

Which Endpoints Can Be Reliably Assessed in Semi-field Pollinator Species Testing without Estimating False Positive or False Negative? MDD's and Replicates Issue

Marco Pompeo Candolfi, Holger Bargaen, Sigrun Bocksch, Olaf Klein, Marco Kleinhenz, Silvio Knaebe and Bronislawa Szczesniak

Eurofins Agrosience Services Ecotox GmbH, Eutinger Str. 24, Niefern-Öschelbronn 75223, Germany

Abstract: Statistical power, number of replicates and experiment complexity of semi-field and field studies on *Apis* and non-*Apis* bee species has become a major issue after publication of the draft European Food Safety Authority (EFSA) Guidance on risk assessment of plant protection products (PPP) on bees (*Apis mellifera*, *Bombus* spp. and solitary bees). According to this guidance document, field studies have to be designed to be able to detect significance differences as low as 7% for certain endpoints such as reduction in colony size. This will require an immense number of replicates which is obviously not feasible. In the present study, key endpoints such as mortality, termination rate and number of brood cells in honeybee studies, cocoon production and flight activity in solitary bee studies and number of gynes in bumble bee studies (just to mention some of the endpoints considered) in semi-field studies were analyzed, with *Apis mellifera*, *Bombus terrestris* and *Osmia bicornis* during the past five years (2013-2017). The results indicate huge differences in the percentage minimal detectable differences (%MDDs) (MDD expressed as median of control value of the endpoint in percent) depending on endpoint and species tested. For honeybee semi-field studies, the lowest %MDDs recorded were between 10% and 15% for the endpoints foraging, number of brood cells and colony strength. The highest %MDDs were observed for the endpoint termination rate, with a %MDD of almost 50%. For the endpoints in bumble bee semi-field studies the %MDDs varied between 17% for bumble bee colony weight and 53% for average mortality during the exposure period in the tunnel. The %MDD for the number of gynes (young queens) was slightly below 25%. For the semi-field solitary bee test system, the %MDDs for the measured endpoints seem to be lower than those for the other two species tested. The %MDDs for the endpoints hatching of offspring, nest occupation and number of cocoons were 8%, 13% and 14%, respectively. Most of the %MDDs were between 10% and 30% indicating clearly that the currently performed experimental design for the semi-field pollinator studies allowed to determine relatively small effects on key study endpoints. The analysis indicated that for all the three bee species tested, the semi-field test design detected low %MDDs for most of the endpoints. It was also observed that detectable differences between the control and PPP treatments were much lower in semi-field test designs than in field studies with these bee species. The “perfect sample size” really does not exist but test design and statistical analysis can be adapted to lower the %MDDs. Measured and simulated (according to Student's *t*-test-distribution) data and results showed that statistical evaluations parameter selection (e.g., alpha value), data transformation (log10) and the number of replicates had a direct effect on the ability of the test design to detect lower or higher %MDD values. It could show that a change of alpha value from 0.05 to 0.1, increases the ability of the studies to detect lower %MDDs. For most of the measured endpoints, increasing the number of replicates e.g., from four to eight, improved the power of the test design by decreasing the %MDD. The reduction magnitude of the %MDD is dependent on the endpoint and selection of statistical parameters such as the alpha value. Parameters that display effects at a biologically relevant scale will be a better indicator for effects than parameters that are able to detect minor differences that are not biologically relevant.

Key words: *Apis mellifera*, *Bombus terrestris*, *Osmia bicornis*, OECD75, minimal detectable difference (MDD), statistical power.

1. Introduction

Risk assessment of plant protection products (PPP)

Corresponding author: Marco Pompeo Candolfi, Ph.D., research fields: ecotoxicology and regulatory risk assessment.

on bees (*Apis mellifera*, *Bombus* spp. and solitary bees) and required studies for European registration are outlined in draft European Food Safety Authority (EFSA) Guidance [1, 2]. According to this guidance document, field studies have to be designed to be able

to detect significance differences as low as 7% for certain endpoints such as reduction in colony size. An analysis presented by Miles [3] showed that, to be able to detect such a small difference of 7% in honeybee field studies, 28 fields 4 km apart with a total of 196 colonies (7 colonies/field) would be required. This is obviously not feasible.

Risk assessment for PPP includes studies of different tiers and scales—from laboratory studies to semi-field and field studies [4, 5]. Laboratory studies use worst case exposure scenarios, therefore researchers would be at the risk of being overprotective in an unrealistic way in risk assessment [6]. Especially for honeybees being highly developed, socially organized insects and other pollinator species, laboratory tests will always face the problem of creating an artificial situation for the test organisms in an extreme way. While laboratory studies help to identify potential adverse effects, field studies help to identify the consequences of these adverse effects in the real world [7].

Under realistic field conditions, adverse effects may possibly not occur if exposure is lower than expected due to avoidance behavior of the honeybees (or pollinators). Potential effects may also not cause any problems on the level of the whole colony, if effects are buffered within the structures of the colony or if the effects are small in relation to biological variability or other adverse effects of natural environmental origin [8].

Extrapolation of laboratory data to forecast the situation in the field is still not a standard procedure and will need verification by testing under field conditions. Statistical analysis of data from semi-field and field studies seems to be a weak aspect for this approach [9]. Due to the realistic environmental context and the high complexity of semi-field and field studies, the precision (low variability and high repeatability) known from laboratory studies is difficult, maybe even impossible, to achieve [1, 10]. Studies under field conditions have to cope with the issue of pseudo-replication [11] as well as a low number of replicates [12].

In this paper an attempt has been made to give insight into the limits and options of data generated for regulatory purposes under Good Laboratory Practice (GLP) in semi-field tunnel studies with different pollinators.

For studies on effects of PPP on honeybee brood, there is a study design available (OECD Guidance Document No. 75) [13] to assess the brood development success of honeybee colonies by exposing honeybees to treated flowering crop within tents, thereby avoiding the problem of pseudo-replication. Honeybee colonies are installed in tunnels and allowed to forage on flowering crop treated with water (control) or test item(s). For practical reasons, the number of replicates is limited for this test design. In addition to the normally measured endpoints such as mortality, colony strength and honeybee flight, the development of eggs until emergence is assessed. Colonies remain within the flight area (tunnel) for 7 d after application, which is approximately the time an egg needs to develop to capped pupae. As one of the endpoints, the termination rate for the selected eggs (expressed as rate of eggs that did not reach state of hatching adult honeybee) is calculated.

For non-*Apis* pollinators (bumble bees and solitary bees), guidelines are not available yet but are in development and ring-tested by an International Commission for Plant Pollinator Relationships (ICPPR) working group. The test design follows in principle the honey bee OEPP/EPPO Guideline No. 170 (4) [14] including necessary adaptations. Due to the lower number of individuals, bumble bees and solitary bees are easier to be kept in the tunnels for a longer exposure period (bumble bees at least for 14 d, solitary bees for their whole activity period).

The calculation of *p*-values and the concept of null-hypothesis significance testing (NHST) have been questioned recently [15, 16]. On the other hand, de Winter [17] points out the usefulness of *t*-test calculations even with small numbers of replicates. An overview is given on NHST statistics for several

endpoints of honeybee brood, bumble bee and solitary bee studies conducted for regulatory purposes. Moreover, the effects of changes in test design on NHST and minimal detectable difference (MDD) statistics are assessed by modeling single parameters of the Student's *t*-test formula. For some key endpoints, also the data from OECD75 honeybee studies were analyzed regarding MDD and power analysis.

The target was to analyze OECD75 semi-field honeybee studies, compare the data to other pollinator semi-field studies and to investigate options for possible changes in study design, to optimize the detection of significant treatment effects for such regulatory higher tier studies. Student's *t*-test statistics was used to keep things simple and to focus on the influence of study design decisions on information content of the data.

2. Materials and Methods

Data from 50 GLP regulatory studies with honeybees (*Apis mellifera*) performed according to OECD75 [13], seven studies with bumble bees (*Bombus terrestris*) and 10 studies with solitary bees (*Osmia bicornis*), all conducted in Germany between 2013 and 2017 from Eurofins Agrosience Services Ecotox GmbH, were used to calculate the MDD for several key study endpoints. All studies were conducted by enclosing the test organisms in tents covering attractive feeding crop (*Phacelia tanacetifolia* or winter oil seed rape) treated with test item(s), reference standard or water. An overview on the test design for each species is presented in Table 1. The endpoints assessed in the studies and evaluated in this publication are summarized in Table 2.

For studies conducted with just one test item treatment group and a water control, Student's *t*-test was used. For data with one control and several test item treatments, only the first treatment group (T1) was used for analyses to achieve similar data structure in all studies. Only the first treatment group (T1) was used, the data of the other treatment groups were not included into the calculation. MDD and percentage MDD (%MDD) calculations were done using SAS 9.3. Boxplots were plotted to visualize distribution of resulting MDD values for several endpoints and pollinator species.

In the next step, only data from OECD75 honeybee studies were taken to have a closer look at the effects of log-transformation of data, change in numbers of replicates, a different threshold for alpha of the statistical analysis and different schemes for rating MDD values. This was achieved by modeling single parameters of the *t*-test and keeping the population variance constant.

Population variance from the actual study data (50 GLP OECD75 honeybee studies) was taken for the calculations which were based on the assumption that a larger number of replicates would improve precision of the parameters without changing the accuracy in a predictable direction.

Being a *post-hoc* approach, with the intention to show what can be achieved from real studies, the power term was excluded from the formula. The aspect of power within the data is considered later on.

The formula for calculation of MDD according to Zar [18] is as follows:

$$\delta = \sqrt{\frac{2sp^2}{n}} \times t(\alpha(i), \nu) \quad (1)$$

Term	Meaning of term	Use in this paper
δ	MDD	Result
sp^2	Population variance	Result from study data—regarded as fix for modeling
n	Number of replicates	Matter of test design—influence of change of number of replicates was considered
ν	Degrees of freedom	
$\alpha(i)$	Alpha threshold indicating highest <i>p</i> -value which would be regarded as statistically significantly different from control	MDD at $\alpha = 0.05$ was compared to MDD for $\alpha = 0.10$

MDD values were expressed as %MDD from the mean value of the control group. The %MDD expresses the size of the effect that can be statistically detected between the control and test item treatment. Benchmarks given by Cabrera *et al.* [19] as well as Brock *et al.* [20] were used to classify the studies, accordingly in Table 3. The scale/effect classes proposed for bumble bees by Cabrera *et al.* [19] focus on classifying more effect categories < 50% while the scale used in aquatic studies proposed by Brock *et al.* [20] focuses on assessing/evaluating more in detail effect classes with %MDD > 50%. Figures were prepared to plot distribution of study data for several endpoints according to these benchmarks and depending on modeled changes in number of replicates, different levels for alpha and transformation of data.

Bar graphs were plotted showing the distribution

of %MDD classes according to Cabrera *et al.* [19] or Brock *et al.* [20] using transformed or untransformed data, different number of replicates or a different level for alpha while preserving population variance as observed in the “real world” studies.

The power of *t*-tests of different endpoints of the studies in relation to the number of replicates, alpha levels and desired %MDD were calculated and plotted as dependency of the desired %MDD. Bar graphs were used to represent the distribution of data of several endpoints depending on desired %MDD, alpha level and number of replicates.

The following formula for calculation of power according to Zar [18] was used:

$$t\beta(1), \nu \leq \frac{\delta}{\sqrt{\frac{2sp^2}{n}}} - t(\alpha(i), \nu) \quad (2)$$

Term	Meaning of term	Use in this paper
β	Probability of the test to fail to indicate a statistically significant difference even though there is one	Used for calculation of power (see below)
δ	Desired MDD	Matter of choice
sp^2	Population variance	Result from study data—regarded as fix for modeling
n	Number of replicates	Test design choice—influence of changing the number of replicates was considered
ν	Degrees of freedom	
$\alpha(i)$	Alpha threshold indicating highest <i>p</i> -value which would be regarded as statistically significantly different from control	The %MDDs were calculated at $\alpha = 0.05$ and compared to %MDD obtained for $\alpha = 0.10$
Power	$1 - \beta$	Result

Bar graphs were plotted showing the distribution of classes of power of tests for brood termination rate on day 22 (BTR22) by varying/modeling the desired %MDD from 20% to 50% of control, as well as changing the number of replicates from $n = 4$ to $n = 8$ or changing alpha from 0.05 to 0.10, but keeping population variance of the “real world” studies.

Actual power of test (%) was plotted against actual MDD as percent of control value keeping population variance found in “real world” studies but modeling the situation for different number of replications and change in alpha level. Two sets of graphs were prepared showing the situation for a desired %MDD of 20% and of 50%.

A range of %MDD was modeled in relation to test design and statistical evaluation of the study (transformation of data, number of replicates, alpha = significance limit) as well as target detection of 20 %MDD or 50 %MDD and displayed with color coding as “heat-map”. Additionally, the power of the statistical design was calculated and expressed in bar charts within the heat-map. These calculations/modeling were done for honeybee, bumble bee and solitary bee studies and the endpoints listed in Table 2.

In theory low %MDDs should indicate high discriminatory ability of the statistical analysis. However, in the “real world”, other factors may be

of importance when analyzing statistically significant effects between two treatment groups. To illustrate this aspect of a study design, two similar endpoints of honeybee brood studies, BTR22 (BTR at the end of an entire brood cycle, an endpoint that is calculated from extremely precise photo assessments) and the “difference between numbers of brood cells day BFD22 \pm 1 to BFD0” (an estimated parameter) were analyzed. Both endpoints indicate the success of the brood development during the exposure phase of the honeybee study but differ in MDDs and %MDDs. For this analysis, data of the control were compared to the toxic reference treatment and the %MDDs plotted against the MDDs in scatter plots. Each data point in these figures represents a single study indicating whether the MDD was higher than the difference between control (C) and reference substance (R) or not.

To address and illustrate the importance of the biological relevance of effects, a constructed data set of two fictive experiments was used; each with four replicates. Study A with absolute MDD of 15.9 and %MDD of 299.1% and study B with an absolute MDD of 2.1 and %MDD in percent to control of 4.7%. One data set had large variation and the other small variation in the data set as well as MDD and %MDD. The intention was to illustrate that the structure of the data as well as the magnitude of the expected effects have also to be considered while evaluating the %MDDs.

3. Results

As shown in Fig. 1, the %MDD calculated for different study endpoints and species tested varies expressively.

For honeybee semi-field tunnel studies (Fig. 1), the lowest median %MDDs were between 11% and 20% for the endpoints foraging, number of brood cells, compensation index (CI) and colony strength. The highest %MDD was observed for the endpoint termination rate, with a median %MDD of 48%.

For bumble bee endpoints (Fig. 1) the median %MDDs in semi-field studies varied between 17% for bumble bee colony weight and 53% for average mortality during the exposure period in the tunnels. The median %MDD for the number of gynes (young queens) was 23%.

For the semi-field *Osmia bicornis* test system, the median %MDDs for the measured endpoints seemed to be lower compared to the other two species tested (Fig. 1). The median %MDDs were all around 10%: for the endpoints hatching of offspring the %MDD was 8%, for nest occupation the %MDD was 13% and for the number of cocoons the %MDD was 14%.

Figs. 2-6 show the results of simulations on the influence of the test design (four or eight replicates) and the choice of statistical parameters (alpha value 0.05 or 0.1; one-or two-sided statistical test; non-or log-transformed data set) on the percentage of studies able to detect specific %MDD values for different honeybee endpoints.

Comparing the results obtained for mortality (Fig. 2), colony strength (Fig. 3), termination rate (Fig. 4) and foraging (Fig. 6), it was apparent that the results differed greatly depending on the endpoints. In most of the selected test design/stats parameter configurations for each of the endpoints, only few cases/studies were able to detect %MDDs < 10%. Considering e.g., the currently used test design for semi-field brood honeybee studies with four replicates, an alpha of 0.05, a one-sided test with log10 data transformation (5th column in each of the Figs. 2-4 and 6), only 0%, 17%, 2% and 61% of the performed studies could detect %MDDs < 10% for the endpoints mortality (Fig. 2), difference of colony strength BFD0 to BFD5 (Fig. 3), termination rate (Fig. 4) and foraging (Fig. 6), respectively. If the number of replicates increased from four to eight, an improvement was observed with 21% to 30% more studies detecting %MDDs < 10% for the endpoints mortality (number of studies increased from 0% to 21%, columns 5-6 Fig. 2), colony strength (number of studies

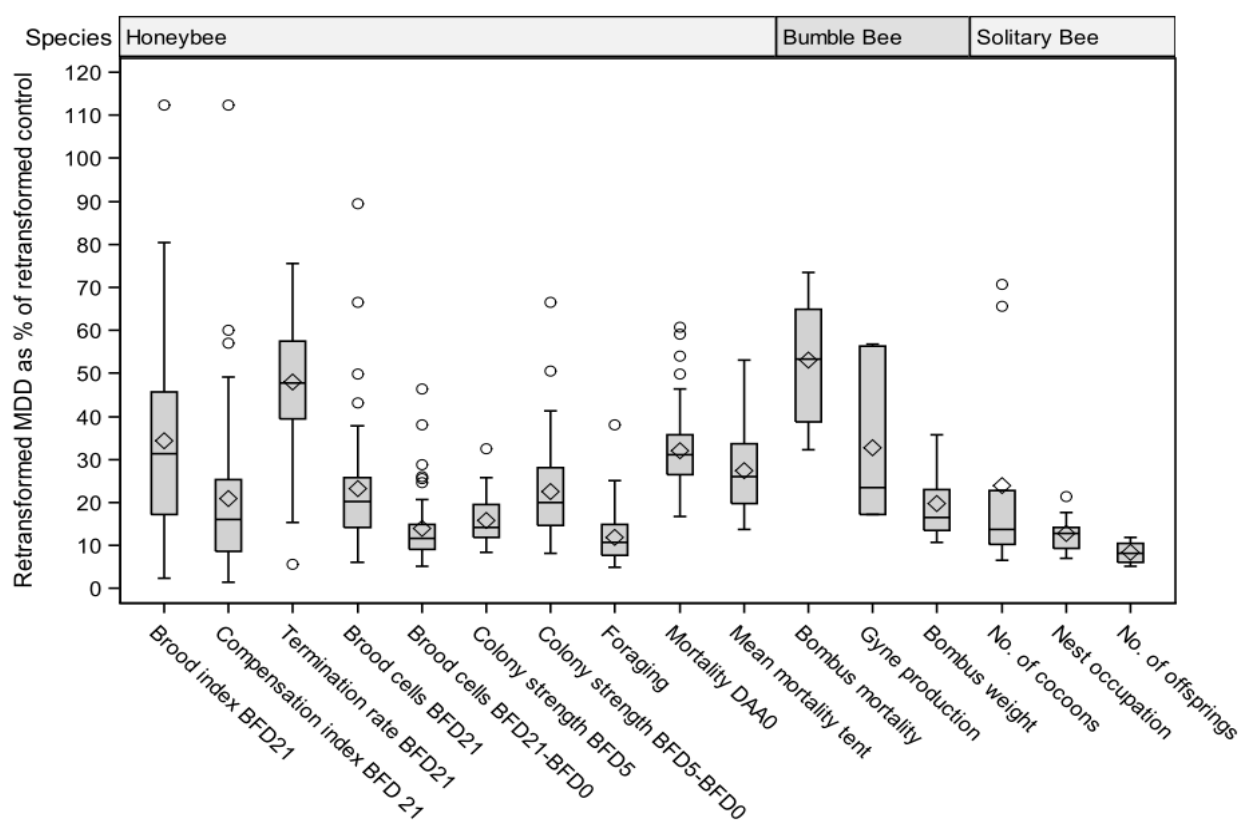


Fig. 1 Box plot of percentage minimal detectable differences (%MDDs) calculated for different endpoints and bee species in semi-field tent studies.

Legend: \diamond indicates mean value; line within the +75th and -25th percentile box indicates the median value.

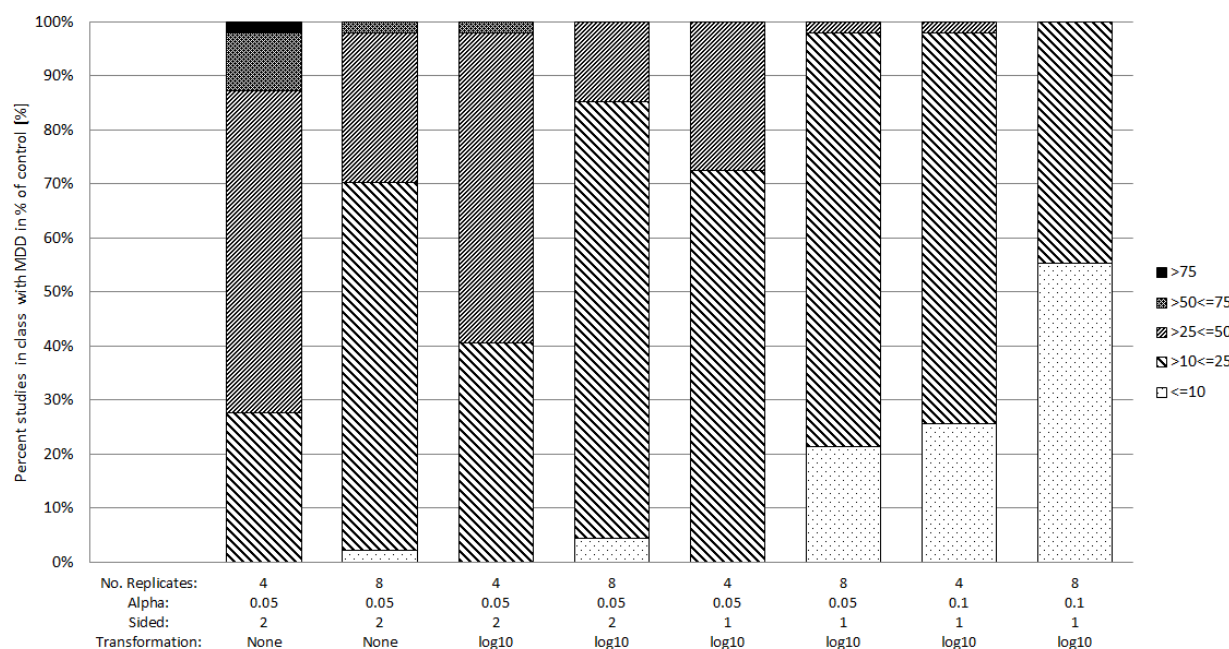


Fig. 2 Influence of test design (x-axis: no. of replicates, alpha and data transformation) on the distribution of honeybee studies (y-axis: expressed as % studies in the respective class) in different %MDD classes as per Cabrera *et al.* [19] (side legend).

Endpoint considered: mortality.

increased from 17% to 47%, columns 5-6 Fig. 3) and flight activity (increase from 61% to 91% of all studies, columns 5-6 Fig. 6) but not for BTR (number of studies increased from 2% to 4%, columns 5-6 Fig. 4). The situation improved also with a detection level of %MDDs $\leq 25\%$: with four replicates, an alpha of 0.05, a one-sided test and with log10 data transformation (5th column in each of the Figs. 2-4 and 6) 72%, 81%, 12%, 98% of the studies detected %MDDs $< 25\%$ and with eight replicates, an alpha of 0.05, a one-sided test and with log10 data transformation (6th column in each of the Figs. 2-4 and 6), approximately 98%, 96%, 36% and 100% of the studies performed detected %MDDs $< 25\%$ for the endpoints mortality (Fig. 2), difference of colony strength BFD0 to BFD5 (Fig. 3), termination rate (Fig. 4) and foraging (Fig. 6), respectively.

Having a closer look at the endpoint difference of colony strength BFD0 to BFD5 (Fig. 3), with a test design of four replicates and performing a two-sided test on non-transformed data at an alpha of 0.05 (Fig 3, 1st column), just two of the studies conducted could

detect a %MDD $< 10\%$. On the other hand, taking the same endpoint difference of colony strength BFD0 to BFD5, with the test design consisting of eight replicates and performing one-tiled test on log-transformed data at an alpha of 0.1 (Fig. 3, 8th last column), almost 77% of the studies conducted could detect a %MDD $< 10\%$. These are extreme cases concerning selected test design/stats parameter configurations, showing the “worst” and so to say the “best configurations”. For the endpoints mortality and foraging, a test design change from four replicates, two-sided test on non-transformed data at an alpha of 0.05 to eight replicates and performing a one-tiled test on log-transformed data at an alpha of 0.1 (1st and 8th last column in Figs. 2 and 6) increased the number of studies with %MDD $< 10\%$ from 0% to 55% and from 50% to 98% for mortality and foraging, respectively. However, this was not valid for the endpoint termination rate (Fig. 4), where even a test design with eight replicates and performing a one-tiled test on log-transformed data at an alpha of 0.1 (8th last column in Fig. 4) did not result in a major improvement

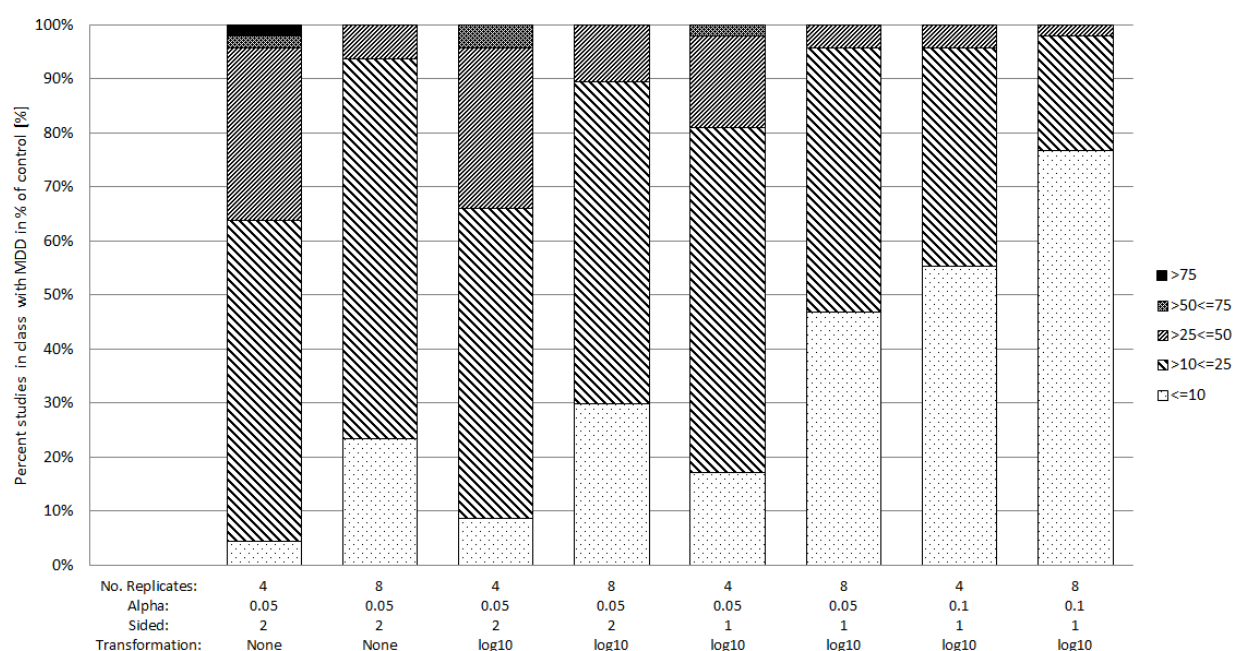


Fig. 3 Influence of test design (x-axis: no. of replicates, alpha and data transformation) on the distribution of honeybee studies (y-axis: expressed as % studies in the respective class) in different %MDD as per Cabrera *et al.* [19] (side legend). Endpoint considered: difference of colony strength BFD0 to BFD5.

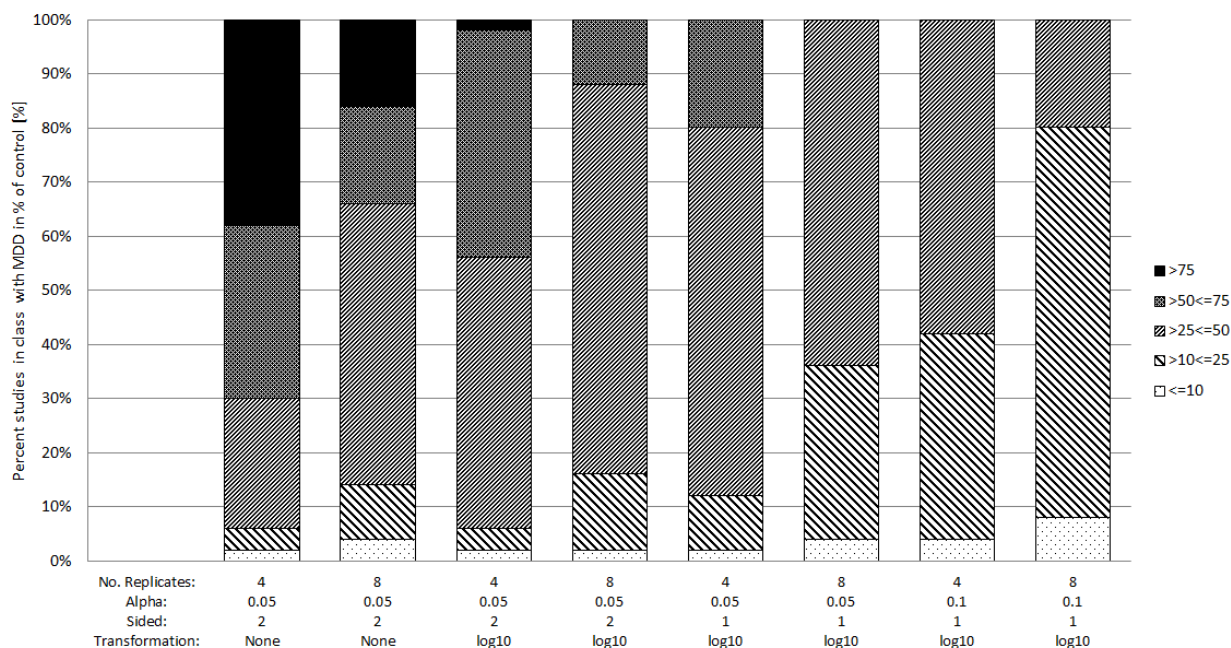


Fig. 4 Influence of test design (x-axis: no. of replicates, alpha and data transformation) on the distribution of honeybee studies (y-axis: expressed as % studies in the respective class) in different %MDD classes as per Cabrera *et al.* [19] (side legend).

Endpoint considered: termination rate.

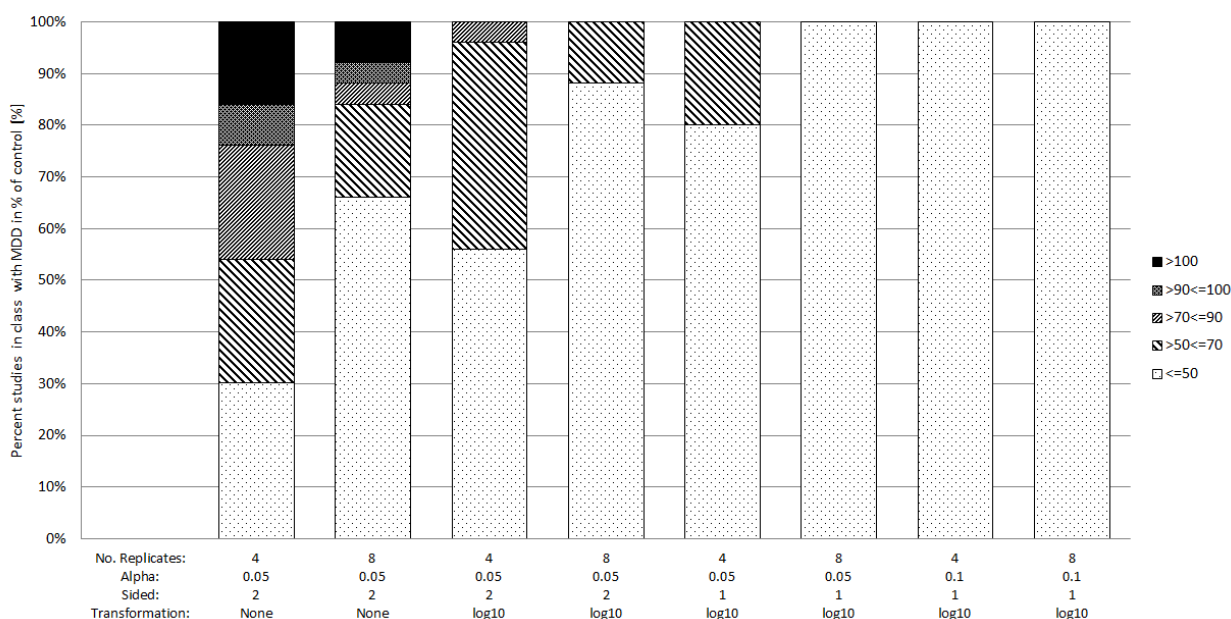


Fig. 5 Influence of test design (x-axis: no. of replicates, alpha and data transformation) on the distribution of honeybee studies (y-axis: expressed as % studies in the respective class) in different %MDD classes as per Brock *et al.* [20] (side legend).

Endpoint considered: termination rate.

in detection of %MDDs < 10% (8% of studies instead of 2% of studies).

The analysis suggests that due to data distribution and expected (negative) effects, in general a log10

data transformation and a one-tailed statistical evaluation should be applied to the data set.

The %MDD class setting according to Cabrera *et al.* [19] or Brock *et al.* [20] are different: the first one

focuses more on the sharper separation on the lower values (many %MDDs classes between 0% and 50%) while the class setting suggested by Brock *et al.* [20], which is often used in aquatic ecotoxicology, focuses more on the differentiation of %MDD values in the range of 50% to 100%. If these two classification schemes are applied to the data set (Figs. 4 and 5), thereby changing the %MDD scale at which the study designs detected differences, a clear different interpretation of the study results was observed. Using the classification of Cabrera *et al.* [19] only few study design combinations are in the lowest classification class ($< 10\%$), while using the Brock *et al.* [20] classification much more study design combinations would be in the lowest classification class ($< 50\%$).

The relationship between %MDD and statistical power will now be analyzed by varying the test design. Fig. 7 summarizes the MDD modeling calculations for different endpoints and species according to different model parameter settings (data transformation, desired MDD, replicate numbers and alpha). Bars inside each cell indicate the power distribution of the *t*-tests for

each endpoint according to the model settings and population variance found in the “real world” data set. Examples are shown in Figs. 8 and 9 and Figs. 10 and 11 for the endpoints mortality and termination rate.

Results of power analysis vary over a wide range, depending on the endpoint that was assessed/modeled. As a general trend in modeling, an increase of the %MDD at which a difference to the control should be detected (namely from 20% to 50% as in Fig. 7 and Figs. 8-11), the %MDDs usually keeps the value/position while the power is lifted up (increased). By increasing the number of replicates, the actual %MDD shrinks while the power slightly rises. When modeling a change of alpha from 0.05 to 0.1, the actual %MDD shrank with a major increase in the experimental power.

A closer look at the key endpoint BTR in OECD75 honeybee semi-field studies indicates that only the increase of %MDD detection class from 20% to 50% had a large impact in the power of the endpoint BTR in honeybee studies, while increasing the number of replicates and the alpha level from 0.05 to 0.1 had a low impact on the power of statistical analysis (Fig. 12).

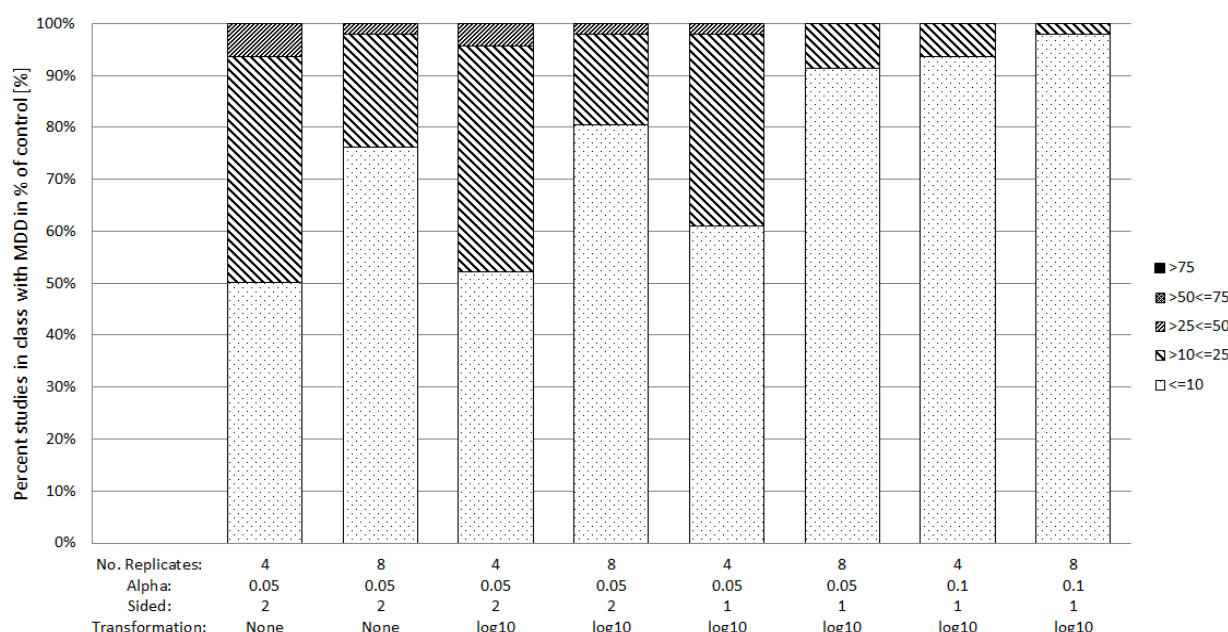


Fig. 6 Influence of test design (x-axis: no. of replicates, alpha and data transformation) on the distribution of honeybee studies (y-axis: expressed as % studies in the respective class) in different %MDD classes as per Cabrera *et al.* [19] (side legend).

Endpoint considered: foraging.

Settings for modelling

Transformation
Desired MDD %C:
Replicates
Significance level alpha

Endpoints of models

Brood Index
Compensation Index
Brood Termination Rate [%]

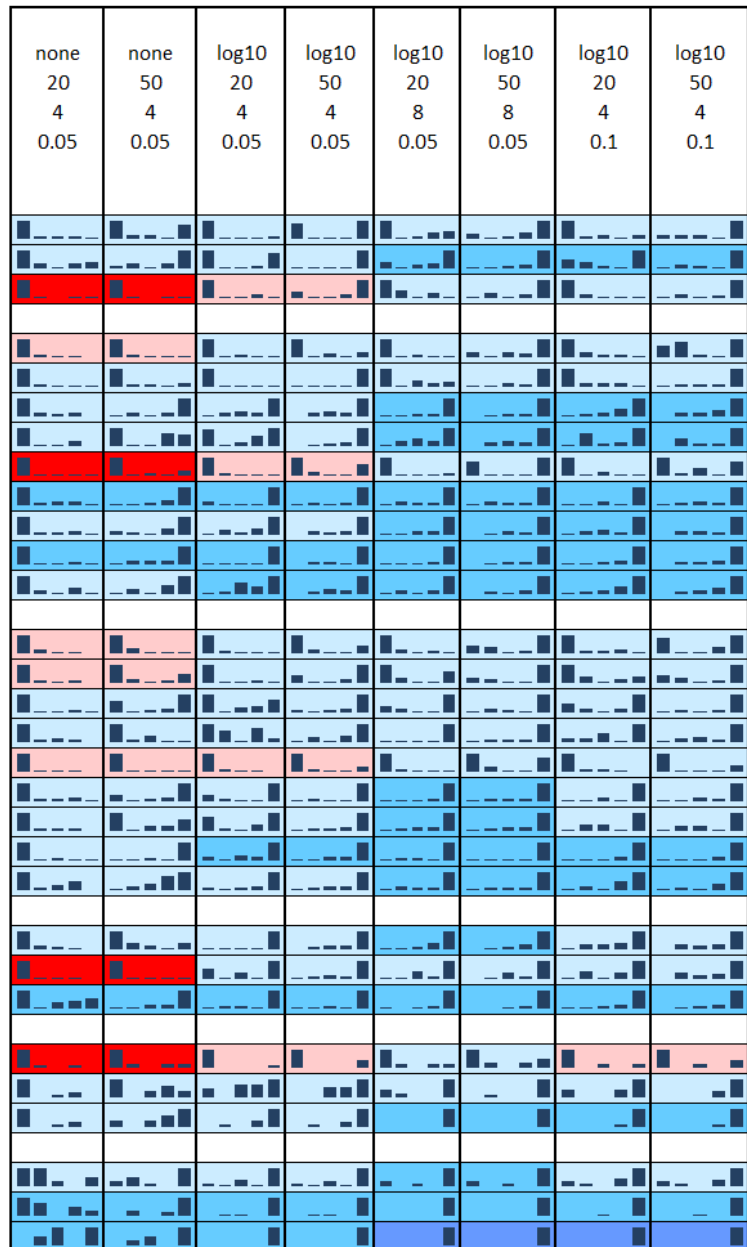
No. of cells with eggs at BFD21
No. of cells with larvae at BFD21
No. of cells with pupae at BFD21
No. of cells with nectar at BFD21
No. of cells with pollen at BFD21
No. of cells with brood at BFD21
No. of cells with food at BFD21
No. of empty cells at BFD21
No of honeybees at BFD21

Difference of No. of cells with eggs at BFD21-BFDO
Difference of No. of cells with larvae at BFD21-BFDO
Difference of No. of cells with pupae at BFD21-BFDO
Difference of No. of cells with nectar at BFD21-BFDO
Difference of No. of cells with pollen at BFD21-BFDO
Difference of No. of cells with brood at BFD21-BFDO
Difference of No. of cells with food at BFD21-BFDO
Difference of No. of empty cells at BFD21-BFDO
Difference of No of honeybees at BFD21-BFDO

Mortality exposure phase
Mortality at day of application
Foraging exposure phase

Bombus mortality
Bombus gyne production
Bombus weight

Osmia cocoon production
Osmia nest occupation
Osmia offspring



Mean actual %MDD of C
from (>) to (<=) colour in map
0 10
10 25
25 50
50 75
75 120

Bars indicate power-classes from left to right
<=60 >60<=70 >70<=80 >80<=90 >90

Fig. 7 Heat-map of MDD (color coding using Cabrera *et al.* [19] scale and power distribution (bars in the cells) for different endpoints and species according to the model selected (table header)).

Heat-map containing %MDD values between 8.3% and 119.0%. Bars in each cell indicate the power distribution of the *t*-tests for each endpoint according to the model settings and the population variance found in the “real world” data set from actual studies.

Which Endpoints Can Be Reliably Assessed in Semi-field Pollinator Species Testing without Estimating False Positive or False Negative? MDD's and Replicates Issue

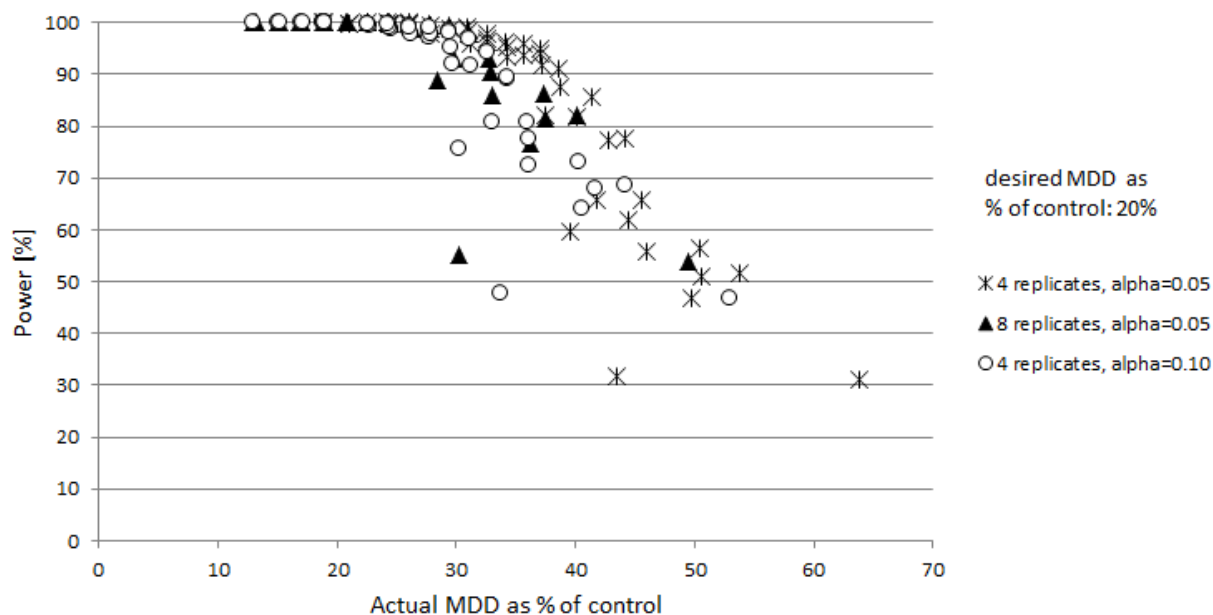


Fig. 8 Power (y-axis) and actual %MDD (x-axis) of single studies depending on number of replicates and level of alpha for honeybee mortality with a desired %MDD of 20% of control value.

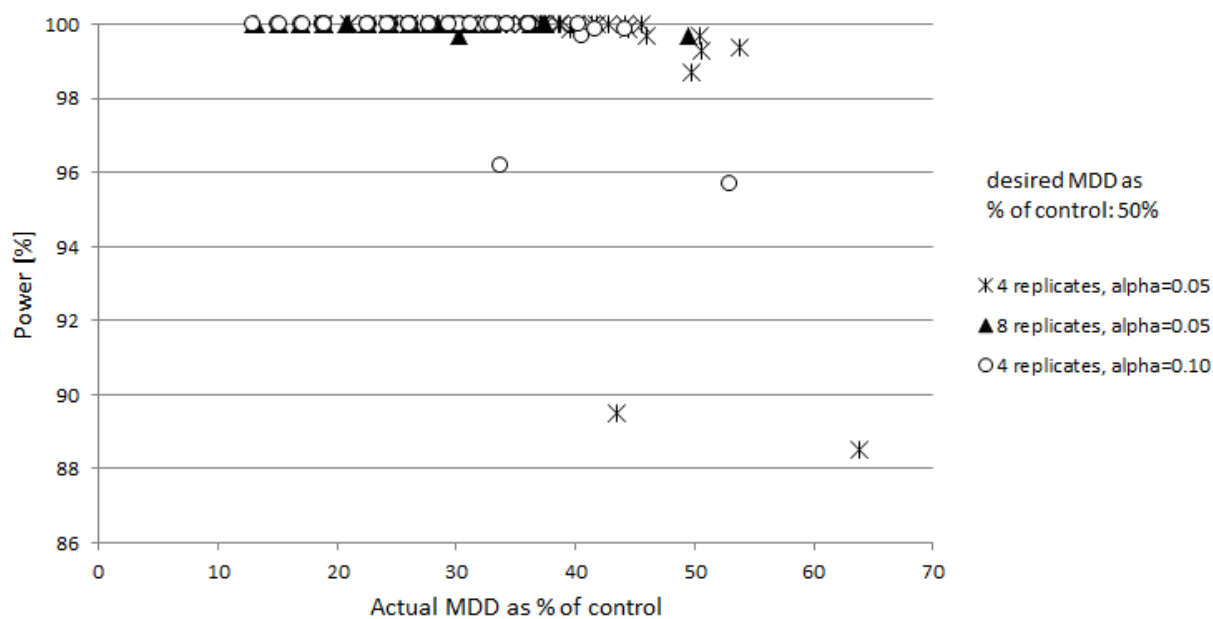


Fig. 9 Power (y-axis) and actual %MDD (x-axis) of single studies depending on number of replicates and level of alpha for honeybee mortality with a desired %MDD of 50% of control value.

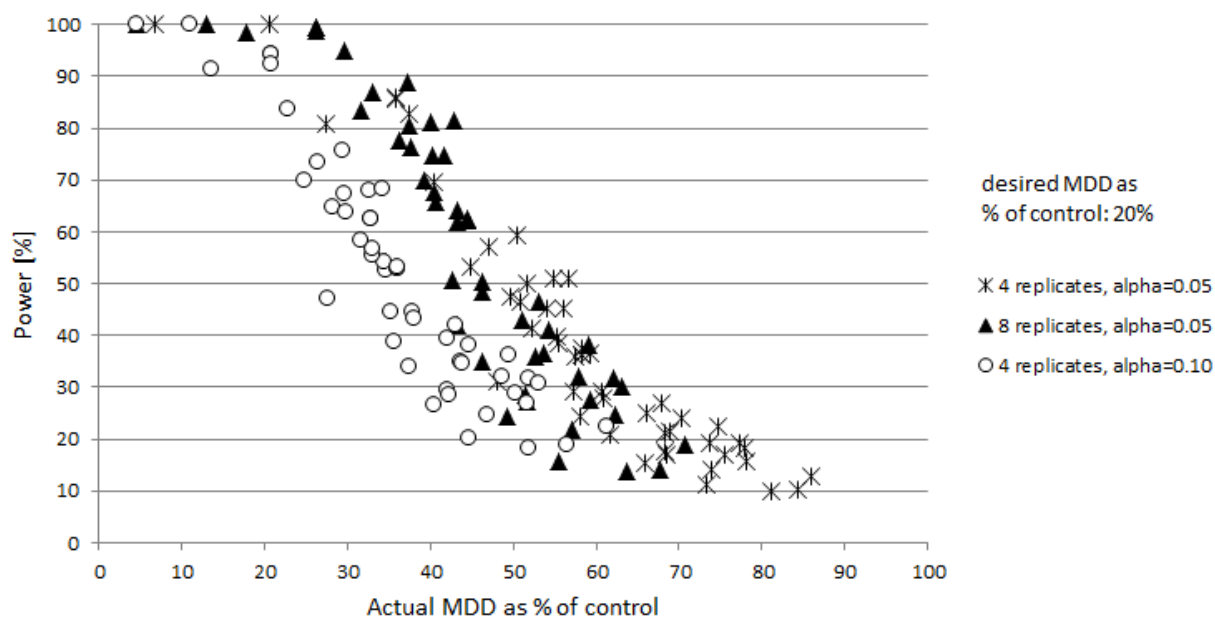


Fig. 10 Power (y-axis) and actual %MDD difference (x-axis) of single studies depending on number of replicates and level of alpha for the endpoint honeybee brood termination rate (BTR) at BFD22 \pm 1 (= 22 \pm 1 d after the initial “brood fixing day” (BFD)) for desired %MDD of 20% of control value.

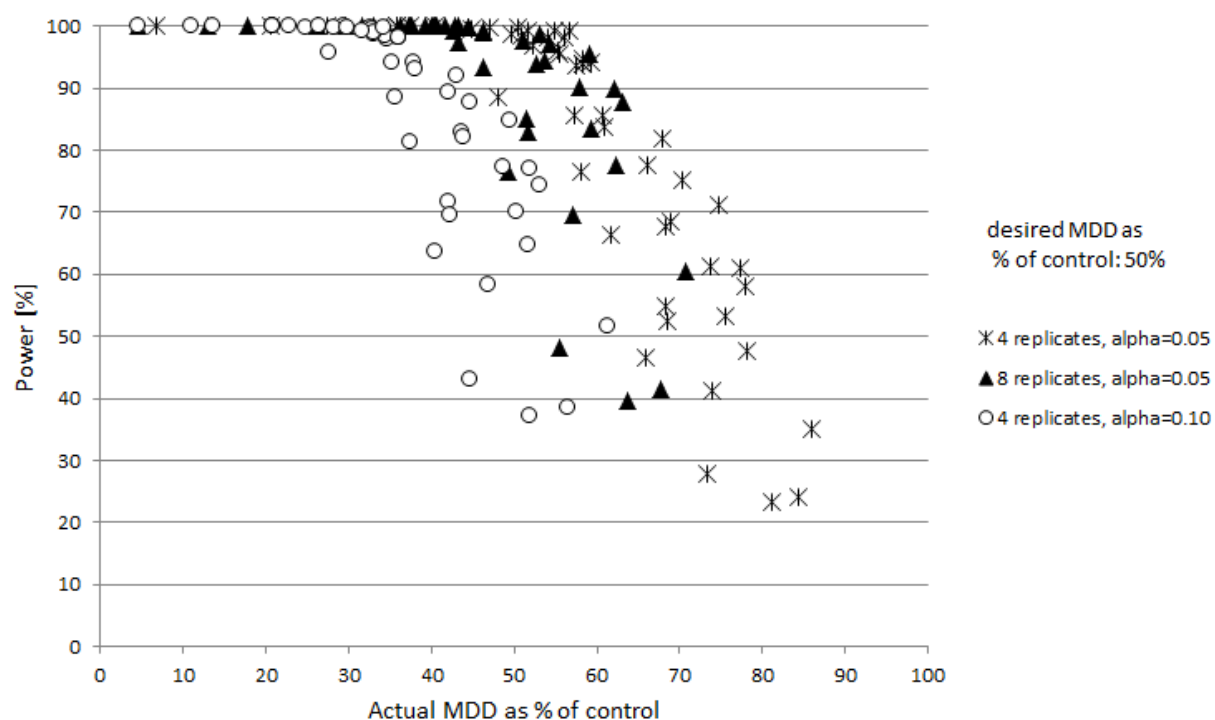


Fig. 11 Power (y-axis) and actual %MDD difference (x-axis) of single studies depending on number of replicates and level of alpha for the endpoint honeybee BTR at BFD22 \pm 1 (= 22 \pm 1 d after the initial BFD) for desired %MDD of 50% of control value.

Which Endpoints Can Be Reliably Assessed in Semi-field Pollinator Species Testing without Estimating False Positive or False Negative? MDD's and Replicates Issue

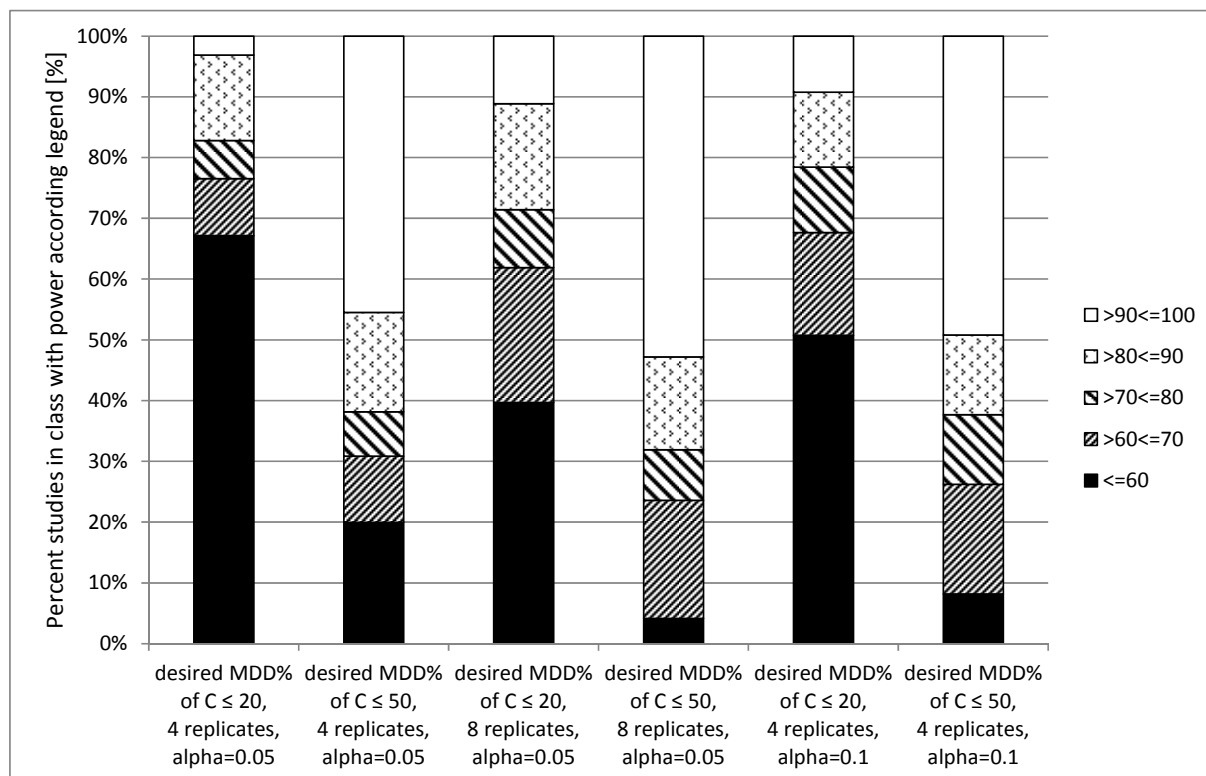


Fig. 12 Distribution of studies (y-axes) to power classes (legend on the right side of the figure) depending on desired MDD in percent of control (20% or 50%), number of replicates (four or eight) and level of alpha (0.05 or 0.1) for endpoint honeybee BTR at BFD22 ± 1.

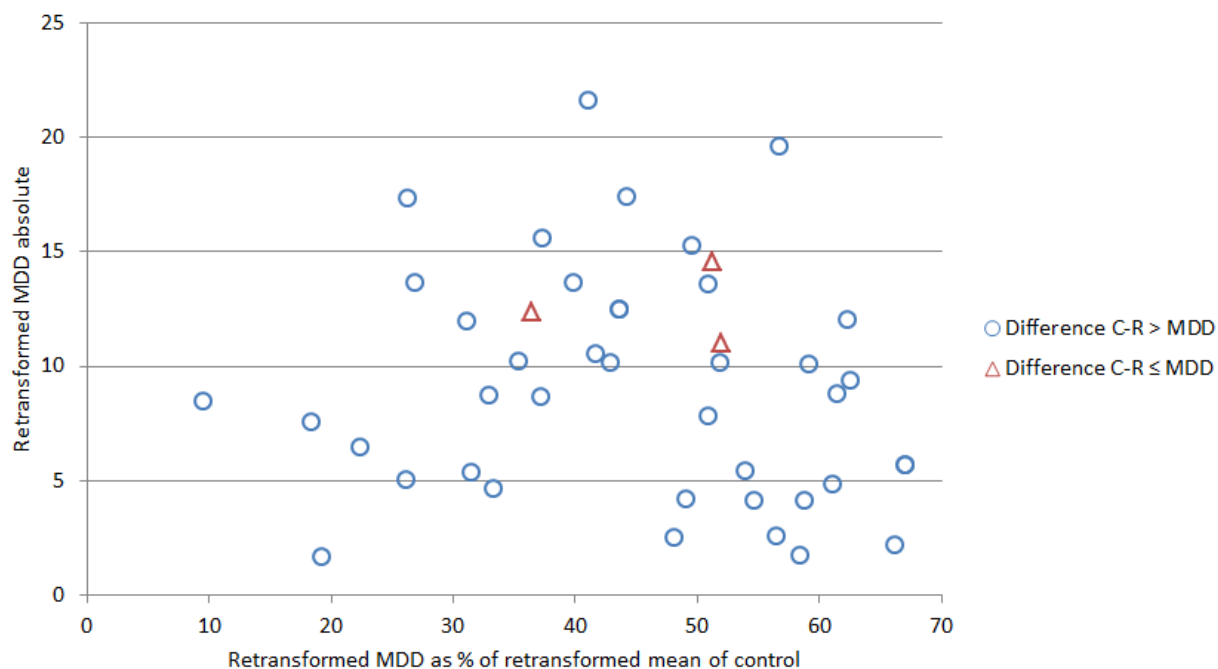


Fig. 13 Retransformed absolute MDD from log-transformation (y-axis) and retransformed MDD as percent of mean of control (x-axis) of single studies, with indication if toxic reference was statistically significantly different from control for endpoint honeybee BTR at BFD22 ± 1.

41 out of 44 data sets showed higher differences between control (C) and reference substance (R) than the MDD.

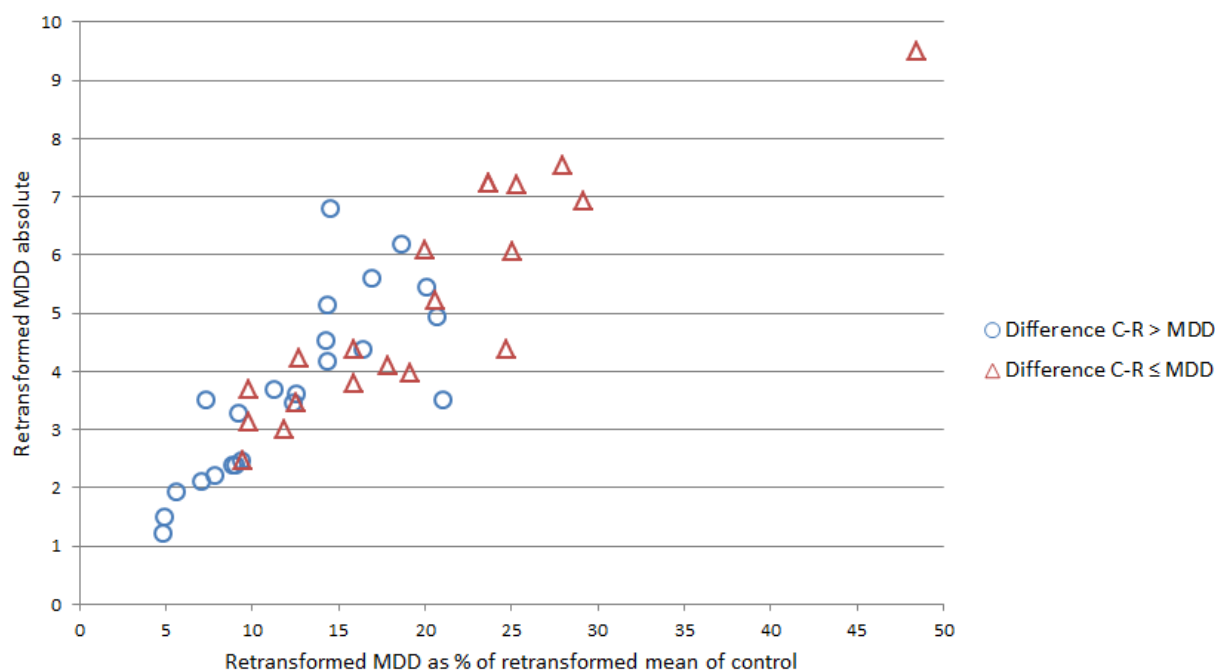


Fig. 14 Retransformed absolute MDD from log-transformation (y-axis) and retransformed MDD as percent of mean of control (x-axis) of single studies, with indication if toxic reference was statistically significantly different from control for the endpoint “difference between number of brood cells day BFD22 \pm 1 to BFD0”.

24 out of 44 data sets showed higher differences between control (C) and reference substance (R) than the MDD.

Table 1 Study design of the semi-field tunnel studies conducted between 2013 and 2017 with *Apis mellifera*, *Bombus terrestris* and *Osmia bicornis* and studies used for statistical analyses presented in this paper.

Test design/test organism	<i>Apis mellifera</i> Honeybee	<i>Bombus terrestris</i> Bumble bee	<i>Osmia bicornis</i> Solitary bee–red mason bee
Crop	<i>Phacelia tanacetifolia</i> or winter oil seed rape	<i>Phacelia tanacetifolia</i>	<i>Phacelia tanacetifolia</i> , or winter oil seed rape
Experimental unit	Tunnel	Tunnel	Tunnel
Size of exp. unit (m ²)	80-100	60	60
No. of replicates	4-6 (Few studies with only three replicates)	4-6	4
Initial colony size: min/max/mean (no. of adults)	2,405/15,665/7,443	32/189/82	-
Initial population size (no. of cocoons)	-	-	60 females/90 males
Exposure period in exp. units (no. of days)	7	Min. 14 up to 29	Min. 20 up to 50
Post-exposure period (no. of days)	Up to 21	Up to 23	6-8 months
No. of conducted studies	50	7	10
Guideline	OECD75 [13]	No guideline, based on Cabrera <i>et al.</i> [19]	No guideline

The three important “shortcomings” of the MDD approach are: (1) absolute and %MDD difference to the control, (2) dependency of the evaluation between absolute and %MDD difference to the control and (3) biological relevance. Two examples from the data set

(Figs. 13 and 14) and theoretical result assumptions (Table 4) were used to outline these points. Point 1) as shown in Fig. 13, not only %MDD values but also absolute values and statistical significance should be plotted to be able to make correct data interpretation

and conclusions (see Figs. 13 and 14 on how results can expressively vary depending on the endpoint). Point 3) the question here is: will study A (absolute MDD 15.9 and MDD in percent to control 299.1%) or

study B (absolute MDD 2.1 and MDD in percent to control 4.7%) show an effect of the reference treatment more clearly? The answer to this last point will be addressed in the discussion.

Table 2 Endpoints assessed in the semi-field tunnel studies (and reported in this paper) conducted with *Apis mellifera*, *Bombus terrestris* and *Osmia bicornis*.

Species	Endpoint assessed and presented in this paper	Definition/method
<i>Apis mellifera</i>	Brood index (BI)	Measure of the brood development assessed via photo-method, at least 200 cells containing eggs per colony, pictures analyzed with HiveAnalyzer® [21].
	Compensation index (CI)	Indicator for colony recovery assessed via photo-method, at least 200 cells containing eggs per colony, pictures analyzed with HiveAnalyzer® [21].
	Brood termination rate (BTR)	Failure of the brood development assessed via photo-method, at least 200 cells containing eggs per colony, pictures analyzed with HiveAnalyzer® [21].
	No. of brood cells	Number of cells with eggs, larvae or pupae estimated according to the Liebefeld method [22].
	No. of storage cells	Number of cells with nectar or pollen estimated according to the Liebefeld method [22].
	No. of empty cells	Number of empty cells estimated according to the Liebefeld method [22].
	Colony strength	Number of bees in the colony estimated according to Imdorf, A. <i>et al.</i> [22].
	Foraging activity	Number of bees foraging on the crop recorded on a 1 m ² area, at different places in each tunnel [13].
<i>Bombus terrestris</i>	Mortality	Number of dead bees collected in the bee traps and on linen sheets laid out in the tunnels [13].
	Mortality	Cumulative number of dead larvae and adults within 7 d after application collected inside the hives and on linen sheets laid out in the tunnels.
	No. of gynes (young produced queens)	Sum of queen larvae, queen pupae and emerged young queens counted after deep-freezing of the colony at the end of the study.
<i>Osmia bicornis</i>	Colony weight	Weight increase of the colony (brood nest and adult bees) during the exposure phase. Hives were weighed every 2-3 d during the exposure phase.
	Nest occupation	The percentage of released females actively nesting. Nesting females were counted at night, when bees rested in their nesting cavities.
	No. of produced cocoons	Number of cocoons produced per nesting female. Cocoons were collected and counted after maturation.
	No. of offspring	The percentage of offspring (2nd generation) emerging from produced cocoons. Cocoons were overwintered and bees emerging from cocoons were counted the following spring.

Table 3 Classes of minimum detectable differences (MDD) as percent of mean of control (%MDD) for benchmarks according to Cabrera *et al.* [19] and Brock *et al.* [20].

MDD as % of the mean of the control for different effect classes					
Class	1	2	3	4	5
Cabrera <i>et al.</i> (2016)	≤ 10	> 10 ≤ 25	> 25 ≤ 50	> 50 ≤ 75	> 75
Brock <i>et al.</i> (2015)	≤ 50	> 5 ≤ 70	> 70 ≤ 90	> 90 ≤ 100	> 100

Table 4 Fictive study results and respective MDD calculations for a study with two treatments (control C and reference R) and four replicates per treatment (C 1-4 and R 1-4).

Study	C1	C2	C3	C4	R1	R2	R3	R4	MDD Actual difference to control	%MDD
A	2.5	3.8	15	0	86	98	52	100	15.9	299.1
B	45	48	43	44	55	58	53	54	2.1	4.7

4. Discussion

The %MDDs observed in honeybee tunnel studies varied greatly depending on the endpoints measured. Some endpoints such as foraging (counted endpoint) and colony strength assessment (estimated endpoint) have very low %MDD (11%-20%) while other endpoints such as brood index and termination rate (calculated endpoints based on extremely precise photo assessments) have much higher %MDDs (32%-48%). This is surprising, since lower %MDDs were expected where the assessed endpoints were measured very precisely (e.g., termination rate with photo measurements) and not on endpoints where the assessment was made by estimating and not precise counting or measurement of the endpoint (e.g., colony strength where the number of bees was calculated by estimating the % coverage of the frames with bees) [22].

Evaluation of honeybee field monitoring studies [23] reported %MDD for colony strength in the range of 15%-21%, which is in the same range as recorded in the semi-field brood studies. For the assessment of the amount of brood, Rolke *et al.* [23] reported %MDD of 30%-40%, while the average %MDD in the 50 honeybee studies presented in this research was much lower, around 20%. In Rolke *et al.* [23] as well as in this study, colony strength as well as the amount of brood assessment endpoints were evaluated with the same methods, therefore the results are directly comparable, showing the influence of the test design on the %MDD detectable (field monitoring vs. tunnel semi-field study design).

None of the honeybee endpoints had %MDDs < 10%, confirming the critics of Bakker [10] to the 7% threshold value proposed by EFSA [1, 2] for such biological systems. However, most of the %MDDs were between 10% and 30%, indicating clearly that the experimental design of the semi-field honeybee study allowed for the determination of relatively small effects on key study endpoints.

The MDDs for bumble bee semi-field tunnel study endpoints, varied between 17% and more than 53% depending on the endpoint, thus being in a similar range as the honeybees %MDDs. The number of gynes (produced young queens), which is probably the most important endpoint in these types of studies, was slightly below 25%. This is a very low value compared to the %MDD observed in bumble bee field monitoring studies conducted in Germany by Sterk *et al.* [24], where the %MDD was 45%. This difference may indicate an effect of the test design on the resulting MDDs. The %MDDs calculated in this study were around 20% (with exception of bumble bee mortality with a %MDD of > 50%), indicating clearly that the experimental design of the semi-field bumble bee study allowed for the determination of relatively small effects on key endpoints.

Solitary bee semi-field studies seem to have the lowest %MDDs compared to honey and bumble bee endpoints. The %MDDs for the measured endpoints were all near 10%, which is a very low value for these types of semi-field tunnel study designs. In contrast to honeybees and bumble bees, there is no social structure for the solitary bees influencing behavior and/or exposure. For *Osmia bicornis* field monitoring studies conducted in Germany [25], reported much higher %MDDs for the same endpoints as in the semi-field studies presented in this research with %MDD between 27%-39% and 53%-64%, for emergence and nest occupation, respectively.

In general it can therefore be stated that semi-field bee species test designs detect much lower differences between the control and PPP treatments when compared to field studies. This is certainly due to the fact that more test-design specific parameters can be standardized in semi-field trials when compared to field trials [4, 5], which has a direct effect on obtained results variability and ability to detect treatment dependent effects.

Measured and simulated data and results showed that statistical evaluations parameter selection (e.g.,

alpha value), data transformation (log10) and the number of replicates had a direct effect on the ability of the test design to detect lower or higher %MDD values. However, the magnitude of this effect depends on the endpoint.

It is very much debated if this kind of PPP side effect studies or in general biological data should be evaluated at an alpha level of 0.05 or other alpha values [26]. It could be shown that change of alpha value from 0.05 to 0.1, increases the ability of the studies to detect lower %MDDs. By increasing the alpha more false positive will occur, so it has to be decided if this can be acceptable or not. To avoid this, dose-response studies are certainly an improvement in performing such studies. Example of dose-response field studies were described by Overmyer *et al.* [27] where a new test design for honey bee colony field feeding study was described to better link dose and effects in higher tier pollinator studies.

Log10 transformation of the data helps especially in case non-normal distributed data [28]. A clear effect could be seen for e.g., for the endpoints BTR, number of honeybee hives cells filled with pollen and bumble bee mortality (Fig. 7). For other endpoints the transformation is so to say “less efficient” but still brings lower %MDDs. It is always suggested to transform the data before analysis. Type of transformation should be based on the distribution of data (normal, log-normal or others). A one-sided statistical evaluation/test has to be preferred compared to a two-sided evaluation. In principle negative effects of the PPP are expected and therefore, one-sided effect detection statistics is the correct approach. This decreases the %MDDs observed in all assessed endpoints, which is desirable.

For most of the measured endpoints, increasing the number of replicates e.g., from four to eight, slightly improves the power of the test design by decreasing the %MDD. The reduction magnitude of the %MDD is dependent on the endpoint and selection of statistical parameters such alpha (e.g., 0.05 or 0.1).

The modeling presented considered two replicate scenarios: four and eight replicates. These two test designs were selected on the bases of commonality (usually such studies are conducted with four replicates, the tunnel being the replicate) and practicability (eight replicates is probably the absolute maximum of replicates that can be handled considering the man power needed to run such trials, often conducted with several treatments). The selection of the test design is finally dependent on the %MDD desired in the experiment.

Interpretation of results depends extremely on the scale used to assess and interpret the %MDDs, e.g., using the scale/effect classes proposed for bumble bees by Cabrera *et al.* [19] which focuses on separating more effect categories < 50% or the scale used in aquatic studies proposed by Brock *et al.* [20] that focuses on assessing/evaluating more in detailed effect classes with %MDD > 50%. Since it appears that the test designs currently used for semi-field pollinators' studies are able to detect low differences, it is suggested to use the Cabrera *et al.* [19] scale to categorize %MDD.

Data endpoints in ecotoxicology (semi-) field studies are of very diverse structure. Parameters that display effects in a biologically relevant size will be a better indicator for effects than parameters that are able to detect minor differences at a biologically not relevant scale. In other words, it is always key to discuss also the biological relevance of the effects and not blindly accept minor statistical effects as biologically relevant, e.g., for the survival and performance of the bee colonies. Cox [29, 30] stated that statistical significance is quite different from scientific significance. Therefore, estimation of the magnitude of effects is essential, regardless of whether statistically significant departure from the null hypothesis is achieved. The importance of the statistical analysis in evaluating bee semi-field or field studies [1, 2] is at the risk of being overestimated. The data set should always be plotted to select optimized

settings for statistical analysis. There is much more than just p -values and power in a data set.

If good knowledge of the biology of the test organism, the ecology of its environment and the possibilities and limitations of statistical data analysis are combined, most (semi-) field studies can add significant knowledge to the risk assessment of PPP. Expertise and knowledge on all of these aspects is necessary to create study designs that allow for generation of meaningful data, selection of relevant endpoints and appropriate interpretation of the results. As shown in this publication, each endpoint has its very own assets and limitations for the reliable detection of risks, for dealing with the variability of measured data and for evaluation of the impact of toxic signals on the size of the effect. Thus, the requirements for appropriate statistical methods, the minimum number of replicates and the maximum acceptable %MDD of the control should be defined individually per endpoint to allow for conclusions that are both scientifically sound and biologically relevant.

5. Conclusions

The MDDs are a good statistical tool to evaluate and interpret semi-field pollinator studies. However, assessment of both scientifically sound and biologically relevant data needs to be done for a proper interpretation of the result.

The %MDDs observed in OECD75 honeybee studies varied depending on the endpoint assessed. Most of the %MDDs were, however, between 10% and 30%, indicating clearly that the experimental design of the semi-field honeybee study allowed for the determination of relatively small effects on key study endpoints. None of the honeybee endpoints had %MDDs < 10%, confirming the critics to the 7% threshold value proposed by EFSA for such biological systems.

The measured and simulated honeybee data and results showed that statistical evaluations parameter selection (alpha value e.g., 0.05 or 0.1), data

transformation (log10) and the number of replicates had a direct effect on the ability of the test design to detect lower or higher %MDD values. The magnitude of this effect depends on the endpoint. For most of the measured endpoints, increasing the number of replicates e.g., from four to eight, slightly improves the power of the test design by decreasing the %MDD.

The MDDs for bumble bee semi-field tunnel study endpoints, varied between 17% and more than 50% depending on the endpoint, thus being in a similar range as the honeybees %MDDs.

Solitary bee semi-field studies showed the lowest %MDDs compared to honeybees and bumble bee endpoints. The %MDDs for the measured endpoints were all near 10%, which is a very low value for these types of semi-field tunnel study designs.

The *O. bicornis* semi-field test system seems to be the most robust study type when compared to *A. mellifera* and *B. terrestris*.

Acknowledgments

The authors would like to thank Jane Sanotra for reviewing the manuscript.

References

- [1] European Food Safety Authority (EFSA). 2013. "EFSA Guidance Document on the Risk Assessment of Plant Protection Products on Bees (*Apis mellifera*, *Bombus* spp. and Solitary Bees)." *EFSA Journal* 11 (7): 3295.
- [2] European Food Safety Authority (EFSA). 2014. "Guidance on the Risk Assessment of Plant Protection Products on Bees (*Apis mellifera*, *Bombus* spp. and Solitary Bees)." *EFSA Journal*. Accessed January 27, 2015. <http://www.efsa.europa.eu/en/efsajournal/pub/3295>.
- [3] Miles, M. 2013. "Bee Guidance Documents: An End Users View." *Special SETAC Symposium on Pesticide Risk for Pollinators: Testing Methodologies, Risk Assessment and Risk Management*. Accessed March 1, 2017. http://sesss08.setac.eu/embed/sesss08/Mark_Miles_Bee_guidance_documents_end_users_view.pdf.
- [4] Candolfi, M., Bigler, F., Campbell, P., Heimbach, U., Schmuck, R., Angeli, G., Bakker, F., Brown, K., Carli, G., Dinter, A., Forti, D., Forster, R., Gathmann, A., Hassan,

- S., Mead-Briggs, M., Melandri, M., Neumann, P., Pasqualini, E., Powell, W., Reboulet, J. N., Romijn, K., Sechser, B., Thieme, T. H., Ufer, A., Vergnet, C. H., and Vogt, H. 2000a. "Principles for Regulatory Testing and Interpretation of Semi-field and Field Studies with Non-target Arthropods." *J. Pest Science* 73 (6): 141-7.
- [5] Candolfi, M. P., Blümel, S., Forster, R., Bakker, F. M., Grimm, C., Hassan, S. A., Heimbach, U., Mead-Briggs, M. A., Reber, B., Schmuck, R., and Vogt, H. (eds.) 2000b. *Guidelines to Evaluate Side-Effects of Plant Protection Products to Non-target Arthropods*. IOBC, BART and EPPO Joint Initiative, IOBC/WPRS Publisher, 158.
- [6] Carreck, N. L., and Ratnieks, F. L. W. 2014. "The Dose Makes the Poison: Have "Field Realistic" Rates of Exposure of Bees to Neonicotinoid Insecticides Been Overestimated in Laboratory Studies?" *Journal of Apicultural Research* 53 (5): 607-14. doi: 10.3896/IBRA.1.53.5.08.
- [7] Romeis, J., Hellmich, R. L., Candolfi, M. P., Carstens, K., de Schrijver, A., Gatehouse, A. M. R., Herman, R. A., Huesing, J. E., McLean, M. A., Raybould, A., Shelton, A. M., and Waggoner, A. 2011. "Recommendations for the Design of Laboratory Studies on Non-target Arthropods for Risk Assessment of Genetically Engineered Plants." *Transgenic Res* 20: 1-22.
- [8] Henry, M., Cerrutti, N., Aupinel, P., Decourtye, A., Gayraud, M., Odoux, J-F., Pissard, A., Rueger, C., and Bretagnolle, V. 2015. "Reconciling Laboratory and Field Assessments of Neonicotinoid Toxicity to Honeybees." *Proc. R. Soc. B* 282 (1819): 20152110. Accessed November 23, 2017. <http://dx.doi.org/10.1098/rspb.2015.2110>.
- [9] Heimbach, F., Schmuck, R., Grünwald, B., Campbell, P., Sappington, K., Steeger, T., and Davies, L. P. 2017. "The Challenge: Assessment of Risks Posed by Systemic Insecticides to Hymenopteran Pollinators: New Perception When We Move from Laboratory via Semi-field to Landscape Scale Testing?" *Environ. Toxicol. Chem.* 36 (1): 17-24.
- [10] Bakker, F. 2015. "Design and Analysis of Field Studies with Bees: A Critical Review of the Draft EFSA Guidance." *Integr. Environ. Assess. Manag.* 12: 422-8. doi:10.1002/ieam.1716.
- [11] Hurlbert, S. H. 1984. "Pseudoreplication and the Design of Ecological Field Experiments." *Ecological Monographs* 54 (2): 187-211.
- [12] Lemoine, N. P., Hoffman, A., Felton, A., Baur, L., Chaves, F., Gray, J., Yu, Q., and Smith, M. D. 2016. "Underappreciated Problems of Low Statistical Power in Ecological Field Studies." *Ecology* 97: 2554-61.
- [13] OECD. 2007. "Guidance Document on the Honeybee (*Apis mellifera* L.) Brood Test under Semi-field Conditions." *OECD Environment, Health and Safety Publications, Series on Testing and Assessment No. 75, ENV/JM/MONO*, 22.
- [14] OEPP/EPPO. 2010. "Guideline for the Efficacy Evaluation of Plant Protection Products–Side Effects on Honeybees." *Bulletin OEPP/EPPO Bulletin* 40, 313-19.
- [15] Cumming, G. 2014. "The New Statistics: Why and How." *Psychological Science* 25 (1): 7-29. Accessed November 16, 2013. <http://www.sagepub.com/journalsPermissions.nav>.
- [16] Wasserstein, R. L., and Lazar, N. A. 2016. "The ASA's Statement on *p*-Values: Context, Process and Purpose." *The American Statistician* 70 (2): 129-33. doi: 10.1080/00031305.2016.1154108.
- [17] de Winter, J. C. F. 2013. "Using the Student's *t*-test with Extremely Small Sample Sizes." *Practical Assessment, Research & Evaluation* 18 (10). Available online: <http://pareonline.net/getvn.asp?v=18&n=10>.
- [18] Zar, J. H. ed. 1999. *Biostatistical Analysis*. 4th edition. Englewood Cliffs: Prentice Hall.
- [19] Cabrera, A. R., Almanza, M. T., Cutler, G. C., Fischer, D. L., Hinarejos, S., Lewis, G., Nigro, D., Olmstead, A., Overmyer, J., Potter, D. A., Raine, N. E., Stanley-Stahr, C., Thompson, H., and van der Steen, J. 2016. "Initial Recommendations for Higher-Tier Risk Assessment Protocols for Bumble Bees, *Bombus* spp. (Hymenoptera: Apidae)." *Integr. Environ. Assess. Manag.* 12: 222-9. doi:10.1002/ieam.1675.
- [20] Brock, T. C., Hammers-Wirtz, M., Hommen, U., Preuss, T. G., Ratte, H. T., Roessink, I., Strauss, T., and van den Brink, P. 2015. "The Minimum Detectable Difference (MDD) and the Interpretation of Treatment-Related Effects of Pesticides in Experimental Ecosystems." *Environ. Sci. Pollut. Res. Int.* 22: 1160-74.
- [21] Höferlin, B., and Höferlin, M. 2015. HiveAnalyzer. Version 1.32. <http://hiveanalyzer.visionalytics.de>.
- [22] Imdorf, A., Bühlmann, G., Gerig, L., Kilchenmann, V., and Wille, H. 1987. "Reviewing the Estimation Method for Determining the Brood Area and the Number of Female Workers in Free Flying Bee Colonies." *Apidologie* 18: 137-46.
- [23] Rolke, D., Fuchs, S., Grünwald, B., Gao, Z., and Blenau, W. 2016. "Large-Scale Monitoring of Effects of Clothianidin-Dressed Oilseed Rape Seeds on Pollinating Insects in Northern Germany: Effects on Honey Bees (*Apis mellifera*)." *Ecotoxicology* 25: 1648-65.
- [24] Sterk, G., Peters, B., Gao, Z., and Zunkier, U. 2016. "Large-Scale Monitoring of Effects of Clothianidin-Dressed Oilseed Rape Seeds on Pollinating Insects in Northern Germany: Effects on Bumble Bees (*Bombus terrestris*)." *Ecotoxicology* 25: 1666-78.
- [25] Peters, B., Gao, Z., and Zunkier, U. 2016. "Large-Scale Monitoring of Effects of Clothianidin-Dressed Oilseed

Rape Seeds on Pollinating Insects in Northern Germany: Effects on Mason Bees (*Osmia bicornis*).” *Ecotoxicology* 25: 1679-90.

- [26] Harcum, J. B., and Dressing, S. A. 2015. “Technical Memorandum #3: Minimum Detectable Change and Power Analysis.” Developed for U.S. Environmental Protection Agency by Tetra Tech, Inc., Fairfax, VA, 10.
- [27] Overmyer, J., Feken, M., Ruddle, N., Bocksch, S., Hill, M., and Thompson, H. 2018. “Thiamethoxam Honey Bee Colony Feeding Study: Linking Effects at the Level of the Individual to Those at the Colony Level.” *Environmental Toxicology and Chemistry* 37 (3): 816-28.
- [28] Limpert, E., Stahel, W. A., and Abbt, M. 2001.

“Log-Normal Distributions across the Sciences: Keys and Clues: On the Charms of Statistics, and How Mechanical Models Resembling Gambling Machines Offer a Link to a Handy Way to Characterize Log-Normal Distributions, Which Can Provide Deeper Insight into Variability and Probability-Normal or Log-Normal: That is the Question.” *BioScience* 51 (5): 341-52. Accessed April 11, 2018. [https://doi.org/10.1641/0006-3568\(2001\)051\[0341:LNDATS\]2.0.CO;2](https://doi.org/10.1641/0006-3568(2001)051[0341:LNDATS]2.0.CO;2).

- [29] Cox, D. R. 1977. “The Role of Significance Tests.” *Scandinavian Journal of Statistics* 4: 49-70.
- [30] Cox, D. R. 1986. “Some General Aspects of the Theory of Statistics.” *International Statistical Review* 54: 117-26.