Agricultural and Forest Entomology

Agricultural and Forest Entomology (2015), **17**, 164–172 DOI: 10.1111/afe.12092

The power of statistical tests using field trial count data of nontarget organisms in environmental risk assessment of genetically modified plants

Hilko van der Voet and Paul W. Goedhart

Biometris, Plant Research International, Wageningen University and Research Centre (WUR), PO Box 100, 6700 AC Wageningen, Netherlands

Abstract

1 Publications on power analyses for field trial count data comparing transgenic and conventional crops have reported widely varying requirements for the replication needed to obtain statistical tests with adequate power. These studies are critically reviewed and complemented with a new simulation study.

2 The reasons for the different reports are elucidated and can be classified as additional (but hidden) replication, selection of favourable endpoints with low variation, and reporting at an unusual scale.

3 A new simulation study was performed to investigate the relationship between statistical power and replication under a variety of data-generating and analysis methods. Approximately 60 replications should be sufficient to detect a 50% (two-fold) decrease in taxon numbers, provided that the coefficient of variation in the counts does not exceed 100%.

4 Replication can be accomplished not only by using multiple blocks in a single trial, but also by repeating the experiment in multiple years and/or at different sites. With other (e.g. agronomic) treatment factors in the field trial, without interaction with variety, the effective replication can be increased by investigating the main variety effect summed over the other treatment factors. Repeated measures may also increase the power if the expected difference is equal over time and the time points are sufficiently spaced.

Keywords Comparative assessment, data transformation, genetically modified organisms, overdispersed Poisson distribution, randomized blocks, repeated measurements, simulation, split-plot design.

Introduction

The guidance from the European Food Safety Authority (EFSA) for environmental risk assessment (ERA) of genetically modified plants recommends that applicants provide an analysis estimating the power for each difference test on each measurement endpoint, based on the stated effect size and assuming a 5% type I error rate, and that the analysis shall be performed at the planning stage of the study (EFSA, 2010). Recently, several studies aimed to provide more detailed statistical guidance for ERA field experiments for genetically modified organisms (GMOs). For example, Perry *et al.* (2009) stressed the need for a prospective power analysis based on treatment effect sizes of interest. The ideas with respect to prospective power analysis and choosing statistical models for counts and quantal data are expressed by Semenov *et al.* (2013) who provided several decision trees and a checklist to assist the interpretation of statistical analyses of field trials.

Goedhart *et al.* (2013, 2014) summarized statistical models that could be useful in the analysis of ERA field experiments. For count data, the Poisson distribution is the basic distribution, although it was noted that over-dispersion and/or excess zeroes imply the need for more advanced distributions, such as the overdispersed Poisson (OP), negative binomial (NB) or Poisson log-normal (PL) distribution. Excess zeroes could be handled assuming an additional spike of structural zero counts in addition to the other data (which may still contain incidental zero values). Such zero-inflation models can be analyzed directly (mixture models) or in a two-step procedure (hurdle models). Goedhart *et al.* (2014) provided a simulation tool to generate dummy field trial datasets based on any of these distributions. For the analysis of such data, these advanced models can also be used

Correspondence: Hilko van der Voet. Tel.: +31 317480811; fax: +31 317483554; e-mail: hilko.vandervoet@wur.nl

to estimate parameters such as the difference between the GMO and its comparator. Alternatively, despite all of the complex modelling options, a simple data transformation such as the logarithm followed by normal-theory modelling is also an option for analysis, and this is commonly used in practice (Goedhart *et al.*, 2014).

More specific guidance on sample size calculation for ERA field trials has been given by Perry *et al.* (2003), Duan *et al.* (2006), Prasifka *et al.* (2008) and Comas *et al.* (2013). These studies, however, arrive at very different conclusions. For a two-fold change in nontarget counts (i.e. +100% or −50%) Perry *et al.* (2003) conclude that 60 replicates provide a power of at least 85%, provided that mean count levels are larger than 5 and the coefficient of variation is smaller than 100%. By contrast, Duan *et al.* (2006) observed that combining data, as obtained from a 2-year field study employing a split-plot design with only four block replications, with the genetically modified plant and its comparator at the main plots and four insecticide treatment regimes at subplots provided at least 80% power to detect a 50% difference in 24 out of 28 comparisons. Similarly, for 80% power to detect a −50% change Prasifka *et al.* (2008) found that, on average, less than six replicates would be sufficient in their datasets. Comas *et al.* (2013) appear to need only three replicates to detect, with power 80%, impacts varying between 13% and 29% relative to the comparator's mean for field tests using yellow sticky traps. By contrast to these findings, the study by Perry *et al.* (2003) suggests that, in many cases, it will be very difficult to detect impacts of approximately 30% with sufficient power.

In the present study, we first describe the issues of interest in our critical review of previous work and the new simulation study. Next, the results of our critical review are presented, highlighting differences between the reported studies. Subsequently, this is complemented with results from the new simulation study, focussing on the robustness of different statistical models. Finally, the results from all of the studies are discussed.

Materials and methods

In a critical review of four previous studies (Perry *et al.*, 2003; Duan *et al.*, 2006; Prasifka *et al.*, 2008; Comas *et al.*, 2013), we describe how conclusions on the power for given designs and/or required sample sizes to obtain a given power were obtained. We consider the specific issues:

- Experimental design, including multiple treatment factors, split plots, multiple years and repeated measurements;
- assumptions on the statistical distribution of count data;
- mean count levels and variability used;
- method of statistical analysis for which the power analysis is performed; and
- whether or not a simulation study was performed.

Subsequently, we complement the findings of these previous studies with results from a new simulation study. Almost all of the power analysis results in the four reviewed papers are based on a normal approximation for transformed data, with different transformations $\log_e(C+1)$, $\log_{10}(C+1)$ or square-root, plus a power calculation based on the *t*- or noncentral *t*-distribution in relation to a linear model on the transformed scale. Whether such calculations are appropriate and sufficiently accurate for more realistic distributions of counts, such as the NB, can only be examined by simulation. Perry *et al.* (2003) performed such simulations using Taylor's power law (Taylor, 1961) for the variance. Unfortunately, only randomization tests based on such models were examined and not the power for standard linear or generalized linear models.

Table 1 'Assessing and monitoring the impacts of genetically modified plants on agro-ecosystems' (AMIGA) simulation study (see Supporting information, Doc. S1)

Abbreviation	Description	Used for simulation	Used for analysis
OP	Overdispersed Poisson	x	x
NB	Negative binomial	x	x
PL	Poisson log-normal	x	—
PO	Power model with $\beta = 1.5$	x	x
GM	Gamma	—	x
LN	$\log_e(C+1)$ transformation	—	x
SQ	Square-root(C) transformation	—	x

Models and transformations used in simulating and analysing count data. An explanation of the models is provided by Goedhart *et al.* (2014). In the present study, we report only the results for the NB simulation model and the OP, LN and SQ analysis models.

One of the aims of the European Union-funded project 'Assessing and monitoring the impacts of genetically modified plants on agro-ecosystems' (AMIGA) is to provide protocols on how to perform field studies, as well as on how to analyze data obtained from such studies. To address the lack of simulation studies for the power of tests on count data, a new study was set up with multiple models for data generation and multiple ways to analyze such data (see Supporting information, Doc. S1). A summary of the models used for data generation and data analysis is given in Table 1. The general idea was that we do not know the true distribution behind field trial count data, although it would be good to select a method of analysis that performs well under four different plausible models used for data generation, as described by Goedhart *et al.* (2014). These models are the OP, the NB, the PL and the Power model (PO). It may be noted that the PO only specifies a variance-to-mean relationship and not a distribution. In the simulations, a negative binomial distribution was used with the same variance function as the specified Power function (Goedhart *et al.*, 2014).

Simulated data were analyzed by fitting six different models (Table 1). All models were fitted using standard facilities in the statistical package GENSTAT (VSN International, 2012). Tests for transformed data [log-normal (LN) and square-root normal (SQ) methods] were *t*-tests. Tests for untransformed data [OP, NB, PO and gamma (GM) methods] were based on fitting generalized linear models (GLMs; McCullagh & Nelder, 1989) with a log link function, and were likelihood ratio tests (NB) or Pearson-scaled likelihood ratio tests (OP, PO, GM). The dispersion parameter of the negative binomial distribution was bounded to the interval (0.001, 1000) to avoid numerical problems. Because the gamma distribution can not handle zero observations, zeroes were replaced by 0.001. For extreme cases of simulated data, it was sometimes not possible to fit a model

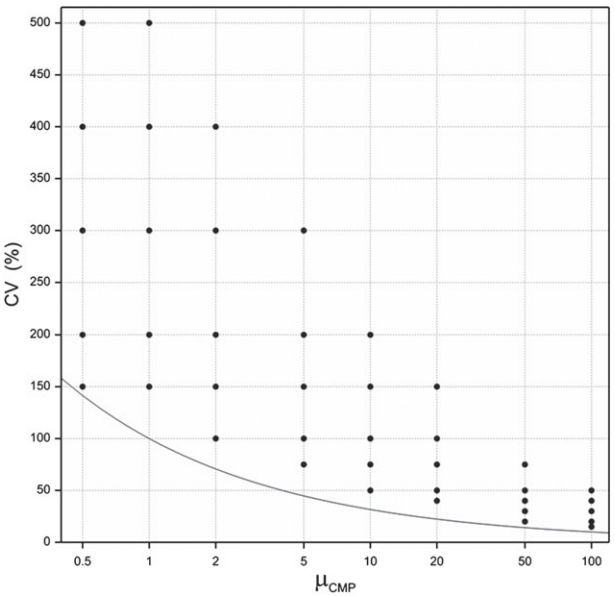

Figure 1 The line represents pure Poisson variation, expressed as coefficient of variation (*CV*), as a function of mean count level μ_{CMP} ($CV = \mu_{CMP}^{-0.5}$). Dots represent chosen combinations of comparator mean and *CV* in the 'Assessing and monitoring the impacts of genetically modified plants on agro-ecosystems' (AMIGA) simulation study.

and so adaptations were made, For example, with simulated datasets with only zero counts (as obtained when simulating at low count levels), there is no indication of a difference and, in these cases, the test *P* value was set to 1.

In the present study, we focus on the comparison with previous literature, and therefore single out the results using the NB model for data generation (also used in a slighly different way by Perry *et al.*, 2003), as well as three methods of analysis: (i) the OP model (McCullagh & Nelder, 1989), which was suggested but not investigated for power by Perry *et al.* (2003); (ii) the LN model used by Perry *et al.* (2003), Prasifka *et al.* (2008) and Comas *et al.* (2013); and (iii) the SQ model used by Duan *et al.* (2006).

The power of tests for the difference between the GMO and its comparator (CMP) was investigated in a completely randomized design with the two varieties as treatments. The number of replications was set to 4, 6, 8, 10, 15, 20, 30, 40, 60 or 100. The CMP means were varied; in the present study, we report results for CMP mean counts of 2, 5 and 20. For each mean count level, a set of five values for the coefficient of variation (CV) was investigated. The CV values chosen were based on results in the literature, most notably the graphs in Duan *et al.* (2006). It may be noted that pure Poisson variation puts a lower bound on CV values that are realistic at any specific mean count level (Fig. 1). The CV value was used to define the dispersion parameter k of the negative binomial distribution by solving $CV = \sqrt{\mu_{CMP} + k\mu_{CMP}^2}/\mu_{CMP}$. The same dispersion parameter was used for the GMO. This implies that, in case μ_{GMO} is smaller than μ_{CMP}, the GMO has a larger CV than the comparator, which fits the expectation based on Fig. 1.

Simulations were made under the null hypothesis of no difference and for ratios Q of GMO to CMP means of 0.25 (75% or four-fold decrease), 0.5 (50% or two-fold decrease) and 0.75 (25% or 1.33-fold decrease). The nominal significance level was 0.05 in a two-sided test. The number of simulated datasets was 1000 for each simulation setting.

Critical review of previous studies

Perry et al. (2003)

One of the most extensive studies on power analysis of GMO field trials to date has been performed as part of a large 5-year trial in the U.K., termed the Farm-Scale Evaluations (FSE). The power analyses considered a comparison of two varieties (genetically modified herbicide-tolerant versus conventional crop) in a randomized block design with two plots (half-fields) per block (field). Two models were considered for generating the count data *C*.

The first data generating model assumed a normal distribution for transformed counts $\log_e(C+1)$, which was termed the log-normal model. The power of a two-sided paired *t*-test at significance level $\alpha = 0.05$ for $n = 20$, 30, 40, 60 or 90 blocks for detecting 1.3-fold, 1.5-fold and two-fold differences was evaluated using the standard calculation based on the noncentral *t*-distribution

$$Power = T_{n-1}\left(-t_{n-1,\alpha/2}\,;\lambda\right) + 1 - T_{n-1}\left(t_{n-1,\alpha/2}\,;\lambda\right) \quad (1)$$

where $T_{n-1}(\cdot)$is the noncentral *t*-cumulative distribution function with $n-1$ degrees of freedom, $t_{n-1,\,\alpha/2}$ is the critical value of the t-test, and $\lambda = \delta/sed$ is the noncentrality parameter, with δ equal to $\log_e(1.3)$, $\log_e(1.5)$ or $\log_e(2)$, respectively, and *sed* is the standard error of the difference of the means of the transformed counts. The latter quantity was calculated as $sed = \sqrt{(2/n)\log_e(1+CV^2)}$, for coefficients of variation of the counts (*CV*) equal to 50%, 80% or 100%.

In the second data-generating model, a negative binomial distribution was assumed for the counts *C*. The shape parameter of the negative binomial distribution was chosen such that the relationship between the variance V and the mean count level μ conformed to Taylor's power law $V = \alpha\mu^{\beta}$ (Taylor, 1961) for β equal to 1, 1.5 or 2, respectively. The dispersion (shape) parameter was calculated separately for the GMO and the CMP, which implies that for $\beta = 2$ the GMO and CMP have the same coefficient of variation irrespective of their mean count levels. By contrast, in the AMIGA study we used a single dispersion parameter per CV for all mean count levels in a simulation.

The power of three nonparametric randomization tests, corresponding to three statistical models, was calculated. The nonparametric testing methods were Monte Carlo paired randomization tests on three statistics reflecting the three values of β in the variance to mean relationship as given above. The power of the nonparametric tests was evaluated for 180 scenarios obtained by combining all previously mentioned levels of n (20, 30, 40, 60 or 90) and the relative difference (1.3-, 1.5- or two-fold) with 12 specifically chosen combinations of levels of β (1, 1.5 or 2), CV (50%, 80% or 100%) and the mean count level (over the two varieties in the trial) μ (1, 5, 10 or 50). Field effects were added such that mean counts per field were between

10 times lower or 10 times higher compared with the general mean μ.

For analysis of the count data, three parametric methods were mentioned, although the power of these parametric methods was not evaluated in the simulation study. For $\beta = 1$ (variance proportional to the mean), a GLM with Poisson errors and an estimated scale parameter (the OP model) was suggested as an appropriate parametric method of data analysis. For $\beta = 1.5$, a parametric GLM analysis with user-defined error and log-link was suggested but not elaborated. For $\beta = 2$ (constant coefficient of variation), reference was made to the log-normal model as providing an efficient parametric analysis.

The results of the power analysis were summarized as follows: a replication of 60 plots per variety should provide adequate power (at least 80%) to detect 1.5-fold multiplicative differences as long as CV does not exceed 50% and the mean abundance μ exceeds 5. It was suggested that this fairly large replication could be obtained, as in the FSE, by dividing the replications over multiple years (e.g. 20 fields per year over 3 years). Based on the log-normal model, the power was also adequate (87%) for a CV of 80% and almost adequate (75%) for a CV of 100%, although this was not true for most of the scenarios involving the negative binomial simulation model and the randomization tests. It was not clear whether the lower power observed for the randomization tests was a result of the change from a parametric to nonparametric analysis or the negative binomial rather than log-normal data generation. In a small simulation study, we found that the power differences between the parametric and nonparametric tests were minimal with log-normal data. This suggests that a log-normal approximation may over-estimate the power if, in reality, the data are generated by a negative binomial distribution using Taylor's power law to define the over-dispersion. Note, that this is also true in the case $\beta = 2$, when the variance function is the same as for the log-normal model, although the forms of the distributions are still different and the necessary addition of 1 to the counts in the log-normal model leads to distortions. On the other hand, a later statistical assessment of the FSE data found inferences to appear robust to misspecification of the power β (Clark *et al.*, 2006) and so it might not be necessary to have very precise estimates.

Duan et al. (2006)

A retrospective, or post-hoc, power calculation was performed using data from a 2-year field study to evaluate the effects of MON 863 corn compared with a conventional corn line on arthropods. Counts of 28 taxa were reported and analyzed for three trap types (eight arthropod groups in pan traps, eight groups in pitfall traps and 12 groups in sticky traps), and for each year separately, as well as the 2 years combined. The experimental design was a split-plot design with 2 years, four blocks per year, two main plots per block and four subplots per main plot. The two corn varieties were assigned to the main plots and four insecticide treatment regimes to the subplots. Count data were obtained on three to five time points per growing season. The sample sizes per endpoint for the combined 2-year data varied between 128 and 286 subplots (2 years × 4 blocks × 2 main plots × 4 subplots × 3–5 time points per year, and omitting many zero or low results that showed little between-plot variation at each sampling time). The replication for each of the two varieties therefore varied between 64 and 143 in the 2-year experiment.

The model for statistical analysis was an additive mixed model for square-root transformed counts, assuming no correlation between subsequent time points. The means and the pooled standard error of the mean (*sem*) for the two varieties (all on the square-root scale) were estimated from the available data using SAS procedure LSMEANS. The standard error of the difference of means (*sed*) was obtained as $\sqrt{2}sem$. The recalculation of differences on the count scale to differences on the square-root scale also required estimates of the residual standard deviation, which was obtained as $\sqrt{N/2}\ sem$, where N was the total number of data points (excluding all data from time points without any count for any treatment). The number of degrees of freedom of the residual error was obtained as the denominator degrees of freedom for the test of the variety difference in the mixed model analysis. The Satterthwaite approximation method available in SAS procedure MIXED was used to account for unequal variances. Calculated degrees of freedom for the 2-year experiment varied strongly, at between 5.25 (for *Coleomegilla maculata* in sticky traps) and 160 (for *Harmonia axyridis* in sticky traps). Such differences can be explained if different variance components are important in the fitted models for different endpoints, although these variance components were not reported. For example, if the main plot variance and the between-times variance can be ignored relative to the subplot variance, then the four subplots per plot and the multiple time points are effectively replicates. Alternatively, if the main plot variance is the major source of variation, the effective number of replicates is just the number of main plots (eight over the 2 years).

For the power analysis, it was assumed that there were no effects of insecticide regime (no significant interactions between variety and insecticide regime were found in the data analysis). The power was calculated using the noncentral t-distribution (Eqn. 1). Power was reported for a −50% (two-fold decrease) or +50% (1.5-fold increase) relative difference, whichever was lower. These effect sizes were recalculated to differences at the square-root scale, which involved finding the root of a quadratic equation for each of the two cases, and therefore four solutions in total. The smallest absolute value among the four solutions of the two quadratic equations was used as critical effect δ to set the noncentrality parameter for the noncentral t-distribution $\lambda = \delta/sed$, using the degrees of freedom as calculated.

The results reported by Duan *et al.* (2006) can be summarized. The retrospective power for the specific design and analysis was often insufficient to obtain at least 80% power in each of the two single years: among the 28 endpoints, only 12 had at least 80% power in the year 2000 trial and only 10 in the year 2001 trial. By contrast, the combined experiment over 2 years had at least 80% power in 24 of the 28 endpoints. The four failing cases were related to low mean counts (< 3) for the control variety in combination with a high coefficient of variation (CV > 100%) and a relatively low replication (twice 64 and twice 96).

Duan *et al.* (2006) presented very useful graphs showing CV against mean abundance and its potential effect on the power of the difference test. These plots show that high CVs (>100%) occur almost exclusively for low mean abundances (e.g. below 5).

Prasifka et al. (2008)

Guidance for future studies was derived in a power analysis based on five studies (on five different locations in the U.S. corn belt) of possible GMO nontarget effects on arthropods. One of the five studies was the same as the one used by Duan *et al.* (2006) but now included an additional year. Each study was conducted over 3 years with two to four replicates per year, and count data were obtained from one to 11 subsamples per plot, in time series of one to 17 sampling times per year. Counts on a large number of taxons were obtained using three to five sampling methods per location.

The basic data (for the negative control treatments only) were averaged over the subsamples, separately for each combination of taxon, sampling method, location and sampling date, and then means and CVs were calculated. In accordance with the study by Duan *et al.* (2006) who found high statistical power for taxa with CV < 100%, the selection of endpoints (taxon × sampling method combinations) to be used for the power analysis was restricted to those with lower CVs. To be included, endpoints were required: (i) to have been sampled at minimally two locations and (ii) to have, in at least two thirds of the location × year combinations, an observed CV less than 100% for at least two consecutive sampling periods. By applying these criteria, approximately 80% of the candidate endpoints were discarded.

Statistical power was calculated using PASS software (NCSS, 2002), without further specification of the statistical method other than the assumption of the use of a repeated-measures analysis of variance for testing the difference between the GMO and comparator variety. As input for the power calculations, the original counts C were transformed to $\log_e(C+1)$ for the comparator and to $\log_e(0.8C+1)$, $\log_e(0.7C+1)$ or $\log_e(0.5C+1)$ for the GMO, mimicking effect sizes of a 20%, 30% and 50% decrease. The comparator mean and the residual mean square on the transformed scale were obtained from an analysis of variance, accounting for block effects.

Prasifka *et al.* (2008) summarized their results as graphs of power against number of replicates for 12 endpoints (three taxa for each of four ecological roles). One of their conclusions is that some endpoints can be sampled with adequate power to detect large (−50%) changes with only three or four replicates. Two observations should be made. The estimated powers are medians in the sets of six to 15 location × year combinations. Furthermore, each power estimate has different numbers of subsamples and time points underlying each plot. Therefore, the estimate of three or four replicates required for adequate power only represents designs with a similar substructure as was used in the underlying data.

Comas et al. (2013)

Although Comas *et al.* (2013) provide the statistically most explicit paper, the same statistical method was also used in other papers from this group (Albajes *et al.*, 2012, 2013). Comas *et al.* (2013) based their power analysis on data from 20 single-year field trials. All trials were randomized complete block designs with three or four blocks. There were two to 10 treatments, always including the transgenic (GMO) versus near-isogenic (comparator) treatments that were of interest. There were four to eight sampling dates on which 26 arthropod taxa in four functional groups were recorded, using visual counts (seven taxa), pitfall traps (six taxa) or yellow sticky traps (13 taxa). Counts were summed over the sampling dates and subsequently transformed using $y = \log_{10}(C+1)$ to provide transformed counts at the plot level. These were analyzed using analysis of variance accounting for block and treatment effects.

The desired power was fixed at 80% (using a significance level of 5%) and the detectable treatment effect was calculated as a function of sampling method, comparator abundance, and numbers of years, blocks and treatments in the field trial. The detectable treatment effect d_c was expressed on a relative scale for log-transformed data. This is a scale that is difficult to understand. Most work on log-normally distributed variables considers relative effects on the original scale or absolute effects on the transformed scale but not relative effects on the transformed scale. The estimated residual standard error s (also on the log scale) was also expressed as a relative standard deviation $s_c = s/m_c$, where m_c is the overall mean of the log-transformed counts y. Then, for each of the three sampling methods, a power curve $s_c = \theta m_c^{-k}$ was fitted to the available data to predict s_c from m_c. Table 2 of Comas *et al.* (2013) shows, for a given comparator abundance (density or catches), the relative detectable treatment effect calculated as $d_c = s_c SF$, where the statistical factor $SF = \sqrt{2/N}\left(t_{df,\alpha/2} + t_{df,\beta}\right)$ captures the design properties for N block-year combinations, T treatments, and d.f. $= (T-1)(N-1)$ degrees of freedom. It should be noted that, in these results, the visual counts (densities) and trap catches are on the original scale, although the derived d_c values are relative values for the means M on the $\log_{10}(C+1)$ scale. Such values may be difficult to interpret but can be transformed to count ratios (folds) as described below. For a positive deviation of the GMO relative to the comparator, $d_c = (M_{\mathrm{GMO}} - M_{\mathrm{CMP}})/M_{\mathrm{CMP}} = \log_{10}(C_{GMO}+1)/\log_{10}(C_{CMP}+1) - 1$; for a negative deviation, it is $d_c = (M_{\mathrm{CMP}} - M_{\mathrm{GMO}})/M_{\mathrm{CMP}} = 1 - \log_{10}(C_{GMO}+1)/\log_{10}(C_{CMP}+1)$. These equations can be rewritten as a ratio of the two expected counts (or 'fold'): $Fold\,(increase) = \left\{ bcf\left(\mu_{\mathrm{CMP}}+1\right)^{1+d_c} - 1 \right\}/\mu_{\mathrm{CMP}}$ [and similarly for *Fold(decrease)*]. The bias correction factor (*bcf*) equals $10^{b_{GMO}-b_{\mathrm{CMP}}}$, where b_{GMO} and b_{CMP} are the biases of $\log_{10}(\mu_{GMO}+1)$ and $\log_{10}(\mu_{CMP}+1)$ as estimators of the means on the transformed scale, M_{GMO} and M_{CMP}, respectively. It was verified that, at least in the considered cases, the two terms are almost equal, and therefore $bcf \approx 1$. Omitting the bias correction factor we can rewrite the expressions for d_c as a ratio of the two expected counts (or 'fold'):

$$Fold\,(increase) = \frac{\mu_{GMO}}{\mu_{\mathrm{CMP}}} = \frac{\left(\mu_{\mathrm{CMP}}+1\right)^{1+d_c} - 1}{\mu_{\mathrm{CMP}}} \qquad (2)$$

$$Fold\,(decrease) = \frac{\mu_{\mathrm{CMP}}}{\mu_{GMO}} = \frac{\mu_{\mathrm{CMP}}}{\left(\mu_{\mathrm{CMP}}+1\right)^{1-d_c} - 1} \qquad (3)$$

For example, Comas *et al.* (2013), in their table 2, report a detectable treatment effect d_c of 0.29 for an experiment conducted in 1 year, with three blocks and two treatments with 35 yellow sticky trap catches. They state that, for example, a value

Table 2 Expected *Fold(decrease)* values of the capacity of field tests to detect negative effects of a genetically modified crop on nontarget organisms for different numbers of blocks (3–6) and treatments (2–6) when the test is conducted for a different number of years (1–3)[a]

		Visual counts				Pitfall traps				Yellow sticky traps			
			Number of years				Number of years				Number of years		
Number of blocks	Number of treatments	Density	1	2	3	Catches	1	2	3	Catches	1	2	3
3	2	1	3.38	1.57	1.38	1.5	∞[b]	2.35	1.83	35	3.00	1.65	1.45
3	2	1.5	2.83	1.51	1.35	3	7.88	2.07	1.69	70	3.07	1.67	1.47
3	2	3	2.46	1.46	1.32	9	4.98	1.97	1.65	140	3.16	1.70	1.48
3	4	1	1.91	1.46	1.34	1.5	3.69	2.03	1.72	35	1.97	1.53	1.40
3	4	1.5	1.79	1.42	1.31	3	2.84	1.84	1.61	70	2.01	1.55	1.42
3	4	3	1.70	1.38	1.29	9	2.56	1.78	1.58	140	2.05	1.57	1.43
3	6	1	1.80	1.44	1.34	1.5	3.21	1.99	1.70	35	1.88	1.52	1.40
3	6	1.5	1.70	1.40	1.31	3	2.59	1.81	1.60	70	1.91	1.53	1.41
3	6	3	1.63	1.37	1.29	9	2.37	1.75	1.56	140	1.94	1.55	1.43
4	2	1	2.02	1.43	1.31	1.5	4.32	1.94	1.63	35	2.08	1.50	1.36
4	2	1.5	1.88	1.39	1.28	3	3.13	1.77	1.54	70	2.12	1.51	1.38
4	2	3	1.77	1.36	1.26	9	2.76	1.72	1.51	140	2.16	1.53	1.39
4	4	1	1.65	1.37	1.28	1.5	2.65	1.79	1.58	35	1.74	1.44	1.34
4	4	1.5	1.58	1.34	1.26	3	2.26	1.67	1.50	70	1.76	1.45	1.35
4	4	3	1.53	1.31	1.24	9	2.12	1.63	1.47	140	1.79	1.47	1.36
4	6	1	1.61	1.36	1.28	1.5	2.49	1.77	1.57	35	1.69	1.43	1.33
4	6	1.5	1.54	1.33	1.26	3	2.16	1.65	1.49	70	1.71	1.44	1.34
4	6	3	1.49	1.31	1.24	9	2.04	1.61	1.47	140	1.74	1.46	1.36
6	2	1	1.57	1.31	1.23	1.5	2.35	1.63	1.46	35	1.65	1.36	1.28
6	2	1.5	1.51	1.28	1.21	3	2.07	1.54	1.40	70	1.67	1.38	1.29
6	2	3	1.46	1.26	1.20	9	1.97	1.51	1.38	140	1.70	1.39	1.30
6	4	1	1.46	1.28	1.22	1.5	2.03	1.58	1.43	35	1.53	1.34	1.26
6	4	1.5	1.42	1.26	1.20	3	1.84	1.50	1.38	70	1.55	1.35	1.27
6	4	3	1.38	1.24	1.19	9	1.78	1.47	1.37	140	1.57	1.36	1.28
6	6	1	1.44	1.28	1.22	1.5	1.99	1.57	1.43	35	1.52	1.33	1.26
6	6	1.5	1.40	1.26	1.20	3	1.81	1.49	1.38	70	1.53	1.34	1.27
6	6	3	1.37	1.24	1.19	9	1.75	1.47	1.36	140	1.55	1.36	1.28

[a]For number of years > 1, the degrees of freedom from a model with blocks nested within years was used instead of the years + blocks main effects model in the original study.

[b]Corresponding to a d_c value of 1.00 and therefore a decreased genetically modified organisms (GMO) mean of 0. However, the normal approximation cannot be realistic in this case because the power analysis assumes a normal distribution on the $\log_{10}(C+1)$ scale, where 0 is the lowest possible value. Reworked version of table 2 in Comas *et al.* (2013).

of 0.25 would mean that 'the test can detect impacts higher than 25% in the comparator's mean density or activity', although this statement is not true. Rather, the value of 0.29 corresponds, for the given settings, to a $(36^{1.29} - 1)/35 = 2.9$-fold increase (or an increase with 190%) or to a $35/(36^{0.71} - 1) = 3.0$-fold decrease (or a decrease with 67%). A reworked version of table 2 in Comas *et al.* (2013), expressing the results as *Fold(decrease)* rather than d_c, is provided in Table 2 in the present study.

AMIGA simulation study

An initial, unexpected result of the AMIGA simulations (not shown here) is that the empirically found type I error of some tests was clearly larger than the nominal significance level of 0.05. As an example, it was between 0.07 and 0.10 for a test using the OP model of data generated from a negative binomial distribution with a mean of 2, a CV of 200% and replication of less than 30. In general, this occurred most often for the OP model at low mean count levels or at low replication. For such cases, it would be misleading to present power results because it is always possible to increase power by increasing the size of the test. Therefore, subsequently, we suppress the reporting of powers if the observed test size was higher than 0.067, which is the 99% upper limit of a binomial distribution with true probability 0.05 and a binomial total of 1000 (the number of simulations).

Power curves for the remaining cases are shown in Fig. 2. The empirically found power results were rather similar for the three methods of analysis where these could be compared.

The most important result from the perspective of the present study is that the CV at a fixed mean count level influences the power more than the mean count level at fixed CV. It can be seen that the results for a CV of 100% are rather similar for the three mean count levels 2, 5 and 20. In all cases, at least 60 replications were found to be required to detect a 50% (two-fold) decrease with 80% power. For a CV of 150%, the required replication would be around or above 100. On the other hand, for a CV of 75%, a replication of 30 could be sufficient. Note that a CV of 75% is not expected (and therefore was not simulated) for a mean count level of 2 because this would mean the absence of

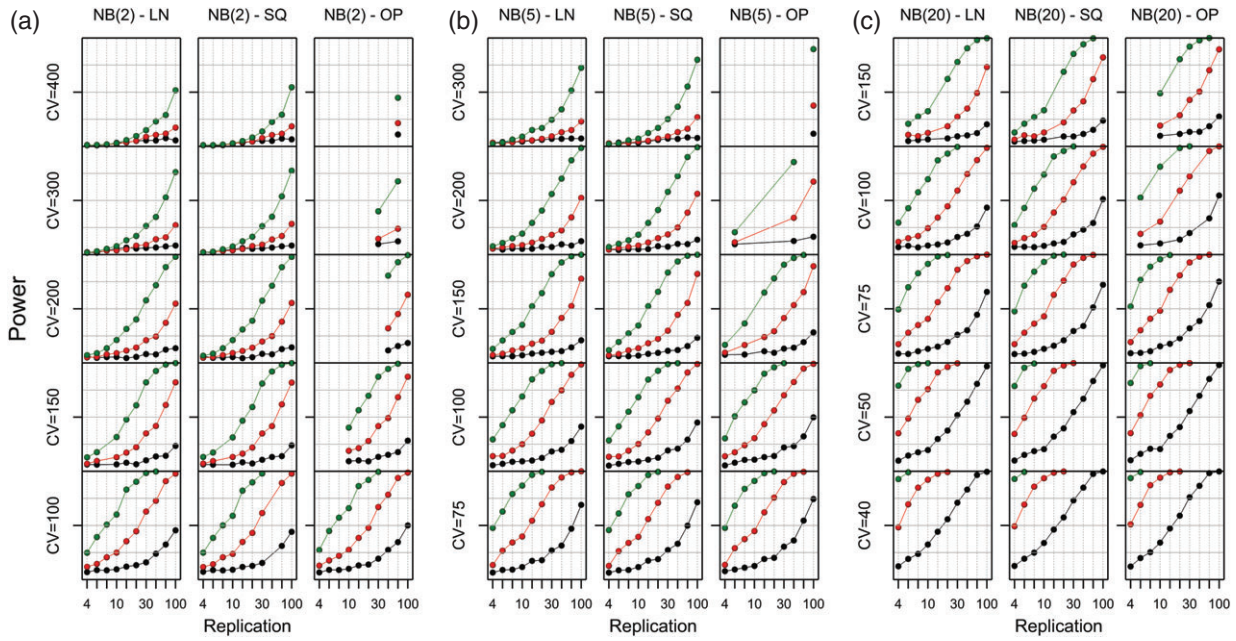

Figure 2 Power of difference tests for negative binomial (NB) data with a mean 2, 5 or 20 (a–c, block of graphs), and varying coefficients of variation (*CV*) (rows) and three methods of analysis [columns log-normal (LN), square-root normal (SQ), overdispersed Poisson (OP)]. For each graph, the power (axis runs from 0 to 1) is plotted against replication (4, 6, 8, 10, 15, 20, 30, 40, 60, 100) for an effect size of $Q = 0.75$ (lower black dots and line), $Q = 0.5$ (middle red dots and line) and $Q = 0.25$ (upper green dots and line). Power values for progressive tests (i.e. when the size of the test is significantly larger than the nominal value 0.05) are not displayed.

almost any extra-Poisson variation above the pure Poisson CV of $\sqrt{2}/2 = 71\%$. Reducing the CV even further (i.e. to 50%) shows that a replication of 15 is sufficient, although such a low CV level is hardly expected for count levels below 10, and was realized for a only a limited number of endpoints in practical studies (Duan *et al.*, 2006).

Discussion

In the present study, we critically reviewed the literature on power analysis for GMO environmental safety field trials and compared the results with those from a new simulation study. We showed that reported recommendations for replication have to be interpreted with care.

In a simple completely randomized design with only the GMO and a comparator, the minimum number of replicates required to achieve 80% power for a 50% (two-fold) decrease in counts was shown to be around 60 in the AMIGA simulation study for data with a CV of 100%. This result was not much dependent on the method of analysis, and similar results were obtained by Perry *et al.* (2003), even when their simulations differed from ours in some of the details, and they analyzed the data using randomization testing.

We could identify reasons why other publications could suggest much lower replication numbers, in the range 3–6. The main reasons found were additional 'hidden' replication (Duan *et al.*, 2006; Prasifka *et al.*, 2008), selection of endpoints to favourable cases with low coefficients of variation (Prasifka *et al.*, 2008), and reporting of results on an unusual scale (Albajes *et al.*, 2012, 2013; Comas *et al.*, 2013).

Experimental designs of field trials may contain additional treatments and/or repeated samplings. Additional treatments are, for example, insecticide spraying or other agronomical regimes. An additional treatment may be randomized at the same level as the varieties, or it may be randomized at a higher or lower level in a split-plot design. Repeated sampling may refer to subsamples of the same plot or to repeated measurements in time series.

If an additional factor is randomized at a lower level than the crop varieties, as in Duan *et al.* (2006), it should first be considered whether the risk assessment (GMO versus comparator) should be made averaged over the additional treatments in the design. If this can be positively answered (as was the case in Duan *et al.*, 2006), then the other factors will add replication to a randomized (block) design. The gain in effective replication will be large if the residual variation at the subplot level of the design is much larger than the residual variation at the main plot level. Alternatively, if averaging over the additional treatments is not intended, then specific comparisons will have to be identified as being relevant for risk assessment. For example, a comparison of a GMO without spraying and a near-isogenic variety under conventional management may be of first interest, although the available replication will be lower than for the averaged comparison.

A second form of additional replication occurs with repeated sampling on the same plots, both within the growing season and over multiple years. Again, the relevant question is whether the GMO to comparator differences are expected to be the same across time points (at some appropriate scale, typically the logarithmic scale). Another relevant question is then whether the time in between sampling time points is sufficiently large to assume

independence between counts. If both questions are answered positively, then time points are effectively replications. Then, the true replication of the study described in Duan *et al.* (2006) can be seen to be 4 (block replications) × 4 (insecticide treatments) × 2 (years) × ~4 (time points per year) = ~128 for each of the two crop lines. Note that the effective replication would be much lower if the correlation between counts at successive time points is large. Similarly, the number of replicates in the power curves presented by Prasifka *et al.* (2008) refer only to blocks and replications over year. Their prospective power analysis assumes that future studies have similar numbers of subsamples (between 1 and 11) and sampling periods (between 1 and 17) as the data sets that they analyzed, and that there are no interactions with other treatments in the design, and that repeated measures are independent.

It is an open question how count data should be analyzed for designs other that the most simple ones. If there is additional 'hidden' replication, as discussed above, then one possibility is to include all other factors (e.g. insecticide treatments and time periods) in an appropriate model and analyze the counts at the observation or trap level. This enables testing of various assumptions such as the absence of an interactions with other treatment factors and or independence across time. If such assumptions can be made, the model itself, with these assumptions incorporated, provides the aggregation. It is then unnecessary to aggregate the data. However, this might involve complicated statistical models such as generalized linear mixed models, which are not always easy to handle.

Another possibility is to sum the counts over the additional replications (e.g. subplots, subsamples, time points) and analyze the resulting sums in a simpler design. Counts should not be summed over different plots because extra-Poisson variation is usually expected, and therefore it is needed to estimate the overdispersion or CV from the data. Over-dispersion has to be modelled at the plot level. The number of residual degrees of freedom to estimate this variation should be sufficiently large, which would be doubtful with, say, just four replicates and two treatments.

Summation of counts will not only give a higher mean count μ and a lower CV (both generally associated with better power), but also it will lead to a lower replication (associated with less power). The balance between these would be a good subject for further study.

A second reason for low recommended replication was found in the selection of available data to the less variable taxa. Prasifka *et al.* (2008) omitted approximately 80% of their candidate endpoints before they calculated their power results. This is not necessarily wrong because, in any large analysis, there will be taxa for which counts are too low (and therefore CV too high) to allow any sensible analysis. However, the rules for selecting endpoints should be clear and agreed with the risk managers.

A third reason for the apparently low values of the required number of replications was found to be unusual scales of reporting. The 'field test capacities of detection' reported by Albajes *et al.* (2012), Albajes *et al.* (2013) and Comas *et al.* (2013) have a complex interpretation. In the present study, we have presented the results of Comas *et al.* (2013) in another form, as the fold increase or fold decrease on the original count scale rather than the log transformed scale, which shows that the required number of replications for 80% power to detect a two-fold decrease would need to be larger than suggested. Future studies on power analysis should report findings on the original count scale rather than on some transformed scale.

A conclusion from the AMIGA simulation study is that linear models applied to transformed counts [e.g. by the $\log_e(C + 1)$ transformation] are surprisingly robust, and even may show better performance on the size of the test than generalized linear models, such as the OP model. The use of linear models on untransformed data with skew distributions should, however, be avoided (Wang & Riffel, 2011).

In the reviewed studies, the effect sizes, expressed as the ratio of GMO mean to CMP mean, were chosen to be 0.5, 0.67, 0.77, 1.3, 1.5 and 2 (Perry *et al.*, 2003); 0.67 and 2 (Duan *et al.*, 2006); 0.5, 0.7 and 0.8 (Prasifka *et al.*, 2008); and 0.25, 0.5 and 0.75 (AMIGA simulation study). Identification of the relevant effect sizes for a specific case requires further study, although a good starting point would be to investigate the power at an effect size equal to the limit of concern (LoC). EFSA (2010) defined the LoC as the minimum relevant ecological effect that is deemed biologically significant, and also is deemed of sufficient magnitude to cause harm. In food-feed risk assessment, a procedure has been developed to derive LoCs from the variation between reference varieties in the same field trials (van der Voet *et al.*, 2011). Perry *et al.* (2009) found this approach to be less appropriate for ERA, also noting that the direct setting of LoCs appears to be more feasible in ERA than in food-feed risk assessment. Until now, specific LoCs have not been set by official bodies, although the range of values reported above may summarize the practical experience of many ecologists and serve as first approximations.

Perry *et al.* (2009) noted that the null hypothesis of a GMO risk assessment test should be that of non-equivalence, with the equivalence/non-equivalence boundary being defined by the LoC. Practical procedures for equivalence tests have been described (Schuirmann, 1987). This would imply that power analysis should also be performed for equivalence tests. The power of equivalence tests was also investigated in the AMIGA simulation study (see Supporting information, Doc. S1).

The required replication of field tests will depend on the properties of the taxa that need to be assessed. Taxa at low count levels will have high CVs and therefore the required number of replications will become high. Some sort of threshold for inclusion of taxa is needed in a prospective power analysis. A possible approach would be to restrict the risk assessment to taxa with a sufficiently low CV (e.g. 100%). Because of extra-Poisson variation, this would restrict the analysis to mean abundances of at least some factor above 1 (at which level the pure Poisson CV is already 100%). The CV to mean graphs of Duan *et al.* (2006) suggest that this would even imply to require mean abundances larger than 5 for some taxa.

In conclusion, a prospective power analysis for field tests is a nontrivial exercise for any but the simplest experimental design. Standard off-the-shelf software for power analysis typically requires the assumption of a normal distribution at some scale. Assuming a CV of at most 100% and an abundance of at least 5, such a prospective power analysis for a simple log count plus 1 transformation is likely to give a good indication of the required replication for any particular design. However, the simulations of

both Perry *et al.* (2003) and ourselves show that the true power may be different under alternative distributions, most notably at low count levels (< 5) or high variation (CV $> 100\%$). We also found that the type I error of some tests was sometimes larger than the nominal significance level of 0.05. Furthermore, the proper replication and variation estimates have to account for all aspects of practical field tests that might influence the results (such as hidden replication, additional treatments and repeated samplings). Therefore, for the design of future experiments, it would be useful to develop power analysis software that addresses these issues, as will be pursued in the AMIGA project. A critical review of previous research and a new simulation study suggest that approximately 60 replicates (e.g. 20 replications in each of 3 years, as was used in the FSE) would be sufficient to detect a 50% (two-fold) decrease in the abundance for a taxon, provided that the CV is not higher than 100%.

Acknowledgements

This is publication No. 5 produced within the framework of the project Assessing and Monitoring the Impacts of Genetically Modified Plants on Agro-ecosystems (AMIGA), funded by the European Commission in the Framework programme 7, theme KBBE.2011.3.5-01. We thank the authors of the described studies for their helpful answers to our questions about their work. HvdV designed the study, analyzed previous research and wrote the first draft of the paper. PWG performed and analyzed the results of the simulation study. Both authors reviewed the draft and agreed on the final version submitted for publication.

Supporting information

Additional Supporting information may be found in the online version of this article under the DOI reference: 10.1111/afe.12092

Doc. S1. Environmental risk assessment of genetically modified organisms: set-up of a simulation study to investigate properties of difference and equivalence tests.

References

Albajes, R., Farinós, G.P., Pérez-Hedo, M. *et al.* (2012) Post-market environmental monitoring of Bt maize in Spain: non-target effects of varieties derived from the event MON810 on predatory fauna. *Spanish Journal of Agricultural Research*, **10**, 977–985.

Albajes, R., Lumbierres, B., Pons, X. & Comas, J. (2013) Representative taxa in field trials for environmental risk assessment of genetically modified maize. *Bulletin of Entomological Research*, **103**, 724–733.

Clark, S.J., Rothery, P. & Perry, J.N. (2006) Farm scale evaluations of spring-sown gentically modified herbicide-tolerant crops: a statistical assessment. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, **273**, 237–243.

Comas, J., Lumbierres, B., Pons, X. & Albajes, R. (2013) Ex-ante determination of the capacity of field tests to detect effects of genetically modified corn on nontarget arthropods. *Journal of Economic Entomology*, **106**, 1659–1668.

Duan, J.J., Jiang, C., Head, G.P., Bhatti, M.A. *et al.* (2006) Statistical power analysis of a 2-year field study and design of experiments to evaluate non-target effects of genetically modified *Bacillius thuringiensis* corn. *Ecological Entomology*, **31**, 521–531.

EFSA (2010) EFSA Panel on Genetically Modified Organisms (GMO). Guidance on the environmental risk assessment of genetically modified plants. *EFSA Journal*, **8**, 1879. [111 pp], doi:10.2903/j.efsa.2010.1879.

Goedhart, P.W., van der Voet, H., Baldacchino, F. & Arpaia, S. (2013) *Environmental Risk Assessment of Genetically Modified Organisms: Overview of Field Studies, Examples of Datasets, Statistical Models and a Simulation Tool. Deliverable 9.1*, AMIGA project [WWW document]. URL http://www.amigaproject.eu/documents/deliverables/ [accessed on 27 October 2014].

Goedhart, P.W., van der Voet, H., Baldacchino, F. & Arpaia, S. (2014) A Statistical Simulation Model for Field Testing of Non-Target Organisms in Environmental Risk Assessment of Genetically Modified Plants. *Ecology and Evolution*, **4**, 1267–1283.

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

NCSS (2002) *PASS 2002: power analysis and sample size for Windows user's guide - II*. NCSS, Kaysville, Utah.

Perry, J.N., Rothery, P., Clark, S.J., Heard, M.S. & Hawes, C. (2003) Design, analysis and statistical power of the farm-scale evaluations of genetically modified herbicide-tolerant crops. *Journal of Applied Ecology*, **40**, 17–31.

Perry, J.N., ter Braak, C.J.F., Dixon, P.M. *et al.* (2009) Statistical aspects of environmental risk assessment of GM plants for effects on non-target organisms. *Environmental Biosafety Research*, **8**, 65–78.

Prasifka, J.R., Hellmich, R.L., Dively, G.P., Higgins, L.S., Dixon, P.M. & Duan, J.J. (2008) Selection of nontarget arthropod taxa for field research on transgenic insecticidal crops: using empirical data and statistical power. *Environmental Entomology*, **37**, 1–10.

Semenov, A.V., van Elsas, J.D., Glandorf, D.C.M., Schilthuizen, M. & de Boer, W.F. (2013) The use of statistical tools in field testing of putative effects of genetically modified plants on nontarget organisms. *Ecology and Evolution*, **3**, 2739–2750.

Schuirmann, D.J. (1987) A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics*, **15**, 657–680.

Taylor, L.R. (1961) Aggregation, variance and the mean. *Nature*, **189**, 732–773.

van der Voet, H., Perry, J.N., Amzal, B. & Paoletti, C. (2011) A statistical assessment of differences and equivalences between genetically modified and reference plant varieties. *BMC Biotechnology*, **11**, 15.

VSN International (2012) *GenStat for Windows*, 15th edn. VSN International [WWW document]. URL www.GenStat.co.uk [accessed on 27 October 2014], UK.

Wang, M. & Riffel, M. (2011) Making the right conclusions based on wrong results and small sample sizes: interpretation of statistical tests in ecotoxicology. *Ecotoxicology and Environmental Safety*, **74**, 684–692.

Accepted 5 October 2014
First published online 20 November 2014