufficient data, the parametric Pearson chi-squared test, available in many packages (Table 1), 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. 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 residual 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 supplementary file with instructions and an example for applying dispersion tests using the DHARMa package.

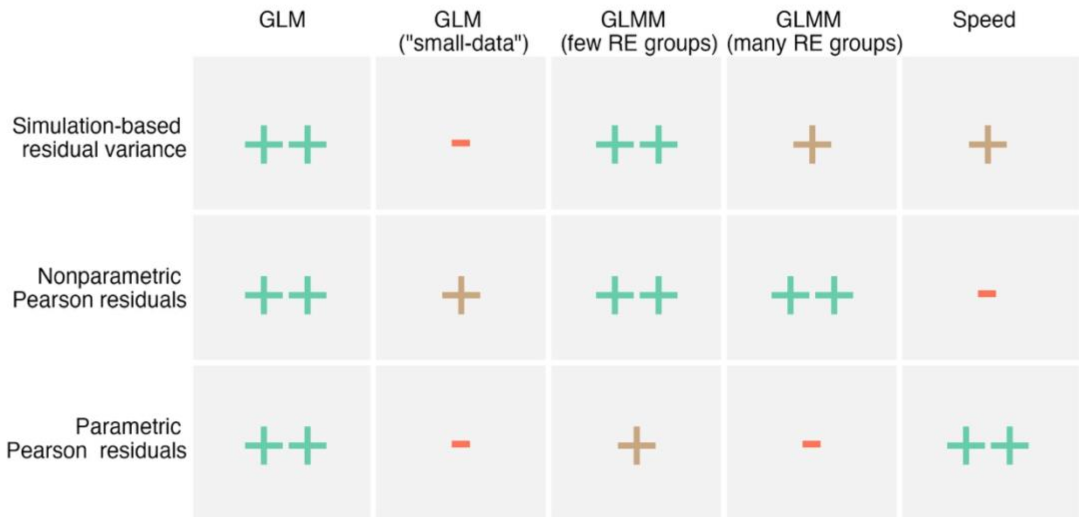


Figure 6. Performance comparisons of the dispersion tests evaluated for each "dimension" for Poisson and binomial models: GLMs in general, GLMs with small sample size or intercept ("small data"), GLMMs with one random effect with few groups/levels, GLMMs with many groups/levels in a random effect, and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

Although our simulation examples focused on overdispersion, the tests considered in our study can also be used to detect underdispersion 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 ecological data analysis when using dispersion tests

For interpretation and applied ecological data analysis, we stress that a significant over- or underdispersion result does not necessarily indicate that the distribution must be changed. First, hypothesis tests evaluate statistical rather than ecological 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 checking for problems beyond the distribution, such as heteroscedasticity, missing predictors, an incorrect link function, excess 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 solution to the underlying problems.

Finally, after ruling out potential model misspecifications leading to under-/overdispersion, we should consider changing the model's distribution, as we may be facing an 'intrinsic' under-/overdispersion problem, likely due to the nature of ecological data (Box 1). A traditional and flexible solution is to use the 'quasi' distributions (Wedderburn, 1974), which essentially correct p-values but do not represent an explicit data-generating process with an associated likelihood, precluding, for example, simulation from the fitted model. A second alternative for adding dispersion is to use observation-level random effects (Bolker et al., 2009; Elston et al., 2001; Harrison, 2014; Ozgul et al., 2009). While often a reasonable solution, excessive use of random effects can create problems in calculating other statistical indicators (such as p-values) that we would rather avoid. For that reason, we consider the best solution to address 'intrinsic' under-/overdispersion is to switch to the corresponding variable-dispersion distributions. For overdispersed count data, the most used is the negative binomial (see S1). However, other distributions have been used in ecology to handle both over- and underdispersion, such as the generalised Poisson, the Conway-Maxwell-Poisson, the Double Poisson, and the Good distributions (Agis et al., 2024; Brooks et al., 2019; Lynch et al., 2014). For discrete proportions data, the beta-binomial distribution (Harrison, 2015) is considered the most appropriate for overdispersed binomial models (Harrison, 2015). Regardless of the approach, an "over-/underdispersion-free" GLM/GLMM is essential for better interpretation of ecological models and for facilitating sound scientific discoveries.

References

Agis, D., Tur, J., Moriña, D., Puig, P., & Fernández-Fontelo, A. (2024). good: An R package for modelling count data. *Methods in Ecology and Evolution*, *15*(12), 2192–2197. https://doi.org/10.1111/2041-210X.14387

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, Special Issue: Emerging Data Analysis*, *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

Brooks, M. E., Kristensen, K., Darrigo, M. R., Rubim, P., Uriarte, M., Bruna, E., & Bolker, B. M. (2019). Statistical modeling of patterns in annual reproductive rates. *Ecology*, *100*(7), e02706. https://doi.org/10.1002/ecy.2706

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

Campbell, N. A., & Reece, J. B. (2005). *Biology* (7th ed.). Pearson Benjamin Cummings.

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

Korner-Nievergelt, F., Roth, T., Felten, S. von, Guelat, J., Almasi, B., & Korner-Nievergelt, P. (2019). *blmeco: Data Files and Functions Accompanying the Book "Bayesian Data Analysis in Ecology using R, BUGS and Stan"* (Version 1.4) [Computer software]. https://cran.r-project.org/web/packages/blmeco/index.html

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

Lesnoff, M., Lancelot, R., & Siberchicot, A. (2024). *aods3: Analysis of Overdispersed Data using S3 Methods* (Version 0.5) [Computer software]. https://cran.r-project.org/web/packages/aods3/index.html

Lindén, A., & Mäntyniemi, S. (2011). Using the negative binomial distribution to model overdispersion in ecological count data. *Ecology*, *92*(7), 1414–1421. https://doi.org/10.1890/10-1831.1

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.

McMahon, S. M., & Diez, J. M. (2007). Scales of association: Hierarchical linear models and the measurement of ecological systems. *Ecology Letters*, *10*(6), 437–452. https://doi.org/10.1111/j.1461-0248.2007.01036.x

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

Nakagawa, S., Ortega, S., Gazzea, E., Lagisz, M., Lenz, A., Lundgren, E., & Mizuno, A. (2026). Location–scale models in ecology and evolution: Heteroscedasticity in continuous, count and proportion data. *Methods in Ecology and Evolution*, *17*(2), 554–566. https://doi.org/10.1111/2041-210x.70203

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

Puig, P., Valero, J., & Fernández-Fontelo, A. (2024). Some mechanisms leading to underdispersion: Old and new proposals. *Scandinavian Journal of Statistics*, *51*(1), 245–267. https://doi.org/10.1111/sjos.12677

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

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

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

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

Watson, S. I. (2024). *Generalised Linear Mixed Model Specification, Analysis, Fitting, and Optimal Design in R with the glmmr Packages* (arXiv:2303.12657). arXiv. https://doi.org/10.48550/arXiv.2303.12657

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

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

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

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

Supporting information for:

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

S1. Trend analysis and current ecological literature practices on dispersal issues

To understand the extent of ecological studies relying on GLMs/GLMMs for count and discrete proportion data and those that address dispersion issues, we conducted a text analysis of the ecological literature over the past 20 years. We used the R package 'europepmc' (v0.4.3, Jahn, 2023) to search for articles in the PubMed and Medline NLM databases from 2005 to 2025. We used combinations of words (Table S1.1) to retrieved the annual records for: (1) the percentage of ecological papers using GLMs/GLMMs for count and discrete proportion data (Figure S1.1a), (2) the percentage of those papers that mention dispersion terms in general (Figure S1.1b), (3) the percentage of ecological papers using GLMs/GLMMs for count data mentioning dispersion terms (Figure BOX 1, main text), and (4) the percentage of ecological papers using GLMs/GLMMs for discrete proportion that mentioning dispersion terms (Figure BOX 1, main text).

Table S1.1. Word combinations used for the literature review on ecological practices for count and discrete proportion data analysed with GLMs/GMMs and dispersion issues.

Terms	Words combination
1. Ecology:	
	"ecology" OR "ecolog*"
2. Generalised linear models for count and discrete proportion data:	
	"count data" OR "poisson" OR "negative binomial" OR "generalized poisson" OR "generalised poisson" OR "conway-maxwell poisson" OR "binomial" OR "beta-binomial" OR "binomial proportion"
3. Generalized linear models for count data only:	
	"count data" OR "poisson" OR "negative binomial" OR "generalized poisson" OR "generalised poisson" OR "conway-maxwell poisson"
4. Generalized linear models for discrete proportion data only	
	"binomial" OR "beta-binomial" OR "binomial proportion"
5. Dispersion terms:	
	"overdispersion" OR "over dispersion" OR "over-dispersion" OR "underdispersion" OR "under dispersion" OR "under-dispersion" OR "dispersion"

The percentage of papers that mention count or proportion data in the context of GLM/GLMM analysis increased 4-fold over 20 years, but appears to have stabilised since 2015 (Figure S1.1A). For those papers, there is an increasing trend in mentioning dispersion terms, reaching almost 25% in 2025 (Figure S1.1B). However, it means that 3/4 of ecological papers that mentioned GLMs/GLMMs for analysing count and/or discrete proportion still don’t report checking for dispersion problems.

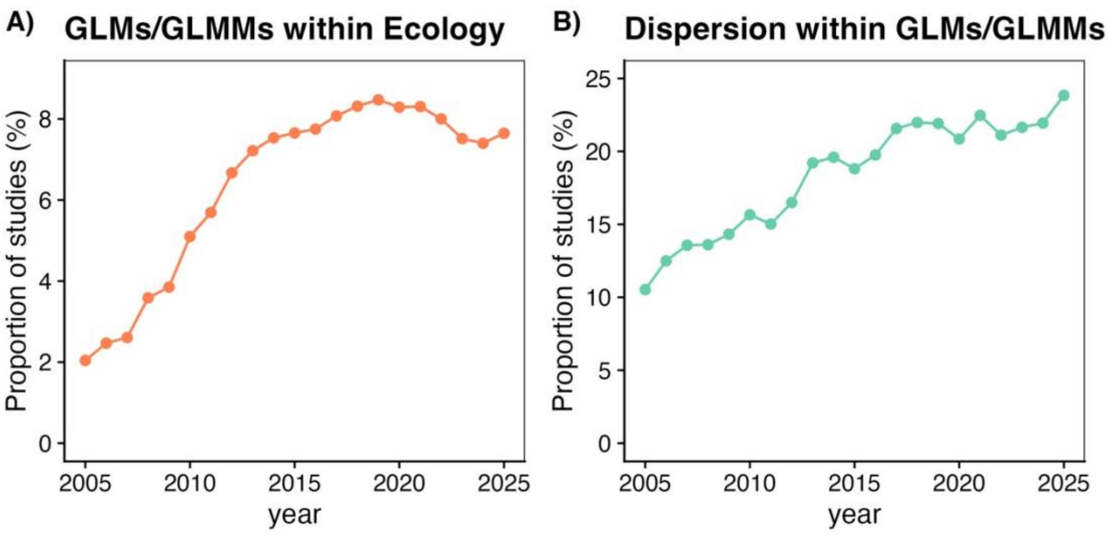


Figure S1.1. Trend analysis from the last 20 years of (A) ecological papers mentioning GLM/GLMMs for count and/or discrete proportion data, and (B) ecological papers that use GLM/GLMMs and mention dispersion terms.

We then summarised the current practices in dispersal issues for the ecological studies using GLMs/GLMMs for count and discrete proportion by searching for papers with the combination of words of the groups 1, 2 and 5 (Table S1.1). The query retrieved 7634 articles; we further selected only open-access articles from journals that publish ecological papers in 2025. From the subset of 457 articles, we randomly selected 200 papers for detailed information searches and retrieved the first 100 papers within the scope (ecology) that used count or discrete proportion data. To reach 100 papers, we read 155 papers; 33 were out of scope, and 22 did not use count or discrete proportion data analysis. Among them, 89 papers explicitly mentioned a dispersion issue in the methods section; 81 papers mentioned overdispersion, 4 mentioned underdispersion, and 4 mentioned both (tested for both issues). A total of 69 papers explicitly reported checking for dispersion, whereas only 40 reported testing for dispersion problems or comparing model fits using AIC.

Of the 40 papers explicitly testing dispersion, 25 reported using the DHARMa R package (Hartig, 2024), and 5 reported using the performance package (Lüdecke et al., 2021). However, almost all of them didn't mention which test. Model comparison using AIC was reported in 5 papers, and the Pearson Chi-squared test (Pearson parametric residuals test) in 4, including one paper that used GLMMs and reported underdispersion in many models (Laumer et al., 2025). This recent literature review shows an increasing number of ecological studies examining dispersion problems, underscoring the importance of appropriate tools for their detection and testing.

Additionally, we found that the most common approach to address dispersion issues in count data was to switch from the Poisson distribution to the negative binomial, or starting with the negative binomial in the first place (46 out of 78 records, 59%). Only 3 papers used the generalised Poisson distribution, and 1 paper reported using the Conway-Maxwell-Poisson for underdispersed data. The quasi-Poisson approach was reported in 7 papers (9%), the use of an observation-level random effects in a Poisson GLMM was reported in 5 papers (6%), and the use of a zero-inflated (Poisson or negative binomial) model was reported 12 times (15%).

For discrete proportion data, we identified 7 papers that report alternative modelling to account for overdispersion. The quasi-binomial approach and the beta-binomial distribution were reported 3 times each. The use of an observation-level random effects in a binomial GLMM was reported in just one paper.

S2. 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 S2.1 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 S2.2A).

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 S2.1B). 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).

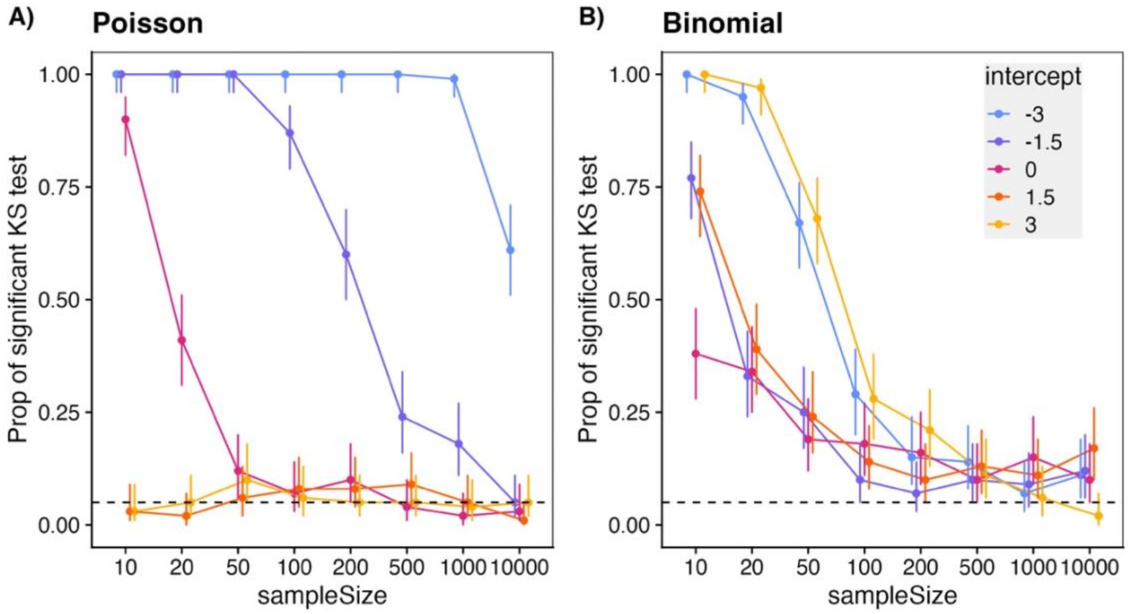


Figure S2.1. Proportion of significant Kolmogorov-Smirnov adherence tests between the empirical distribution of 1000 simulations of the Pearson statistics and a Chi-squared distribution with the same residual degrees of freedom for A) Poisson and B) binomial GLMs. Proportions were calculated from 100 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.

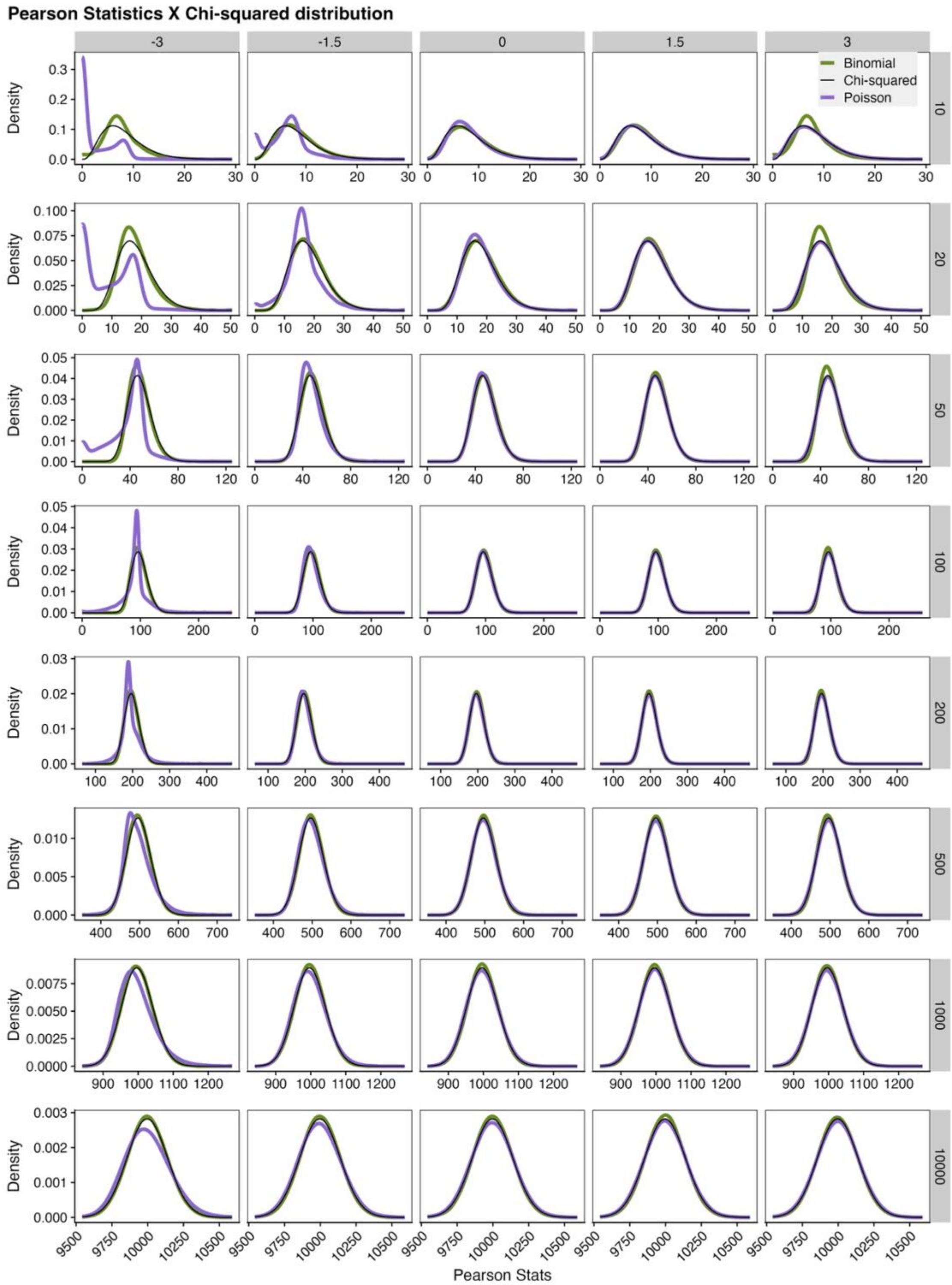


Figure S2.2. Mean Pearson statistics distribution (from 100 simulated curves) for the binomial (green) and Poisson (purple), and the Chi-square distribution in black.

S3. Type I error rates for the GLMs

Figures S3.1 and S3.2 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 S3.1), the simulation-based residual 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 S3.2), the p-values distribution of the simulation-based residual 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.

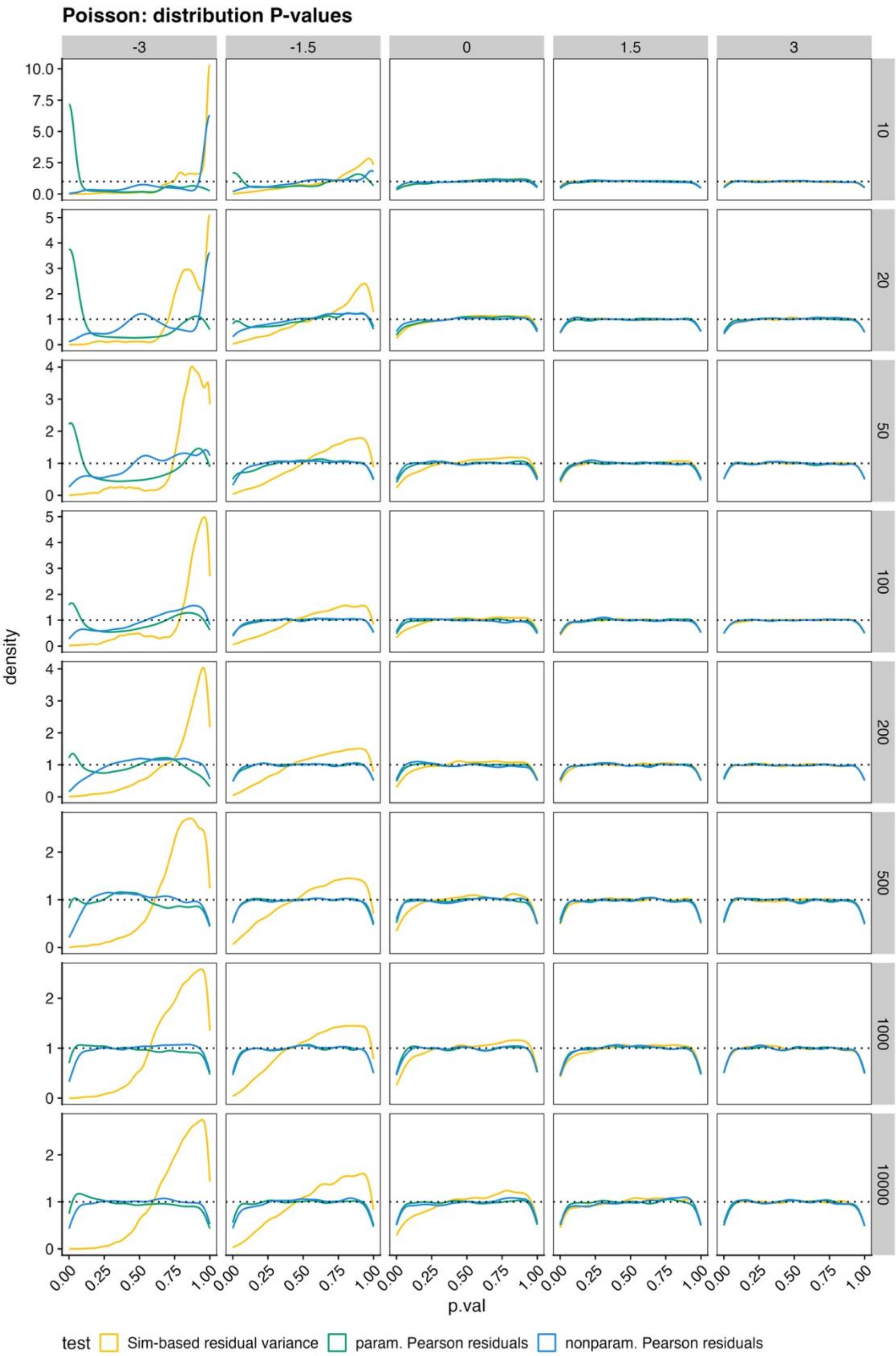


Figure S3.1. Distribution of p-values for the Poisson GLMs for each dispersion test. 10,000 simulations per simulation set (intercept x sample size).

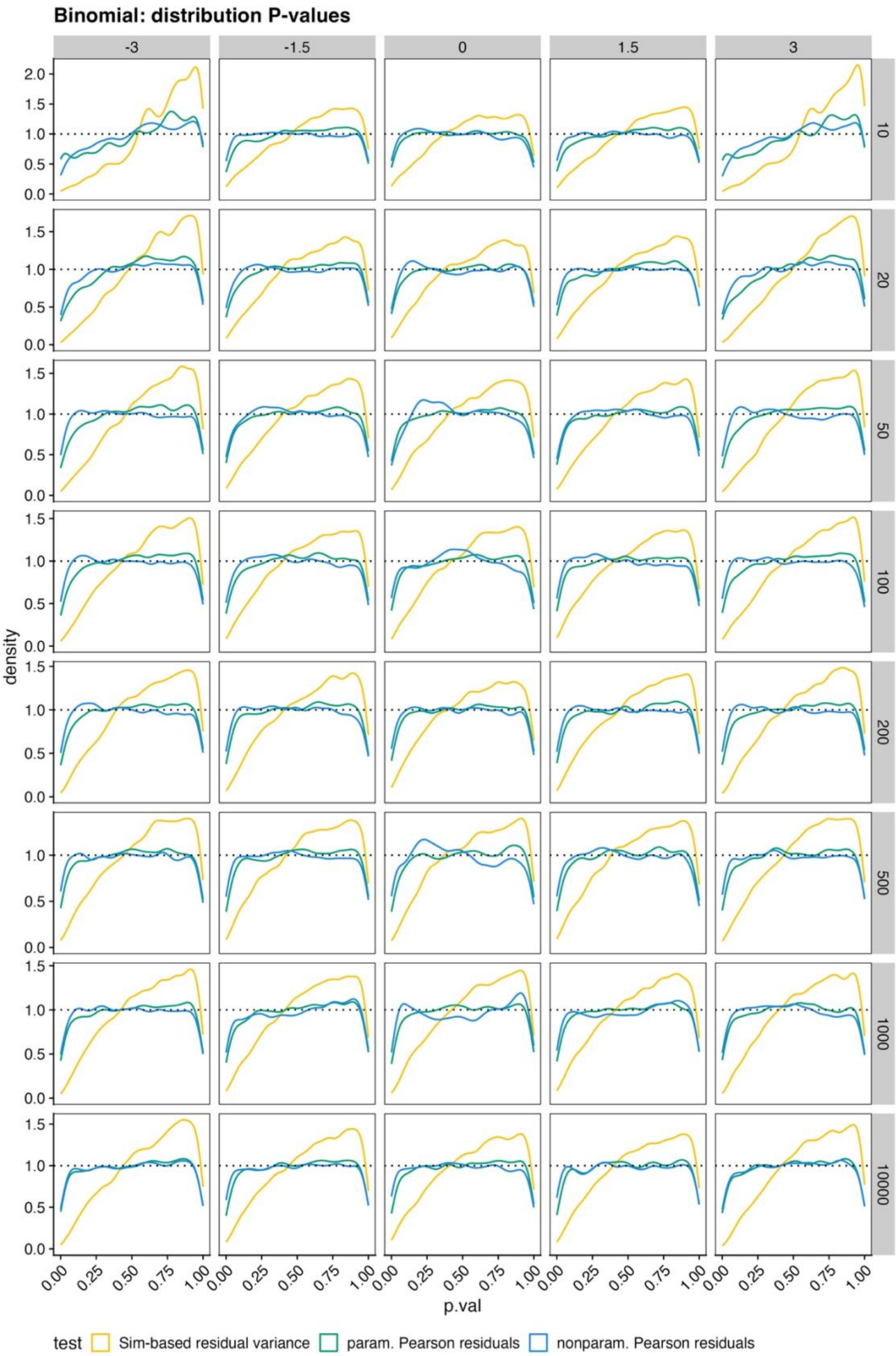


Figure S3.2. Distribution of p-values for the binomial GLMs for each dispersion test. 10,000 simulations per simulation set (intercept x sample size).

S4. 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 S4.1). 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 S4.2 and S4.3), we found that both Pearson-based dispersion statistics presented similar values. In contrast, the dispersion statistic of the simulation-based residual 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 S4.4 and S4.5). Moreover, we found out that the dispersion statistics of the simulation-based residual variance test depend heavily on the slope parameter of the simulated data (Figure S4.6).

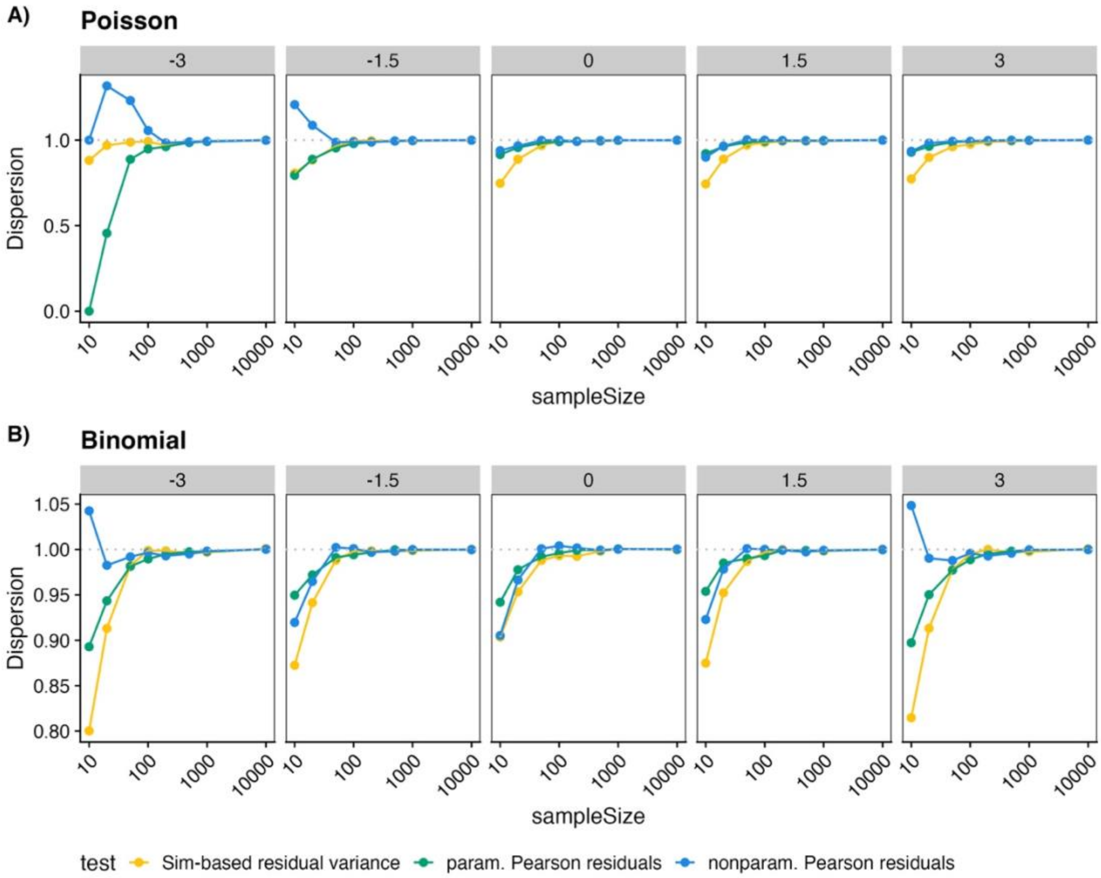


Figure S4.1. Median of the dispersion statistics of the tests for A) Poisson and B) binomial GLMs, simulated without overdispersion for different intercepts (panels) and sample sizes (x-axis) for the three dispersion tests: parametric Pearson test, nonparametric Pearson test, and simulation-based residual variance test. The dotted horizontal line indicates the ratio of 1. Values below the line are considered underdispersion, and above the line are overdispersion. For all simulations, the slope was fixed at 1.

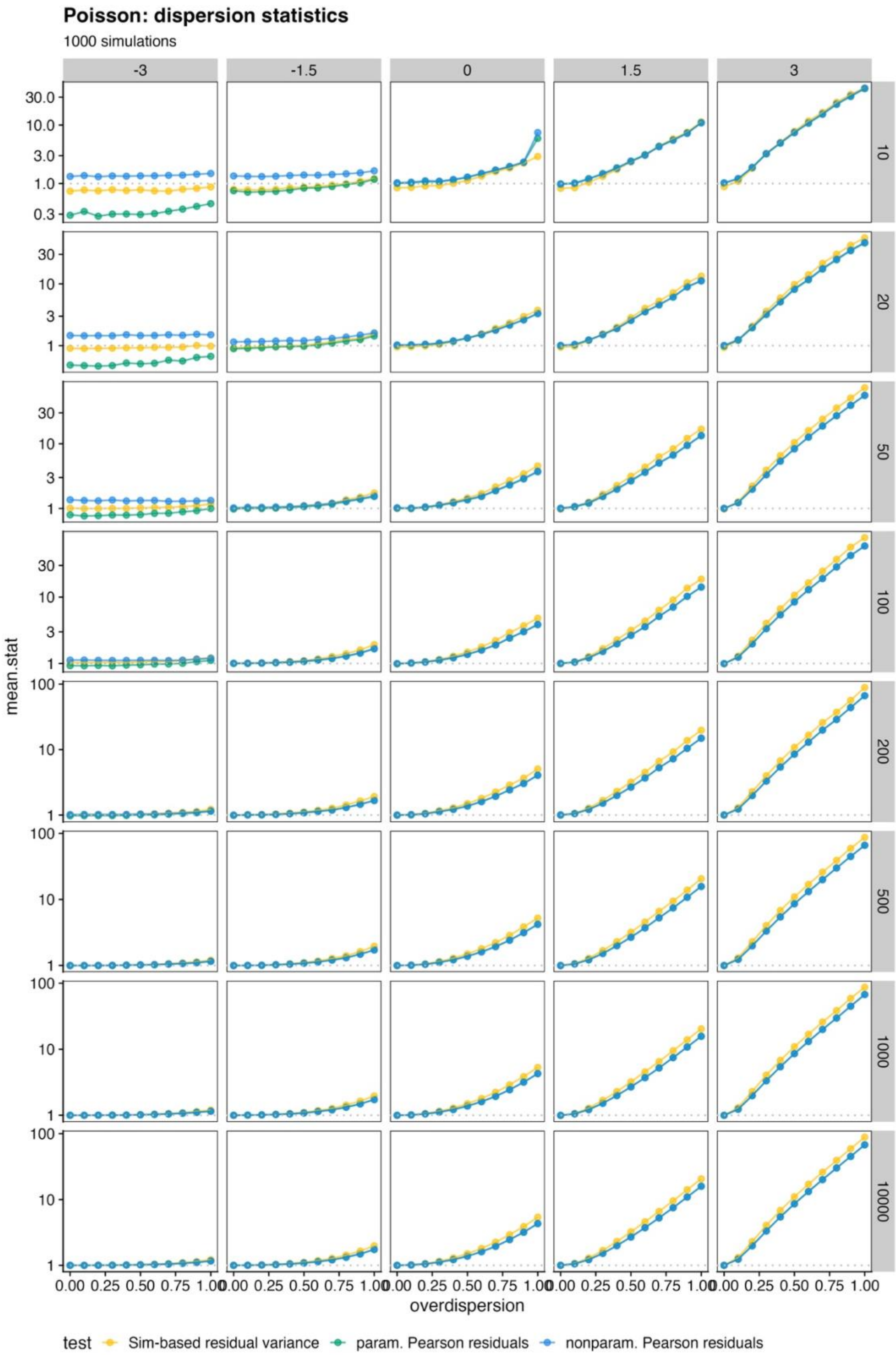


Figure S4.2. Dispersion statistics (median) for GLM Poisson. Notice the different y-axis scales across sample sizes.

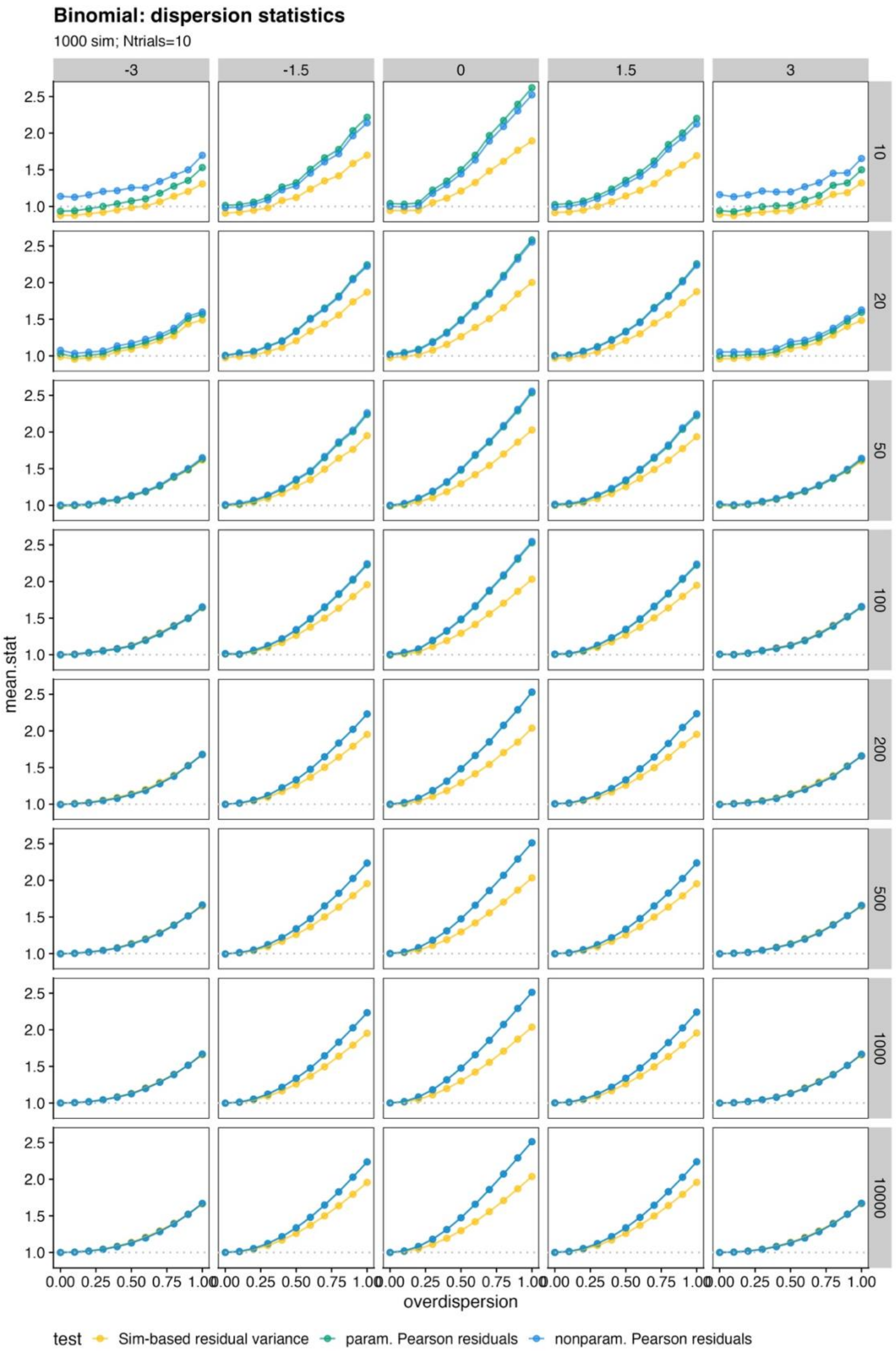


Figure S4.3. Dispersion statistics (median) for GLM binomial.

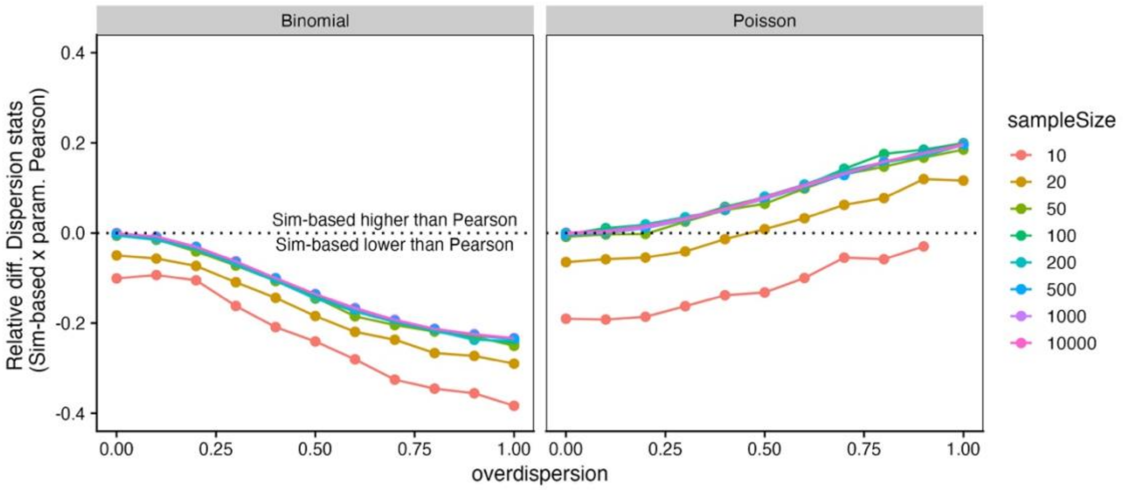


Figure S4.4. The dispersion statistics of the simulation-based residual variance test are smaller than the parametric Pearson test statistics for all binomial models and for small sample sizes in Poisson models. The differences between the two dispersion statistics decrease with increasing sample size (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.

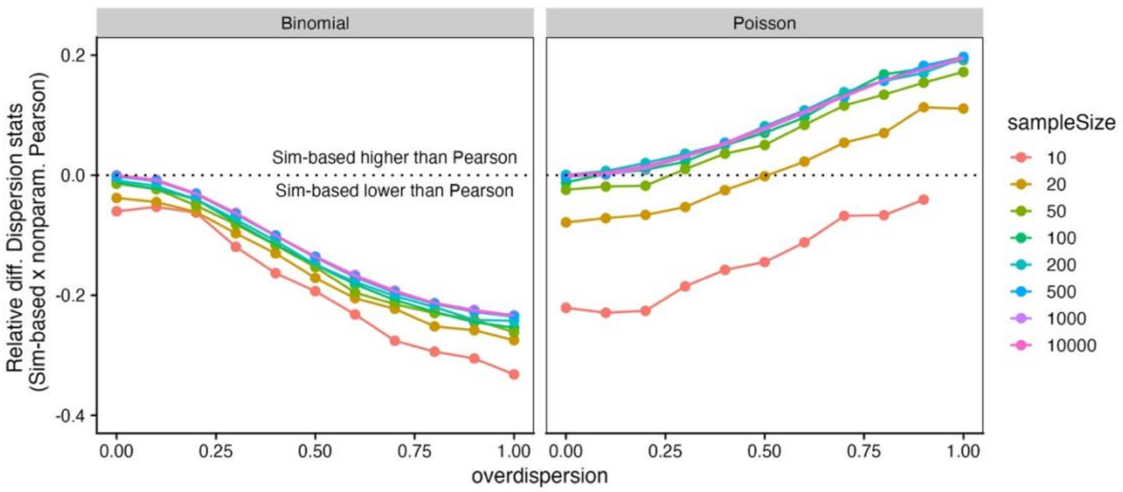


Figure S4.5. The dispersion statistics of the simulation-based residual variance test are smaller than nonparametric Pearson dispersion statistics for all binomial models and for small sample sizes in Poisson models. The differences between the two dispersion statistics decrease with increasing sample 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.

S5: 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 S5.1, we investigated the power and dispersion statistic for datasets simulated with different slopes (the default slope in all other simulations was 1). In Figure S5.2, we investigated the effect of varying the number of trials on the binomial GLMs in terms of power, type I error, and dispersion statistics.

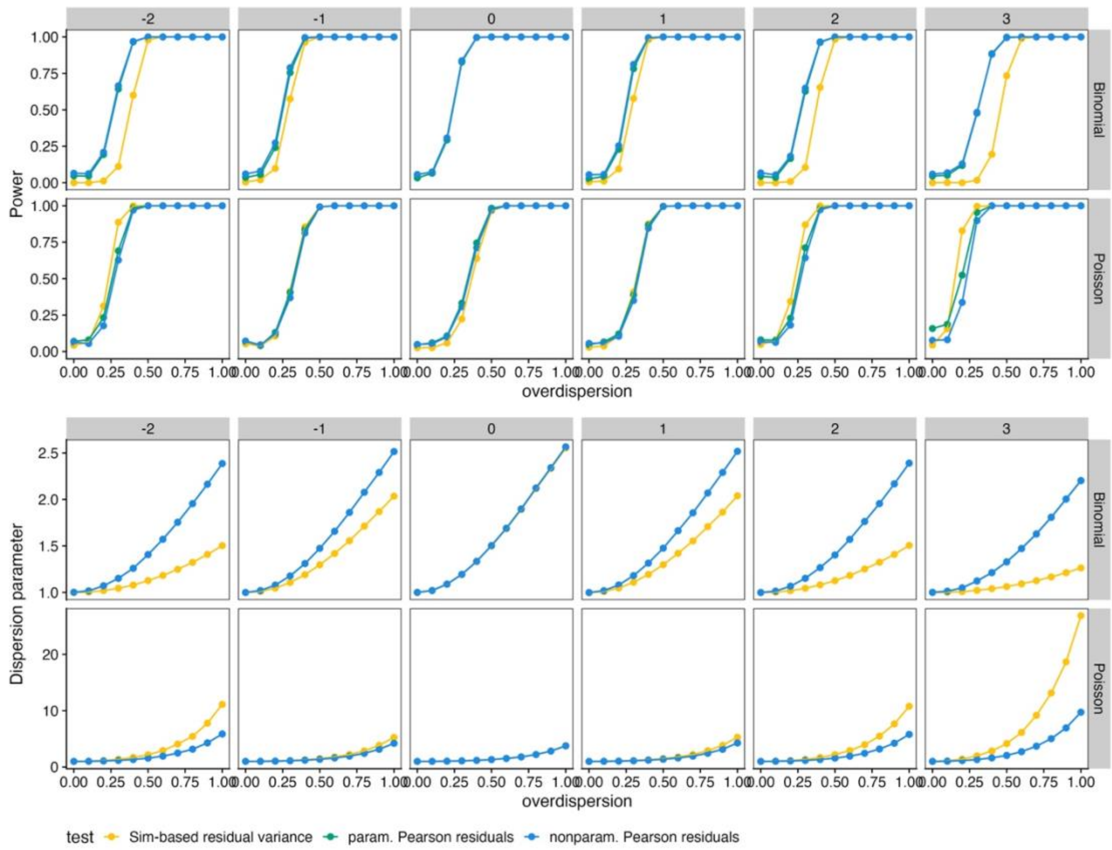


Figure S5.1. Power and dispersion statistics for simulations with different slopes (panel columns) for binomial and Poisson GLMs. Number of simulations = 500; intercept = 0, number of trials for the binomial = 10.

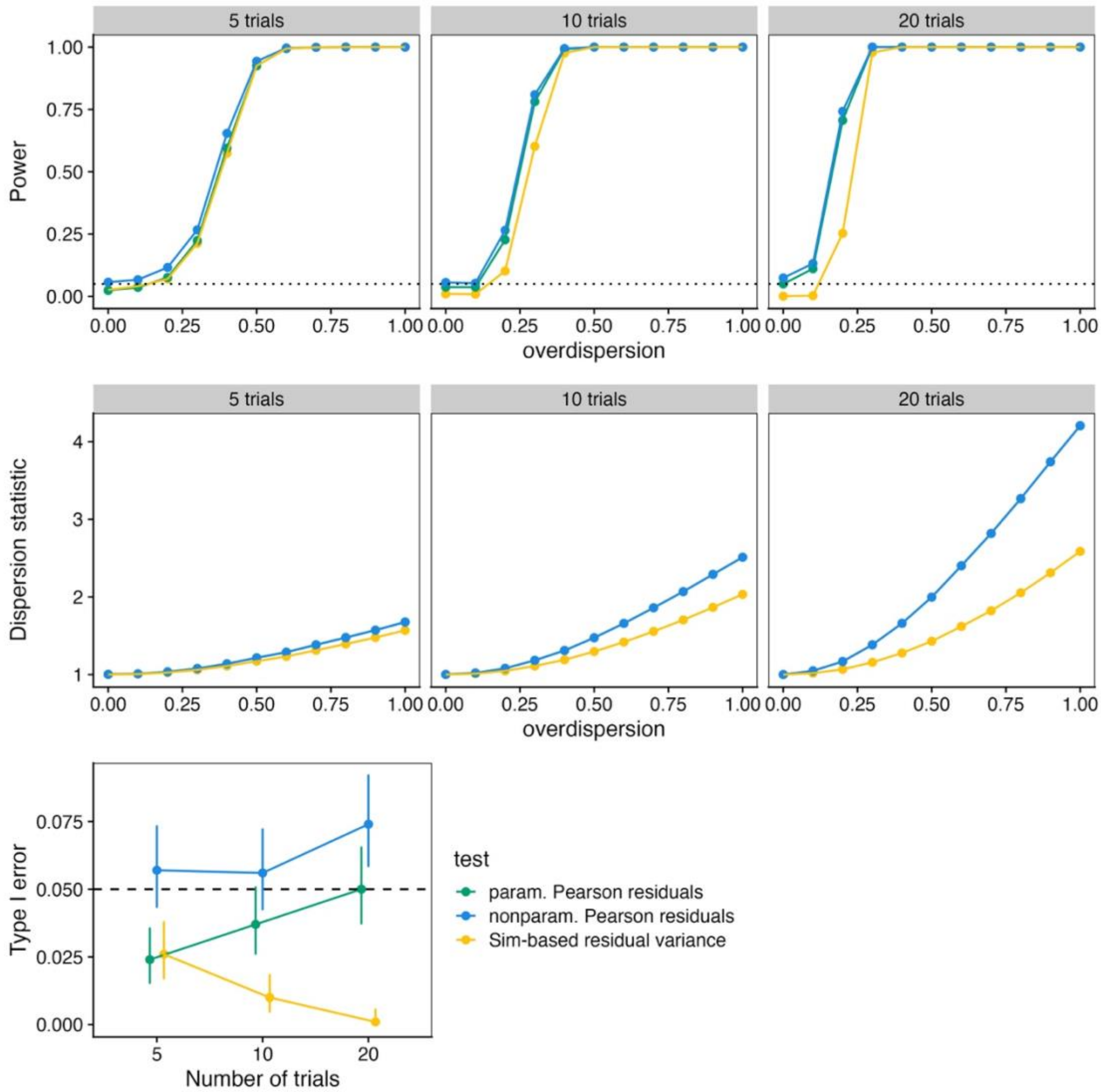


Figure S5.2. Power, dispersion statistics, and type I error of dispersion tests for binomial data simulations with different numbers of trials (panel columns). The fixed parameters are: intercept = 0, sample size = 500, slope = 1. Results for 1000 simulations.

S6. Power for the GLMs

Power calibration

To investigate if the lower power of the simulation-based residual 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 S3.1 and S3.2). This method should provide an estimate of differences in power, controlling for type I error rate (Luke et al. 2017). Figures S6.1 and S6.2 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.

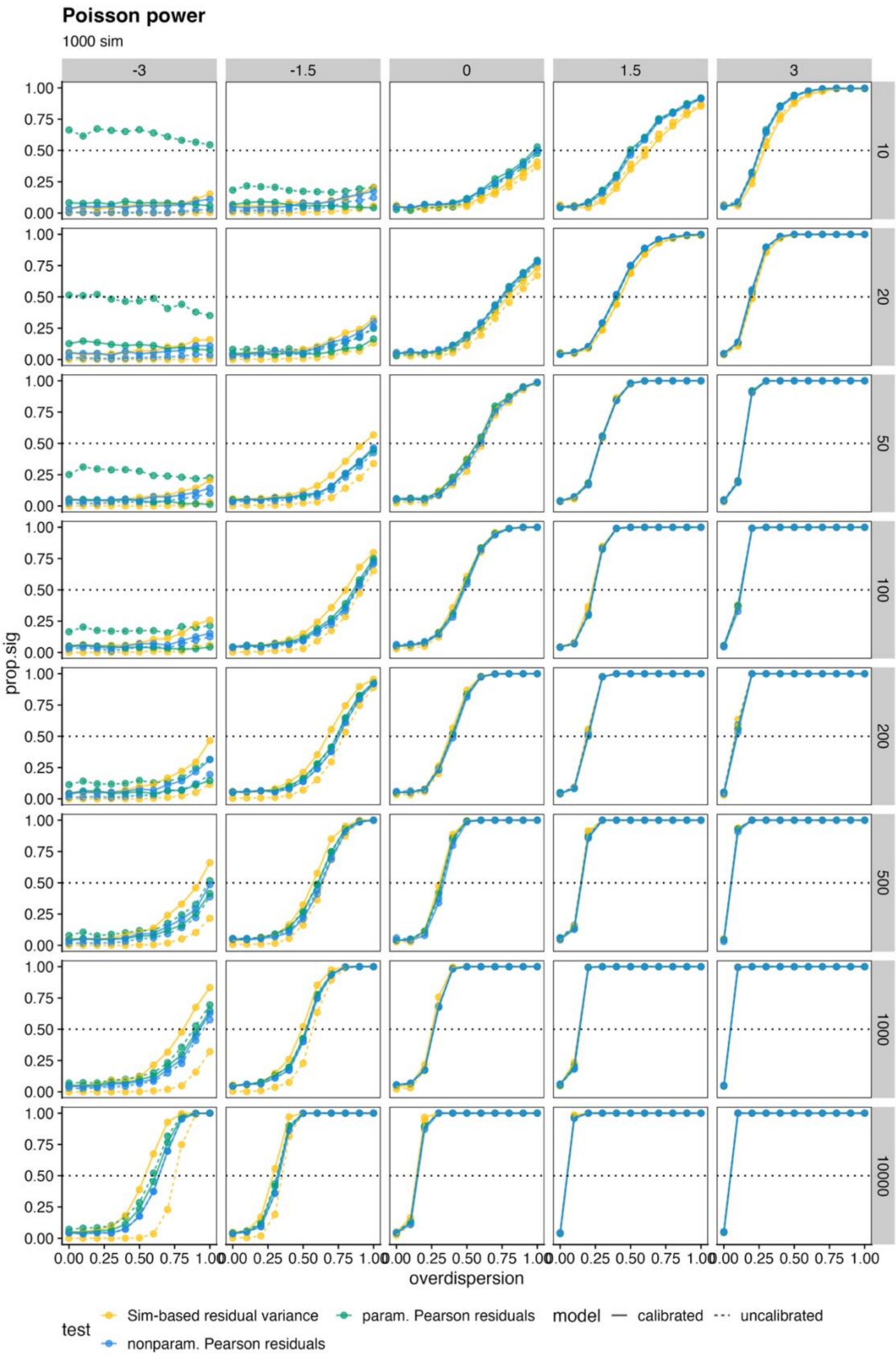


Figure S6.1. Power for GLM Poisson.

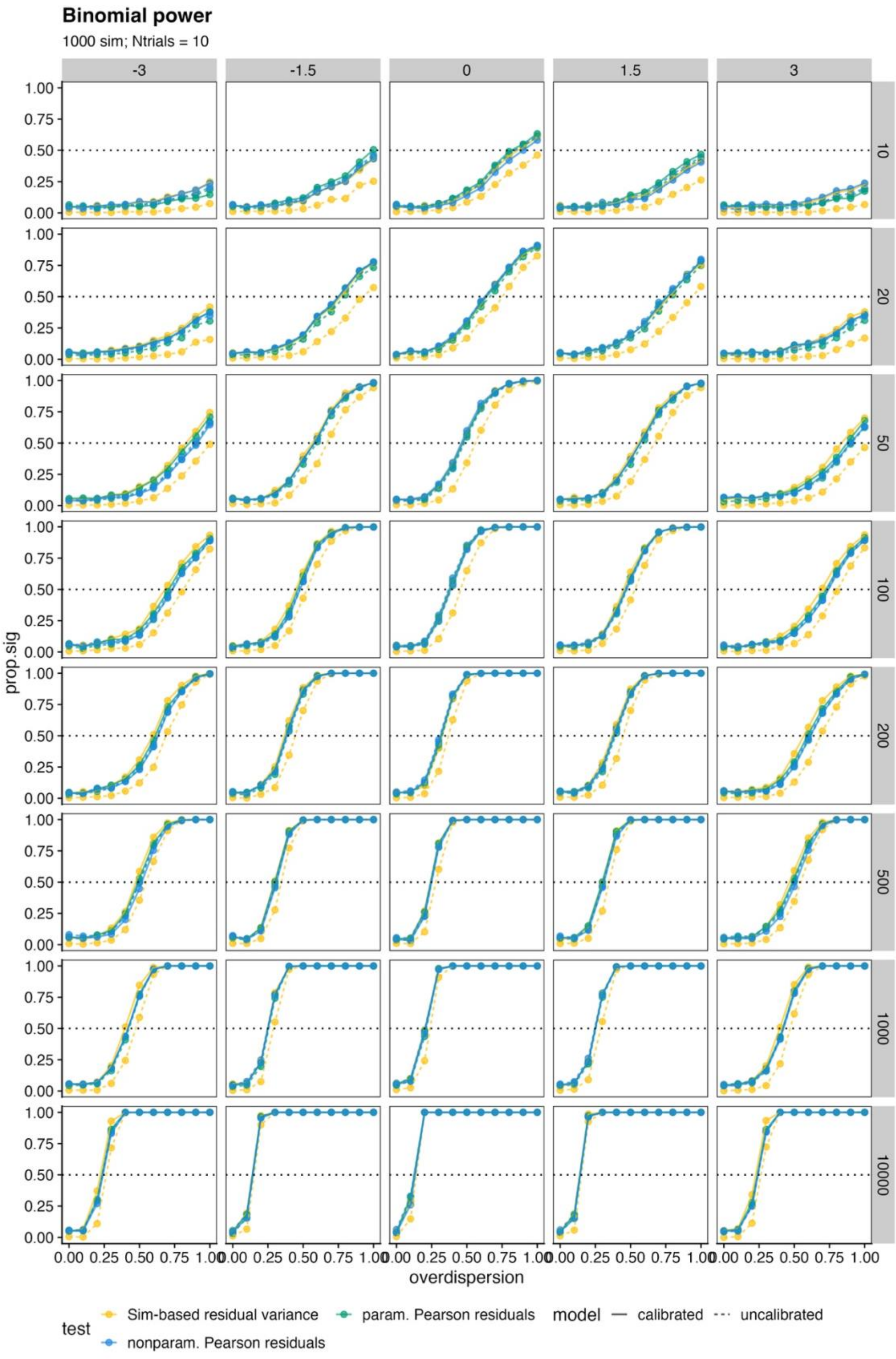


Figure S6.2. Power for GLM binomial.

S7. Additional GLMM results

Type I error rate of the alternative dispersion tests

In Figures S7.1 and S7.2, 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.

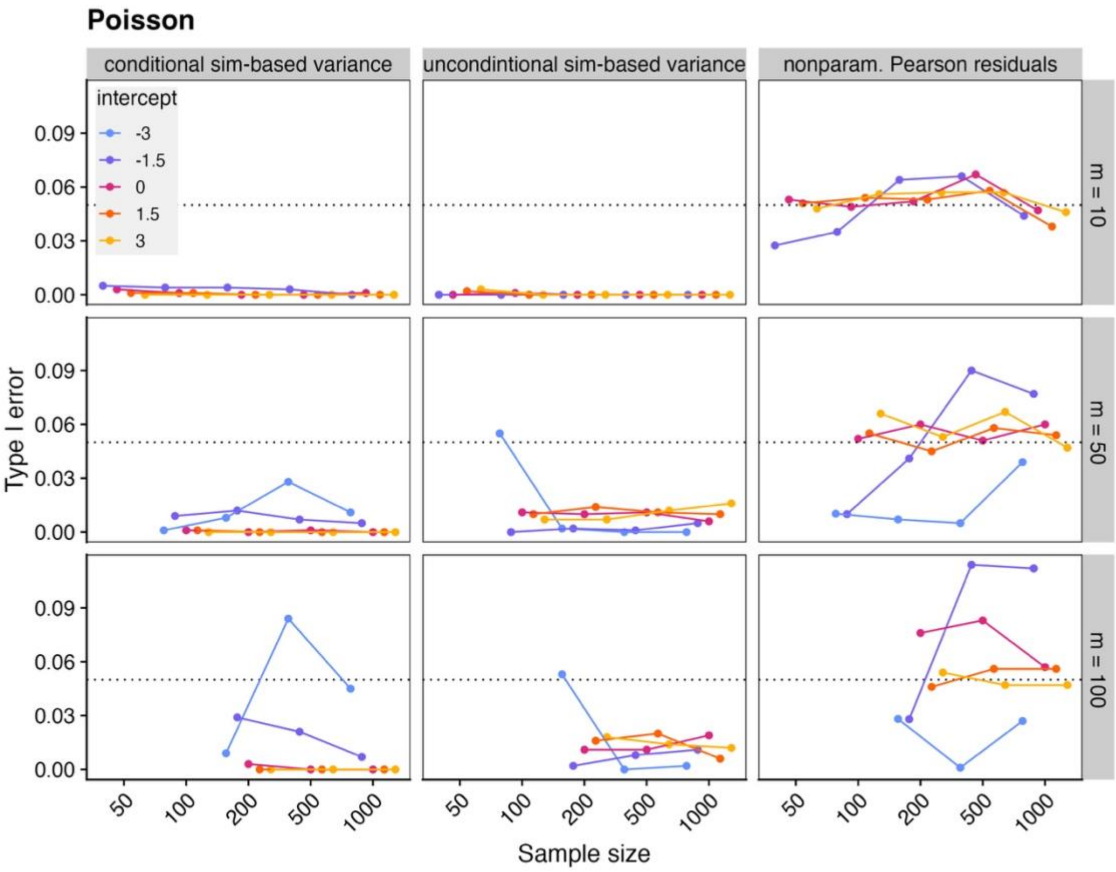


Figure S7.1. Type I error rate for the three alternative dispersion tests for the Poisson GLMMs. 1000 simulations for each parameter set. To improve visualisation of the different intercept lines, the x-axis values were slightly displaced to align with the sample size values.

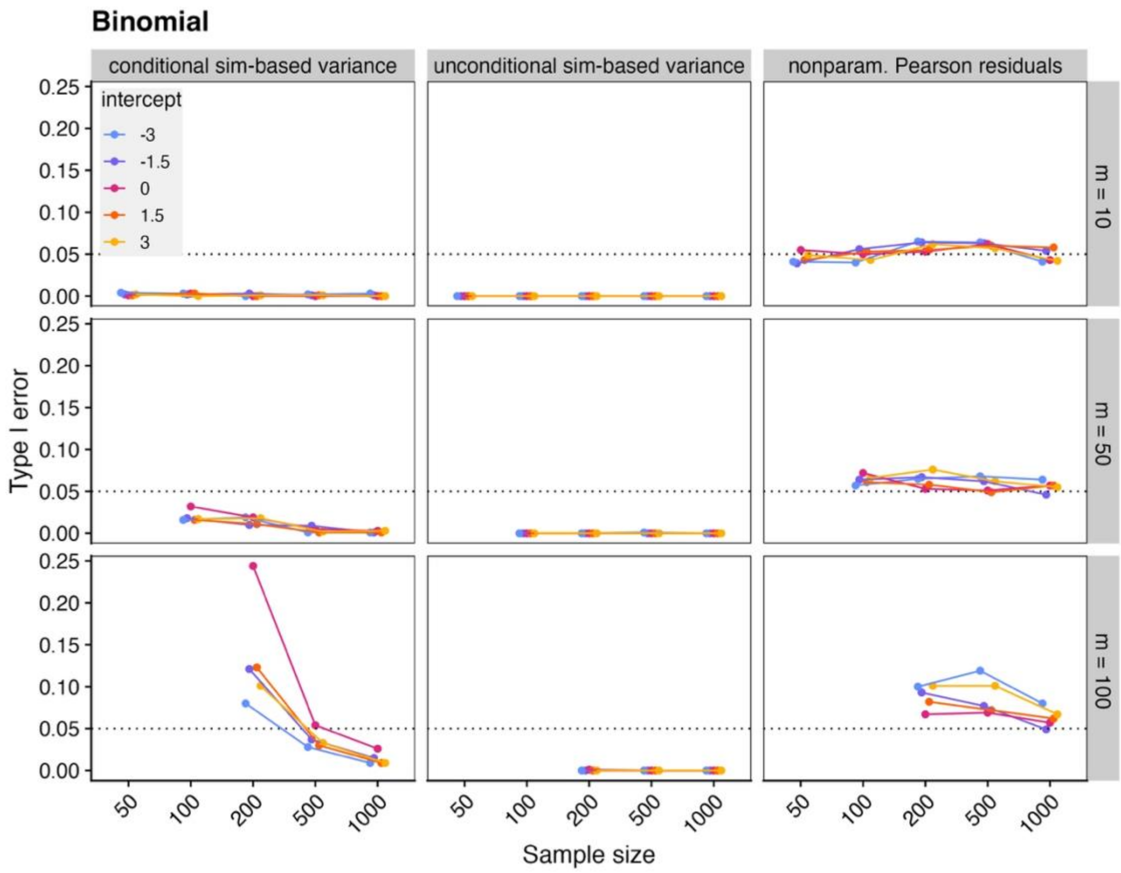


Figure S7.2. Type I error rate for the three alternative dispersion tests for binomial GLMMs. 1000 simulations for each parameter set. To improve visualising the different intercept lines, the values in the x-axis were slightly displaced around the sample size values.

Power of the alternative dispersion tests

In Figures S7.3 and S7.4, 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.

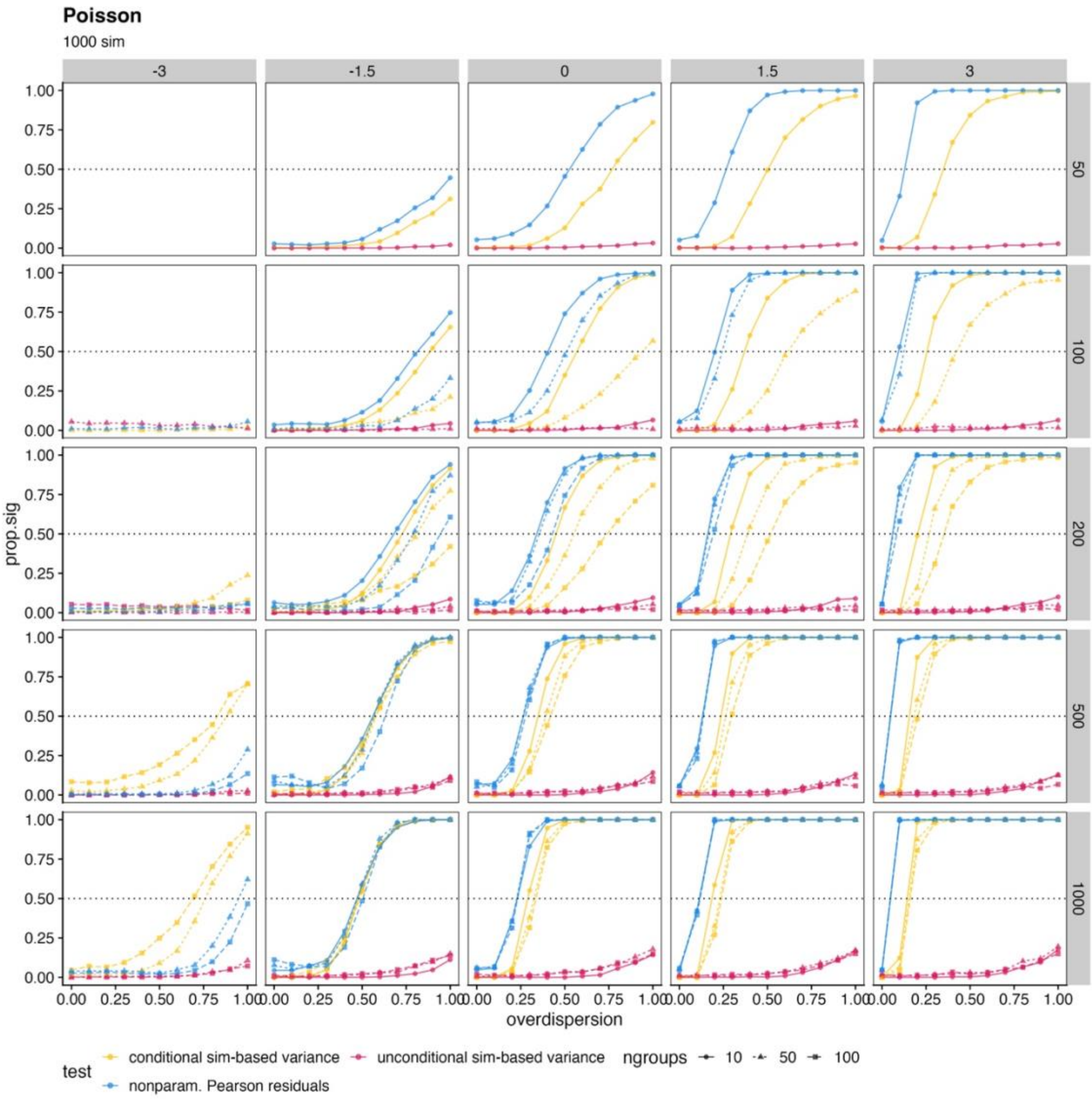


Figure S7.3. Power of the three alternative dispersion tests for the Poisson GLMMs, with different sample sizes (rows), intercepts (columns), and number of groups for the random intercept (line types). The missing lines for the first panel (intercept = -3 and sample size = 50 are due to simulation errors for some tests. For each parameter set, we ran 1000 simulations.

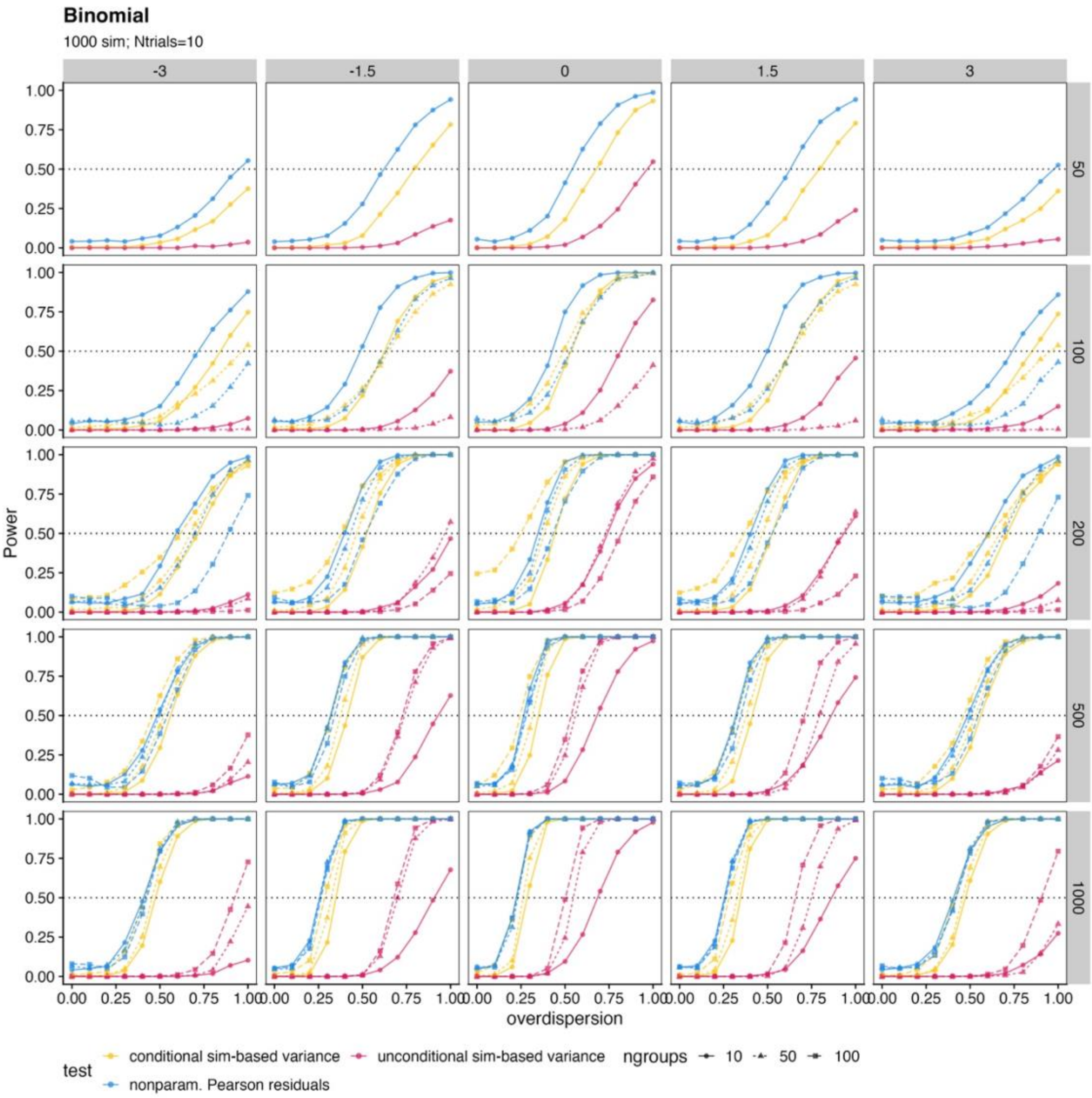


Figure S7.4. Power of the three alternative dispersion tests for binomial GLMMs, with different numbers of observations (rows), intercepts (columns), and number of groups for the random intercept (line types). 1000 simulations for each parameter set.

Dispersion statistics of the alternative dispersion tests

In Figures S7.5 and S7.6, we show the dispersion statistics 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.

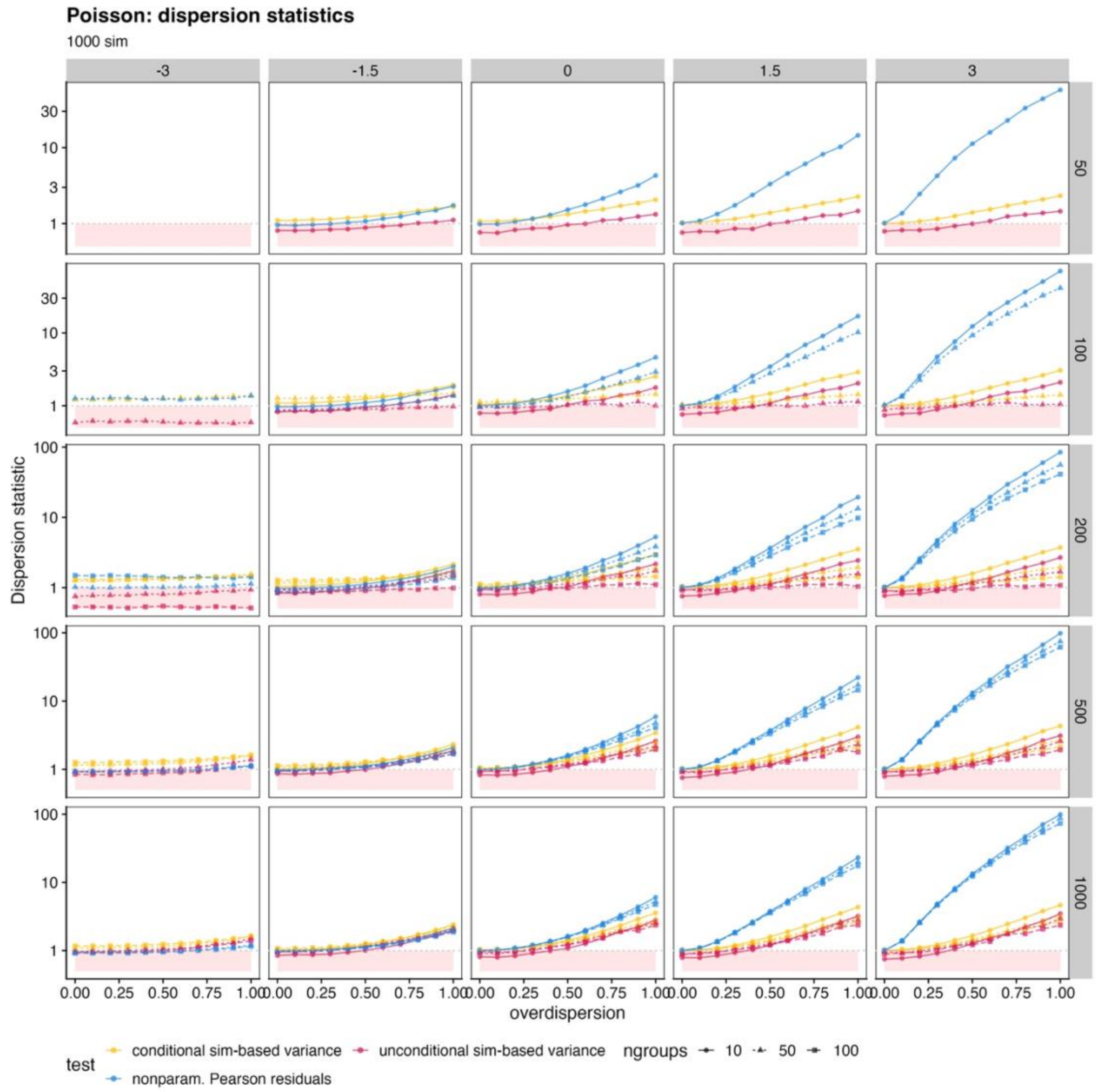


Figure S7.5. Dispersion statistics of the three alternative dispersion tests for the Poisson GLMMs, with different numbers of observations (rows), intercepts (columns) and number of groups for the random intercept (line types). The missing lines for the first panel (intercept = -3 and sample size = 50 are due to simulation errors for some tests. For each parameter set, we ran 1000 simulations.

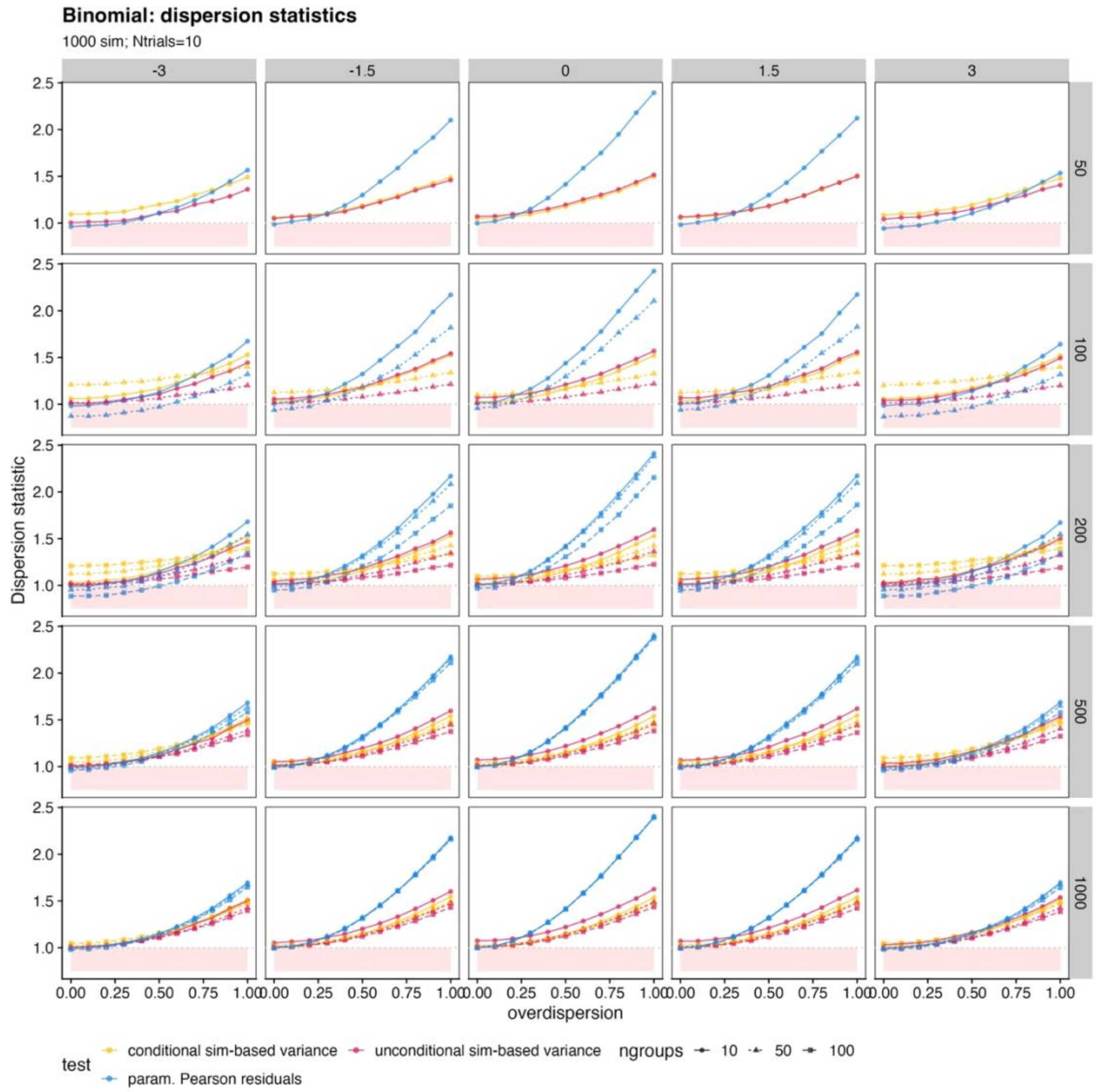
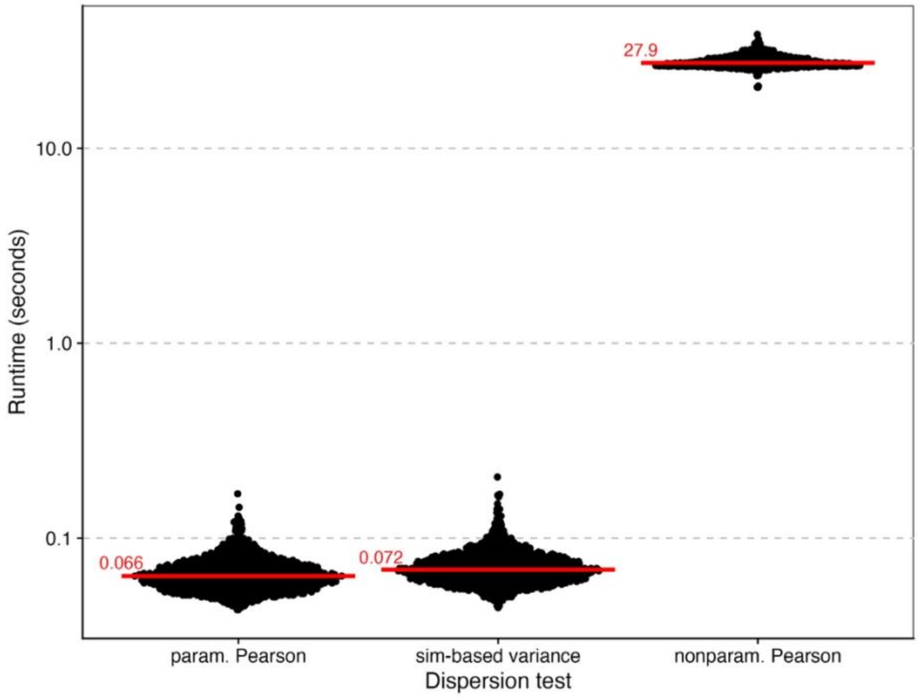


Figure S7.6. Dispersion statistics of the three alternative dispersion tests for binomial GLMMs, with different numbers of observations (rows), intercepts (columns), and number of groups for the random intercept (line types). 1000 simulations for each parameter set.

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 residual variance test with conditional simulations (Figure S7.7). 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 residual 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 S7.7. 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.

S8: Alternative simulation-based residual variance 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 S8.1 and S8.2). We evaluated and compared this test for Poisson and binomial GLMs and GLMMs (conditional simulations only), as we did for the other tests.

$$Approx. Pearson\ observed\ residuals:\ r_i = \frac{(y_i - \hat{\mu})}{var(y_{is})} \qquad \text{(Equation S8.1)}$$

$$Approx. Pearson\ simulated\ residuals:\ r_{is} = \frac{(y_{is} - \hat{\mu})}{var(y_{is})} \qquad \text{(Equation S8.2)}$$

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 S8.1 and S8.2). 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.

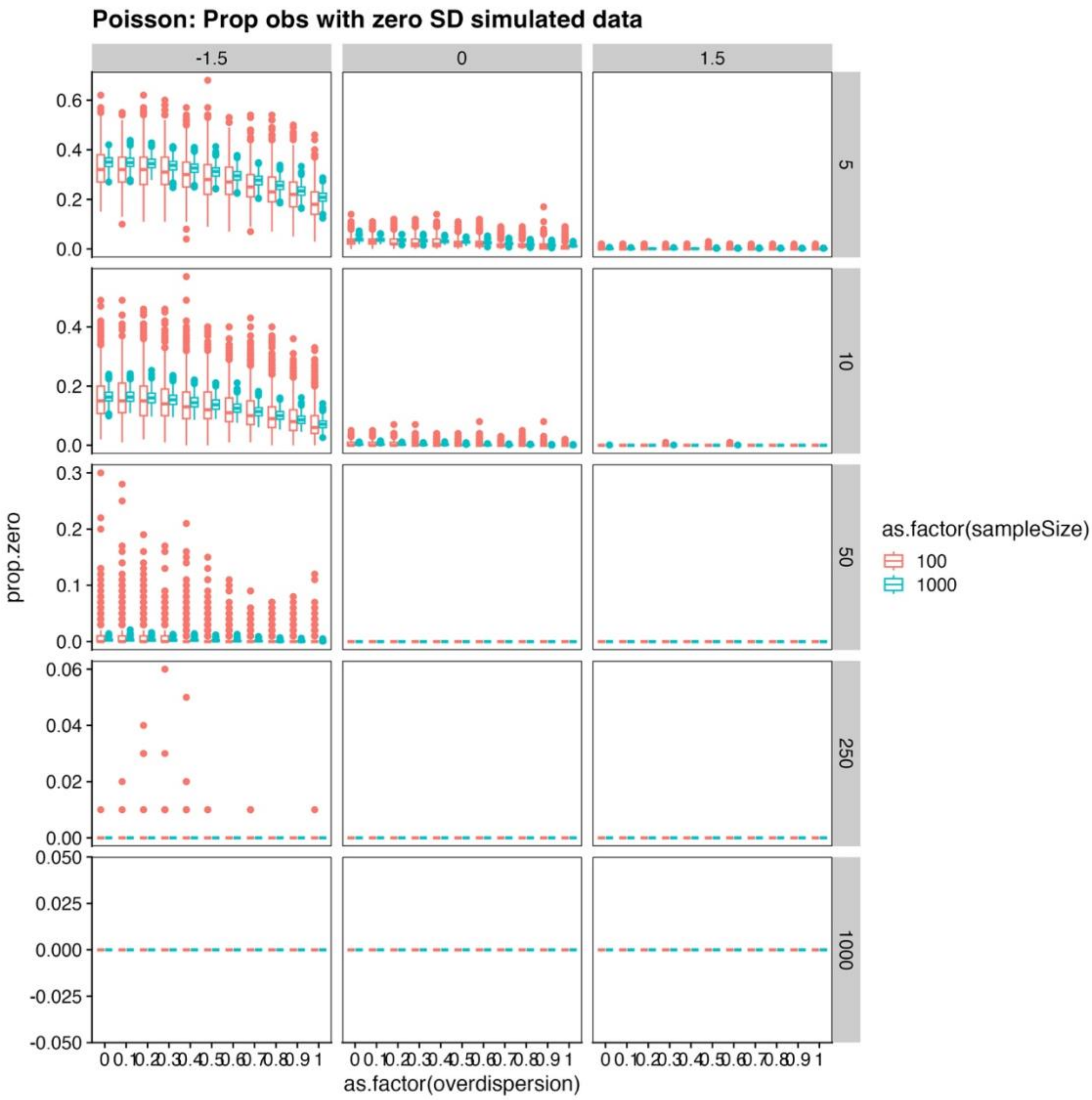


Figure S8.1. Poisson GLM: Proportion of observations with simulated zero variance in the dataset for different combinations of intercept (columns), number of simulations (rows) and sample sizes (colours).

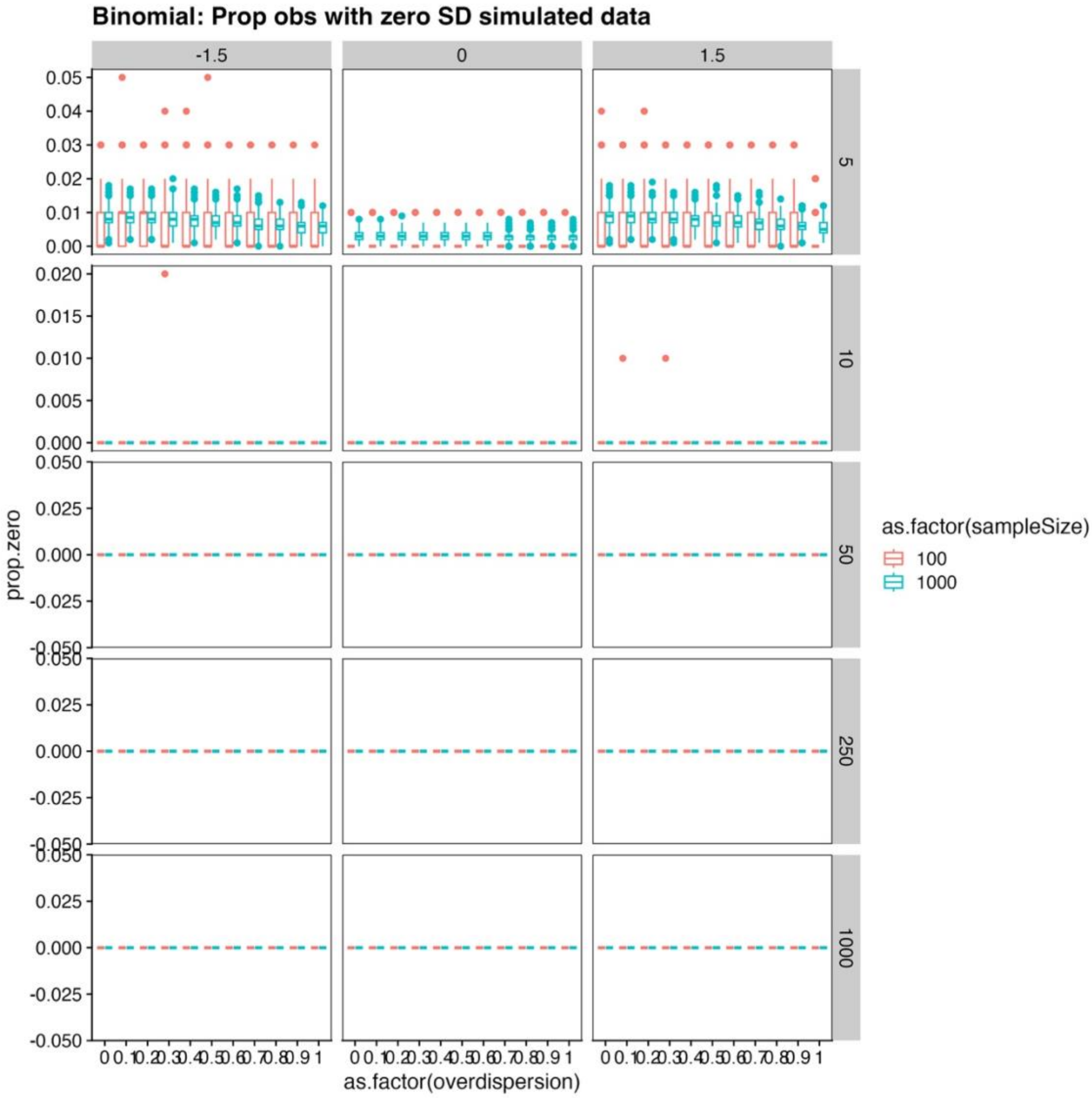


Figure S8.2. Binomial GLM: Proportion of observations with simulated zero variance in the data set for different combinations of intercept (columns), number of simulations (rows) and sample sizes (colours). The number of trials of the binomial was set to 10 in all simulations.

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 S8.3). 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.

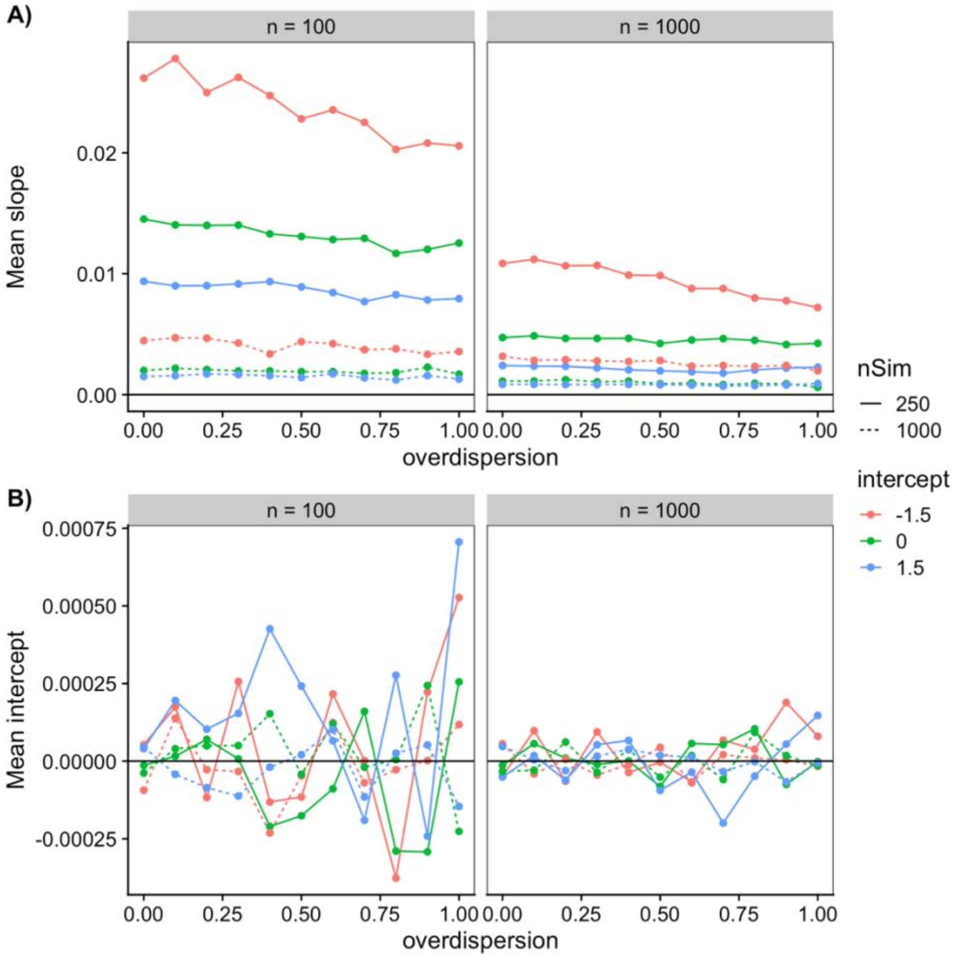


Figure S8.3. Mean slope (A) and intercept (B) of the regression of the difference between the Approximate Pearson residuals and Pearson residuals as response variable and the Pearson residuals as predictor for the Poisson GLMs.

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 residual variance test for the Poisson model. They tended to be conservative for small intercepts (Figure S8.4). 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 S8.4).

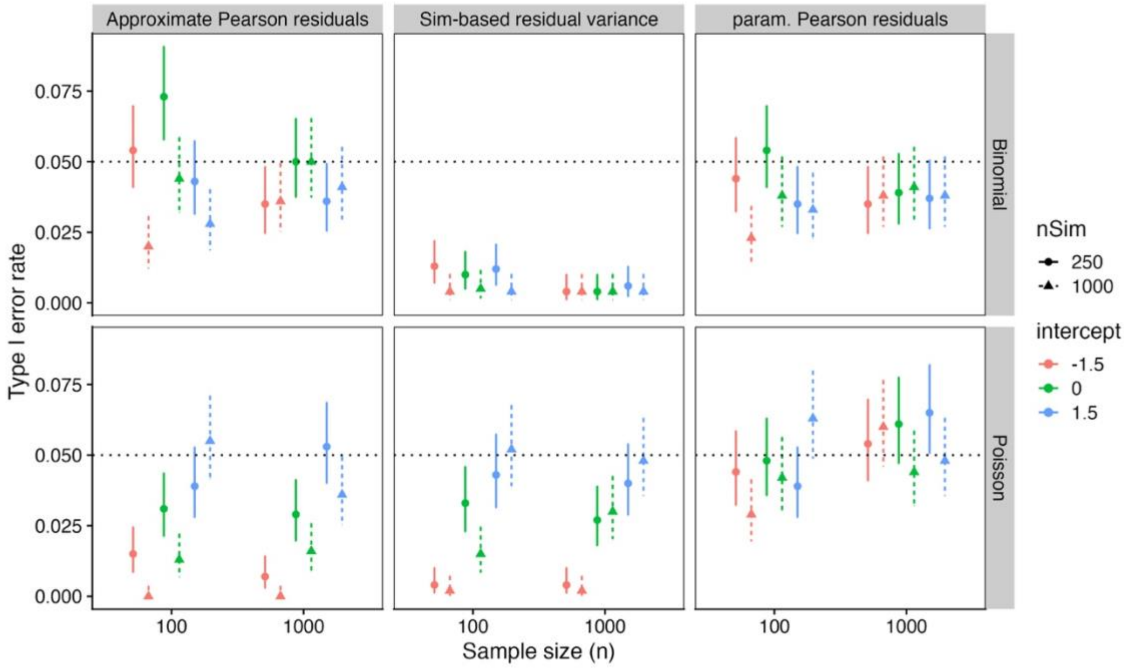


Figure S8.4. Type I error rates for GLMs comparing the parametric Pearson residuals tests, the simulation-based residual variance test and the simulation-based approximate Pearson test.

The dispersion statistics for the alternative simulation-based residual 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 S8.5). Power was very similar among the tests for the Poisson GLM (Figure S8.6). For binomial GLMs, the power of the alternative simulation-based residual test was high and similar to the parametric Pearson residuals test.

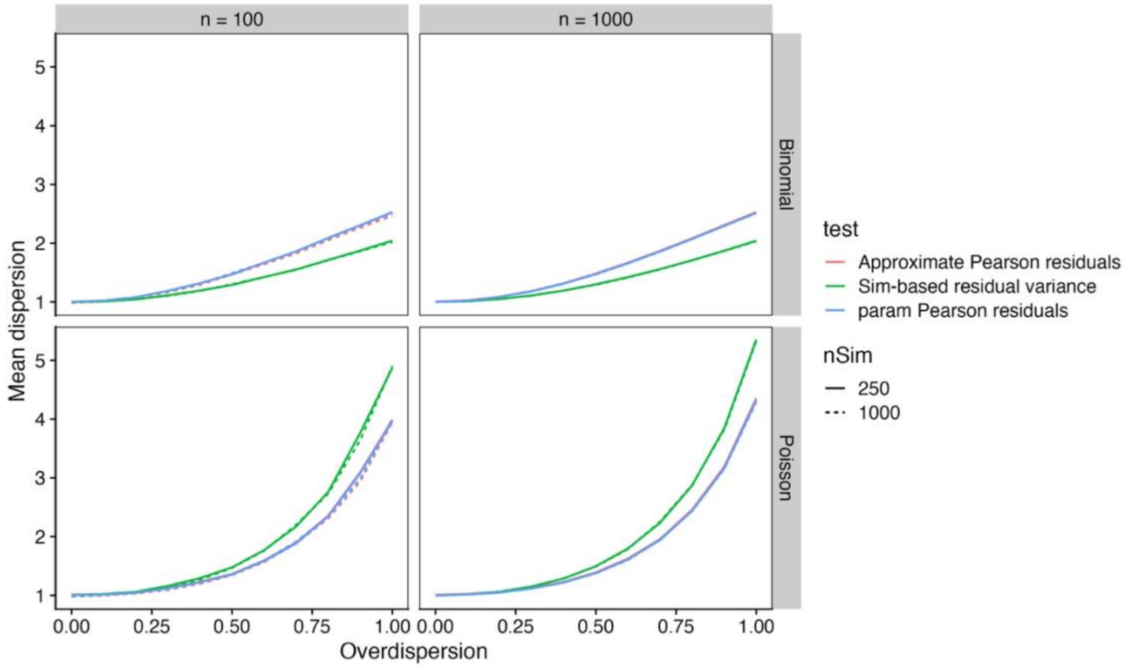


Figure S8.5. Dispersion statistics GLMs. Simulation set with intercept = 0.

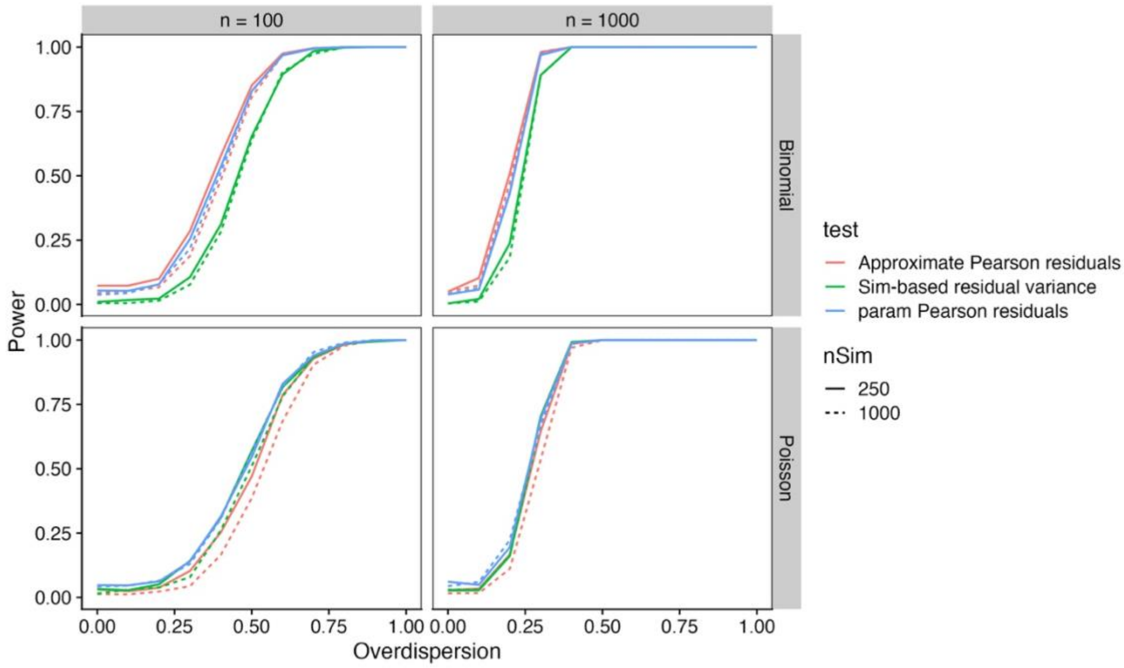


Figure S8.6. Power GLMs. Simulation set with intercept = 0.

For the GLMM simulations, we fixed the number of groups at 100 and the number of simulations at 250 to compare with the cases where the Pearson Chi-squared 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.

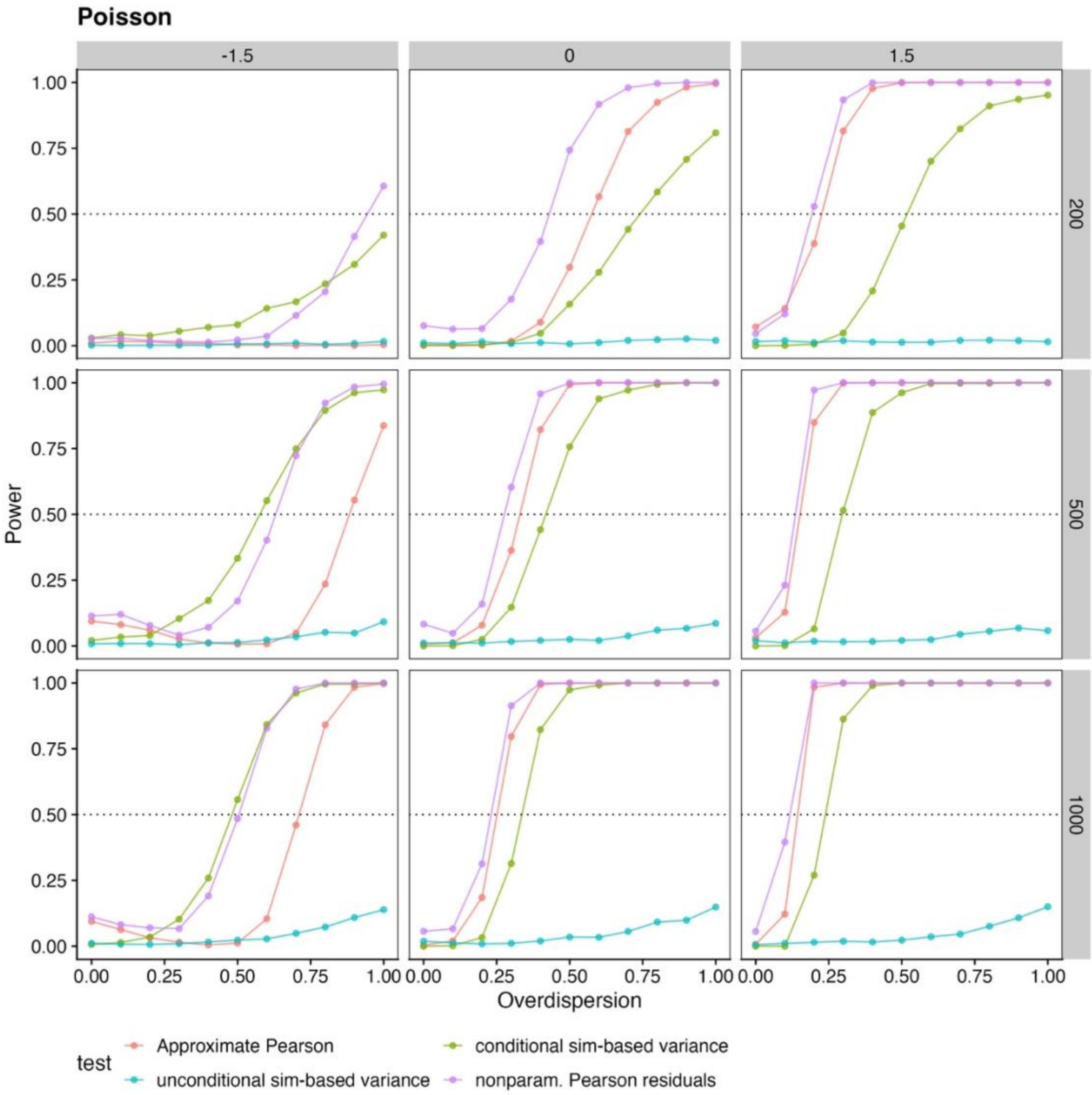


Fig S8.7. Power for Poisson GLMMs for the alternative simulation-based test using an approximation for Pearson residuals compared with the other tests assessed in the study. 1000 simulations for each parameter set: intercept (panel columns) and sample size (panel rows). The fixed parameters are slope = 1, number of groups = 100, and random effects variance = 1.

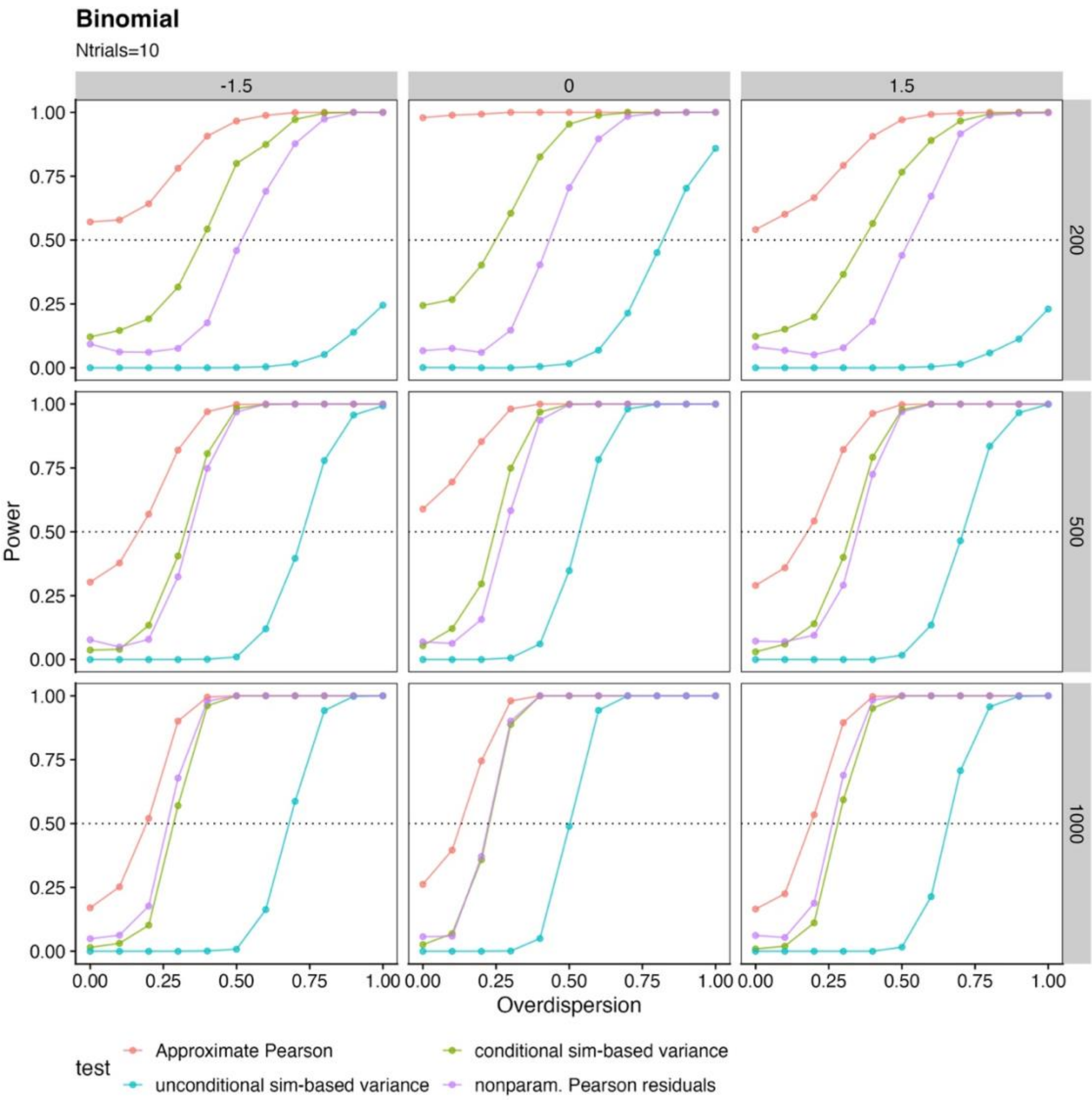


Fig S8.8. Power for binomial GLMMs for the alternative simulation-based test using an approximation for Pearson residuals compared with the other tests assessed in the study. 1000 simulations for each parameter set: intercept (panel columns) and sample size (panel rows). The fixed parameters are slope = 1, number of groups = 100, random effects variance = 1, number of trials = 10.

S9. 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 S9.1), which resulted also in very similar test results (e.g., Figure S9.2 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 residual variance test with conditional simulations). In Figure S9.3, 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 S9.4), especially for very small intercepts and sample sizes.

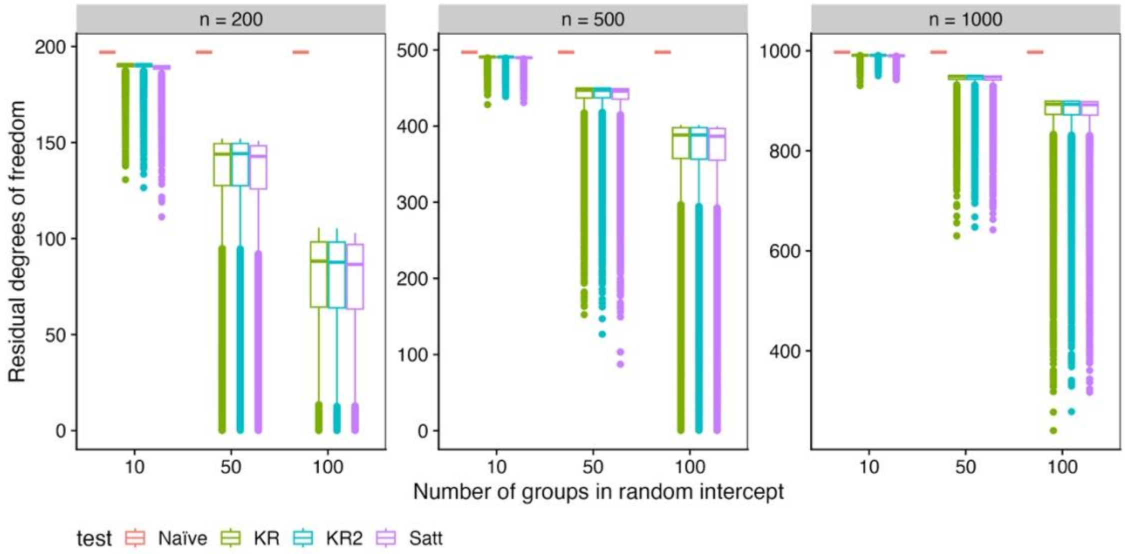
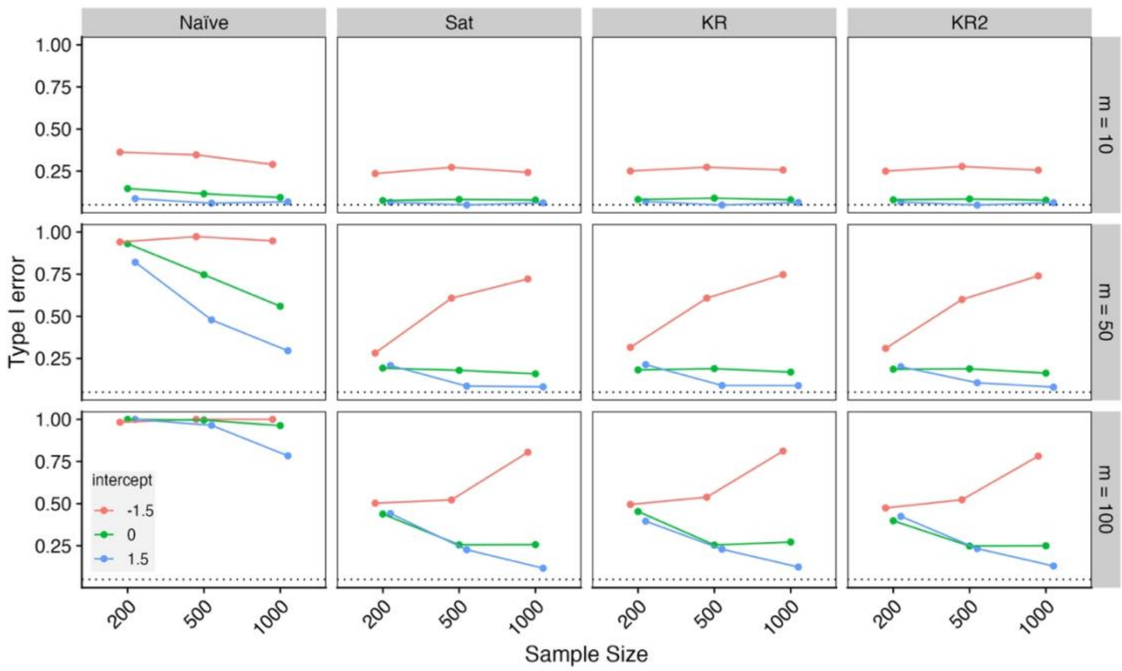


Figure S9.1. Residual degrees of freedom for the different correction methods for Poisson GLMMs with different numbers of groups in the random intercept (x-axis) and sample sizes (panel columns). Please refer to the main text above to relate to each applied correction. 1,000 simulations for each parameter setting, slope = 1, random intercept variance = 1.



Figures S9.2. 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.

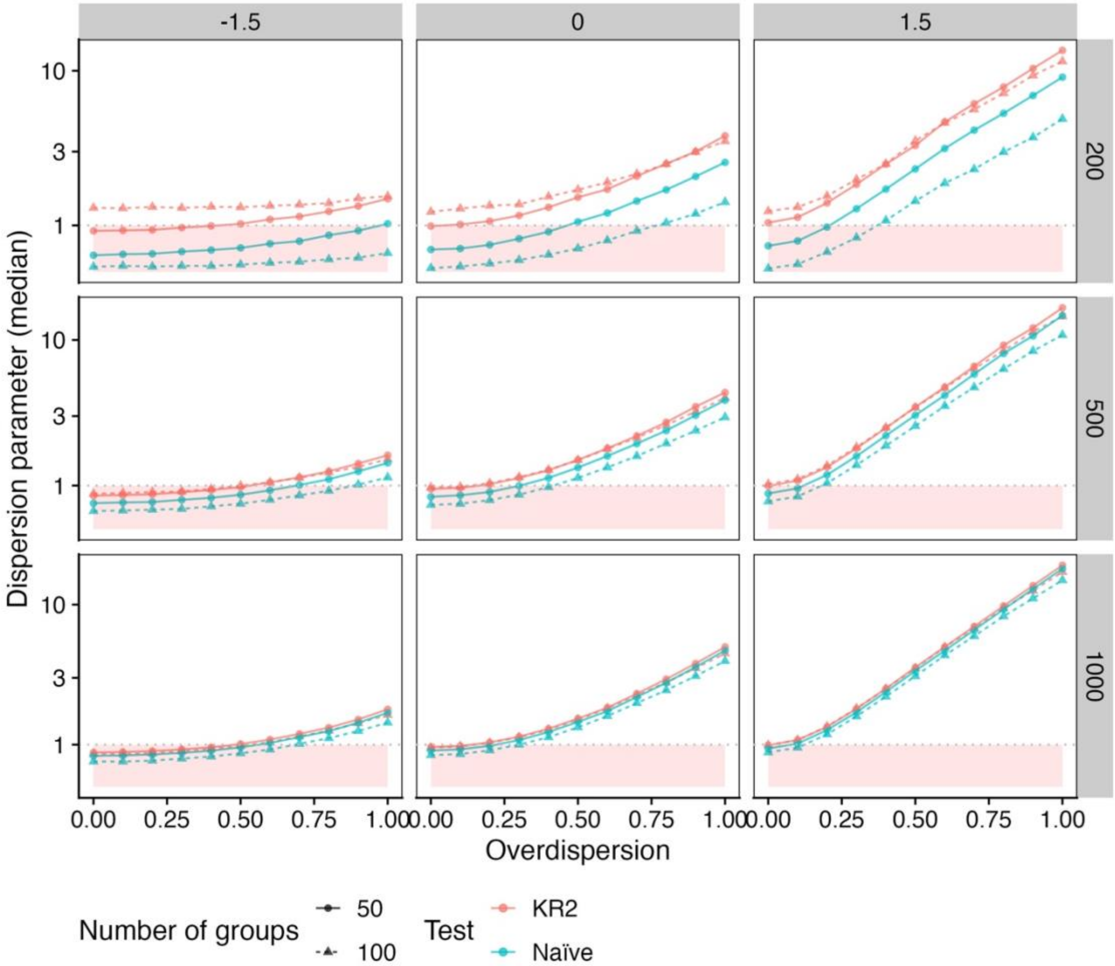


Figure S9.3. Dispersion parameters for the parametric Pearson test for Poisson GLMMs performed with different corrections for the residual degrees of freedom (colours), number of groups in the random intercept (linetype and shape), sample size (panel rows), and intercept (panel columns). Please refer 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.

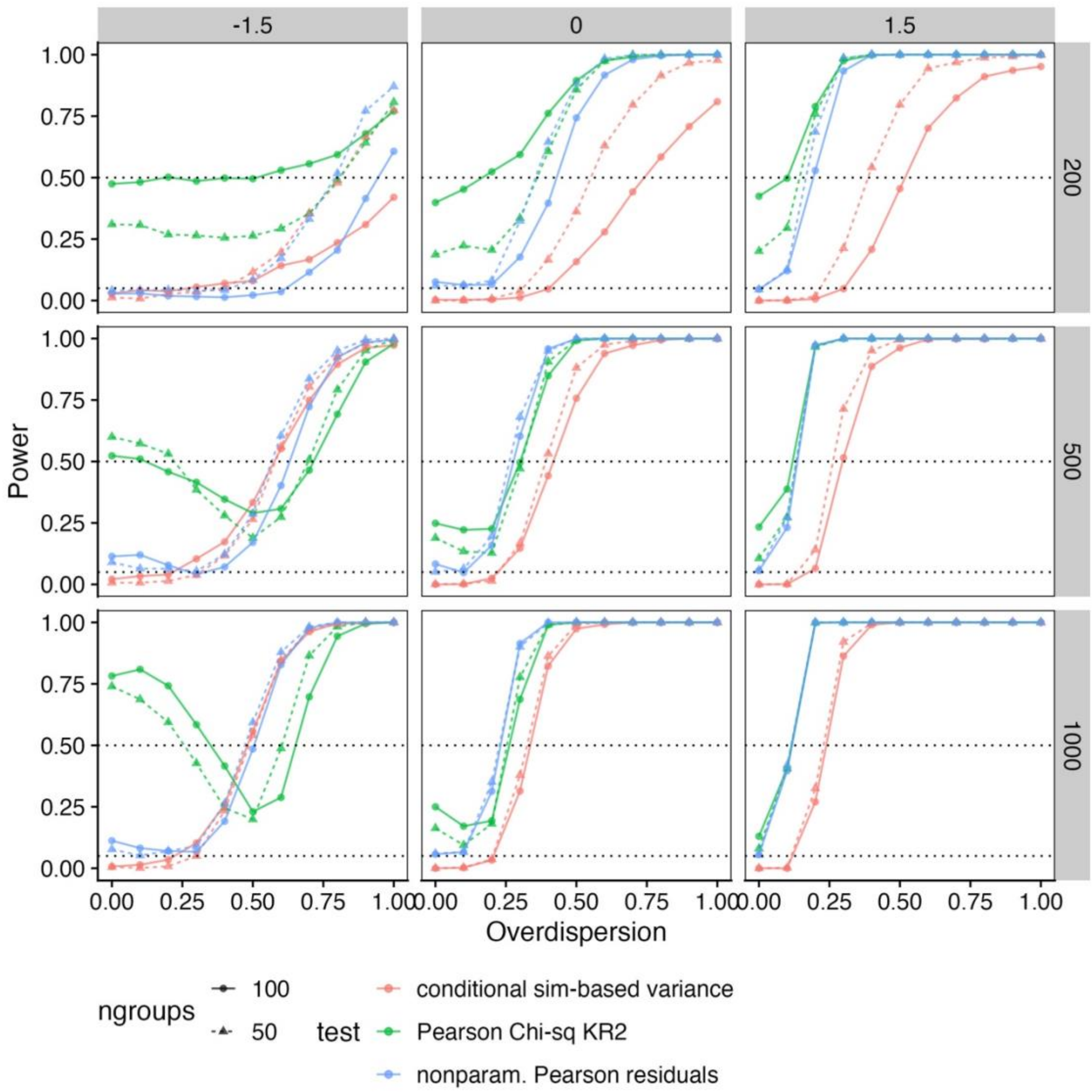


Figure S9.4. Power of dispersion tests for Poisson GLMMs (colours) performed with different numbers of groups in the random intercept (linetype and shape), sample size (panel rows), and intercept (panel columns). Please refer to the main text above to relate to the applied correction for residual degrees 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.

References

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

Jahn, N. (2023). *europepmc: R interface to the europe PubMed central restful web service* (Version 0.4.3) [Computer software]. https://CRAN.R-project.org/package=europepmc

Laumer, I. B., Kansal, S., Van Cauwenberghe, A., Rahmaeti, T., Setia, T. M., Mundry, R., Haun, D., & Schuppli, C. (2025). Wild and zoo-housed orangutans differ in how they explore objects. *Scientific Reports*, *15*(1), 14853. https://doi.org/10.1038/s41598-025-97926-z

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