Annex to:

EFSA (European Food Safety Authority), Adriaanse P, Arce A, Focks A, Ingels B, Jölli D, Lambin S, Rundlof M, Süßenbach D, Del Auguila M, Ercolano V, Ferilli F, Ippolito A, Szentes C, Neri F M, Padovani L, Rortais A, Wassenberg J and Auteri D, 2023. Revised guidance on the risk assessment of plant protection products on bees (*Apis mellifera*, *Bombus* spp. and solitary bees). EFSA Journal 2023;21(5):7989, 76 pp. doi:10.2903/j.efsa.2023.7989

Annex C – Recommendations for higher tier effect studies

1. Introduction, general considerations

When the lower tier risk assessments indicate a concern, higher tier effect study/studies should be conducted. In this guidance document, three types of higher tier effect studies are considered: field studies, semi-field studies and colony feeder studies. However, the colony feeder study only applies to social bees that form colonies. In this Section, it is explained under which conditions a certain study methodology is recommended alongside the main considerations to be fulfilled when the study is conducted. The Working Group has largely chosen existing protocols derived from internationally agreed and adopted guidelines and included recent developments in this area since the publication of the previous guidance document (EFSA (European Food Safety Authority), 2013). Nevertheless, as in the previous guidance document, substantial modifications are recommended to ensure that the studies address the regulatory requirements that are driven by the specific protection goal (SPG).

In all cases, the aim is to demonstrate that colony strength or population abundance is not impacted more than the magnitude dimension of the SPG after a field realistic worst-case exposure to the plant protection product (PPP) for the use under evaluation. The dynamics of the colony strength/population abundance under field realistic exposure conditions can only be studied in field studies in typical agricultural settings. Therefore, the general rule is that if lower tiers indicate concerns, conducting a field study would be the next step. However, there are other study types that could be considered under specific circumstances, which are described in the specific Sections below. Despite the fact that multiple higher tier tests are possible, field studies remain the option with the highest ecological realism and, as such, the results from field tests will generally overrule the results of any other test.

In all cases, the outcome of certain measurements (i.e. endpoints) from the exposed group(s) is compared with control group(s). All conditions of the exposed and control groups should be sufficiently similar with the only important difference being that the exposed group is indeed exposed to the PPP under evaluation, while the control group is not.

It is important to document the process, considerations and outcomes of the study design and execution, as well as to document, explain, and interpret the importance of deviations from the original study plan. Due to the complexity of the higher tier effect studies, the applicants are recommended to submit the specific protocols to the national competent authority for review prior to initiating any kind of effect study.

2. Statistical methodology for higher tier studies

This Section describes the statistical methods that can be used for the analysis of higher tier studies. EFSA (European Food Safety Authority) (2013)The 'test of equivalence' is considered as the most appropriate approach for the analysis, replacing the standard 'test of difference' in EFSA (European Food Safety Authority) (2013). In the Section 2.1, the principles underlying the test of difference and the related limitations are presented; the principle of equivalence testing is presented to show how it can solve the issues inherent to the test of difference. This is followed by a series of examples that show how the interpretation of a study changes when shifting from the old 'hypothese assumptions' to the new. Section 2.2 explains the methodology to be used for an undefined threshold. Section 2.3 includes a summary and some additional recommendations.

Some definitions are crucial for the rest of this Section and it is convenient to summarise them here. An *effect* is defined as a detrimental impact of the pesticide, measured by a change of the endpoint of interest in the 'unsafe' direction (here, always positive by convention); the *effect size* is the magnitude of the change. The presence of an effect (a 'risk') corresponds to any effect size greater than zero; 'no effect' ('no risk') corresponds to an effect size equal to or smaller than zero. If there is a defined SPG with associated magnitude Δ (threshold of acceptable effects) an existing risk can be further classified as 'low' if the effect size is smaller than Δ and 'high' if it is larger. As this Section is intended to cover all higher tier studies and bee groups, any specific reference to the endpoint of interest will be avoided; accordingly, Δ (when defined) should be intended as the difference between test and control measured on an appropriate (unspecified) scale. The focus of the statistical analysis (comparison of the test with a control) is on possible relevant effects of the pesticide, with the minimum effect of interest (when defined) being equal to Δ.

2.1. Difference and equivalence testing

The considerations in this Section are general: they are relevant for all higher tier studies and bee groups. However, for ease of presentation, we will refer to the case of a pre-specified safety margin Δ. Hence, the content of the Section is directly applicable only to honey bee field studies, where there is a defined threshold of acceptable effects. With an undefined threshold, the method requires a few changes explained in Section 2.2.

The statistical approach recommended by EFSA (European Food Safety Authority) (2013) is based on the use of a test of difference which compares the test and control for significant differences (effects) with zero as a baseline. The null hypothesis is that the effect size is zero or negative ('no effect'; no risk); the alternative hypothesis is that there is an effect (a risk):

H_0: effect size ≤ 0 (no risk)

H_1: effect size > 0 (risk)

Note that the outcome of this test (significant/not significant) is not enough to support a conclusion. A significant outcome in favour of H_1 would show that a risk exists, but not whether the risk is 'high' or 'low'. A second step is needed, possibly by comparing the magnitude of the estimated effect size with Δ (the threshold of acceptable effects).

This approach has been defined as a 'proof of effect' and it is appropriate when testing for the existence of a risk ('proof of hazard'). When used to test for safety, however, it has well-known intrinsic limitations (Hoenig and Heisey, 2001; Perry et al., 2009; Schumi and Wittes, 2011). A basic principle of hypothesis testing is that the burden of the proof rests on the alternative hypothesis H_1 (effect) and the experiment is designed to build evidence against the null hypothesis H_0 (no effect). If the experimental evidence in favour of an effect is not strong enough, the null hypothesis that the effect size is zero (or negative) cannot be rejected. With this approach, the absence of an effect (a 'no risk' conclusion) can be rejected but *cannot be statistically supported*. Consistently with the approach, the statistical error that can be directly controlled in the analysis is the 'type I' error, or false positive, which occurs when assessing a no-risk substance and wrongly concluding that there is a risk. The rate of type I error, or false positive rate, is denoted by α_D (the significance level of the test) and is set before the analysis (most commonly $\alpha_D = 0.05$).

It has been noted (Perry et al., 2009) that when testing for safety, the primary concern should not be for false positives but rather for false negatives ('type II' errors), which occur when we wrongly conclude that a high-risk substance (effect size $> \Delta$) has no effect (effect size ≤ 0). The type II error rate is denoted by β_D, and $1 - \beta_D$ is defined as the power of the test (the probability of detecting the effect size of interest, here Δ, when such an effect exists). It is difficult, however, to control β_D and impossible to set it to a pre-defined value: the value of β_D is determined not only by parameters that are set prior to the experiment (such as the magnitude of the effect of interest and the sample size) but also by the variability observed in the experiment, which is not known *a priori*. For this reason, a study resulting in a non-significant effect (no risk) has multiple possible interpretations: it may be that the true effect size is zero (so that there is truly no risk); or that the true effect size is undetected but small ($\leq \Delta$, so that there is a risk, but it is low); or, finally, that there is a true effect size of concern ($> \Delta$, high risk) but this was not detected in the experiment, possibly because the statistical power was not high enough. The approach followed in EFSA (European Food Safety Authority) (2013) to deal with this issue for field studies consisted in performing a prospective power analysis, based on plausible estimates of variability, to identify an experimental design (sample size and replication) that could be used to obtain the desired statistical power. There is no guarantee, however, that the level of variability observed in a specific study will be the same as estimated in these calculations. For this reason, EFSA (European Food Safety Authority) (2013) clarified that it is up to the applicant to conduct a *post-hoc* analysis and support that the experiment had an adequate statistical power. This is a standard request for an analysis based on difference testing, and a post-hoc power calculation can provide useful retrospective information once the experiment is completed. Unfortunately, it has been shown that there are severe limitations in the use of such calculations for the interpretation of negative outcomes (Hoenig and Heisey, 2001). It is worth considering, as an example from risk assessment for aquatic organisms, the methods based on the minimum detectable difference (MDD; Brock et al. (2015)) because regulatory guidelines for pesticide risk assessment recommend their use as an indicator of the reliability of non-significant results

(EFSA PPR Panel (EFSA Panel on Plant Protection Products their Residues), 2013). Mair et al. (2020) have shown that the ability of the MDD to discriminate between true and false negatives is in general much lower than the desired statistical power, and that it cannot be interpreted as an upper bound for the plausible effect size. The reason for this, common to many other post-hoc methods (Hoenig and Heisey, 2001), is that the MDD is calculated disregarding the effect size observed in the experiment. Mair et al. (2020) advocated the use of confidence intervals (CIs) to find upper bounds, as a confidence interval combines information about the effect size and its uncertainty: this idea will be used extensively in the rest of this Section.

The issues in the interpretation of non-significant results, and the lack of control of the relevant error rate, are intrinsic to the statistical approach of difference testing. If the main effect size of interest is Δ, to use 'no effect' as a baseline for the comparison can be misleading: the outcome of a test (whether significant or not significant) contains no information about Δ. More generally, the approach is not entirely consistent with the purpose of the study from the point of view of the risk assessor. The goal of a higher tier study should not be to prove that an effect of the pesticide (of *any* magnitude) exists, but to prove with reasonable confidence that the effect (if any) is smaller than Δ. Small significant effects are of no interest. A fit-for-purpose methodology should be able to discriminate between relevant and irrelevant effects, and to support with good confidence a conclusion on the effect size in comparison with Δ. These considerations led the Working Group to consider an approach based on equivalence testing.

In equivalence testing, the aim is not to find a difference between test and control, but rather to test whether the treatment and control are similar (equivalent). 'Similar' means that the difference lies within a pre-defined range of acceptable values (depending on the problem, 'acceptable' may indicate effects that are not biologically relevant or not a concern for safety). An equivalence-based approach (Wellek, 2010) has been identified as the most convenient methodology in many areas of the risk assessment, where the concern is often that the key characteristics of a potential hazard should not differ too much from those of a reference that has been established as safe. It is currently recommended by several regulatory authorities (US FDA (US Food and Drug Administration), 2001; EMA (European Medicines Agency), 2010) as part of the drug approval process: for example, for the analysis of clinical bioequivalence trials comparing two formulations (new and original) of the same drug (Shao et al., 2000). Equivalence testing is also required by EFSA for the safety assessment of genetically modified (GM) plants: in this case, the characteristics of a GM plant are formally compared with an equivalence interval extracted from a set of non-GM commercial reference varieties (EFSA GMO Panel (EFSA Panel on Genetically Modified Organisms), 2010, 2011). In addition, the EFSA Scientific Committee (2011) discussed the advantages of equivalence testing and recommended that "*less emphasis should be placed on the reporting of statistical significance and more on statistical point estimation and associated confidence interval*". Examples of its application are reported in the EFSA scientific Committee Guidance Document (2017), in the EFSA FEEDAP Panel (EFSA Panel on Additives Products or Substances used in Animal Feed) et al. (2017) as well as in Engel and van der Voet (2021). Although this approach is common for the GM assessment, it was never implemented in the area of pesticides. However, in consideration of the criticism around EFSA (European Food Safety Authority) (2013) on higher tier studies, the working group considers it valuable for the evaluation and design of the higher tier studies on bees. In fact, it offers more flexibility for the study design, being potentially highly demanding, e.g. in terms of number of replicates, for PPPs of high concerns while it may be less demanding for low risk PPPs.

Equivalence testing, in the strictest sense, implies the comparison of the effect with an interval (a lower and an upper limit). The appropriate approach for the present problem is the one-sided version of the test, where the effect is compared with an upper limit only. This is usually defined as a 'non-inferiority' test in clinical trials (Laster and Johnson, 2003; Schumi and Wittes, 2011; Walker and Nowacki, 2011). In the present document, however, the term 'equivalence' is preferred. From now on, when not specified, 'equivalence test' is intended as the one-sided test. The null hypothesis of this test states that the effect size is larger than a given minimum magnitude Δ (equivalence limit); the alternative hypothesis is that the effect size is equal to or smaller than Δ:

H_0: effect size $> \Delta$ (high risk)

H_1: effect size $\leq \Delta$ (low risk)

Compared with the test of difference, the null and alternative hypotheses are reversed. The alternative hypothesis – where the burden of the proof rests – is now that of low risk (equivalence of the effect). The goal of a study is to collect evidence against the null hypothesis of high risk; if the evidence is not strong enough, a high risk cannot be ruled out. In contrast with the test of difference, the outcome of an equivalence test can be directly used to conclude on high or low risk. For these reasons, this method is sometimes called a 'proof of safety' as it is designed to show the *absence* of relevant effects.

The meaning of type I and II errors is also reversed. The definition of type I error (false positive) is the same as for the difference test: it consists in wrongly rejecting the null hypothesis. As before, the rate α_E of such error can be directly controlled (it is set before the analysis). For an equivalence test, however, such error occurs when we assess a high-risk substance (effect size $>\Delta$) and wrongly conclude that there is low risk (effect size $\leq\Delta$). Hence, in this case, it is the rate of the error of primary concern that can be directly controlled. The type II error (false negative; rate β_E) consists in failing to reject H_0 when H_1 is true: here, this corresponds to concluding that there is high risk (effect size$>\Delta$) when the true risk is low (effect size $\leq\Delta$). Hence, from the perspective of safety, β_E is of secondary importance. As for the difference test, β_E cannot be fully controlled, as its value depends on experimental conditions: it can only be estimated with a power analysis. However, as there is no primary interest in β_E from the point of view of the regulator, there is no need for risk assessors to control β_E; hence, there is no need for risk assessors to identify an experimental design adequate for this purpose or to recommend a post-hoc power analysis as in EFSA (European Food Safety Authority) (2013). Controlling β_E is now a matter of primary importance for the applicant/study director, who should be interested in obtaining low false negative rates. This last point will be discussed more specifically in the Section on field studies.

The main differences between the two approaches are summarized in **Table 1**.

Table 1: Main differences between the difference test and the equivalence test for the analysis of higher tier studies

	Difference test	**Equivalence test (one-sided)**
Aim	To prove that there is a risk	To prove that there is no high risk
Null Hypothesis	H_0: effect size ≤ 0 (no risk)	H_0: effect size $>\Delta$ (high risk)
Alternative hypothesis	H_1: effect size >0 (risk)	H_1: effect size $\leq\Delta$ (low risk)
False positive (type I error)	When the pesticide has no associated risk (effect size ≤ 0) but the conclusion is that there is a risk (effect size >0). Error rate: α_D Pre-set in the analysis (e.g. $\alpha_D = 0.05$)	When the pesticide is high risk (effect size $>\Delta$) but the conclusion is that it is low risk (effect size $\leq\Delta$) Error rate: α_E Pre-set in the analysis (e.g. $\alpha_E = 0.2$)
False negative (type II error)	When the pesticide has an associated risk (effect size >0) but the conclusion is that there is no risk (effect size ≤ 0). Error rate: β_D Partially controlled with power analysis targeting a desired value (e.g. $\beta_D = 0.2$)	When the pesticide is low risk (effect size $\leq\Delta$) but the conclusion is that it is high risk (effect size $>\Delta$) Error rate: β_E Partially controlled with power analysis targeting a desired value

It is worth pointing out that the most common statistical methods used to test for equivalence (Schuirmann, 1987) are essentially the same as for difference testing. If a difference test is done by applying a *t*-test for the mean difference between test and control (to compare the effect with 0), the equivalence test is also done with the same *t*-test, with the only change that the mean difference is shifted by $-\Delta$ (to compare the effect with Δ). Hence, the change of paradigm in the analysis does not imply the use of new statistical tools.

Shifting the focus from the detection of effects to a proof of similarity, however, has a significant impact on the way a study will be assessed. The extent of such impact will be explained with a series of

examples. For a better understanding, it is convenient to use a graphical representation based on the confidence interval (CI) of the effect size for the test (a two-sided $100(1-2\alpha_D)\%$ CI for the difference test; a two-sided $100(1-2\alpha_E)\%$ CI for the equivalence test). There is a well-known correspondence between statistical testing and the construction of CIs, so that the outcome of a test (significant/not significant) can be obtained by comparing the CI with the baseline effect of interest (0 for the difference test, Δ for the equivalence test). This is shown in **Figure 1**, where the level of significance is set for simplicity to the same value $\alpha_D = \alpha_E = \alpha$ for both tests.

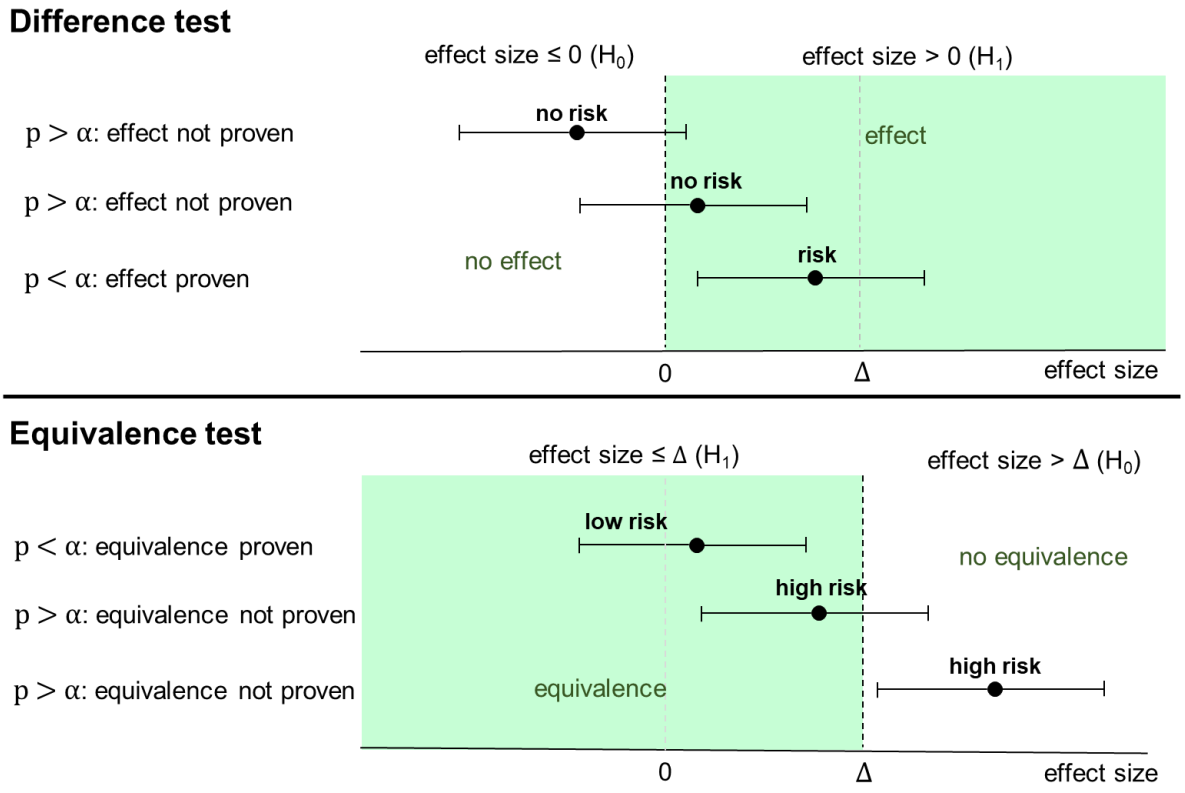


Figure 1: Confidence intervals to test for difference/equivalence. For both tests, a significant outcome (p<α) is obtained when the confidence interval for the effect (bars) falls entirely within the rejection region (green shaded area).

The green shaded area is the 'rejection region' for the test (**Figure 1**), identified by the alternative hypothesis H_1; the null hypothesis is rejected when the CI falls entirely within this region. A test of difference (0 as a baseline) is significant ($p \leq \alpha$; effect proven) when the CI lies completely to the right of 0; it is not significant ($p > \alpha$; effect not proven) when it lies at least in part to the left of 0. A test of equivalence (Δ as a baseline) is significant ($p < \alpha$; low risk proven) when the CI lies entirely to the left of Δ and not significant ($p \geq \alpha$; low risk not proven) when it lies at least in part to the right of Δ.

The graphical representation of the results is used in the following examples to show what would change when analyzing the same study with the two different methods. **Figure 2** shows the results of six independent, hypothetical studies on four substances (A, B, C and D; substances C and D being tested twice). As it is easier to look at numerical examples, we consider the case of honey bee field studies where there is a defined SPG with associated magnitude Δ=10% (decrease in the treatment relative to the control). The result of each study is represented by a point estimate (e.g. the mean) of the effect size and its CI. The levels of significance are again set to a common value $\alpha_D = \alpha_E = \alpha$ for the two tests, so that the same CI can be used for both (as in **Figure 1**), by comparing it with the '0% effect' line for the test of difference and with the '10% effect' line for the test of equivalence. In the scientific literature, when the two tests are applied together, the levels often differ (usually, $\alpha_D \leq \alpha_E$) and the

two CIs have different widths; this choice was avoided here for ease of understanding. The fact that the CIs are symmetrical is another simplification; in general, the symmetry will depend on the choice of scale. A final assumption, which should make the interpretation more straightforward, is that the width of a CI (the precision of the result) depends only on the sample size, with wider CIs corresponding to smaller sizes; this is equivalent to assuming that the variability is very similar across all studies. The experiments in **Figure 2** are discussed in two groups (substances A and B; substances C and D). For each group, the discussion is in three steps: a possible assessment based on a test of difference; a discussion of the issues related to such assessment; and a possible assessment based on the test of equivalence.

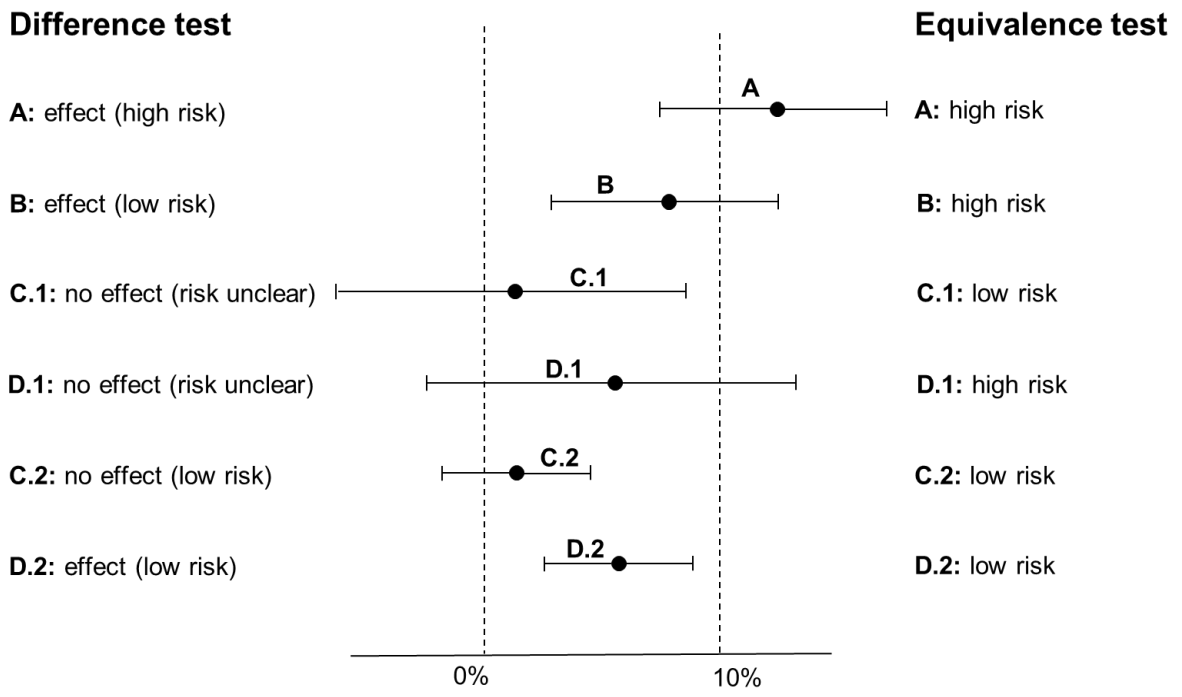


Figure 2: Comparison of the two approaches. Results of the analysis of six hypothetical experiments. Centre: point estimates of the effect (filled circles) and confidence intervals (CIs) (bars). Left: outcomes of the test of difference (effect/no effect), found by comparing the CIs with the 0% vertical line; in brackets, a possible conclusion on risk. Right: outcomes of the test of equivalence (high risk/no risk), found by comparing the CIs with the 10% vertical line.

Substances A and B

Substances A and B are tested in two separate studies with similar sample sizes (the resulting CIs are equal in length).

Difference-based approach. An assessment based on the test of difference would be as follows. As both CIs lie to the right of the '0% effect' line, the test concludes that there is a significant effect (a risk) for both. Once the existence of a risk is established, the level of risk can be assessed by comparing the point estimate (the mean effect) with the 10% level. Based on this comparison, the conclusion would be that substance A presents high risk (effect >10%), while substance B presents low risk (effect <10%).

Comment on the difference-based approach. The conclusion for B is problematic. There is considerable uncertainty on the effect size (expressed by the CI). Such uncertainty has been used (in the test of difference) to detect a risk but it has been ignored in the second step, when concluding on the entity of the risk. The danger is that because of that uncertainty, the confidence we should have in the 'low risk' conclusion might actually be low. A way to account for the uncertainty is to consider the result of an additional test of difference, comparing the effect with 10% instead of 0. The CI is the range of

values of the true effect that would not be refuted by the observed data (Hoenig and Heisey, 2001). In the case of B, the CI includes effects >10%. This means that a one-sided test of difference comparing the effect size with 10% would not exclude that the effect size is ≥10%: in this case, the appropriate conclusion would be 'high risk'. This example shows that the choice of 0% as a baseline, combined with the subsequent assessment of the point estimate, can lead to problematic and even contradictory outcomes.

Equivalence-based approach. An assessment based on the test of equivalence would be the following. In both cases, part of the CI lies to the right of the '10% effect' line and the null hypothesis of non-equivalence cannot be rejected: the conclusion is identical, 'high risk' in the two cases. Note that with this approach, the uncertainty on the estimated effect (the CI) has been appropriately used in the comparison with the 10% baseline; as opposed to the difference test, the point estimate of the effect did not play any role. It might be noted that a possible follow-up experiment to improve the estimation for B would need a higher replication: some considerations on the sample size for honey bee field studies are proposed in Section 3.4.

Substances C and D

Substances C and D are tested a first time in experiments C.1 and D.1, with the same relatively small sample size (identical, relatively wide CIs) and subsequently in experiments C.2 and D.2 with a much larger sample size (narrow CIs).

Difference-based approach. Starting with the first round of experiments, an assessment based on the test of difference would be as follows. For both C.1 and D.1 the CI lies partly below the zero line and the outcome of the test is 'not significant'. Such an outcome is considered ambiguous (EFSA (European Food Safety Authority), 2013): it could mean that the substance has no real effect or that the design of the experiment was too poor to detect a relevant effect. For such cases, as explained above, EFSA (European Food Safety Authority) (2013) recommends a post-hoc power analysis to prove that a 10% effect would have been detected (if it existed) with these experimental conditions. Suppose that such a post-hoc analysis is done, and that the power is shown to be inadequate for both C.1 and D.1: it is necessary to repeat the experiments. Substances C and D are tested again in experiments C.2 and D.2, using a much larger sample size with smaller resulting CIs (the point estimates of the effect for the same substance are identical in the two rounds of experiments, which is unrealistic: the choice was made for the sake of display.) For C.2, the outcome of the test of difference is 'not significant' and in this case, the required post-hoc analysis shows that the power of the test was adequate: hence, it is safe to conclude that there is low risk. For D.2, the outcome is 'significant'; as the estimate of the effect is <10%, the conclusion it that there is low risk.

Comment on the difference-based approach. A possible objection to the assessment of the first round of experiments is that C.1 should raise less concern than D.1. A CI is the range of values of the true effect that would not be refuted by the observed data (Hoenig and Heisey, 2001): for D.1, it includes effect sizes >10%; for C.1 it only includes effect sizes <10%. As for substance B (see above), it would be useful to perform an additional one-sided test of difference for C.1 and D.1 comparing the effect size with 10%. Such test would be significant for C.1 (null hypothesis rejected: effect size<10%) and not significant for D.1 (null hypothesis not rejected: effect size≥10%), and the conclusion would be 'low risk' and 'high risk', respectively. Hence, there are strong indications that the two cases are substantially different. It is worth noting that a standard post-hoc analysis would fail (as it did in this example) to discriminate between two such studies. The two studies have identical sample size and variability (width of the CIs) and differ only in the mean effect size (position of the CI); the techniques for post-hoc power analysis are based on sample size and variability and not on the estimated effect size Mair et al. (2020).

Equivalence-based approach. An assessment based on the test of equivalence would be the following: For the first pair of experiments (C.1 and D.1), there would be two different outcomes, (1) low risk for C.1 (CI entirely to the left of 10%) and (2) high risk for D.1 (part of the CI above 10%). With this method, there would be no follow-up needed for C.1, as low risk has already been proven. For D.1, the follow up would be the same as for the difference-based method, i.e., to repeat the experiment with a larger sample size. The conclusion for the larger experiments C.2 and D.2 is of low risk in both cases.

2.2. Approach with an undefined threshold

For bumble bees and solitary bees, risk managers have not defined a threshold of acceptable effects. Having an 'undefined threshold' has a strong impact on the design and analysis of higher-tier studies for both the methods described in Section 2.1, based on difference and equivalence testing. The implications from the point of view of difference testing ('proof of effect') have been discussed therein. Among them, the experiments cannot be designed to detect a specific magnitude of effects as such a magnitude does not exist. A partial solution proposed in EFSA (European Food Safety Authority) et al. (2022) is to rely on statistical power and the significance of possible effects (see for example the MDD in Section 2.1). While justified within a 'difference testing' paradigm, this solution suffers from all the practical and conceptual issues described in Section 2.1. From the point of view of equivalence testing, the impact of an undefined threshold is even stronger: without a threshold to be used as equivalence margin, a test of equivalence as described in Section 2.1 is simply impossible. It is possible, however, to define an approach appropriate for this case, modifying the methodology for equivalence testing to account for the intrinsic limitations of an undefined threshold. This is shown in the rest of this Section. Such a modified approach is considered preferable to difference testing, as it avoids the issue of non-detected effects and the need for power analysis, and produces a quantitative outcome in terms of effect size in all cases.

As a starting point, it is useful to summarize the approach to equivalence testing explained in Section 2.1, with reference to the bottom of **Figure 1**. The approach is in two steps: the first step is the calculation of the confidence interval (CI) of the effect size. The CI represents the values of the true effect size that are consistent with the observed data. Hence, the upper limit of the interval represents the 'highest plausible level of risk': it is the largest effect size that (with the given confidence) is consistent with the experimental results. The second step is the comparison of the CI with the equivalence limit Δ (threshold). Equivalence is established if the CI lies entirely to the left of Δ or, which is the same, if the highest plausible level of risk is less than Δ (bottom of **Figure 1**). Assume now that for the same study there is no defined threshold. The second step of the process is impossible because Δ does not exist; the 'rejection region' in the bottom of **Figure 1** disappears. The first step, however, is unaffected because the construction of the CI does not depend on the existence of a threshold. The result of the first step is the same, and in particular, the upper limit of the CI still has the meaning of highest plausible level of risk. The recommendation of this document is to use this value for the assessment and to further classify it into pre-defined categories using a modified version of the second step.

A classification is possible because, even in absence of a defined threshold, we can still define a series of effect sizes that, based on the current knowledge, may be considered meaningful from the point of view of risk assessment. Instead of using a single equivalence limit, the upper limit of the CI can then be compared with a *sequence* of increasing equivalence limits. Specifically, it may be possible to perform a series of tests of equivalence using e.g. the following values of the equivalence limit:

1%, 3%, 5%, 7%, 10%, 20%,…

The sequence can be extended to values above 20% by systematically multiplying the last term by two; the choice of values will be discussed shortly. An example of the procedure is shown in **Figure 3**. For small equivalence limits, starting from 1%, the outcome of the test is 'not equivalent' as the upper limit of the CI is always larger than Δ_i; the last equivalence limit for which this occurs is 10%. For the next value in the sequence, 20%, the upper limit is lower than the equivalence limit and equivalence is proven; equivalence is also proven for all the limits higher than 20%. The outcome of the procedure in this case is that equivalence is proven with a 20% limit but not with a 10% limit, or that the 'level of risk' is between 10% and 20% (see **Figure 3**: the upper limit of the CI is in the shaded area between 10% and 20%).

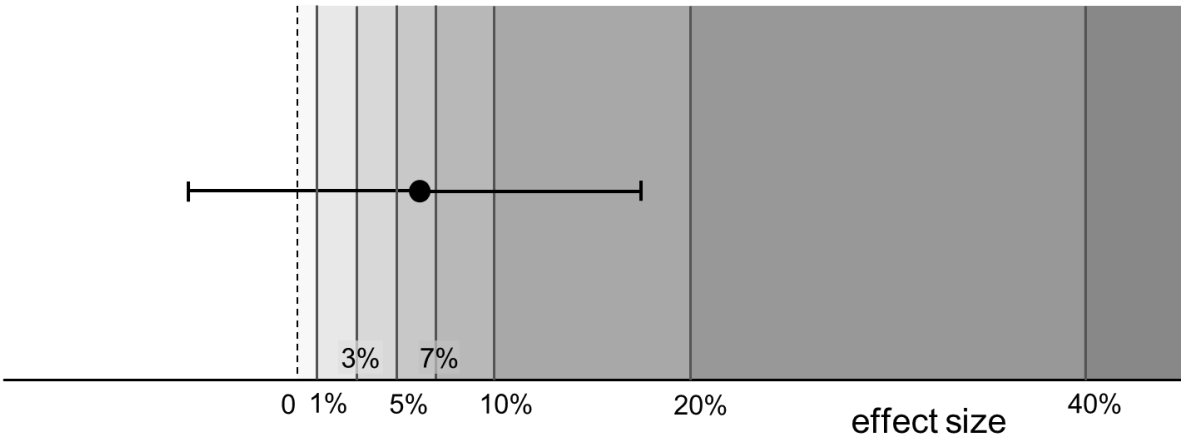


Figure 3: Classifying risk with an undefined threshold. The results of a study are represented by mean and confidence interval (CI) of the effect size. The upper limit of the CI is compared with different ranges of effect size (risk levels: shaded areas). The risk level for this study is between 10% and 20%.

It is worthwhile to note that none of the equivalence limits in the sequence has the meaning of a threshold, as it cannot be used to discriminate between acceptable and unacceptable effects. The values are used for the purpose of classification. The effect observed in an experiment is assigned to a specific risk level, with different risk levels corresponding to different ranges of the effect size. Based on the sequence of equivalence limits, the levels of risk correspond to the following ranges:

0% – 1%

1% – 3%

3% – 5%

5% – 7%

7% – 10%

10% – 20%

(the sequence proceeds by systematically multiplying the last range by two)

Thus, for the purposes of the conclusion of the risk assessment, the assessor will have both a "highest plausible effect" (quantitatively) as well as a translation of what that may mean (qualitatively) as far as actual level of risk to colonies/populations of bumble bees/solitary bees. This can then be communicated to risk managers, along with the full weight of evidence (see Chapter 10.6 of the GD) and scope of the conclusion (e.g., location, season, crop(s), etc.).

In summary, the approach recommended here produces as an outcome for each study: (i) an absolute measure of the highest plausible risk, that is the upper limit of the CI of the effect size; (ii) a categorisation of such measure into different levels of risk (ranges of effect size). This approach has several advantages compared with the standard 'proof of effect' method (EFSA (European Food Safety Authority) et al., 2022): notably, that it will always produce a quantitative, unambiguous outcome (the highest plausible risk). The categorisation into risk levels is also considered useful from practical point of view. A risk level does not have a meaning *per se*: that is, it cannot be considered large or small as the current knowledge does not allow for this distinction (EFSA (European Food Safety Authority) et al., 2022). However, it has advantages for example for the purpose of comparison: risk levels from different studies (and for different substances) can be compared even in absence of a defined level of protection. Hence, risk levels can be used to obtain a more harmonized evaluation of different substances, thus addressing an issue pointed out in EFSA (European Food Safety Authority) et al. (2022).

The levels of risk have been chosen with consideration for what could be magnitudes relevant for the risk assessment. The 10% level is included in the sequence of ranges as it is the threshold of acceptable levels chosen for honeybees and it has been considered as an option by EFSA (European Food Safety Authority) et al. (2022) for the other bee groups. Below 10%, effects differing by a few percentage

points might be considered qualitatively different; accordingly, the interval 0-10% is further partitioned into five distinct ranges. Small differences above 10% are considered less relevant, and increasingly so as the effect size increases; for this reason, the ranges above 10% double at each step. The choice of ranges may be reviewed later in time based on newly acquired knowledge. As remarked above, a qualitative classification of the levels of risk (e.g. small, medium or high) is not currently possible; however, it would be desirable to have such qualitative criteria for the assessment. For that purpose, it may be useful to consider the results expressed as standardized effect sizes (see recommendations in Section 2.3).

2.3. Conclusions and recommendations

Equivalence testing is considered to be the most appropriate approach for the analysis of higher tier studies. With equivalence testing, the burden is on proving the safety of a substance; the assessor can conclude directly on the existence of a relevant risk and can control directly the primary error of interest (of not detecting a high risk). The techniques for equivalence testing are very similar to those used for standard difference testing.

The 'high risk' hypothesis should be tested, for honey bees, using a one-sided equivalence test with significance level α_E=0.2. This corresponds to comparing the upper limit of the two-sided $100(1-2\alpha_E)\% = 60\%$ confidence interval of the effect size with the appropriate magnitude Δ (threshold of acceptable effects). It is acknowledged that the value α_E=0.2 is unusually high; the level used for equivalence testing in the literature is usually set to 0.1 or lower. The present choice corresponds to a 20% false positive rate and an 80% level of confidence in a 'low risk' conclusion; this is equivalent to the 80% power recommended in EFSA (European Food Safety Authority) (2013) for the difference test. The relevant change is that, while the nominal 80% value is the same, the trust in the actual value is different. With the new methodology, the most relevant error rate is under control; in contrast with the uncertainty on the actual 80% power in EFSA (European Food Safety Authority) (2013), the 80% value for the confidence level can be trusted. Based on these considerations, the Working Group did not deem it necessary to also increase the level of confidence with respect to EFSA (European Food Safety Authority) (2013). For bumble bees and solitary bees, the upper limit of the 60% confidence interval should be compared with a series of effect sizes, as specified above (which is the same as performing a series of tests with different limits) and the outcome should be an interval of effect sizes.

The statistical model for the analysis should be chosen by the applicant/study director, with the only condition that it should allow for equivalence testing based on standard methods (Schuirmann, 1987; Wellek, 2010). The models discussed in the present document are usually simple examples; they refer to normally distributed data (possibly after transformation) and the use of standard *t*-tests for the comparison of the effect size with Δ (or multiple values of Δ for an undefined threshold). More generally, it is reasonable to expect that an appropriate choice for the analysis could be a (generalized) mixed model with inclusion of 'treatment' and 'control' (or the difference thereof) as fixed effects. Based on standard techniques, the calculation of confidence intervals of the effect size for equivalence testing should then be straightforward. As there is usually a temporal component in the study (repeated measurements on the same unit, or 'assessments'), it might be convenient to combine the data for all time points (assessments) in the analysis. This could be done for example by including the assessment as an additional fixed effect in the model and testing for a significant interaction between assessment and treatment. If this is not significant, the test of equivalence can be applied only once, to the effect averaged over all assessments; if it is significant, the test of equivalence has to be applied for each assessment independently. These are to be considered as minimal recommendations, to be adapted as needed to the type of study and the data. Deviations from these general guidelines, if substantial (for example the use of nonparametric tests), should be adequately justified.

The results of the equivalence tests should be reported including (at least) *p*-values, effect sizes and confidence intervals; a graphical presentation of the results (see Figure 2) should also be provided. Standardized effect sizes and confidence intervals (dividing the original results by the standard deviation between colonies/populations) should also be provided to aid the interpretation of the results. While difference testing is not considered the most appropriate method to conclude on safety based on a higher tier study, it is of course possible to provide and discuss the results of difference tests as complementary information.

Further considerations on the statistical methodology are provided in this document for specific studies when needed.

3. General considerations for field studies

3.1. Study aim

The aim of a field effect study is to determine the potential effect of a particular use of PPP on bees under fully realistic conditions while maintaining control over potentially confounding factors, such as local field conditions and landscape land use, which may influence PPP exposure or effects through expression in pollen and nectar and/or bee foraging choices (Sponsler et al., 2019). To fulfil this aspect of control in a variable world, it is important to select a suitable study design. Such designs include replicated Before-After-Control-Impact (BACI) and Randomised Controlled Trial (RCT), which are particularly suitable for monitoring anthropogenic impacts on organisms under ecologically realistic conditions (Christie et al., 2019).

For specific considerations and the setup of the bee colonies or population, see the specific sections on honey bee, bumble bee and solitary bee field study guidance (4.2, 5.1, 6.1, respectively).

3.2. Definitions of terms

Treatment: the experimental treatment applied to the fields, e.g. control treatment or application of a PPP treatment (the test item(s)).

Field: a spatially well-defined unit on which the test crop/plant is grown. A field can be either treated or untreated (control) with the PPP, i.e. it is appropriate to refer to a 'control field' (EFSA (European Food Safety Authority), 2013).

Site: a field or group of adjacent fields that have the same treatment, while also including the surrounding landscape context. Sites should be independent, which means that there should be no systematic exchange of bees between sites.

Landscape: does not have exact boundaries, but can be described as a more or less heterogeneous area composed of different habitats, connected by the organisms that utilize them, and can be characterised by the composition and configuration of land covers and landscape elements (With, 2019).

Block: unit of replication, consisting of paired sites in similar landscape contexts receiving the different treatments.

It should be noted that the above definitions are not fully consistent with the terms used in Annex B as a result of the possible differences in the study design between the exposure studies and the effects studies (e.g. absence of bee hives/colonies/nests in residue studies where pollen/nectar are collected directly from flowers).

3.3. Study design principles

The field effect study is the only method to assess effects of PPP use on bees under fully realistic ecological and agronomical exposure conditions (EFSA (European Food Safety Authority), 2015) this is also why field studies are used as the reference tier in environmental risk assessment (ERA). These studies should be experimental, i.e. with controlled exposure and standardized study organisms, so that it is possible to make causal links between exposure and effects. In the context of ERA and bee effect studies, field studies are conducted on (model) crops grown outdoors and the (model) bees are free foraging and not confined by any enclosure (Pettis et al., 2014).

The study system consists of sites of treated and control fields in similar landscape contexts in matched pairs, which together form a block, centered on the bee hives/colonies/nests at the edge of the fields (Figure 4).

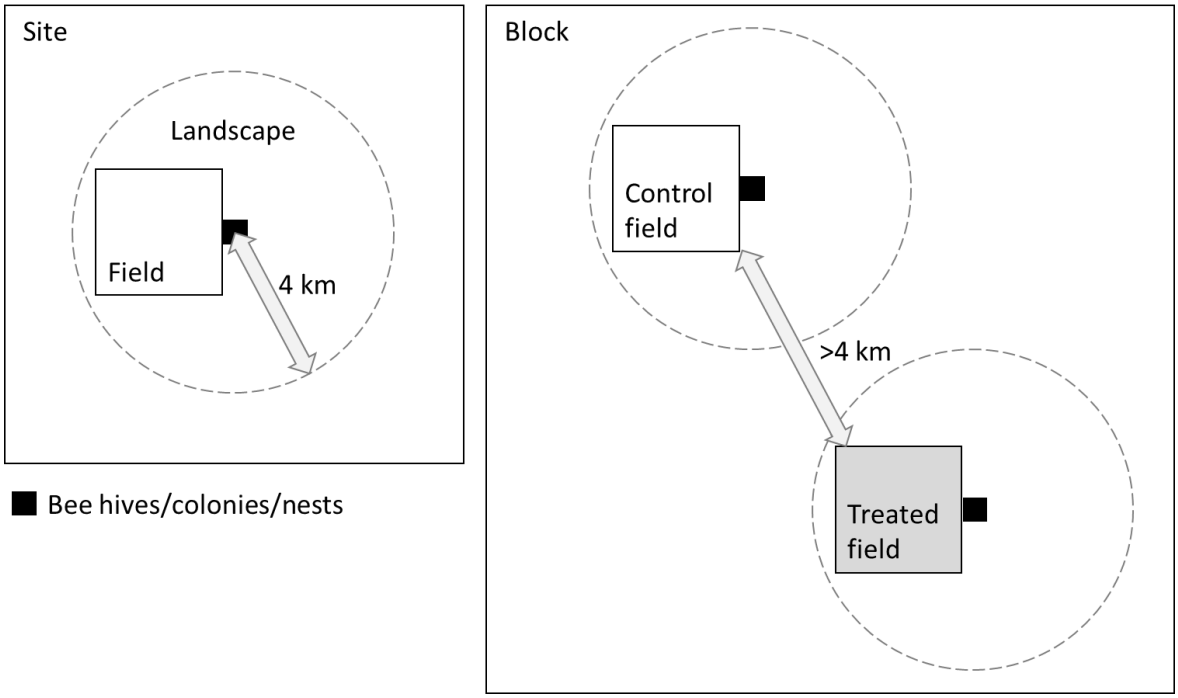


Figure 4: The spatial study design elements relevant for field effect studies. A site includes one or more adjacent fields, with bee hives/colonies/nests at the edge, surrounded by the landscape (r = 4 km). Sites belong to blocks, where the treatments are applied to the fields. Fields within a block are separated by at least 4 km (see Appendix A). Blocks of treated and control fields should have similar surrounding landscapes but be distant enough to exclude systematic exchange of bees.

3.4. Replication and study planning

This Section provides an assessment of the possible level of replication (in terms of number of fields and colonies) that might be needed for field effect studies and indications on how to generate the data in case the level is very large. The approach followed here is very similar to Appendix O of EFSA (European Food Safety Authority) (2013) (see also EFSA (European Food Safety Authority) et al. (2021)): an estimation of field study requirements done with a power analysis, using plausible estimates of variability as parameters. The point of view, however, is very different. The statistical approach recommended in EFSA (European Food Safety Authority) (2013) was based on standard difference testing: the power analysis was done to ensure that, if a relevant effect existed (a high risk) the experiment would be able to detect it with good confidence. As remarked in Section 2, the false negatives (wrong 'no effect' outcomes) are the main concern of the risk assessor in this case; the power analysis is the standard way to control the false negative rate. Here, the statistical approach is based on the test of equivalence, which reverses the null and alternative hypothesis and the meaning of 'false positive' and 'false negative' (Section 2). As the null hypothesis is now that of high risk, the main concern for the risk assessor is now the 'false positive' error (to obtain a wrong 'low risk' outcome for a high-risk substance). This error can be directly controlled (rate α_E) so that an inadequate study will result in a 'high risk' outcome (non-equivalence) with the desired level of confidence ($1 - \alpha_E$) independently of the number of replicates. Hence, there is no need for the risk assessor to perform a power analysis to estimate (and recommend) an adequate level of replication. A power analysis, however, could provide useful indications for the applicant/study director. Such an analysis targets the rate β_E of 'false negatives', which (for an equivalence test) consist in of wrongly concluding 'high risk' for a low-risk substance. This rate (of secondary interest for the risk assessor) could be relevant for an applicant /study director planning an experiment, as it measures the risk of *not* proving the safety of a substance, and should be kept at acceptable levels using a reasonable amount of resources. In this Section, a prospective power analysis for field studies is carried out to understand the possible resource burdens considering different conditions and requirements. The results of this analysis will constitute nothing more than an indication of the possible necessary level of replication; recommendations, as pointed out

above, are not needed. The analysis refers to the case of a defined threshold and a single equivalence limit. Extension to the 'undefined threshold' approach with multiple equivalence limits (Section 2.2) is straightforward.

The model used here is very similar to Appendix O of EFSA (European Food Safety Authority) (2013). We consider a field study with n_B blocks, each block consisting of a pair of fields (control and treated), with a number n_H of colonies (or hives, or nests) in each field. The total number of colonies is then $N = 2n_B n_H$. It would be possible to consider time-aggregated data in this analysis, however, as a single time point should provide the worst-case statistical power, only the single time point is considered here. It is assumed that the relevant endpoint, E, is measured on a scale where all the effects are linear and random effects are normally distributed; the endpoint (as the bee group) is not specified here (an example is provided at the end of this Section). The equivalence limit (threshold of acceptable effects) measured on the same scale is Δ. The mean value of the endpoint is μ_C and μ_T for the control and the treatment, respectively. It is assumed that there is indeed an effect of the pesticide, resulting in an average difference d between the test and the control, $\mu_T - \mu_C = d$, but that this effect is below the upper limit: $d < \Delta$. An experiment should then be able to correctly prove that $d \leq \Delta$ (rejecting the null hypothesis $d > \Delta$); otherwise, the outcome is a false negative. The purpose of this calculation is to estimate the rate β_E of such false negatives based on the other parameters of the model.

The variability is described in the model as follows. Consider a specific block b, with control and treated fields: the mean value of the endpoint in the control field is $m_C(block\ b)$ and the value for one of the n_H control colonies is:

$$E_C = m_C(block\ b) + \varepsilon$$

Where $\varepsilon \sim N(0, \sigma^2)$ is the random error. The simplest assumption for the treated colonies in the same block is that the endpoint, E_T, is described in a similar way, with a mean value shifted by the effect size d:

$$E_T = d + m_C(block\ b) + \varepsilon$$

With this definition, the difference between the mean values of treatment and control is equal to d in every block. It is reasonable, however, to hypothesize that the effect might also change between blocks, so that it is higher than d for some blocks and lower for some others (the average across all blocks being d). This could occur, for example, because the landscape characteristics of the two fields are not exactly identical, or because there is a (possibly unknown) interaction of the pesticide with the landscape, field type or management practices. Accounting for this block-dependent random difference, or 'field effect' f, the equation for the treated field becomes:

$$E_T = d + f(block\ b) + m_C(block\ b) + \varepsilon$$

The model is completed by assuming that the mean value in the control fields is described by a' random block-dependent term, $m_C(block\ b) \sim N(\mu_C, \kappa^2)$ where μ_C is the overall mean; note that $m_C(block\ b)$ is the same for the treated and control field in the same block.

In full detail, the complete model is:

$$E_{tbh} = \mu_t + B_b + f_{tb} + \varepsilon_{tbh}$$

Where the meaning of the indices is as follows: t is the treatment ($t = C,\ T$), b is the block ($b = 1 \ldots n_B$), the pair of indices tb indicates one of the (treated or control) fields in block b and h indicates an individual colony. E_{tbh} is the value of the endpoint for a colony, μ_t ($= \mu_C,\ \mu_T$) is a fixed effect; $B_b \sim N(0, \kappa^2)$ is the random block effect; $f_{tb} \sim N(0, \tau^2/2)$ (with the constraint $f_{Cb} + f_{Tb} = 0$ for every b) is the random interaction between block and treatment (field effect); and $\varepsilon_{tbh} \sim N(0, \sigma^2)$ is the random error.

The starting point for the power analysis is that the substance is low risk: $d < \Delta$. Consider the data generated from a study with n_B blocks and n_H colonies per field; the data are analysed with a one-sided test of equivalence with null hypothesis $d > \Delta$ and with significance level $\alpha_E = 0.2$. The question is: What is the likelihood that, based on these data, the null hypothesis is rejected and the correct conclusion ($d \leq \Delta$) is reached? The probability of obtaining the correct conclusion is the power of the equivalence test, $1 - \beta_E$. This probability can be calculated with standard methods (Julious, 2004). Here, we assume that the variance parameters (σ^2 and τ^2) are known and that the number of colonies per

field n_H is fixed. Based on this, it is possible to calculate the minimum number of blocks n_B needed to obtain a given power $1-\beta_E$:

$$n_B = 2\frac{\left(Z_{\alpha E} + Z_{\beta E}\right)^2}{(d-\Delta)^2}\left(\frac{\sigma^2}{n_H} + \tau^2\right)$$

Where:

$Z_{\alpha E}$ is the α_E-quantile of the standard normal distribution ($\alpha_E = 0.2$)

$Z_{\beta E}$ is the β_E-quantile of the standard normal distribution

d is the true effect size

Δ is the upper equivalence limit

n_B is the number of blocks (total number of fields: $2n_B$)

n_H is the number of colonies

σ^2 is the between-hive variance

τ^2 is the variance of the field effect

It is immediately clear from the equation that n_B depends on the true effect size, d, and (for constant power $1-\beta_E$) it increases as the difference $d-\Delta$ becomes smaller.

To illustrate this using a concrete example, the power analysis for a honey bee field study with $n_H = 6$ colonies per field, is considered. The endpoint is colony size, S and multiplicative effects are assumed so the model is applied to log-transformed data: $E = \ln(S)$. It is not possible to estimate τ^2 with the currently available data, as it is related to the interaction between site and treatment and is expected to be a substance-specific property. Given the lack of relevant data, the estimate of the variance between sites (blocks) is used here as a proxy for τ^2. The relationship between the 'block' variance and τ^2 is unknown and possibly weak, but this choice was considered better than using completely arbitrary values. Estimates for σ^2 and τ^2 (defined as the variance between sites) were calculated in EFSA (European Food Safety Authority) (2013) in the form of coefficients of variation (standard deviation divided by the mean), also based on the assumption of lognormally-distributed data. The estimates were $CV_F = 5\%$ (variation between fields, with $\tau^2 = \ln(CV_F^2 + 1)$) and $CV_H = 15\%$ (variation between colonies, with $\sigma^2 = \ln(CV_H^2 + 1)$). The Working Group considered whether the data collected since 2013 could confirm these two estimates. For this purpose, data from control colonies (colony strength measured at several time points) were extracted from 30 recent honey bee field studies. The data were filtered by keeping only the first 42 days for each study. Eight of those studies had more than one control field, so that it was possible to calculate a CV_F value per study and time point. The set of CV_F values thus calculated was small and could not be considered strong evidence; still, it was consistent with the 5% value proposed in EFSA (2013), which the Working Group decided to confirm. A CV_H value was calculated for each field and assessment time (independently for different fields in the same study). Fields with poorly equalized colonies, that is with too high inter-colony variability at the beginning of the experiment ($CV_H > 15\%$ at the first assessment) were excluded from the analysis. Mean and maximum CV_H across measurements were calculated for the remaining fields. The mean variation in a field over the first 42 days was on average $CV_H = 17\%$; the maximum variation over the same period was on average $CV_H = 25\%$. The Working Group decided that there was support for a higher estimate of inter-colony variability than in EFSA (European Food Safety Authority) (2013) and chose the value $CV_H = 20\%$.

The numerical values of the parameters in the equation are:

$Z_{\alpha E} = -0.84$ i

$n_H = 6$

$\sigma^2 = \log(1 + CV_H^2) = 0.04$

$\tau^2 = \log(1 + CV_F^2) = 0.0025$

$\Delta = |\log(1 - 0.1)| = 0.105$

The free parameters left in the equation are $Z_{\beta E}$ (hence, β_E) and d. The value of n_B was calculated from the equation for several possible combinations of β_E and d (the latter calculated based on a % decrease, using the same logarithmic transformation as for Δ). The results are shown in **Table 2**: the number of fields, $2n_B$, is tabulated against the true effect size (reported as % decrease) for several possible values of the statistical power, $100(1-\beta_E)$.

Table 2: Number of fields needed to conclude (with a given power) that there is low risk for a substance with true effect <10%. The number of colonies per field is set to six.

True effect size	Number of fields[(a)] needed to obtain a given statistical power based on the true effect size			
	power = 90%	power = 80%	power = 70%	Power = 50%
0%	16	10	8	4
1%	18	12	8	4
2%	24	16	10	4
3%	30	20	14	6
4%	40	26	18	8
5%	56	36	24	10
6%	88	56	36	14
7%	152	96	64	24
8%	338	212	140	54
9%	1336	840	554	210

[(a)]The number of fields is twice the number of replicated blocks, including control and treated fields

As remarked above, the chance of correctly concluding that there is low risk (with a given level of replication) depends on the hypothetical true effect size of the substance. If an experiment is planned an experiment on a specific pesticide and chooses the number of fields based on this power analysis (e.g. with target power 80%), they have to hypothesize a 'true effect' of the substance. A 0% effect is the most optimistic assumption in terms of resources: in this example, it corresponds to a design with 10 fields (five blocks; see Table 2). If the effect is truly null, the correct 'low risk' conclusion will be reached in 80% of the studies ($1-\beta_E = 0.8$). However, the power is considerably reduced if the true effect is larger. If the effect consists of a 2% decrease, for example, the power is 70% (and 16 fields would be needed to reach again 80% power). If the effect consists of a 5% decrease, the power is 50%, which means that every other study would reach an incorrect high-risk conclusion (36 fields would be needed in this case to reach 80% power).

The challenge of performing large and well-replicated field effect studies could be eased by performing the study in sets of blocks distributed over several countries or several years (Flores et al., 2021). In such cases, it is essential to follow the same study plan in all countries/regions/years/months so that measurements and results can be considered together. The whole study, including multiple sets of blocks replicated either temporally or spatially, should ideally be planned together even if the sets of blocks are conducted in different countries/regions/years/months. However, it is also possible to plan for a specific set of blocks, evaluate the outcome and plan for additional sets of blocks based on the central tendency and variability of the endpoint.

It is convenient for the applicant/study director to reduce inter-colony variability (parameter σ^2) as much as possible throughout the study, so to decrease the number of replicated blocks corresponding to a given effect size (**Table 1**). For this purpose, it is recommended that the applicant/study director aim to obtain well equalized colonies at the beginning of the experiment. Variability can be measured, as in the example above, by the coefficient of variation defined on a log scale: the condition that $CV_H \leq 15\%$ at the first assessment is considered a good criterion.

3.5. Site selection and setup

The sites should be representative of the region(s) for which authorization is sought. As regards location of the control and treated fields within a single block, it is recommended that they should be as similar

as possible in terms of size and surrounding landscape (for example see Rundlöf et al. (2015),Woodcock et al. (2017)). As with field studies refining exposure at higher tiers (Annex B of the GD), considerations of focal crop, focal field size, test item application and landscape surrounding the effect study fields are necessary.

3.5.1. Spatial separation

Sites (i.e. fields that receive the treatments) within a block should be located in a similar landscape with similar soil conditions (see further under Section 3.5.2). However, different blocks can be located in different landscapes with different soil conditions. Replicate blocks should be separated by at least 10 km and sites within a block should be separated by at least 4 km in order to (a) avoid and (b) minimize the exchange of bees between sites, based upon honey bee foraging ranges (see Appendix A of this document). These distances are based on honey bee foraging ranges since they generally have the larger foraging ranges among the bee groups ((Zurbuchen et al., 2010), (Kendall et al., 2022)). Exchange of bees between sites can be further avoided by providing sufficiently large areas of bee attractive treated crop/plant at each site, since increased availability and decreased distance to high-reward resources can reduce foraging distance (Beekman and Ratnieks, 2000; Steffan-Dewenter and Kuhn, 2003). In addition to practically avoiding exchange of bees, the 10 km separation of blocks also capture the variability in soil and landscape characteristics.

3.5.2. Site and landscape characteristics

Spatially explicit land cover data, capturing the major land uses and crops grown should be used to demonstrate that the landscape contexts do not differ systematically between treatments over the sites (see Box 1, below). There are several land cover/land-use sources and classification systems that could be used (see Box 1). It is recommended to use a recent version of such a classification system and an intermediate level of detail. Values for the individual sites as well as ranges and averages of these categories for the different treatments over all the sites at 4 and 1.5 km radii around the treated fields (see also Annex B of the guidance document) as well as statistical tests confirming the lack of differences between treatment groups should be reported. In addition, location, areas and development stage (BBCH) of bee-attractive mass-flowering crops could be mapped and be used to select sites to reduce bees foraging elsewhere than in the focal field(s). Control and test item treatments are subsequently allocated randomly to the sites within each block.

Box 1: Standardizing landscape characteristics between sites within a block is necessary in order to make conditions other than the treatment as comparable as possible between groups. In order to do this standardization, landscape characteristics need to be quantified. That can be done based on spatially explicit land cover/land use (LCLU) information.

The example below illustrates the process of considering and accounting for LCLU information in the study design and site selection. The aim is to have an intended number of blocks of sites, with two comparable sites in each block, to eventually form a control group and treatment group of sites that are comparable in LCLU. The process is based on trial hosts that are experienced crop growers and extraction and analysis of spatial explicit LCLU information. It is recommended to identify a surplus of potential study sites because of constraints relating to the matching into blocks or occurrence of mass flowering that could distract the bees from the test field(s) and therefore pre-study decisions to exclude sites.

Crop growers are contacted, for example through a farmer organization, and asked if they are willing to grow a (model) crop following Good Agricultural Practices specified in a protocol. If yes, the geographical coordinates (latitude and longitude in decimal degrees) of the potential location of bee hives/colonies/nests or the center of the experimental field(s) is extracted either using a GPS in the field or from digital maps. Using a Geographical Information System (GIS) software, LCLU information is extracted at the 1.5 and 4 km radii from the coordinates. There are several LCLU sources and classification systems that could be used, for example CORINE Land Cover data (Büttner, 2016), LUCAS (Pflugmacher et al., 2019) and the Integrated Administration and Control System, possibly complemented by national data sources, aerial photos and/or ground truthing. For example, if CORINE Land Cover data is used, this could include the level 1 categories

(artificial surfaces, agricultural areas, forest and semi natural areas, wetlands and water bodies) as well as the level 2 categories for agricultural areas (arable land, permanent crops, pastures, heterogeneous agricultural areas) and if IACS in combination with a national land cover source is used this could include agricultural land, annually tilled arable land, semi-natural grassland, mass flowering crops, forest and urban, including an explanation of how these categories were defined.

Geographically close (but still at least 4 km apart) sites with similar LCLU are matched into blocks. This is repeated until the intended number of blocks are reached (see Section 3.4). Within a block, test item and control treatments are then randomly allocated to the two comparable sites. The table below provides an example of how to report the averages and ranges of the categories for the different treatments over all the sites at one scale as well as statistical tests confirming the lack of differences between treatment groups. It should be noted that the information as reported in the table below would not be sufficient to appropriately characterise the LULC types and their cover of flower resources to meet the requirement of less than 10% alternative flower resources set in the higher tier exposure studies (see Section 1.3 in Annex B of the GD).

Table 3 Land cover/land use (LCLU) in the 1.5 km landscape surrounding the planned location of the bee hives/colonies/nests and test of difference between treatments:

LCLU variable	Control sites Mean (range)	Treated sites Mean (range)	Test of difference F (df = 1, 7), P
Agricultural land (%)	58 (10-88)	56 (6-83)	0.29, 0.61
Annually tilled arable land (%)	39 (3-71)	34 (0-75)	0.64, 0.45
Semi natural grassland (%)	3 (0-7)	4 (0-9)	0.16, 0.70
Mass flowering crops (%)[1]	8 (0-24)	8 (1-18)	0.01, 0.93
Forest (%)	25 (2-75)	24 (1-67)	0.23, 0.63
Urban (%)	3 (0-9)	3 (0-9)	0.53, 0.49

[1]Mass-flowering crops include all such crop even if not currently in bloom: oilseed rape (46%), potato (28%), pea (18%), bean (4%), fruit and berry cultivation (4%), and herbs and seeds (<1%).

3.5.3. PPP use history

The test fields (both treated and control) are located in agricultural areas usually with intensive farming practice. This is important to consider both in the selection of test fields and in the choice of (model) crop plant, planned pest management and exposure characterization.

It is expected that each field will have its own history of PPP use, also within the same season (e.g. chemical weed control at early crop stages). Therefore, some PPP residues – other than the test item – might contaminate pollen and nectar of the test fields (both the treated and the control). It is, however, expected that the contamination from earlier pesticide uses (e.g. performed earlier in the season or in the previous year) is small compared to the PPP application(s) made during the flowering. Nevertheless, all PPP applications on the test crops should be avoided as far as possible and no PPP applications should be done between BBCH 50-69. If some PPP applications were unavoidable (before BBCH 50), these should be the same in the treated and the control field within a block and must be well-documented and reported (PPP that had been used, application method, application rate, time of application, BBCH of the crop). Also, it is recommended that PPPs with low persistence and with low toxicity profile to bees should be selected.

Effect studies in which the test item is planned to be applied in early growth stages (e.g. seed treatment, spray application in early grows stage) are special cases. This is because, if other PPP are also used on the same crop in early growth stage (again only before BBCH 50), the level of contamination in pollen and nectar can be in the same order of magnitude as the test item itself. Also, the pollen and the nectar in the control field could be contaminated similarly. This can make the interpretation of the results difficult, particularly if it is expected that the other PPP may induce some adverse effects on the studied endpoint. Therefore, if other PPPs are used, it must be demonstrated that this/these other PPPs do not have an effect on the studied endpoint, and therefore the control field can still be considered as a real control.

The need for PPP uses other than the test item could potentially be reduced by selecting test (model) crops that generally have few pest problems and thus require little PPP use (but see 3.5.4).

3.5.4. Test (model) crop/plant

It is preferable to carry out effect studies with the proposed crop outlined in the GAP, but it may also be possible to use a highly attractive model plant such as *Phacelia tanacetifolia* or oilseed rape *Brassica napus* and extrapolate the study findings to a range of crops. The key issue in selecting a suitable crop is to ensure that the residues, and hence the exposure to bees, are environmentally relevant in line with the exposure assessment goal (ExAG). Exposure field studies that will be used to define the realistic worst-case PPP mass uptake per bee per time for each GAP, as described in Annex B of the Guidance Document, are pre-conditions of the effect field studies. Alternatively, Tier 1 exposure estimates should be considered.

The common and scientific name and variety of the tested crop, planting date, seasonal growth cycle of the crop, calendar date and time of the start and the end of flowering period should be specified. The phenological growth stages and BBCH identification should also be recorded during the study period and at all sampling dates. It is recommended to select the variety based on their potential for flowering and pollen production. The crop/plant variety should be the same within a block. Plant density should be documented by recording 1) the number of plants within a square meter and 2) the plant percentage coverage of the soil at 10 representative locations (i.e. avoiding tramlines and areas with poor establishment) within the field(s) at each site, to ensure that the crop density is similar between control and treated fields.

When using a bee attractive test crop/plant, the area should be sufficiently large to sustain the bee colonies or populations (see Annex B of the guidance document). More generally, EFSA (European Food Safety Authority) (2013) recommended a minimum of 2 ha to provide sufficient flowers and support exclusive foraging while Medrzycki et al. (2013) recommends a minimum of 5 ha area of the treated crop/plant to represent a major nutritional source for the colonies during the flowering period.

Placement of bees should be timed to the flowering stage of the crop and the type of test item application. For test item application during flowering, hives/colonies/nests should be present at the edge of the field before application. The same consideration can apply for test investigating the effect of PPP drifting off-crop. For pre-flowering applications, bees should be placed at the fields at early flowering (e.g. BBCH 61-63).

3.5.5. Creation of colonies, initial conditions and location of bee hives/colonies/nests

There will always be variation between bee colonies and local bee populations, however it is the responsibility of the applicant/study director to ensure that the starting conditions for any colonies and populations are as similar to each other as possible at the start of the experiment. Guidance on this, for the three bee groups, can be found in (Hodge, 2019) and in the following sections for the specific bee groups. With increased variability between colonies, populations and sites at the start of the study, the more blocks will be needed to demonstrate equivalence (see Section 3.4).

The methods for standardising honey bee and bumble bee colonies differ, however they both rely on an initial standardisation of the colony size for honey bees or colony weight for bumble bees which can then be followed by a paired matching of the colony endpoint (see Box 2, below). For solitary bees, experience from *Osmia bicornis* suggests that providing a larger starting population reduces variability (EFSA (European Food Safety Authority) et al., 2022).

Box 2: Randomly assigning colonies to a treatment group can often result in well equalized, comparable groups. However, it is also possible that some larger or smaller colonies may be assigned to a one treatment group by chance, thereby increasing the variation between group, even if the mean mass of each group is comparable.

In the example below, 16 bumblebee colonies with between 20-50 individuals were assigned to two groups. The colonies ranged in mass between 51.2-100.25 g. In the first method "Random", the colonies were randomly assigned to a group as either "Control" or "Treatment". In the second method the colonies were ordered from the lightest to the heaviest and assigned into pairs. Each colony within a pair was then randomly assigned to a treatment group. In both cases, the group sizes are comparable based on a comparison of the mean colony size using a t-test (Random: Control = 80.40g, Treatment = 69.65g, t = 1.313, p = 0.22; Paired: Control = 75.75g, Treatment = 74.32g, t = 0.16, p = 0.87).

Here, for the colonies that were randomly assigned, the control group received more of the larger colonies and also contains the smallest and largest of the colonies, this both increases the mean size of the control group, and also means that there is more variability in the control relative to the treatment group (see the figure just below). Ranking the colonies by weight and randomly splitting similarly sized colonies into different treatment groups results in two groups where not only is the mean colony size very similar, but also the variation in colony size between the groups is equalized.

Therefore, the working group recommend always equalising the variability between treatment groups. This method is valid for both bumblebee and honeybee colonies.

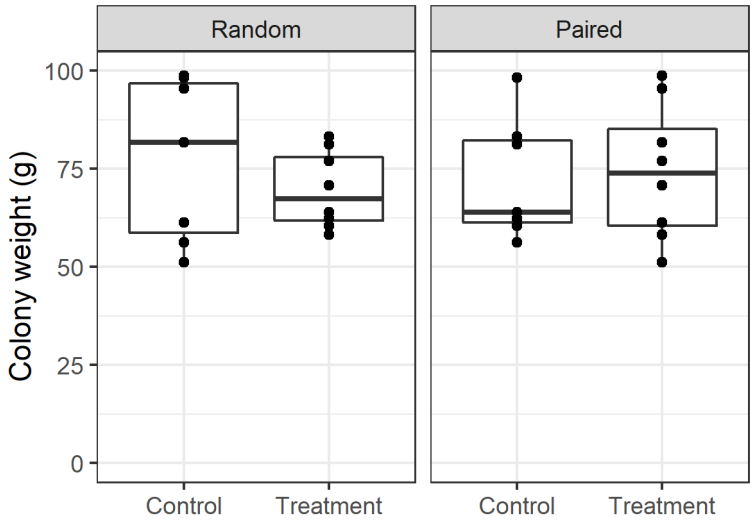


Figure 5 Outcome of assigning colonies to a treatment group randomly (left) or by ranking the available colonies by size into pairs and randomly assigning a member of each pair into a group (right). The mean values for each group for each method are not significantly different from each other however the variability (measured as SD) is almost twice as high in the control relative to the treatment group (Control SD = 8.49, Treatment SD = 4.2) in the randomly assigned groups whereas the variability is almost equal in the paired group (Control SD = 7.24, Treatment SD = 7.14). Boxes show the median, lower and upper quartiles of the data, whiskers show the minimum and maximum values, circles represent the individual datapoints.

Bees and colonies used should be free of signs of disease and of high quality (see also separate Sections on the species groups).

Bee hives/colonies/nests should be placed at the edge of the test fields. Care should be taken to find locations that fulfil requirements of the different (model) bee species (see separate Sections on the species groups) and are easy to access in order to be able to easily place the bees there and conduct the assessments.

3.5.6. Test substance and pesticide application

The test item should be clearly identified and characterised by providing: identity (common name, IUPAC name and CAS number), state or form, source, batch number and date of certificate of analysis of the formulation, purity of the substance for test conducted with the technical material, concentration of the active ingredient(s) for studies conducted with formulations and storage conditions of the test substance. The representative formulated end-product should be used or if a different formulation is used or there is reliance on effects data from another formulation, a case should be provided to justify the extrapolation.

3.5.7. Verifying alignment with ExAG

3.5.7.1. Dietary exposure vs contact exposure and comparison exposure in field effect study vs 'realistic worst-case' exposure in the environment

In order to conduct a valid field effect study an important requirement is that the exposure in the higher tier effect study represents not an average situation, but a so-called 'realistic worst case' environmental exposure, as mentioned in EU Regulation 1107/2009. For dietary exposure the 'realistic worst case' environmental exposure has been defined as the 90th spatial percentile of the 'Ecotoxicologically Relevant Exposure Quantity', the EREQ, which is the PPP residue intake of an individual, taken as an average for the bee colony located at the edge of the treated field (expressed as mass per bee per time unit).

The 90th spatial percentile residue intake for dietary exposure is represented by the PEQdi calculations in the defined exposure assessment tiers (Chapter 5 of the guidance document), with the screening tier and Tier 1 having stricter, i.e. higher predicted residue intake than Tier 2, where residue levels are measured in field exposure studies (Annex B of the Guidance Document). Methods described in Annex B demonstrated that the 90th spatial percentile residue intake can be obtained with reasonable certainty on the field with the highest residues found in nectar and pollen out of (i) 5 'minimum alternative forage' fields or (ii) 15 'randomly selected landscape' fields. Chapter 5 explains that the residue intake can be calculated by multiplying the measured residues in pollen and nectar by the estimated nectar and pollen consumption. In the higher tier field effect studies residue intakes need to be calculated by using measured residues from captured bees/bee traps. So, in order to conclude that the exposure in the field effect study is sufficient, the measured residue intake of the field effect study (the estimated exposure dose, EED) should be equal or higher than the predicted residue intake (i.e., PEQ_{di}) determined according to Chapter 5. The comparison should be carried out with the PEQ based on independent measured residue trials (Tier 2), but if not available, the PEQ from the lower tier exposure assessment will be used.

3.5.7.2. Calculating the EED in a field effect study

In the field effect study, exposure has to be verified by measuring the concentration in nectar and pollen collected by the bees. Field effect studies are done in a block design, with each block containing treated and untreated (control) fields. The residue level in pollen and nectar collected by bees has to be measured in the treated fields of each block. As described in Annex B of the Guidance Document, the residue level is measured in triplicate at several timepoints in each treated field. At each timepoint, the geometric mean concentration of the three replicate samples has to be calculated. The highest geometric mean concentration of the timepoints of each field, called $CONC_{max}$, is then used to calculate the EED for that field, as shown in Figure 6.

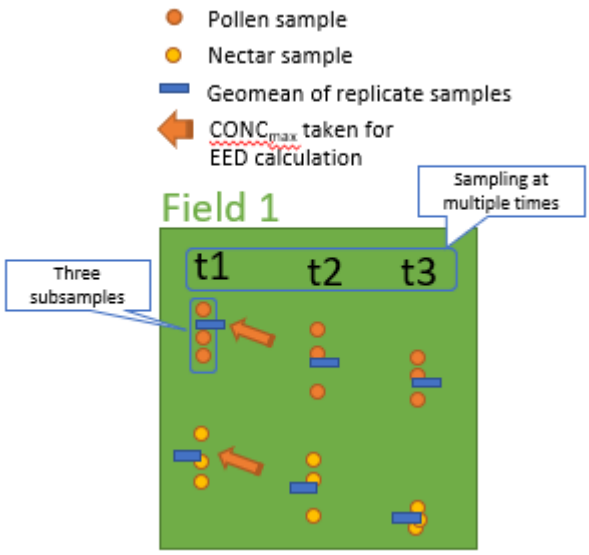


Figure 6 CONCmax derivation in a single field of a field effect study

Like the PEQ_{di}, the EED is calculated using the equations described in Chapter 5. However, most of the parameters in these equations are not relevant for the EED, e.g., the GAP dependent parameters. This is because in a field effect study, the concentration in pollen and nectar is directly measured in the study and not derived with the application rate and the default RUD distribution. The EED is derived by adding up the intake via pollen and the intake via nectar. For the acute EED, the maximum measured concentration in nectar and pollen ($CONC_{max}$) and the (measured or default) sugar content in nectar are used. For the chronic EED, these parameters are used together with the (measured or default) dissipation in nectar and pollen if application is overspray during flowering, and the time window (default). Note that for field effect studies with pre-flowering applications only (even if spray application is used), dissipation in nectar and pollen is not relevant and only the maximum concentration should be used.

The equations to calculate the EED are:

$$EED = \frac{1}{1000}\left(CONC_{po,j}\ CMP_{po,j}\right) + \left(CONC_{ne,j}\frac{CMP_{su,j}}{SN_{(field)}}\right)$$

$$CONC_{po,j} = f(CONC_{po,max} DT50_{po} w)$$

$$CONC_{ne,j} = f(CONC_{po,max} DT50_{po} w)$$

where:

EED is the estimated exposure dose in the field effect study in µg a.s./bee.

$CONC_{po,max}$ is the highest concentration in pollen collected by bees of the timepoints in a field effect study, expressed in mg a.s./kg; it is a geometric mean of triplicate samples at one timepoint

$CONC_{ne,max}$ is the highest concentration in nectar collected by bees of the timepoints in a field effect study, expressed in mg a.s./kg; it is a geometric mean of triplicate samples at one timepoint

$CONC_{po,j}$ is the concentration in pollen that is relevant for a risk case j, expressed in mg a.s./kg (note that it is the same as $CONC_{po,max}$ for acute, and may be a time-weighted for chronic adult and larvae)

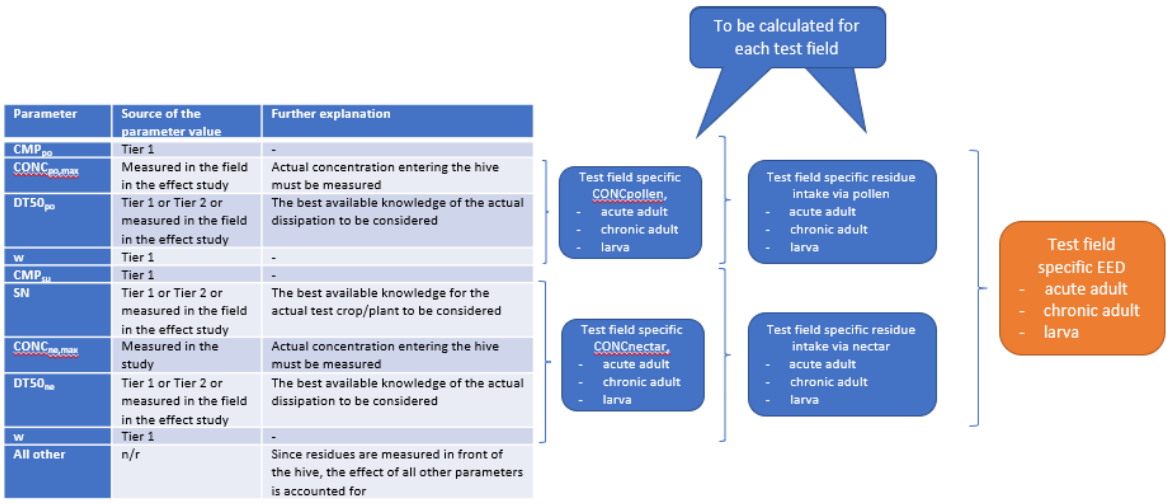
$CONC_{ne,j}$ is the concentration in nectar that is relevant for a risk case j, expressed in mg a.s./kg (note that it is the same as $CONC_{ne,max}$ for acute, and may be a time-weighted for chronic adult and larvae)

The other parameters are described in Chapter 5.

As shown in Figure 7, the value of the relevant parameters is either a default (Tier 1) value, a refined Tier 2 value (i.e., measured in a field exposure study), or a value measured in the field effect study itself. The EED is calculated for the three risk cases for each of the treated fields separately.

The 90th percentile EED can be calculated without a Monte Carlo analysis, since both $CONC_{max}$ and the DT50 are single values. SN is also a single value in Tier 1 but may be a range in Tier 2. Then, to arrive at a 90th percentile EED, the 90th percentile SN must be used in the calculation.

Note that there is no need to consider separate scenarios here. The EED is (normally) derived in a study on an attractive flowering crop and is aimed at covering all scenarios.



Parameter	Source of the parameter value	Further explanation
CMP_{po}	Tier 1	-
$CONC_{po,max}$	Measured in the field in the effect study	Actual concentration entering the hive must be measured
$DT50_{po}$	Tier 1 or Tier 2 or measured in the field in the effect study	The best available knowledge of the actual dissipation to be considered
w	Tier 1	-
CMP_{su}	Tier 1	-
SN	Tier 1 or Tier 2 or measured in the field in the effect study	The best available knowledge for the actual test crop/plant to be considered
$CONC_{ne,max}$	Measured in the study	Actual concentration entering the hive must be measured
$DT50_{ne}$	Tier 1 or Tier 2 or measured in the field in the effect study	The best available knowledge of the actual dissipation to be considered
w	Tier 1	-
All other	n/r	Since residues are measured in front of the hive, the effect of all other parameters is accounted for

Figure 7 EED derivation in a single field of a field effect study

The three EED values are calculated for each treated field of the field effect study. In this way, three ranges of EED values are derived, for the three risk cases, and the median and mean values of these ranges have to be calculated, as shown in Figure 8.

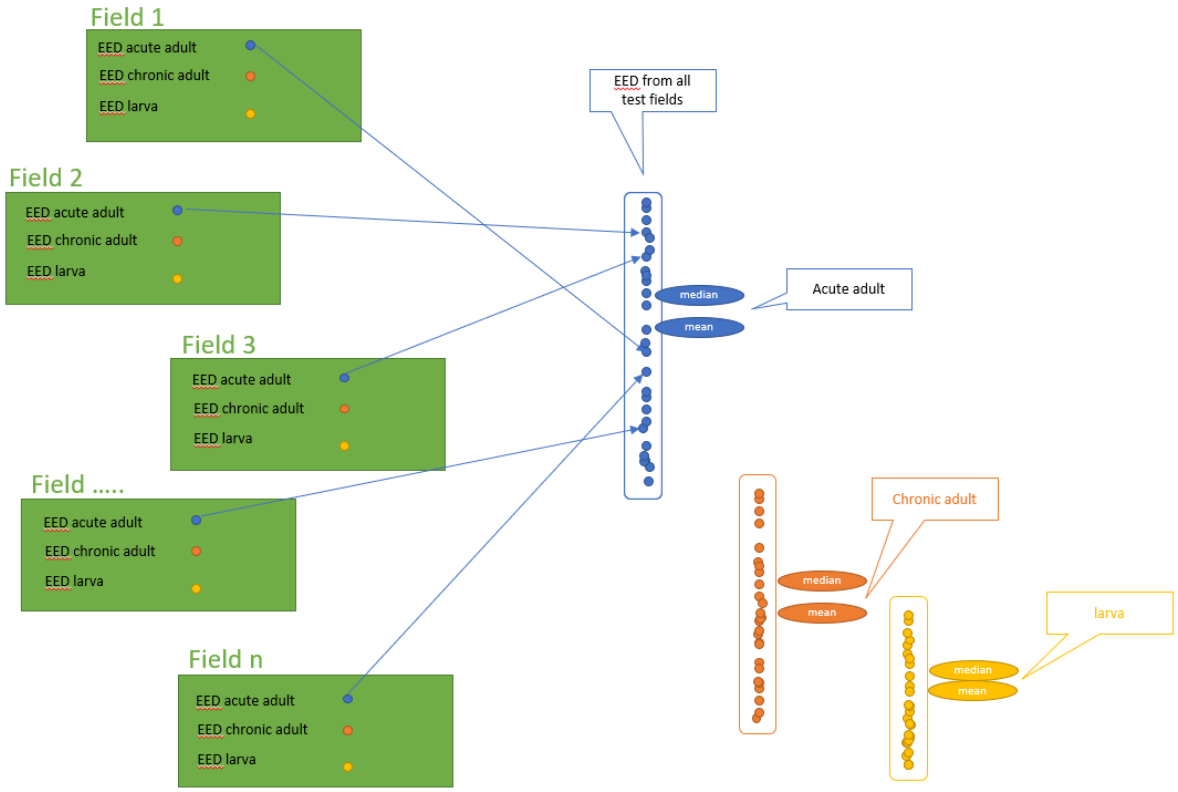


Figure 8 Overall EED derivation in a field effect study

3.5.7.3. Comparing the PEQ and the EED

Finally, the estimated exposure for the GAP under evaluation must be compared to the achieved exposure in the field effect study. To be able to conclude that sufficient exposure was reached, both the mean and the median value of the EED for each risk case must be equal to or higher than the PEQdi. This is shown in Figure 9.

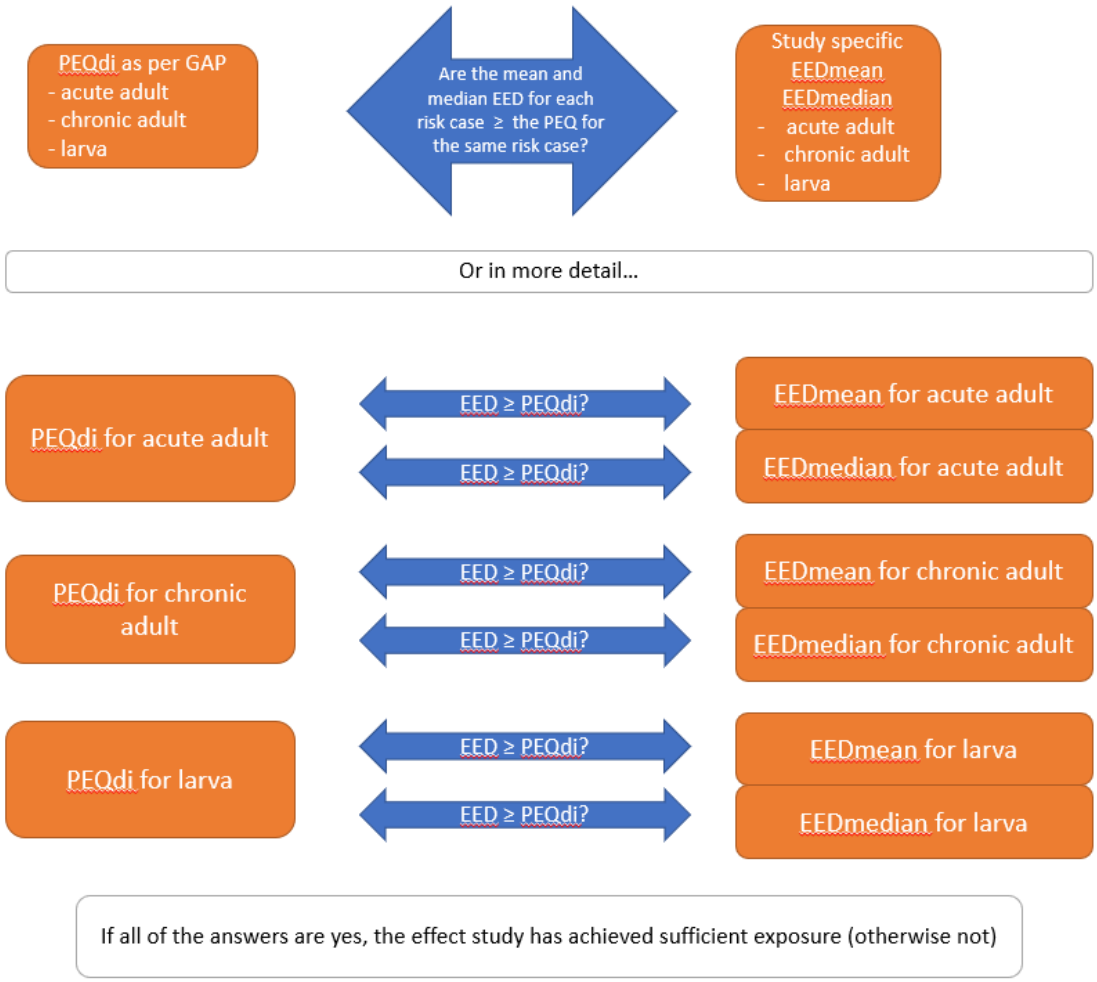


Figure 9. Verifying whether the exposure in the effect study meets the exposure assessment goal.

3.5.7.4. Considerations for extrapolation between bee groups

Preferably, for bee-collected Tier 2 measurements (i.e., in field exposure studies) of the residue level in nectar and pollen, honey bees are used for refinement of the honey bee exposure, bumble bees for bumble bee exposure and solitary bees for solitary bee exposure. However, if samples from bumble bees and/or solitary bees are not available, honey bee collected samples can be used to refine PEQ_{di} estimations for bumble bees and solitary bees, provided that the exposure field study was done on a honey bee-attractive mass-flowering crop.

In higher tier effect field studies, in principle residues must be collected by the bee groups separately. The exception to this is nectar for solitary bees since experience with sampling nectar from solitary bees is very scarce. In EED calculations for solitary bees, the Tier 1 RUD_{ne} should be used. Alternatively, it is possible to use the lowest measured concentration in nectar as collected in the field effect study by either honey bees or bumble bees (the lowest concentration will lead to the lowest EED and is therefore the conservative choice), provided that the solitary bees were kept in the same field as the other bees.

3.5.7.5. Contact exposure verification

If the risk stems primarily from contact exposure, then semi-field effect studies (in cages or tunnels) are to be performed at an application rate that covers the proposed GAP. Exposure is verified via

foraging activity measurements, and there is no need to measure the contact exposure level on the bees.

Alternatively, it is possible to refine the concentration on the bees (PEQ_{cn}) that is predicted in Tier 1, as described in Annex B and Chapter 5, however, experience with such studies is scarce.

If Tier 1 or 2 calculations indicate no dominant risk case for the overall colony/population risk and field effect studies are performed, it is considered sufficient to verify exposure via dietary intake (as described under 3.5.7.3), i.e., additional measures to verify exposure via contact are not needed.

3.5.7.6. Crop and application pattern with worst-case exposure of GAP vs used crop and application rate in the field effect study

Generally, the risk assessment starts by identifying the crop and associated application pattern that are expected to result in the highest exposure in the submitted GAP. Ideally, the crop, application timing (with respect to BBCH crop stage) and application rate in the field effect study coincides with the crop, application timing and rate identified as generating the highest exposure of all possible uses described in the GAP. Note that to get 'realistic worst-case' exposures it may be necessary to use an application rate that is higher than the one mentioned in the GAP. Multiple application, in field effect studies, may be converted to one single application that is expected to comply with the ExAG (unless repellence at the higher application rate is observed/expected).

However, the applicant might not follow the GAP in terms of application rate in the field effect study but apply the test item during the flowering even if the GAP is for pre-flowering application. Similarly, the applicant may replace the crop in the GAP with a more bee attractive crop (See 3.5.4) and additionally make the application during flowering. In such cases, the measured residue intake resulting from the application in the field effect study will, in many cases, be higher than the required 'realistic worst-case' exposure, as calculated in the dietary exposure assessment tiers. Thus, in such cases a higher tier field exposure study (Annex B) may not needed, because the lower tier exposure is relatively low: e.g., in case of pre-flowering applications of more than 15 days before flowering, a PFF factor of 0.33 or lower is applied to calculate the residue intake, PEQ_{di}. Thus, with other (bee-attractive) crops and an application during flowering (not requested for in the submitted GAP and/or not identified as the use resulting in the highest exposure) the residue intake in the field effect study is expected to often result in 'realistic' worst-case exposures or higher. This needs to be confirmed by the PEQ (of Tier 1, or Tier 2) being lower than the estimated effect dose (EED) in the higher tier field effect study.

4. Higher tier studies for honey bees

4.1. Study types and selection of study types for honey bees

Three types of studies are considered to be suitable by the Working Group to address specific regulatory question(s). These three types, namely field study, semi-field study and colony feeder study, are discussed in detail in the following Sections.

As explained in the above Sections, in all cases, the aim is to demonstrate that the colony strength (colony size = the number of adults that form the colony) is not impacted more than the magnitude dimension specified in the SPG after a field realistic worst-case exposure. This is best addressed in field studies. Nevertheless, in specific circumstances, other endpoints from specific types of higher tier effect studies could be used as a surrogate. This is the case when:

1) the lower tier risk assessments indicate that the impact on colony strength is clearly dominated by the risk arising from the contact exposure. In those cases, a semi-field test investigating forager mortality might be conducted.

2) the lower tier risk assessments indicate that the impact on colony strength is clearly dominated by the risk arising from the effect on larvae. In those cases, a colony feeder test investigating brood development might be conducted.

Otherwise, only field tests can address the risk indicated by the lower tier risk assessments. Those cases are when the impact on colony strength/population abundance is indicated to arise form:

- a combination of different routes of exposure

- a combination of effects on the adults and larvae
- primarily dietary risk to adults

The Working Group has defined when to consider that the overall risk is clearly dominated by either the contact risk or by the risk on larvae:

1) when the contribution of the risk originating from the contact route of exposure is at least 95% as quantified by the relative contribution Δ_j of an individual risk case j, then it is considered that the contact route of exposure clearly dominates the risk (on colony strength/population abundance). This also means that the dietary route of exposure contributes not more than 5% in the overall risk. The Working Group acknowledged that it will be the case only for those pesticides of which contact toxicity is considerable higher than the toxicity through the dietary route.

2) when the contribution of the risk originating from the effect on larvae is at least 95% as quantified by the relative contribution Δj of an individual risk case j, then it is considered that the risk to larvae clearly dominates the overall risk (on colony strength/population abundance). This also means that the risk to adults contributes not more than 5% in the overall risk. The Working Group acknowledged that it will be the case only for those pesticides of which are considerably more toxic to brood than towards adult forms (e.g. IGRs).

The relative contribution Δ_j of an individual risk case *j* (contact, acute dietary, chronic dietary, larva) can be derived from the calculations for the overall risk following the formula for the quantification of the contribution of a risk case to the overall predicted effect as given in Section 7.1.4 of the guidance document.

If the semi-field or feeder study indicates high risk, a field study can be performed to investigate the effect on colony strength under full field conditions. However, note that the overall conclusion will be drawn based on all available higher tier information by performing a Weight of Evidence (see GD Chapter 10).

The recommendations for study type selection for honey bees are illustrated in

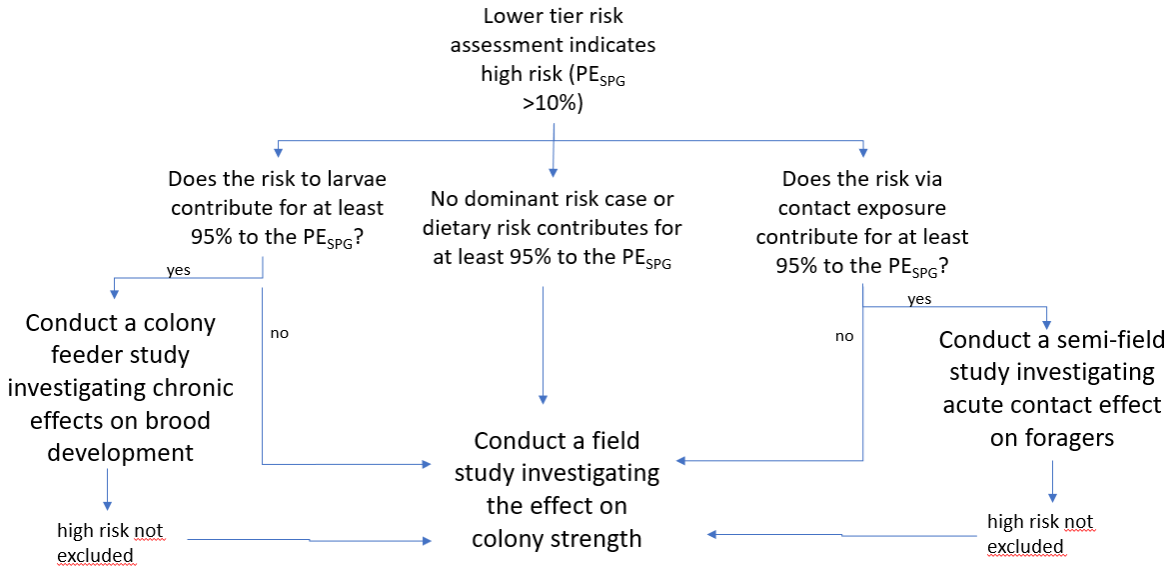


Figure 10.

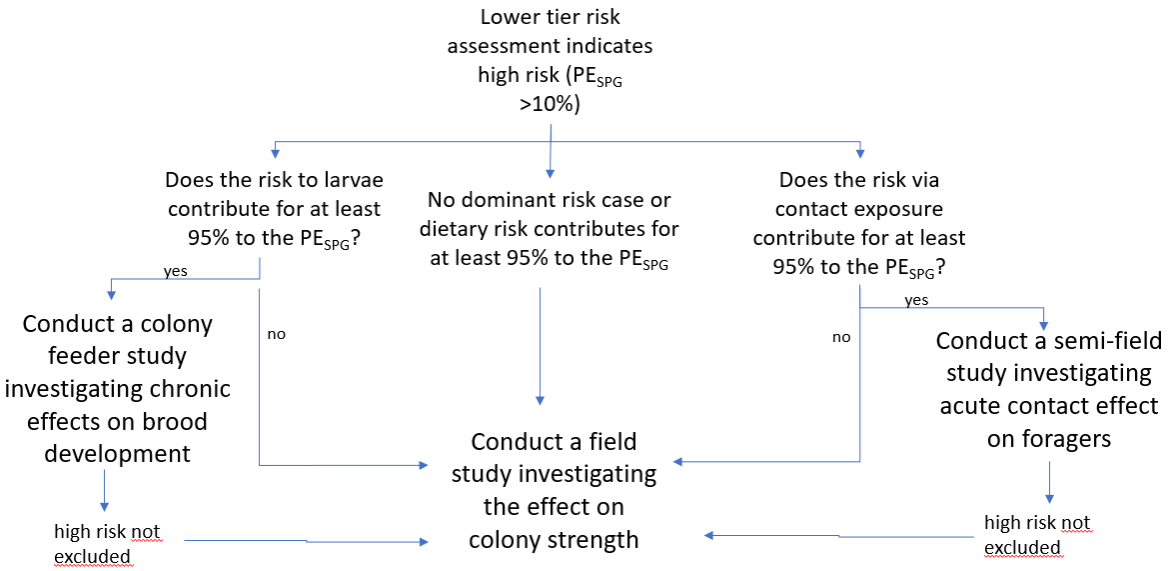


Figure 10: Study type selection for honey bees

In addition, both semi-field and field studies may be used when concerns for sublethal effects on foraging behavior are identified in the lower tier risk assessment conducted for sublethal effects (see chapter 9, and 4.2.4 and 4.3.3, below).

4.2. Honey bee field study

4.2.1. Study aim and setup

The study aim and setup are detailed in the general field effect study design guidance (see Section 3, above). In summary, honey bee colonies are placed at the edge of test fields with flowering plants and it is ensured that bees forage mainly from plants on the test fields.

4.2.2. Creation of colonies and initial conditions

The honey bee colonies should be healthy and created to be as similar as possible at the start of the study. Colonies should be created at least one week before the start of the study (Lückmann and Schmitzer, 2019) and given a unique identification number that should be attached to all data collected on/from this colony. Colony creation could preferably be done by one beekeeper (or team of beekeepers working together). If multiple beekeepers are used, they should be balanced between treatments, i.e., the same bee origin and beekeeper within a block. Queens should preferably be sisters and less than two years old. If there are several queen lineages and ages, these should be evenly distributed between control and treatment sites.

Colonies should be of equal strength (as much as possible, with regard to number of adult bees, amount of brood in different stages and food stores) initially and allocated to a treatment groups (control, exposed) at random or structured way (see Box 2 in 3.5.5). Although the choice of colony size is up to the applicant, OEPP/EPPO (2010) and Lückmann and Schmitzer (2019) recommend using colonies containing at least 10000 adult bees, which may be adapted to follow regional and seasonal weather and beekeeping practices. However, other effect studies have successfully used smaller colonies, starting with 3000 adults bees (Rundlöf et al., 2015; Woodcock et al., 2017) which provides advantages in both handling the colonies and controlling variability and exposure. Colonies should consist of multiple brood combs and brood in all stages as well as adult bees. There should be limited food stored, to motivate foraging on the treated fields.

The most important consideration is that the beekeeper(s), queen linage(s) and colony creation, size and initial conditions are as similar as possible for colonies at sites within a block. However, it may reduce variability to maintain similar condition also among blocks within a study.

Colonies should be cared for using good beekeeping practices. Colonies should not receive any medical treatment within four weeks before the start of the study and there should be no visible signs of disease (e.g., Medrzycki et al. (2013)). It is recommended, but not required, to provide analytical confirmation of the disease-free status. Colonies should be placed at the edge of the test field in a wind-protected location and without excess disturbance, with easy access for placement and assessments. Colonies should be placed at least three days before assessments starts, to allow acclimatisation to the local conditions (Lückmann and Schmitzer, 2019).

4.2.3. Determination of exposure

In any effect study, it is important to determine the ecotoxicological relevant exposure so that it can be confirmed how it relates to the realistic worst-case exposure expected in the area of use of a PPP (exposure assessment goal - ExAG, i.e., the 'realistic worst case' pesticide mass uptake per bee per time unit). In the field effect study, it needs to be demonstrated that the bees at the treated sites have been exposed to the test item at or above the ExAG and that the bees at the control sites have not been exposed. The determination of exposure, and lack of exposure, to the test item is done following the methods of the field exposure study that includes bees (see Annex B). However, in field effect studies, more hives per field will be present compared to higher tier exposure studies (in which only one or two hives per field are used for sampling). While in the effect field studies determining exposure to all the hives in a field would be best, this likely has practical limitations. Therefore, it is recommended to sample bees from at least half of the hives in each test field, with an absolute minimum of two hives per field. Samples from bees of different hives must be pooled. More details on the comparison of the exposure level in the effects study with the field exposure studies (EED vs PEQ) is described in Section 10.5 of the Guidance Document.

4.2.4. Determination of effects

The most valuable endpoints in effect studies are those that can be directly linked to the SPG. The primary assessment endpoint directly linked to the SPG is for honey bee colony size measured as the number of adult bees in a colony (EFSA et al., 2021). Honey bee colony strength can be estimated using two methods: 1) the Liebefeld method (Imdorf et al., 1987; Dainat et al., 2020) and 2) digital photography combined with image analysis (Wang et al., 2020; Bozek et al., 2021), with only minor disturbance to the colony from opening the hive and taking out frames for inspection or photography.

The Liebefeld method is based on one or more observers estimating the number of adult bees in a colony and sometimes also includes the brood and food stores cells (Imdorf et al., 1987; Dainat et al., 2020). The observer visually estimates the number of adult bees and/or the area or proportion of the comb sides covered with bees and/or capped (covered) brood cells, which in the case of area and proportion can be translated to the number of bees and capped cells (Imdorf et al., 1987; Dainat et al., 2020). Training and continuous calibration is important for consistent and precise colony strength estimates (Dainat et al., 2020) and experienced observers can achieve an accuracy of 96% for adult bee number and 99% for capped brood estimations (Imdorf et al., 1987). Using two observers with assistance for taking notes (either an additional person or audio recording) would increase the quality of the estimations. The assessments should be made early or late in the day when most bees are in the colony (Imdorf et al., 1987; Dainat et al., 2020). However, in large experiments with many colonies (~100), it is usually not possible to restrict the assessment timepoint like this. In that case it is important to alter the assessment timepoint among experimental groups.

The digital photography combined with image analysis method builds on similar principles as the Liebefeld method. However, instead of a visual inspection of all comb sides of the frames, digital photos are taken (Wang et al., 2020). The number of bees is then estimated from the photos visually or using an image analysis software (Wang et al., 2020).

Automated monitoring of honey bee colonies is under development (Marchal et al., 2020). The method is promising, but there is currently insufficient information and experience to provide detailed guidance.

There may be other relevant computational methods for estimating colony strength in the future, at which time additional guidance on their use can be considered.

In addition, sublethal effects on foraging behaviour can be studied in field studies. The most relevant endpoint is the amount of pollen collected per flight, or, as a proxy for this endpoint, the number of bees returning with pollen. The duration of a foraging flight is a useful endpoint to refine concern from the homing flight study. Chapter 9.5 of the guidance document gives further details about these endpoints.

4.2.5. Study duration

The study should assess at least two brood cycles (42 days) to ensure that a large part of the brood is exposed to any residues that are brought into or stored within the colony. Colony size estimations should be performed approximately every 10 days, as a compromise between tracking changes in colony strength and minimizing disturbance to the colony. The first assessment is done on the day before (d-1) or on the day (d0) of spray application and the day before or on the day of placement of colonies at the sites in case of soil or seed treatment or pre-flower spray application. Thereafter, assessments are done at 10 ± 1, 21 ± 1, 31 ± 1, 42 ± 1 days after application or placement at the sites (the latter in case of seed/soil treatment/pre-flower spray application).

In case if the PPP use under evaluation indicated a concern for the winter bee scenario related to time-reinforced toxicity (TRT, see chapter 8 of the guidance document) that were not solved in the lower tier assessments (therefore higher tier assessment was triggered), then the study should not stop at day 42 ± 1 after application or placement at the sites. In those situations the colonies (both from the control field and the treated field) should be moved to a common place for overwintering. It is recommended that there should be minimal pesticide use within 4 kms of the over-wintering apiary,. Such studies should not be started later than September in order to ensure that the observation period covers at least half a year (see also in 8.2.4 of the guidance document). After the winter, at the beginning of the foraging season in spring, at least one additional colony assessment must be performed and reported.

4.2.6. Environmental conditions

The environmental conditions during the whole study period should be measured at the sites and recorded (at least daily min/max air temperature, daily rainfall, daily min/max air humidity, wind conditions and directions, daily hours of sunshine and daily total solar radiation). Alternatively, data from weather stations no more than 20 km distance from the experimental fields should be used. The conditions in treatment and control sites should be comparable at the beginning of the study. It should also be described whether they were comparable during the study and if the weather conditions during the study are expected to have influenced the study, e.g., rainy and/or cold conditions that may have influenced the foraging activity of the bees. Additionally, it should be clearly documented whether the weather conditions at the sites are representative of the usual climate in the respective area of use.

4.2.7. Data analysis and interpretation of the results

The analysis should follow the recommendations in Section 2. A one-sided equivalence test ($\alpha = 0.2$) should be applied with an equivalence limit corresponding to a 10% decrease in the treatment with respect to the control. The statistical model and scale used for the analysis should be chosen appropriately (see also the discussion in Section 3.4). One option is to analyse the data separately for each assessment and test for equivalence at each different timepoint. It is also possible to analyse all timepoints together with a single equivalence test, provided that the interaction between treatment and timepoint is included in the model and tested for significance. If equivalence is proven, that is if the relative decrease in the test is smaller than 10%, then it is possible to conclude that there is low risk. Otherwise, it cannot be concluded that the SPG is respected.

4.3. Honey bee semi-field study

4.3.1. Background

The WG propose using semi-field studies when, based on lower tier risk assessment, a low risk to colony strength is not excluded and is driven by contact exposure or, if based on lower tier tests, a concern for sublethal effects is indicated. The WG decided to limit the use of semi-field tests to these two scenarios as, whilst semi-field studies have been used to investigate side effects of chemicals to honey bees for decades, there are multiple characteristics of these tests that limit their usefulness in risk assessment. In keeping with EFSA, 2013a, the WG recognize that semi-field studies on honey bees have a range of limitations; the studies are short term, the foraging area is limited and the test colonies are often much smaller and vary in composition from typical colonies. Additionally, the behaviour of bees can be altered by long term confinement, which further limits the length of the studies and precludes extended studies on bee development and, whilst it is possible to assess colony strength as part of a semi-field study, it will not be possible to assess an impact on the development of the colony. Nevertheless, the tests are suitable for monitoring forager bees as they perform foraging activity on the crop for several days. Considering the above, the test is only considered useful to investigate forager mortality, flight activity, and foraging behaviour on the crop.

The proposed method is largely based on the guidance given in EPPO PP 1/170 (4). Nevertheless, for some aspects, other available guidance for semi-field test, such as OECD Guidance document No. 75 (OECD 75) and proposed recommendations for its improvement (Pistorius et al., 2011), were also considered. It is noted that OECD 75 focuses on brood development as primary endpoint, however brood development is not a relevant endpoint in this context.

4.3.2. Limitations and usability of the test

Under adverse climatic conditions, the foraging activity might be low (for more details, see Section 7 of OECD 75). Only one pesticide application during flowering is foreseen in this test.

4.3.3. Scope and aim of the test

The use of the semi-field studies in the context of this guidance document are twofold:

- Inform potential effect on colony strength by investigating forager mortality
- Inform the assessment of sub-lethal effects

Investigating potential effect on colony strength by investigating forager mortality

The test may be used when the lower tier risk assessments indicate that the impact on colony strength is clearly dominated by a risk arising from contact exposure (see the Section 4.1).

Since contact risk is considered to arise from acute exposure, the test must cover an exposure event during or shortly after (within hours) the PPP application. The test should be able to detect an increase in mortality within a couple of days of the exposure event. If no effects are observed within that time frame, it is assumed there are no delayed effects induced by this exposure. Contact risk arising from pesticide spraying during the flowering period of the treated crop, weeds or field margin may be addressed by this test. The test can be used to investigate effects of daytime pesticide applications (when bees are actively foraging) and pesticide applications outside the daily flight activity of the foragers. The test is usually performed for spray applications. Conceptually, exposure arising from dust deposition may also be addressed, however experience with such a contamination route in the context of this test is scarce, therefore no specific guidance can be given.

Investigating sub-lethal effects

The test may be used when a potential concern for sublethal effects from contact and/or dietary exposure has been identified in the lower tier assessment for sublethal effects (see Chapter 9 of the Guidance Document).. The most relevant endpoint is the amount of pollen collected per flight, or, as a proxy for this endpoint, the number of bees returning with pollen. Chapter 9.5 of the guidance document gives further details about these endpoints. Note that the duration of a foraging flight cannot be refined in a semi-field study.

4.3.4. Method

For the aim as defined above, the study design described in the "Semi-field tests" Section of EPPO PP 1/170 (4) is considered suitable. However, since the scope and the aim of the test in the context of this guidance document is more limited than in the EPPO document, further descriptions are given below in order to amend and clarify some of the aspects of the study design and interpretation.

Cage size

Although the working group generally recommends following the EPPO PP 1/170 (4) as to cage size, it is noted that a minimum crop area of 40 m^2 is given, whereas follow-up publications have proposed larger areas (e.g., Pistorius et al. (2011)). This is assumed to be feasible, as reports of semi-field studies with significantly larger areas (100-120 m^2) are commonly submitted to support pesticide dossiers. The WG therefore recommends enclosures sizes to contain 80 m^2, or more, of crop area.

Water-permeable sheets

Following the EPPO PP 1/170 (4), water-permeable sheets should be placed in the cage to collect dead bees. The Working Group recommend that the following additional requirement should be implemented:

- At least 15% of the surface area of the cage should be covered by water-permeable sheets
- The ends of the cage, including the corners, should be covered by water-permeable sheets at least 0.5 m wide
- The front of the hive (the entrance of the hive) should face the water-permeable sheets

These requirements were set to maximize the probability of collecting a large proportion of bees that have died outside of the hive.

Crop

The Working Group considered that the crop specified in the GAP should be the first choice for a test crop. However, if the crop specified in the GAP is not considered suitable due to its limited attractiveness to bees (note that a high level of exposure must be demonstrated), or due to agronomical reasons or risk envelope considerations, then other options may be acceptable. Semi-field tests are often conducted with highly attractive flowering crops, e.g., *Phacelia tanacetifolia*, which the Working Group would consider a suitable surrogate crop for these studies.

Colony size

EPPO PP 1/170 (4) recommends small test colonies with approximately 3000–5000 bees. The Working Group considered that smaller colonies than this range would not be appropriate to study forager mortality, but larger colonies might be considered. However, the size of the colonies (with consideration also to the brood) should be appropriated for the size and the amount of food available in the cage.

Controls

EPPO PP 1/170 (4) recommends the use of two control groups: a negative and a positive control. The crop in the negative control cages should be treated with the same method as the crop in the test item treated cages, but with water (e.g., over-sprayed with similar volume of tap water). The crop in the positive control cages should also be treated with the same method, but with a pesticide known to be toxic to bees. The application rate of that pesticide should be high enough to cause considerable forager mortality. The pesticide mentioned in the EPPO document is dimethoate. The Working Group acknowledges the experience gathered with dimethoate but notes that other pesticides might also be suitable positive controls.

The role of the positive control is to demonstrate that the test system is suitable to study forager mortality and that there is a cause-effect relationship (exposure-effect relationship). This is rather straightforward for tests with pesticide applications during the flight activity of the bees, but less obvious for applications outside the active period. If the test item is applied outside the active flight period (e.g., at night), but the positive control is applied during flight activity (e.g., during the daytime), the positive control demonstrates the general sensitivity of the system, but cannot demonstrate that the system is able to adequately capture an exposure-effect relationship of the test item, as different processes may affect exposures and effects during flight activity vs outside of flight activity. In order to prove that the system is generally sensitive and that it is capable of capturing an exposure-effect relationship, should

it exist, the WG recommends that the positive control always be applied in the same way as the test item.

Replication

EPPO PP 1/170 (4) mentions that the number of replications should be sufficient to allow appropriate risk assessment, which is normally the minimum of three, but a lower number may be appropriate. However, considering the proposed endpoint and aim of these studies, fewer than three replicates for the treatment and negative control groups is not sufficient. If a positive control is also included, at least two replications for the positive control are required. Increasing the number of the replications will make the observations more robust. It is noted that Pistorius et al. 2012 proposed at least four replications. All cages should have the same dimensions and same set-up, and the hives should be randomly allocated to the cages.

The conclusion was supported by a power analysis for the equivalence test carried out on ten semi-field studies. Despite the limitations of the data (not discussed here) the analysis clearly showed that three replicates should be sufficient to prove equivalence within the 10% limit (using a conservative estimate of variability) provided that the true effect size is zero or relatively small. As was also shown for field studies, a higher number of replicates would be required in order to determine whether the effect is above or below 10% when the actual effect is rather close to that value. As for field studies (Section 3.4), it is possible to analyze how many replicates are sufficient based upon effect-level assumptions in order to determine an adequate test set-up, evaluate the outcome and plan for additional replicates based on the results.

Assessment of the exposure

As in every toxicity test, sufficient exposure of the test organisms must occur. EPPO PP 1/170 (4) recommends an observation method to demonstrate that a sufficient number of bees are foraging on the treated test crop. For semi-field tests aiming to study forager mortality or foraging behaviour, the Working Group considered that this method is suitable to demonstrate that the exposure of the test bees is comparable with field realistic worst-case situation (i.e., least a 90th percentile exposure situation). The Working Group highlighted that the relative time between the pesticide application and the foraging activity is crucial. For pesticide applications made during the active foraging period, a sufficient number of foragers should be present on the crop during and shortly after the pesticide application. For pesticide applications made outside the active foraging period enough foragers should be present on the crop at the earliest potential exposure time after application (likely the next morning).

Further details on those observations are available in sections further below.

Feeding

EPPO PP 1/170 (4) mentions that feeding of the colonies may be necessary. Since feeding the colonies may reduce foraging on the crop, the WG does not recommend feeding during the test (which is anyway relatively short).

Observations

1) To inform potential effect on colony strength by investigating forager mortality, the following observations as described in EPPO PP 1/170 (4) must be performed:

- number of dead bees in the dead bee trap
- number of dead bees on the water permeable sheets
- foraging activity on the crop
- weather conditions
- colony size (number of adult bees)

2) To inform the assessment on sub-lethal effects, the endpoints as detailed in section for sublethal effects (chapter 9) must be investigated. In addition, it must be demonstrated that bees were sufficiently exposed, as described above.

The WG noted, that in the recent years novel methodologies had been developed or are under developments (e.g. bee counter or video recording at the hive entrance, RfID technology, digital

photography, hive weighing, etc.). The WG investigated whether it would be possible to recommend any/some of those methods to replace conventional observation methods, however, although several are very promising, due to a lack of experience needed to set best-practices, the Working Group does not recommend immediate use. Further research and experience is recommended.

Additional notes

Any observations other than those listed above are considered as a plus, but they are not considered essential.

It must be noted that the crop area, the colony size, stored food in the hive, attractiveness of the crop, the potential exposure (number of bees exposed during or shortly after the application) and therefore the sensitivity of the test are interrelated parameters.

If semi-field conditions are selected to assess sub-lethal effects, particular attention should be paid to the colony composition (such as adult-to-brood ratio, availability of stored food) since significant differences between the treatment and the negative control groups at the beginning of the test may also cause significant differences in some endpoints (e.g. amount of pollen collected).

4.3.5. Interpretation of the results

Validity criteria

EPPO PP 1/170 (4) has no validity criteria, nevertheless, it includes qualitative criteria for mortality in the negative and positive controls. The Working Group considered that the demonstration of a sufficient level of exposure is also a key issue.

Exposure (foraging activity on the crop)

The foraging activity on the crop shortly before the treatment should be in line with the prescription of EPPO PP 1/170 (4)) (e.g., 5 per m^2 if the test crop is *Phacelia*). If the number of bees foraging on the crop are less than prescribed, then it will be assumed that the exposure of the foragers during and shortly after the spray applications was not sufficient and the test will be considered not to be compliant with the exposure assessment goal. It is noted and accepted that for applications made outside the flight activity of the bees, this observation cannot be performed. In those cases, the focus should be on the observation made at the earliest potential exposure time following application (likely the next morning).

In addition to the initial assessments, the Working Group recommends performing foraging activity observations at least twice every day throughout the test (at least for five days after the start of the exposure).

If, at a later time point (i.e., some hours after the application or on the next day), the flight activity is considerably decreased in the treatment group compared to the negative control, this may be interpreted as a repellent effect of the test item. If at a later time point the foraging activity in both the treatment group and the negative control is lower than prescribed (e.g., due to the weather conditions) then it should be noted that the potential exposure time was shorter than in many field realistic conditions. In both cases, this information should be taken into consideration for the risk assessment. In general, two days exposure with intensive flight activity should be considered as sufficient to study acute forager mortality.

Weather conditions

The weather conditions are not directly used to judge the results of the study, but will be used to interpret potential anomalies, if observed.

Forager mortality

The Working Group proposes two methods to assess the dead bee counts. The second must be used only if, based on the first, low risk cannot be concluded.

It is noted that, although the endpoint is called "forager mortality", it will not reflect solely the mortality of the forgers, but also the mortality of the in-hive bees. This is because counts from dead bee traps are used. Nevertheless, significant mortality of the in-hive bees is not expected, therefore the mortality count still largely represents the forager mortality.

Both methodologies rely on an estimate of the total cumulative mortality (all foragers that die in the cage) of the first 5 days of the exposure, calculated for each replicate and treatment. It is noted that the total number of dead foragers includes not only the bees found in dead bee traps and in sheets (which are counted), but also the bees dying in the cropped area, which are not observed. It is therefore necessary to estimate a dead count for the cropped area. For this purpose, the Working Group proposes a linear extrapolation based on the proportion of the area of the sheet and the cropped area. The linear extrapolation (that assumes that the same number of bees dies in 1 m^2 of the crop as in 1 m^2 of the sheet area) was considered as a worst-case approach. This is because of the well-known phenomena of the 'cage effect' which indicates that a considerable proportion of bees dies at the edges of the cages.

Extrapolation example: the ratio of the crop area and sheet area is 4:1. In a given (treatment) cage, a total of 20 dead bees were found in the dead bee traps and 80 bees were found dead in the sheet area in the first 5 days after the start of the exposure. Since the sheet area is five times lower than the cage area, it is assumed that five times more bees died in the total area of the cage, than were found within the sheet area. In this case a total of 400 (5 x 80) dead bees can be estimated based upon the bees found on sheets, and 20 bees were found in the dead bee traps, resulting in a total bee mortality of 420 dead bees.

1) mortality in the test cages only

A low risk can be concluded if the cumulative mortality is not higher than 10% of the total number of the adult bees in the colony, based on the colony strength estimation at the beginning of the test. In the example above, assuming that the initial colony strength assessment estimated 5000 adult bees, the extrapolated total mortality of 420 dead bees would be <10% of the initial adult population. As the total mortality in all the test item treated cages is less than 10%, the risk is low because the difference with the mortality in the control cages is never going to be larger than 10%.

Should the extrapolated total mortality in any of the treatment cages exceed 10% of the initial total number of adult bees, a low risk has not been shown based on this methodology and the second methodology must be considered.

2) mortality in the test and control cages

The mortality observed in the treatment cages should be compared with the mortality observed in the negative control. The daily and the five-day cumulative mortality in control cages should be calculated with the same method as for treatment cages. A one-sided equivalence test ($\alpha = 0.2$) should be applied with an equivalence limit corresponding to a 10% increase in the average mortality in the treatment with respect to the control. If the test indicates that the difference between the treatment and the negative control is smaller than 10%, then it is possible to conclude that there is low risk. Otherwise, it cannot be concluded that the SPG is respected.

Colony size (colony strength)

EPPO PP 1/170 (4) prescribes an initial colony assessment (before the exposure) and another one at the end of the exposure. For the scope and the aim of the test, the only requirement is to estimate the number of adult bees of the colony. The results of these assessments are considered only as additional information that help to interpret the forager mortality counts and they cannot be used for direct comparison with the SPG (i.e., a less than 10% decrease in colony strength in the treatment compared to the negative control based upon the colony assessments cannot lead to the conclusion that the SPG is met). The first colony assessment will be used to calculate how many dead bees would represent the 10% of the colony. The second assessment will be used to compare with the initial assessment and then compare the trend observed in the treatment with the trend in the negative control. The WG considers that only qualitative comparison is necessary to check whether the trends are in line with the trends seen in the assessment of forager mortality. If the assessment of forager mortality indicates that the SPG is respected, but significant depopulation is observed in the treatment compared with the control, then the conclusion from the assessment of forager mortality might be overruled. No detailed guidance can be given regarding what should be considered as "significant depopulation", therefore a case-by-case assessment is required for the judgement of this endpoint.

4.4. Honey bee colony feeder

If the lower tier tests indicate that the larval risk to larvae is the dominant risk (see Section 4.1, above) of a PPP, then a colony feeder study may be used to refine the risk. The colony feeder intends to mimic the exposure of a colony to a PPP via spiked food offered directly in the hive for nine days. This chronic exposure period covers the pre-imaginal brood stages and ensures that all larval stages are exposed.

The primary advantage of the colony feeder study is that the concentration and duration of the PPP exposure are under the control of the investigator; also, herbicides can be tested without causing adverse effects on crop plants as may occur in semi-field and field studies (Lückmann and Tänzler, 2020).

For honey bees, the feeder study is only recommended for assessing the risk to brood. This is because the exposure route for the colony is highly artificial and cannot mimic actual foraging activity on pesticide treated crops (US-EPA (Environmental Protection Agency), 2016) and is therefore not considered fully suitable to assess the risk to forager bees. Provided that the study conditions are such that alternative forage is scarce and there is no excess food storage (see below), the spiked food provided to a colony will enter the colony provisions and the brood provisioning should be very similar to natural conditions.

The method described below is largely based on Lückmann and Schmitzer (2019), which is an update of the Oomen method (Oomen et al., 1992) that also covers chronic exposure. Also, the recommendations for colony feeder studies from the US-EPA (US-EPA (Environmental Protection Agency), 2016) were taken into account.

4.4.1. Test procedure

Colonies are kept outdoors under free-flying conditions and provided with food containing a known concentration of pesticides for nine days. The endpoint in the current protocol is the amount of capped brood. This is the closest endpoint to the SPG, as opposed to more traditional brood endpoints such as the Brood termination-rate and the Brood index. While eggs have proven to be the most exposed and sensitive pre-imaginal stage (Lückmann and Schmitzer, 2019), following the amount of capped brood over two brood cycles will also catch a disruption in brood development. Of all brood stages and food cells, capped brood can be best distinguished and can therefore be determined with the highest precision (e.g., Alves et al. (2020)).

Other endpoints such as adult mortality, behavior and colony strength may be noted and considered as supplementary information.

4.4.2. Test period

The test should be performed when the bee brood is growing, which will be only during spring and summer in the central and northern European regions and spring and autumn in the Mediterranean zone. Periods when the brood is decreasing and colonies are preparing for overwintering or oversummering should be avoided.

4.4.3. Test site

The test is performed at a single location.

To maximize exposure and avoid dilution, and to avoid contamination with other pesticides, the site should be chosen such that during the feeding period, the requirements for "minimum alternative forage" field studies as described in Annex B of the guidance document are met.

If colonies are in danger of starvation, additional feeding may be necessary, but only after the second brood assessment (see 4.4.5). If feeding is necessary (with sugar solution and/or pollen patties), all colonies should be treated in the same way. Also, a bee-attractive crop that can offer both pollen and nectar to the test colonies may be planted at the test site; it should only come into flower after the exposure period.

During the whole study period, the study area must be characterized for available forage and potential pesticide contamination.

4.4.4. Preparation and condition of the colonies

The recommendations of Lückmann and Schmitzer (2019) should be followed:

- Colonies are prepared at least 1 week before treatment.
- Colonies should not receive medical treatment at least 4 weeks before the start of the study.
- Queens should be of the same age, in the reproductive phase and not more than 2 years old; sister queens should be used if possible .
- Initial colony strength is recommended to be 10,000 -15,000 bees, but smaller or larger colonies may be accepted, e.g., if relevant for the region or time of year.
- Colonies are recommended to consist of a third to half of the total combs in the brood frame with brood. In hive types with 10 combs per frame, three to five should be combs with brood. In the brood comb(s) of a colony, all brood stages should be present, i.e., eggs, larvae and pupae (capped cells).
- The amount of stored food prior to the exposure period can affect the extent and timing of the exposure to the artificial food source, thus effects on the colony may be delayed due to delayed consumption of treated food. Colonies should not be in nutritional stress but excessive nectar/honey stores and excessive pollen stores should be avoided to ensure that bees are consuming the sucrose solution (and/or optionally pollen substitute) provided during the experiment as much as possible. See e.g. FAO (Food and Agriculture Organisation) et al. (2021) for good beekeeping practices related to nutritional requirements.
- Colonies should be free of visual symptoms of diseases according to good beekeeping practice and without unusual occurrences.
- Colonies should be as homogeneous as practically possible at the start of the study with regard to daily mortality, strength, brood and food stores. It is recommended to have a surplus of colonies prepared and select the most suitable ones based on data collected before the test starts. For more guidance on randomization of the test colonies, see chapter 3.

4.4.5. Setup of the colonies and duration of the study

Colonies should be set up at the test site at least 3 days before the start of the study in order to become familiar with their new surroundings. Accordingly, disturbance and mortality due to translocation will be minimised.

Beginning with Capped Brood Determination Day (CBDD) 0 the study will last 42 days, covering two brood cycles. The study can optionally be extended by assessing additional brood cycles, each lasting a further 21 days.

A study overview is shown below, redrafted from Lückmann and Schmitzer (2019):

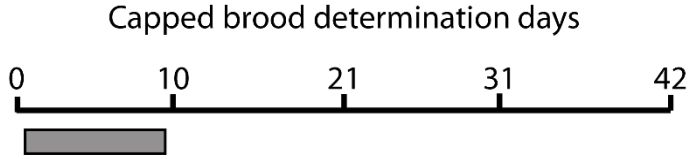


Figure 11: Honeybee colony feeder study overview, redrafted from Lückmann and Schmitzer (2019). The capped brood are measured on the day before the 9 day feeder exposure (grey block) starts, and on days 10, 21, 31, and 42 (± 1 day). The applicant can optionally record mortality or behavioural data daily during the study period.

4.4.6. Treatment groups and replication

The Working Group analyzed capped brood data from single timepoints in control hives from colony feeder studies and found high levels of between-colony variability. An explanation for this high variability could not be found. Therefore, the Working Group cannot provide an indication of the number of colonies that would be needed in treatment and control to prove equivalence within a 10% limit with sufficient power. However, it is also possible to plan for a specific number of colonies, evaluate the outcome and based on this, plan for additional blocks (additional colonies for each treatment group) at the same study site. Potential additional blocks should be preregistered in advance. i.e. reported in the study protocol.

If only a spiked nectar substitute is offered, fenoxycarb must be used as a reference substance. Four hives must be used in the reference group, following Lückmann and Schmitzer (2019).

If pollen as well as nectar are spiked, and significant uptake of the offered food is seen, a reference group is not necessary, since in that situation, maximum exposure is assumed to be reached.

4.4.7. Mode of treatment

The Oomen test uses spiked sugar solution only, and published feeder studies have used either spiked sugar solution (e.g., Faucon et al. (2005)) or spiked pollen substitute (e.g., Pettis et al. (2012), Tsvetkov et al. (2017); see Appendix B). It is noted that for many proposed PPP uses, both nectar and pollen will be contaminated, and both are relevant food sources for brood. Worker larvae are fed royal jelly, a secretion made by the nurse bees, for the first three days of their lives. After this, worker larvae feed for two days on worker jelly, which is a mixture of royal jelly, honey (made from nectar) and beebread (made from pollen). The nurse bees themselves will also consume both nectar and pollen, and many PPP uses will include exposure scenarios that consider both nectar and pollen. Thus, for a fully realistic exposure simulation in the feeder study, both the nectar substitute and the pollen substitute should be spiked with the pesticide. However, little experience has been gained with feeder studies using both a spiked nectar substitute and a spiked pollen substitute at the same time and there are some practical limitations in using a spiked pollen substitute (see 4.4.7.2., below). It is also noted that especially in the earlier days of their development, larvae are most exposed via processed food made mainly from nectar. In view of the lack of knowledge and experience with spiked pollen substitute, the Working Group agreed that for the moment, it is only required to expose bees via a spiked nectar substitute. However, spiked pollen can be added if desired. For some uses, using only exposure via pollen substitute can be a refinement option (i.e., those uses for which only exposure via pollen leads to a risk to brood in the lower tier; in addition, spiking pollen may be an alternative for PPPs that are difficult to dissolve in nectar).

All concentrations must be verified by chemical analysis.

4.4.7.1. Nectar substitute: sugar solution– required

Control colonies are fed with untreated 50% (w/v) aqueous sugar solution. The test and reference item are mixed in 50% (w/v) aqueous sugar feeding solution. The ready-prepared sugar solutions are simultaneously offered in one feeding trough per colony. The feeder is put into an empty magazine on top of the populated bee magazines according to specific hive system and good beekeeping practice. The bees should freely access the feeding solution; drowning of bees in the feeding solution should be avoided by appropriate measures. On each consecutive day the amount of food remaining should be determined by weighing the feeder. Spiking of the feeding solution must be performed shortly before each feeding event, daily. The initial feeding should be performed on CBDD 0 or 1 day later.

Bees should be able to feed ad libitum and the offered amount should thus not be depleted. Lückmann and Schmitzer (2019) recommend administering 0.5 L of sugar solution daily over a period of 9 days. To ensure that feed is always in excess, the amount of sugar solution offered per day may need to be higher.

4.4.7.2. Pollen substitute: pollen patties– optional

Pollen traps must be used during the nine exposure days to limit pollen entry from foraging and stimulate consumption of the offered patties.

Pollen patties should be composed of a mixture of pollen and sugar solution. Artificial pollen substitute can be obtained from commercial providers – if used, information on its composition and nutrient level should be reported. If natural pollen is used, chemical analysis is required to show that there is no contamination with the test substance, other PPPs and contaminants, and to determine the nutrient level. Pollen patties are offered to the bees inside the hive so that bees can freely access them. The pesticide is mixed with the pollen substitute and fed daily to the bees.

Appendix B collects data on pollen consumption. Useful information comes particularly from the study of Pettis et al. (2014), which reported both colony strength and pollen intake and found an intake of 30 g patty/d for colonies of 30,000 bees. Roessink and Van der Steen (2021) calculate a pollen consumption of 21 g/d for a colony of 10,000 bees and confirmed that value in a field study with two colonies. However, it remains the applicant's responsibility to ensure an excess of pollen. Pollen patties should be replaced daily and weighed to determine the amount remaining. Bees should be able to feed ad libitum so the pollen should always be offered in excess.

4.4.8. Treatment concentration

The treatment concentration should be based on a back calculation from the residue intake (PEQ_{di}) estimated for Tier 2 exposure estimation. In lack of Tier 2 exposure estimation, Tier 1 exposure estimation might be considered. It is expected that in most of the cases, the higher of the 90th percentile nectar or pollen concentration should result in a sufficiently high test concentration (i.e., sufficiently high residue intake in the test).

The reference item fenoxycarb should be offered at 42 mg a.s./0.5 L sugar solution per colony per day (calculated from a rate of 300 g a.s./400 L water/ha.) As this dose in sugar solution has been ring-tested, it is not necessary to offer fenoxycarb also in pollen patties in case both nectar and pollen are spiked with the test item. However, the reference item colonies should receive untreated pollen patties.

The control hives receive untreated sugar solution (and untreated pollen patties, in case both nectar and pollen are spiked with the test item) daily.

4.4.9. Data assessment and results

The total amount of covered/capped brood in the whole colony should be checked before treatment and in ~10 days intervals after the start of the exposure (five observations): d-1 or d0 (before exposure), d10 ± 1, d21 ± 1, d31 ± 1, d42 ± 1. Either digital methods (e.g. Alves et al. (2020), Wang et al. (2020)) or the Liebefeld method (Imdorf et al., 1987) can be used to determine the amount of capped brood; see Section 4.2.4 for more details.

Optionally, also mortality, colony conditions and colony development, and bee behaviour can be studied, see details in Lückmann and Schmitzer (2019). These endpoints are not the focus of this study but can be used as supporting information.

4.4.10. Climatic conditions

During the entire test period the air temperature, relative air humidity and precipitation should be recorded at the test site and/or obtained from the nearest weather station.

4.4.11. Endpoints and evaluation of the data

To determine potential effects of a test item on honey bee brood, the amount of capped brood should be compared between the control and the treatment at each measuring day. A one-sided equivalence test should be applied (for example, based on a simple *t*-test) with an equivalence limit corresponding to a 10% relative decrease in the amount of capped brood. A risk is indicated if there is more than 10% decrease at any timepoint.

Optional endpoints can be assessed following the recommendations in Lückmann and Schmitzer (2019):

- mortality
- behaviour
- colony condition

4.4.12. Validity criteria

For the test to be valid, the amount of capped brood should be >10% lower in the reference treatment than in the control at least in one timepoint, and the difference has to be statistically significant at least in one timepoint. In addition, the capped brood presence in the control colonies should show a normal appearance throughout the whole test period, i.e., according to the season and expected colony development.

4.4.13. Recommendations for further research

The endpoint in the honey bee colony feeder study currently is the amount of capped brood. This endpoint has a clear link to the SPG. However, the amount of capped brood was shown to be very variable in a small dataset available to the Working Group, and consequently, it was not feasible to define the number of replicates that would allow the detection of 10% difference between the treated and control group with sufficient power. Thus, the current study protocol includes the requirement of a post-hoc power analysis. In future, based on a larger dataset that allows a better understanding of the sources of variability, it may be possible to amend the protocol and define the required number of replicates *a priori*.

Historically, other endpoints have been used to describe effects on brood, especially the Brood termination-rate (BTR), the Brood-index (BI) and the Compensation-index (CI). These endpoints are determined based on the expected (normal) development time of eggs to larvae, pupae and adults (see OECD 75 for a detailed description of the calculations). The Compensation-index is an indicator for recovery of the colony and not suitable to address the SPG, in which no effects >10% are tolerated at any timepoint. The BTR and the BI are potentially suitable, since they can be derived for every assessment day. However, further research is needed to determine:

- the link with the SPG – is it possible to extrapolate effects on the BTR or the BI to colony strength?
- whether the variability in these endpoints is lower than for the amount of capped brood, which, if so, would imply that study setups can be less demanding than for capped brood. Figure A1 from Luckmann & Schmitzer (2019) suggests that the variability between colonies is very high for the BTR. However, the underlying dataset for chronic exposure studies could be scrutinized further, investigating e.g. the variability at the different assessment dates of the studies.

5. Higher tier studies for bumble bees

Until recently most regulatory experience in the EU was focused on the assessment of PPPs in the honey bee, *A. mellifera*. However, the publication of the previous guidance document (EFSA (European Food Safety Authority), 2013) increased the scope of regulatory tests to include bumble bees and solitary bees. The introduction of bumble bees into the regulatory process has introduced some benefits, e.g., increased representation for wild bees, and challenges, e.g., less experience and fewer standardised protocols. The major differences between the life history traits of bumble bees and honey bees (summarised in Section 1.3 of the Guidance Document) means that it is not possible to simply use the same protocols for bumble bees that are available for honey bees without modification. To ensure an effective risk assessment, EFSA would conventionally use internationally agreed and adopted protocols for the basis of the higher tier risk assessment, however, the number of ring-tested higher tier tests for bumble bees is still limited (but see Klein et al. (2022) for semi-field). Therefore, the working group proposes the following guidelines for three types of higher tier tests in bumble bees; a colony feeder test, a semi-field test, and a field test which could be used until internationally agreed and adopted guidelines are available.

5.1. Study aim

The SPG determines the endpoints for assessment in each of these experiments, although the bumblebee SPG is characterised by an undefined threshold, it is still based on the colony strength and

is defined as the number of adult individuals in a colony. Therefore, the SPG dictates that the number of adult bees in a colony is the relevant attribute that needs to be measured.

Although full colony censuses are possible with bumblebee colonies, a full colony census is an invasive procedure which can be disruptive to a colony and is not something that can be frequently carried out on large numbers of colonies. Given the practical limitations of regularly carrying out an invasive colony census and the strong relationship between colony strength and colony weight, the Working Group propose that a colony census can be carried out prior to the start, and at the end, of the assessment period and that the weight of the colony (bees, brood, nest structure) is used as a proxy for colony strength in periods in-between (See Appendix B, (EFSA (European Food Safety Authority) et al., 2022)).

5.2. Creation of colonies and initial conditions

Whilst testing the effect of PPPs from the start of the colony lifecycle, e.g., from the foundation by a single queen, would be desirable, it is not something that is currently practical to implement. Although some protocols that allow individual queens from wild or commercial sources to rear a nest, they only allow the rearing of small numbers of colonies per year, therefore rearing colonies from individual queens would not be practical for risk assessment. Using small, early developmental stage colonies has become common practice (Gill et al., 2012; Whitehorn et al., 2012; Rundlöf et al., 2015), and these have the advantage of being readily commercially available throughout the year.

Commercial colonies can be variable in size, and it is good practice to minimize the variation within and between treatment groups; therefore, the Working Group recommends that colonies be ordered at a similar developmental stage, i.e., young colonies, containing around 30 individuals, with a maximum range of between 20-50 adult individuals. Similarly, the weight of the colonies in each group should be equalised as per Section 3.5.5.

The colonies should only contain workers and a single queen. The presence of a high proportion of drones in a colony is normally associated with a mature colony that has reached the switch point, when the queen starts to lay haploid eggs to produce male bees. The presence of drones indicates that the colony is either late in the colony lifecycle (Duchateau and Velthuis, 1988; LOPEZ-VAAMONDE et al., 2009) or that the colony displays a relatively rare phenotype in commercially produced bumblebees where the switching point to producing males occurs early (Di Pietro et al., 2022), both of which would be sufficient to exclude the colony from the risk assessment. Similarly, the presence of a queen is required to ensure that the colony remains viable, and any colonies without a live queen at the start of the experiment should be excluded. Both of these scenarios should be relatively rare in young, healthy colonies, however the WG recommends that more colonies are ordered than required (~10%) to account for colonies that may have to be excluded.

Screening for common pathogens: There is no requirement for a pathogen screen, however the colonies should be free from any visible signs of ill health and should have a certificate of health from the breeder.

Treatment and control groups: Each experiment should contain one or more "*treatment*" groups which contains the PPP being tested. Each experiment should also contain a "*negative control*" group which receives the same treatment as the treatment groups, minus the exposure to the PPP. Additionally, the feeder and semi-field experiment should also contain a "*positive control*" group which receives a treatment with a known result, and therefore should show a particular change during the experiment. Dimethoate is of often used as a positive control, however any PPP that induces a clear, measurable, effect that can be detected in at least one time point would be suitable.

Number of colonies: No recommendations are provided here on the number of colonies. As explained in Section 2, as the recommended statistical method is based on the test of equivalence, the statistical error of concern (the 'false positive' error) is directly controlled by the level of significance of the test. Hence, there is no need for the risk assessor to recommend an adequate level of replication: poorly replicated studies will show higher risk than studies with more replicates.

Colony housing: Suitable housing should be provided for the experiment; this may depend on local conditions. However, the working group suggest a well-ventilated housing that is under shade from full sun and provides protection from any local predators.

5.3. Data collection and primary endpoints

The applicant should go through three stages of data collection that occur before, during and after the exposure effect study. These are described below.

Colony census (start SPG measurement): The initial colony weight and colony census can be carried out up to five days before the experiment starts, this can be carried out either in the cage they arrive in from the supplier or when the colonies are transferred to a new colony box. The census may include temporarily removing the adults to count them, identifying the sex of the bees, and confirming the presence of a live queen, all under red light.

Females and males are easiest separated by the female's stinger and the male's clampers, which are located at the end of the abdomen between the last sterna and terga. Worker and queen females are separated by the larger size of the queens. The most reliable method to distinguish between workers and males is to check their genitalia. Males often have a larger gap at the end of their abdomen, between the last sterna and terga, and have two clampers that are usually visible in this gap. If these are not visible the last sterna and terga should be separated carefully with a needle or forceps and clampers are checked inside the abdomen. Workers have a smaller gap and a stinger instead of the clampers. The two also differ slightly in their appearance, males having slightly longer hind legs and a squarer shaped end of their abdomen.

The number of workers, drones and the presence of a live queen should be reported as the output for the initial colony census and colonies which exceed the maximum size, contain drones, or lack a live queen should be excluded.

Colony census (final SPG measurement): The final census is more complex than the initial census and includes an assessment of the total colony size and reproductive output by assessing the cocoons within the nest, this method is based on Rundlöf et al. (2015) and Hodge (2019). On the evening of the last day of the experiment the colonies should be sealed and weighed, followed by colony euthanasia by freezing the colony for at least 24 hours. The colonies can remain frozen until the applicant can carry out a full colony census. As each bee must be classified by sex and caste, the number of workers, drones and queens should be reported.

The applicant should also report a breakdown of the colony production based on the cocoons in the colony. After thawing, the nest structure can be carefully removed from the housing and broken apart, allowing the cocoons to be measured. The separation of queen cocoons from the worker/male cocoons can be done based on the width of the cocoons: usually <12 mm for worker/males and >12 mm for queens in *B. terrestris* (see supplementary figure 1c in Rundlöf et al. (2015)). As each cocoon will need to be assessed the applicant should report the number of:

- eclosed queen cocoons
- eclosed cocoons of workers or drones
- intact queen cocoons
- intact cocoons of workers or drones

Additionally, during the final census it should be confirmed that the colony contains only the focal test species of bumblebee. Any colonies that have been colonised by a parasitic cuckoo bee, which enter and kill the resident queen and take over reproduction in the colony, should be excluded from the analysis. It should be ensured that any cuckoo species present in the test region can be identified.

Colony weight (intermediate SPG measurement): The colonies should be weighed on a weekly basis starting at the beginning of the experiment, ending on the last day. The reported weight should only include the colony, e.g., without supplementary feeders attached, and any non-colony weight should be subtracted, e.g., the empty colony box should be weighed, and that mass should be subtracted from the recorded colony weights.

The Working Group considered two options for the duration of the assessment, either a static end point where the assessment ended after a fixed period, or a dynamic endpoint, where the experiment continued until the colony showed two consecutive weeks of static weight or weight loss to indicate the colony was reaching the end of its life cycle (**Figure** 12). The Working Group propose the use of a static endpoint as, in their opinion, the standardised observation period will make planning, implementing and

analysing the results easier, thus allowing greater levels of experimental replication where needed. However, the test may be extended if the switch point has not been reached after eight weeks.

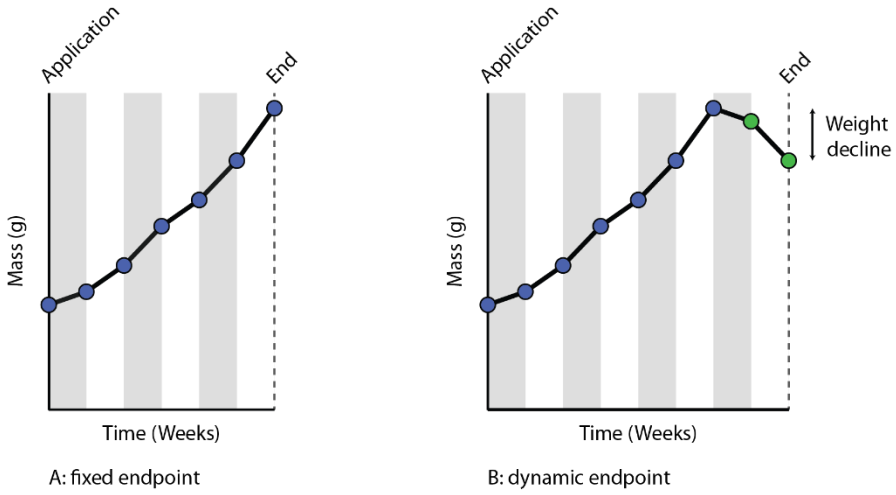


Figure 12: Two options for the higher tier bumble bee experiments. Option A: Fixed endpoint, colonies are maintained for a fixed (eight weeks) period of time from the initial pesticide application. Option B: Dynamic endpoint, the colonies are maintained until they experience two consecutive weeks of weight decline. The WG propose option A.

A complete higher tier test will therefore produce the following output:

- Initial colony census of all adult bees, including identification of each bee's sex and caste, and the colony weight
- Weekly measurements of colony weight between weeks 0 and 8
- Final colony census of all adult bees, including identification of each bee's sex and caste, and the colony weight
- The total number of empty cocoons, split by size (< or > 12 mm)
- The total number of full (intact) cocoons, split by size (< or > 12 mm)

Whilst the data described in the bullet points above should be reported, the endpoints to be analysed are colony strength, via the proxy of colony weight at all timepoints between week 0 and 8, and the reproductive output of the colony, via the number of queen-sized cocoons (eclosed + intact) recorded at the final census. The counting of all cocoons at the end will provide an estimate of total production of bees over the lifetime of the colony. It should be well-correlated to maximum colony weight, but the collection of this data specifically would support the use of colony weight. Whilst the presence of adults of the different castes confirms the colony stage that the weight trajectory should indicate, the cocoon counts (divided by the four aforementioned groups) are the most informative. In short, collecting all the non-weight information supports the interpretation of the weight data.

The statistical analysis of colony weight at each time point and number of queen cocoons in the final census should follow the general recommendations in Section 2.3 and the specific methodology described in Section 2.2 for the 'undefined threshold' approach. For each endpoint/measurement, a two-sided 60% confidence interval (CI) of the effect size (defined on a scale appropriate for the analysis) should be constructed. The upper limit of the CI should be used to calculate the 'level of risk': a categorization of the observed effect based on relevant ranges of effect size. The ranges are those defined in Section 2.2 (see also **Figure 3**) based on the relative difference between test and control; the comparison with the upper limit of the CI should be done on the same scale, with scale transformation of the results if needed. The statistical model and scale used for the analysis should be chosen appropriately; see also Section 3.4.

5.4. Field studies

The aim of these studies is to gain insight into the risks for bumble bee colonies under realistic field conditions. The bumblebee field test should follow the same general principles and study design as discussed in Section 3. Given the similarity between the honey bee and bumble bee designs the WG notes that both types of trials can be run simultaneously at the same sites. Whilst the study design should follow the guidelines in Section 5, the focus of the bumblebee test is still the colony size based on start and end point censuses and colony weight taken at intermediate timepoints.

The field test is the most realistic assessment of a PPP that can be conducted, therefore the results of a well conducted, full field test supersedes lower tier or other higher tier tests.

Choice of crop: The applicant should follow the guidelines outlined in Section 3.5.4.

Number of colonies: The applicant should follow the guidelines outlined in Section 3.4.

Size of field: The applicant should follow the guidelines outlined in Section 3.5

Test length: The test should last 8 weeks from the pesticide application.

Exposure assessment: the exposure assessment should be conducted as outlined in Annex B Section 11.14.2.3

Data to be reported by the applicant: The applicant should follow the guidelines outlined in Section 5.3

Data analysis: the analysis should follow the recommendations in Section 2 and Section 5.3; see also Section 3.4 for considerations on replication and statistical power.

5.5. Semi-field studies

The purpose of this test is to measure the effect of a PPP on the colony under semi-field conditions. This test consists of keeping single bumble bee colonies in an enclosure containing a flowering crop which feeds the colony for two weeks. After two weeks in the cage, the colonies are transferred to an area with minimal pesticide contamination where they can forage freely for six weeks. This test can be used to assess the effect of chronic exposure to a PPP and uses colony censuses at the start and end of the experiment as a direct measure of colony size, and colony weight as a proxy for colony size in between the full colony census. Whilst there is no currently available ring tested protocol, the guidance for the semi-field study for honey bees (e.g., OEPP/EPPO (2010) and OECD (Organisation for Economic Co-operation and Development) (2014)) can be readily adapted to bumble bees and a number of recent publications and preprints provide extensive guidance for adapting this protocol for bumble bees (Tamburini et al., 2021; Klein et al., 2022; Wintermantel et al., 2022).

This test combines multiple realistic elements, including the requirement for bees to actively forage on a treated crop, however, it is not fully realistic, and as such the results of this test may be superseded by the results of a full field tests.

Choice of crop: See Section 3.5.4.

Number of colonies: The applicant should follow the guidelines outlined in Section 3.4.

Number of sites and location of enclosure: The benefit of the semi field test is that multiple enclosures can be placed on the same site. This means there can be only one site per semi field test, although more may be used. The sites should be representative of the region(s) for which authorisation is sought. The number of treatment and control colonies should be equalised between sites, and at each site, the location of the control and treated plots within the site should be decided at random.

Size of enclosure: Size of the enclosure should, as a minimum, follow OEPP/EPPO (2010) and OECD (Organisation for Economic Co-operation and Development) (2014) which specifies tunnels with a minimal size of 40 m^2 crop space, a minimum height of 2.5 m, and a covering gauze with a maximal mesh size of 3 mm. However, a number of recent publications have recommended that larger tunnels with more available crop (Tamburini et al., 2021; Klein et al., 2022; Wintermantel et al., 2022), therefore the Working Group recommend enclosures sizes to contain 60 m^2, or more, of crop space.

Test length: Bumble bee colonies should be exposed to the treated crop for a two-week period. After exposure in a tunnel, colonies should be moved to an area where they have access to foraging resources and minimal exposure to pesticides for a six-week period to continue their development (**Figure** 13).

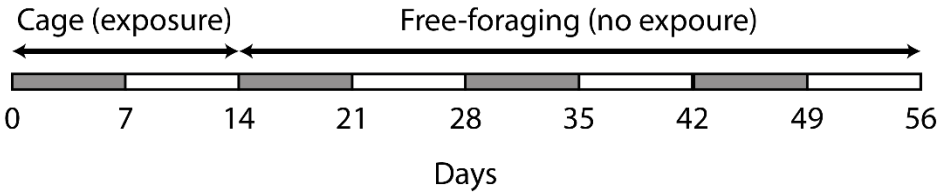


Figure 13: Outline of timeline for the eight-week semi field experiment. Shown is the two-week period the bumblebees are restricted to a cage and exposed to a PPP before being moved to an area where they can forage freely for six weeks. Colony weights should be measured every seven days, as indicated on the diagram. A full colony census should occur shortly before day 0 and after colony euthanasia on day 56.

Data to be reported by the applicant: The applicant should follow the guidelines outlined in Section 5.3.

Exposure assessment: The Working Group recommends that flight activity at the colony entrance (the total number of bees flying in and out) be used as a proxy for the foraging activity in order to show that the bees are actively foraging on the crop and are therefore exposed to the PPP. A recent study by Klein et al. (2022) reported average flight activity at between 4-15.8 bees per 10-minute observation period in control colonies containing between 20-92 individuals.

Therefore, the Working Group propose a minimum flight activity of 4 bees per 10-minute period shortly after application in order for the test to be valid. Both for pesticide applications made during the active foraging period (between dawn and dusk) and for pesticide applications made outside the active foraging period (between dusk and dawn), sufficient traffic at the hive entrance should be recorded within 12 hours of application. Since the first day of the exposure is crucial, a second assessment, later on the same day which demonstrates flight activity is required and should be performed no earlier than 6 hours and no later than 12 hours after the first observations. For example, if an applicant applies a PPP at 08:00 am, they will need to carry out the first observation within 12 hours of application, and the second observation within 6-12 hours of the first. Practically, they would need to record an activity of ≥ 4 bees per 10 minutes observation at any point between 08:00 and 11:00 and again after at any point between 16:00 and dusk to meet these requirements.

The applicant should report both assessments on the day of (or following, for applications done between dusk and dawn) application and need to meet the minimal activity requirements of ≥ 4 bees per 10 minutes observation period.

The flight activity can be conducted by visual assessment (counting bees entering and leaving the hive) or by automatic counting device fit to the hive/box entrance.

The role of a positive control is to demonstrate that an experiment is suitable to study an endpoint of interest through the use of a treatment resulting in a predictable effect. For example, in the case of the semi-field tests, an adequate PPP to use as a positive control would induce high levels of mortality. Given that positive controls are, by design, expected to show large, obvious effects then it is only necessary for the applicant to show a substantial effect in a small number of colonies. Therefore, the Working Group recommend using PPPs of known toxicity than can demonstrate that the experiment is suitable to induce and detect an increased mortality using the methods described for honeybees in section 4.3.5 in the positive control colonies. At least three repetitions must be used.

Data analysis: the analysis should follow the recommendations in Section 2 and Section 5.3.

5.6. Feeder experiments (dietary only)

The purpose of this test is to measure the effect of a PPP on the colony using exposure to pesticides delivered in food in combination with more realistic environmental conditions in the field. The test is based on studies, where a colony is provided with dosed food inside the colony box rather from foraging on treated crops (Gill et al., 2012; Whitehorn et al., 2012; Arce et al., 2017; Botías et al., 2021). During

the exposure period the bees are free to choose to forage from the experimental food or from the surrounding environment, therefore the applicant should monitor how much of the PPP is being consumed by the colony throughout the experiment.

This test can be used to assess the effect of chronic exposure to a PPP and uses colony censuses at the start and end of the experiment as a direct measure of colony size, and colony weight as a proxy for colony size in between the full colony census. Unlike the honey bee feeding test this study is not focused on larval health but on colony-level characteristics such as colony census, queen production and colony weight.

This test requires the least resources of the three tests, however, as the exposure route is the least realistic of the higher tier tests the results of this test may be superseded by the semi-field or the field tests.

Choice of crop: Not applicable. Exposure occurs via food (sucrose solution or pollen patties) supplied to the colony. The applicant should provide the colony with food following the guidelines in Section 4.4.7. There should be sufficient forage available to support the colonies.

Number of colonies: The applicant should follow the guidelines outlined in Section 3.4.

Number of sites and location of colonies: Each experiment should be carried out at one field site. The sites should be representative of the region(s) for which authorisation is sought. Within the field, the control and treated colonies should be placed at a location in a paired design with each location containing a treated and control colony with each pair being a minimum of 10 meters from the next pair. Between the treatment and control sites, there should be a minimum of 2 meters separating the colonies, although greater distances will reduce the chance of workers drifting between colonies (ZANETTE et al., 2014), applicants can also colour code the entrance to the colonies to help the bees recognise their colonies (Gill et al., 2012, Arce et al., 2017). The positioning of the control and treated colonies within a pair should be assigned randomly (See Figure 14).

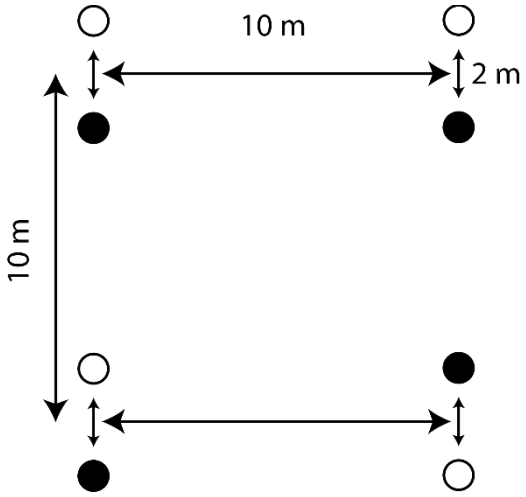


Figure 14: Paired colony design for a feeder test. Colonies are to be separated by 10m from each other, and the treatment and control colonies within each pair should be separated by 2m. The location of the treatment and control within each pair should be randomly assigned. Black/White circles represent different treatment groups.

Test length: Bumble bee colonies should be provided with their relevant treatment for a two-week period, followed by allowing the bees to forage freely for 6-weeks (**Figure** 5).

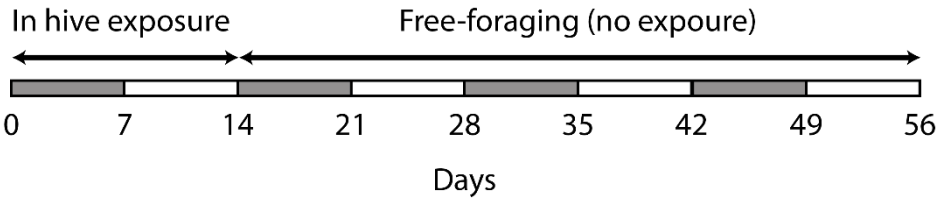


Figure 15: Outline of timeline for the eight-week colony feeder experiment. Shown is the two-week period the bumblebees are provided with access to a PPP within the colony. After the exposure period the PPP treated food is removed from the colony and the colony can continue to forage freely and develop for six weeks. Colony weights should be measured every seven days, as indicated on the diagram. A full colony census should occur shortly before day 0 and after colony euthanasia on day 56.

Data to be reported by the applicant: The guidelines outlined in Section 5.3 should be followed.

Exposure assessment: The actual colony level exposure can be calculated by multiplying the amount of sucrose removed by the colony by the concentration of the PPP in solution. This is later used to calculate the EED_j at the individual level (see 10.5 in the guidance document).

A positive control should be included, i.e., a PPP of known toxicity than can demonstrate that the experiment is suitable to induce and detect an increased mortality in the positive control colonies (see also 4.4.8, above). At least three repetitions must be used.

Data analysis: the analysis should follow the recommendations in Section 2 and Section 5.3.

6. Higher tier studies for solitary bees

The expansion within the EU to consider solitary bees in the regulatory testing of PPPs has already been suggested (EFSA, 2013a). This expansion is relevant because most bee species are solitary (Michener, 2007; Danforth et al., 2019) and because this group of bees may include intrinsically (e.g. Haas et al., 2022) and ecologically sensitive species, since each female is a reproductive unit lacking the buffering capacity of sociality (e.g. Franklin and Raine (2019); Sgolastra et al. (2019)). However, the possibility of conducting higher tier studies with solitary bees to support PPP risk assessment includes challenges such as limited experience and protocols for conducting tests ((Lehmann and Camp, 2021); (EFSA (European Food Safety Authority) et al., 2022)). Until internationally agreed and adopted guidelines are available, the Working Group propose the following guidance on conducting field and semi-field studies with solitary bees based on experiences gained on *Osmia* spp. (e.g. Franke et al. (2021); Stuligross and Williams (2020)). The selection of *Osmia* spp. is motivated by the well-known ecology and experience with rearing of species in this genus (Eeraerts et al., 2020). In a European context, *O. bicornis* and *O. cornuta* are relevant species (Sgolastra et al., 2019), with *O. bicornis* selected as the main model (Franke et al., 2021). There are a limited number of other solitary bee species that have been considered in an ERA context (Kopit and Pitts-Singer, 2018; Sgolastra et al., 2019; Lehmann and Camp, 2021), but those are not considered further here. Further considerations to address the uncertainties around the extrapolation of the results to the numerous EU solitary bee species are strongly encouraged (see also Chapter 15).

As for honey bees and bumble bees, field studies are considered the main method of testing effects of PPPs under realistic conditions. However, the more limited spatial requirements of *Osmia* spp. individuals compared to those of whole colonies (Sgolastra et al., 2019) as well as the intrinsic variability in nesting and reproductive performance in *Osmia* spp. (Lehmann and Camp, 2021; Rundlöf et al., 2022) make the semi-field test relevant as a model ecosystem (c.f. Schmitz (2008)) for exploring population relevant endpoints (Stuligross and Williams, 2020). The proposed methods are largely based on the guidance given in Franke et al. (2021), developed by the ICPPR non-*Apis* working group, complemented with relevant scientific literature. The guidance in Franke et al. (2021) is based on 25 studies produced in ring-testing by nine laboratories in three European countries over two years.

6.1. Principles of the tests and methods

Starting populations and nesting units are introduced to a test crop during the flowering period. Primary endpoints are the reproductive performance of females and the production of offspring in relation to the starting population, thus requiring that the progenies are followed through their whole development cycle. The endpoint outcomes are compared among treatment groups, including at least a negative control and a test item group and for semi-field studies also a positive control.

6.1.1. Initial conditions and starting population

Cocoons of the test species, primarily *O. bicornis* but *O. cornuta* could be considered, are sourced from a commercial supplier. Cocoons should be healthy and free of visible signs of disease and are stored at 1-4 °C until incubation. Incubation for emergence to align with test crop flowering and test item application is done at around 22 ± 2 °C and 60-80% relative air humidity.

Cocoons (or adults) are placed at the test crop along with nesting units where the females can provision brood cells and lay eggs. The timing of placement and the number of cocoons (or adults) included in the starting population is reliant on the food resource availability, which is why the considerations differ between field and semi-field studies. The sex ratio of the cocoons in the starting populations should, however, be the same, with 1 female to 1.5-2 males (Rundlöf et al., 2015; Stuligross and Williams, 2020; Franke et al., 2021). Female and male cocoons can be placed in separate tubes or trays for emergence in the cages, to make it easier to track the emergence of the two sexes. Female cocoons are usually larger than male cocoons, but to definitely separate the sexes, cocoons could be carefully opened at the top using a scalpel, without damaging the bees inside, and inspecting the bees (Roessink I, 2019). *Osmia bicornis* males have white-grey hair on their faces while females have black. Any remaining closed cocoons should be removed before test item application and start of assessments to prevent release of natural enemies (e.g., parasites) and addition of females after treatment.

Nesting units should have a rain protective roof and consist of a high number (so that this does not become a limiting factor) of artificial nesting cavities with a depth of 150 mm and a diameter of 8 mm (*O. bicornis*) or 10 mm (*O. cornuta*). Franke et al. (2021) suggest using medium-density fiberboard (MDF) trays over plastic for the nesting units since it increases nesting activity and aids in fulfilling the quality criteria. The nesting units also include transparent sheets that can be placed over the opened nesting cavities (Franke et al., 2021) or paper linings that can be temporarily removed (Stuligross and Williams, 2020) to track reproduction over time. The nesting units should be placed above ground, with openings facing south-east to promote morning activity. Nesting units need to be supplemented with a source of moist loamy soil that will be used by the females to construct brood cell partitioning and seal the nest entrance.

6.1.2. Choice of crop

Osmia bicornis, like most other *Osmia* species, is polylectic and collects pollen from a range of available sources, with a preference for *Ranunculus* spp. and *Quercus* spp. (Haider et al., 2014). However, the species willingly forage from *Brassica napus* and *Phacelia tanacetifolia* (*(Holzschuh et al., 2013; Schwarz et al., 2022)*), which are the recommended test crops by Franke et al. (2021). It is also possible to provide a diversity of flower resources to, for example, prolong the flowering period (Stuligross and Williams, 2020; Klaus et al., 2021). Selecting *P. tanacetifolia* provides flexibility in order to ensure a continuity in flower resources over a longer time by cutting part of the crop area, see Section 6.1.3, also in a semi-field setting (Schwarz et al., 2022).

6.1.3. Test duration and timing of bees and test item application

The test includes periods of adult and larval exposure as well as a post-exposure period to allow the full life cycle of the species. *Osmia bicornis* females are active for about 2 months during February-May, depending on region. The development cycle of the offspring includes approximately 1 month of larval feeding, 1-2 months of pre-pupal stage and approximately 1 month pupal stage (Sgolastra et al., 2019). Eclosure occurs naturally in late summer, but the adults overwinter in their cocoons and emerge the following spring. An outline of the field and semi-field experiments are presented in Figures 16 and 17.

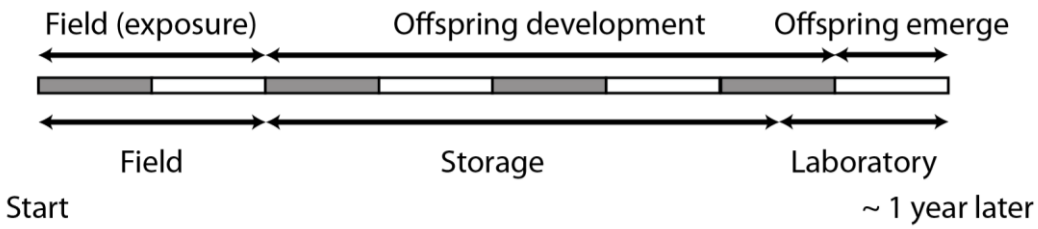


Figure 16: Outline of the solitary bee field study. Solitary bees are allowed to forage freely next to a treated field whilst they build and provision their nests. After the exposure period the cocoons are removed and stored until the offspring emerge, approximately 1 year later.

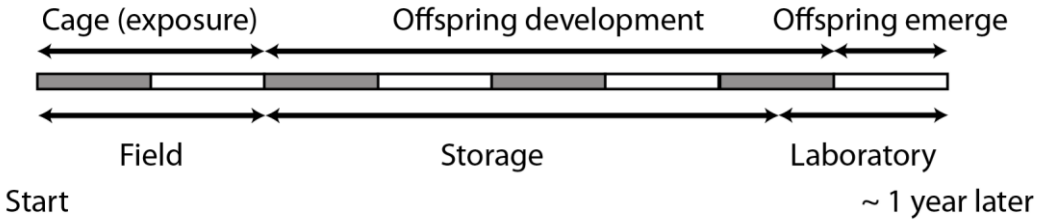


Figure 17: Outline of the solitary bee semi-field study. Solitary bees are restricted to a cage containing treated flowering crops whilst build and provision their nests. After the exposure period the cocoons are removed and stored until the offspring emerge, approximately 1 year later.

Franke et al. (2021) recommend that the test is conducted close to the natural activity period of the species in the region, i.e., during spring or early summer. It is, however, possible to manipulate emergence of adults using temperature regulation and incubation (see Section 7.1.1) but this may reduce pupal emergence rate of the parent generation (Rundlöf et al., 2015). On the other hand, low temperature during early spring could reduce female nesting activity (Franke et al., 2021). After adult emergence, there is a pre-nesting period that should be considered in relation to the timing of the crop bloom, the test item application and the assessment period. Franke et al. (2021) recommend that the test item can be applied when 30% of the females in the starting population have commenced nesting, i.e., started brood provisioning and laying eggs, at the nesting unit(s). For spray application, only brood provisioned after the application should be considered.

It is important to provide flower resources during the whole adult activity period, whether under field or semi-field conditions. The continuity of flower resource provisioning can be prolonged by either using a mix of plant species (Stuligross and Williams, 2020; Klaus et al., 2021) or by cutting part of the crop area to delay the flowering, which can be done for example when using *P. tanacetifolia* (Schwarz et al., 2022). The latter example could also be used when application of the test item is done pre-flowering (see (Schwarz et al., 2022) for an example). The adult exposure phase ends at the end of crop bloom (BBCH 69 for oilseed rape; (Franke et al., 2021)). The crop bloom period could be between 3 and 6 weeks (Franke et al., 2021).

The test item is applied when females have started nesting either during bee flight or at night (see Section 4.3.3 for further details on the application time).

6.1.4. Post-exposure conditions

At the end of crop bloom, the nesting units are covered with a fine mesh to prevent further nesting and access of parasites. The nesting units are left in place for at least 1 month, but up to 2 months, to allow the larvae to develop and pupate (Roessink I, 2019; Franke et al., 2021). Thereafter, the nesting units can be carefully transported to and placed in well-ventilated storage at ambient temperature, protected from rain and natural enemies (e.g., parasites and birds). Thereafter, cocoons can be harvested and should be hibernated at 1-4 °C and 60-80% relative air humidity for approximately 4 months. Full development of the bees can be confirmed by using a scalpel to carefully open the top of 10 cocoons, without damaging the bees inside (Roessink I, 2019). Finally, cocoons can be incubated at around 22 ± 2 °C and 60-80% relative air humidity to assess emergence success of the next generation. When

cocoon numbers are high, a subset of 80 cocoons could be used to assess emergence success in the next generation.

6.1.5. Assessment of exposure

Assessment of exposure is done through residues in larval pollen provisions in the nest for field studies (Rundlöf et al., 2022; Schwarz et al., 2022) and through flight activity before and after application in the semi-field study (Franke et al., 2021) – see respective Sections for further details (Sections 6.2 and 6.3).

6.1.6. Data collection and endpoints

The relevant measurement endpoints are those that relate to population abundance, directly or indirectly, with a preference for those that have low variability to increase measurement precision and consistency (Stuligross and Williams, 2020; Franke et al., 2021; EFSA (European Food Safety Authority) et al., 2022). Franke et al. (2021) suggest the following endpoints:

- Emergence success of the parental generation: count the number of empty female and male cocoons in the starting population in 3–4-day intervals.
- Nesting activity: count the number of actively nesting females at the nesting unit daily or in 3-4 day intervals. This can be done by individually marking the females or counting the number of females at a nesting unit when there is no flight activity (i.e., during the night or early morning). The mean number of nesting females can be determined by averaging the three highest recorded numbers.
- Reproduction – cell production: mark and count the number of brood cells (i.e., pollen provision with egg separated by mud partitioning) produced after application in at least 3–4-day intervals or even daily. This can be done by either taking a picture or making marks on the transparent sheet or the paper linings. All of the cells counted at one timepoint is defined as a cohort.
- Reproduction – cocoon production: count the number of cocoons before hibernation. If the cell counts were done in cohorts, the cocoon production could also be done separately for the cohorts.
- Reproduction – emergence success of the next generation: count the number of females and males emerging from the hibernated cocoons. If the cell and cocoon counts were done in cohorts, the emergence success could also be done separately for the cohorts. Remaining intact cocoons could be checked to determine why it is still closed (i.e. not mature, parasitized, mature but dead).

A population growth rate (λ) of ≥ 1 indicates that the population is stable or growing. For estimating population growth rate in *Osmia lignaria*, closely related to *O. bicornis*, Stuligross and Williams (2020) suggest using the multiplicative product of: 1) nesting probability (i.e., proportion of the females in the starting population that produce at least one offspring), 2) total offspring production (i.e., the number of offspring surviving to emergence after hibernation) and 3) the proportion of female offspring (i.e., the proportion of daughters in the next generation). Based on this suggestion and the endpoints above, population growth rate can approximately be produced by combining information from emergence success in the starting population (number of emerged females), nesting activity (mean number of nesting females) and emergence success of the next generation (offspring production and proportion daughters).

The working group support the measurement and reporting of the five endpoint types suggested by Franke et al. (2021) and listed above, time-resolved when possible, as well as the reproductive endpoints (cell production, cocoon production and emerged adults in the next generation) and the expressed per emerged or nesting female in the starting population ((Franke et al., 2021), (EFSA (European Food Safety Authority) et al., 2022)), aiding the comparison between control and treatment, and using the information to calculate population growth rate for each population (cage or field site) following Stuligross and Williams (2020). The latter is largely equivalent to the number of emerged daughters expressed per nesting female in the starting population, i.e. if 10 females from the starting population nest in the nesting unit in a cage and there are 12 females (daughters of the starting population) emerging from the cocoons produced in that cage, the population growth rate is 1.2 (12 divided by 10).

As a supplement, flight activity at the nesting unit(s) can be recorded throughout the study both before and after application (see Franke et al. (2021)).

6.1.7. Data analysis and interpretation of the results

Data analysis should follow the recommendations in Section 2; see also Section 3.4 for considerations on replication and statistical power. A one-sided equivalence test ($\alpha = 0.2$) should be applied with a sequentially increasing equivalence limit for the treated group in relation to the control group (see Section 2.2). The endpoints to be analysed are nesting probability (i.e., proportion nesting females), cell and cocoon production, emergence success of the next generation and population growth rate. The different endpoints capture different potential impacts of the test item, from adult sublethal and lethal effects to developmental effects throughout the life cycle.

Study quality criteria mainly relate to the nesting initiation of the females in the starting population before application as well as confirming exposure in the test item group. Franke et al. (2021) propose that a study is valid if ≥30% of the females in the starting population have started nesting, i.e., are occupying the nesting unit(s), the night or morning before the application. Nest occupation can be aided by providing suitable nesting units (i.e., MDF rather than plastic) and planning the timing of the study (i.e., avoiding cool and rainy weather). Methods for confirming exposure differ between the field and semi-field studies (see Sections 6.2.4 and 6.3.4). Additionally, flight activity at the nesting units can be used to confirm similarity in pre-application conditions between the treatment groups as well as the maintainance of suitable conditions based on flight activity in the control group.

6.2. Solitary bee field study

The aim of a solitary bee field study is to evaluate the effects of PPPs under field conditions, allowing the bees to move and forage freely. The setup of the solitary bee field test should follow the general principles of field study design detailed in Section 3 and the general principles outlined in Section 6. If multiple bee groups will be evaluated, it is recommended to run the tests simultaneously at the same field sites (e.g., (Rundlöf et al., 2015; Woodcock et al., 2017)). The Working Group considers a population for *O. bicornis* in a field study setting to be the starting population and their offspring in nesting units at a field. If the study is conducted during the natural activity period of the species, it is possible that naturally occurring individuals could use the nesting units. However, the study design of selecting similar study sites within a block that are allocated to either control or treatment groups would make it equally likely for this to occur in both treatment groups. In addition, the comparison is always the outcome in the treatment group in relation to the control group.

6.2.1. Treatment and control groups

The study includes two treatment groups: a negative (water) control group and one or more test items groups.

The level of replication is up to the applicant/study director. Some guidance could be found in the eight studies evaluated in EFSA (European Food Safety Authority) et al. (2022), where the replication ranged from 3 to 8 blocks (i.e., field sites in each treatment group). The analysis of reproduction-related endpoints standardised by the females in the starting population are, however, variable, which may hamper the ability to conclude equivalence. For example, Ruddle et al. (2018) showed that there may however be adaptions that could be made to reduce this variability, both by increasing replication and by increasing the size of the starting population (see Section 6.2.2).

6.2.2. Starting population, field size and flower availability

The size of the starting population is up to the applicant/study director. It should be noted, however, that a larger starting population may be a strategy to reduce endpoint variability (EFSA (European Food Safety Authority) et al., 2022) as well as confirming that ≥30% of the females in the starting population have started nesting (see Section 6.2.3). Studies with 12-25 females in the starting population had higher endpoint variability than studies with 682-1200 females in the starting population. It should be noted that a very large population density is unlikely to occur under natural conditions (Steffan-Dewenter and Schiele, 2008), but it could aid in conducting a field study with lower variability and thus in reaching statistical equivalence. A large starting population may also ensure the possibility of

collecting larval pollen provision to confirm sufficient exposure. The starting population can be divided among multiple nesting units, even if they are still considered to be one population.

Cocoons (or adults) are placed at the test crop before bloom (e.g., BBCH 57 for oilseed rape; Ruddle et al., 2018), along with the nesting units. The nesting units should be covered with chicken wire or something similar to prevent bird predation on the brood (Peters et al., 2016).

Franke et al. (2021) recommends a maximum of 1 female per m^2 in cages with flowering plants, which has generally also been followed for other *Osmia* spp. cage studies (e.g. (Stuligross and Williams, 2020)). This recommendation can also be used as a minimum area of flowering crop that should be provided in relation to the starting population size in a field study.

6.2.3. Validity criteria

As in the semi-field study (see Section 6.3.3), confirming that ≥30% of the females in the starting population have started nesting, i.e., are occupying the nesting unit(s), the night or morning before application would indicate favourable initial conditions.

6.2.4. Assessment of exposure

Exposure is assessed through residues in pollen entering the nest. Examples of how this can be done based on the larval pollen provisions can be found in Ruddle et al. (2018), Rundlöf et al. (2022) and Schwarz et al. (2022) (see Annex B for more details). Residues of the test item found in the pollen provisions are used to calculate the EED and this is compared to the PEQ (for more details see Section 10.5 of the Guidance Document).

6.3. Solitary bee semi-field study

The solitary bee semi-field study is limited by the enclosure of the bees, excluding the possibility to include assessment of the consequences of, for example, finding a mate or navigating between the nesting unit and food resources. The advantages of the semi-field study are that some conditions can be controlled, such as the containment of the population and directing the foraging to the test crop, as well as the availability of a ring-tested protocol for *O. bicornis* (Franke et al., 2021).

6.3.1. Treatment and control groups

The study includes three treatment groups: a negative (water) control group, a positive (toxic) control group and one or more test items groups.

For the positive control group, Franke et al. (2021) evaluated the use of the organophosphate insecticide dimethoate applied at 75 g a.i./ha, which produced significantly reduced nesting activity in 7 of 8 valid studies, in 6 of 7 for flight activity and in 5 of 8 and 4 of 7 for cell and cocoon production, respectively.

The level of replication is up to the applicant/study director. Franke et al. (2021) used a replication of 4-6 to statistically detect differences of <50% between positive and negative controls for most of the endpoints.

6.3.2. Starting population and cage size

It is up to the applicant/study director to choose the starting population and cage size, with a maximum of 1 female per m^2 of cage area with flowering crop. Franke et al. (2021) recommend a starting population of ≥30 females per cage and one nesting unit with a minimum of two artificial nesting cavities per female. Cages should then be a minimum of 30 m2, preferably 60 m^2. However, others have used smaller starting populations and cage sizes, for example 6-8 females in 9 m^2 (Stuligross and Williams, 2020), but then continuously supplemented with more bees to maintain 5 nesting females. The working group recommend having one larger starting population and not supplementing the population.

Cocoons (or adults) are placed in the cage at early crop bloom (e.g. BBCH 59-60 for oilseed rape; Franke et al., 2021), along with the nesting unit.

6.3.3. Validity criteria

The main quality criteria are 1) that ≥30% of the females in the starting population have started nesting, i.e., are occupying the nesting unit(s), the night or morning before application, and 2) that there are statistically significant effects on nesting activity and cell and cocoon production in the positive control.

6.3.4. Assessment of exposure

Exposure is confirmed through statistically comparable flight activity between all treatment groups shortly before application and before and after application for the negative control group as well as confirmation that ≥30% of the females in the starting population have started nesting before application.

Flight activity is recorded by counting the number of bees entering the nesting cavities in the nesting unit for a fixed time interval of at least 5 minutes. Recordings are done at least directly before and just after application, as well as the day after application. Flight activity recordings should be done in parallel between the treatment groups to reduce the influence of external factors such as weather.

6.4. Future developments and alternative methods

In addition to the aforementioned methods, tracking the nesting of individually marked females, and mark-recapture techniques in general, could be explored to track sublethal effects that may have consequences for solitary bee populations. Tracking nesting of individually marked females have been used to study *Osmia* spp. nesting and reproduction and the techniques are already partially developed (e.g. (Tepedino and Torchio, 1982; Ladurner et al., 2008; Sandrock et al., 2014)). Mark-recapture techniques could depart from the limited experience in bees but much more for other organisms in estimating population size and its change (e.g.(Lebreton et al., 1992; Steffan-Dewenter and Schiele, 2004; BRIGGS et al., 2022)). This could also be the foundation for simple population models for the future. Another future development is alternative tests that could evaluate potential impacts on ground nesting solitary bees (Willis Chan and Raine, 2021) since the majority of species in this group are ground nesting (Danforth et al., 2019) and may experience other exposure routes compared to above ground and cavity nesting bees such as *Osmia* spp. ((Kopit and Pitts-Singer, 2018; Sgolastra et al., 2019)).

7. References

Alves TS, Pinto MA, Ventura P, Neves CJ, Biron DG, Junior AC, De Paula Filho PL and Rodrigues PJ, 2020. Automatic detection and classification of honey bee comb cells using deep learning. Computers and Electronics in Agriculture, 170:105244. doi: https://doi.org/10.1016/j.compag.2020.105244

Arce AN, David TI, Randall EL, Ramos Rodrigues A, Colgan TJ, Wurm Y and Gill RJ, 2017. Impact of controlled neonicotinoid exposure on bumblebees in a realistic field setting. Journal of Applied Ecology, 54:1199-1208. doi: https://doi.org/10.1111/1365-2664.12792

Baveco JM, Focks A, Belgers D, van der Steen JJ, Boesten JJ and Roessink I, 2016. An energetics-based honeybee nectar-foraging model used to assess the potential for landscape-level pesticide exposure dilution. PeerJ, 4:e2293. doi: 10.7717/peerj.2293

Beekman M and Ratnieks FLW, 2000. Long-range foraging by the honey-bee, Apis mellifera L. Functional Ecology, 14:490-496. doi: https://doi.org/10.1046/j.1365-2435.2000.00443.x

Beekman M, Sumpter DJT, Sephides N and Ratnieks FLW, 2004. Comparing foraging behaviour of small and large honey-bee colonies by decoding waggle dances made by foragers. Functional Ecology, 18:829-835. doi: https://doi.org/10.1111/j.0269-8463.2004.00924.x

Botías C, Jones JC, Pamminger T, Bartomeus I, Hughes WOH and Goulson D, 2021. Multiple stressors interact to impair the performance of bumblebee Bombus terrestris colonies. Journal of animal ecology, 90:415-431. doi: https://doi.org/10.1111/1365-2656.13375

Bozek K, Hebert L, Portugal Y, Mikheyev AS and Stephens GJ, 2021. Markerless tracking of an entire honey bee colony. Nature Communications, 12:1733. doi: 10.1038/s41467-021-21769-1

BRIGGS EL, BARANSKI C, MÜNZER SCHAETZ O, GARRISON G, COLLAZO JA and YOUNGSTEADT E, 2022. Estimating bee abundance: can mark-recapture methods validate common sampling protocols? Apidologie, 53:1-24

Brock TC, Hammers-Wirtz M, Hommen U, Preuss TG, Ratte HT, Roessink I, Strauss T and Van den Brink PJ, 2015. The minimum detectable difference (MDD) and the interpretation of treatment-related effects of pesticides in experimental ecosystems. Environ Sci Pollut Res Int, 22:1160-1174. doi: 10.1007/s11356-014-3398-2

Büttner G, 2016. CORINE Land Cover Products. European Landscape Dynamics, CRC Press. pp. 85-88.

Christie AP, Amano T, Martin PA, Shackelford GE, Simmons BI and Sutherland WJ, 2019. Simple study designs in ecology produce inaccurate estimates of biodiversity responses. Journal of Applied Ecology, 56:2742-2754. doi: https://doi.org/10.1111/1365-2664.13499

Couvillon Margaret J, Schürch R and Ratnieks Francis LW, 2014. Dancing Bees Communicate a Foraging Preference for Rural Lands in High-Level Agri-Environment schemes. Current Biology, 24:1212-1215. doi: https://doi.org/10.1016/j.cub.2014.03.072

Dainat B, Dietemann V, Imdorf A and Charrière J-D, 2020. A scientific note on the 'Liebefeld Method' to estimate honey bee colony strength: its history, use, and translation. Apidologie, 51:422-427. doi: 10.1007/s13592-019-00728-2

Danforth BN, Minckley RL, Neff JL and Fawcett F, 2019. The solitary bees: biology, evolution, conservation, Princeton University Press.

Danner N, Härtel S and Steffan-Dewenter I, 2014. Maize pollen foraging by honey bees in relation to crop area and landscape context. Basic and Applied Ecology, 15:677-684

Danner N, Keller A, Härtel S and Steffan-Dewenter I, 2017. Honey bee foraging ecology: Season but not landscape diversity shapes the amount and diversity of collected pollen. PLOS ONE, 12:e0183716. doi: 10.1371/journal.pone.0183716

Danner N, Molitor AM, Schiele S, Härtel S and Steffan-Dewenter I, 2016. Season and landscape composition affect pollen foraging distances and habitat use of honey bees. Ecological Applications, 26:1920-1929

Di Pietro V, Ferreira HM, Van Oystaeyen A, Wäckers F, Wenseleers T and Oliveira RC, 2022. Distinct Colony Types Caused by Diploid Male Production in the Buff-Tailed Bumblebee Bombus terrestris. Frontiers in Ecology and Evolution, 10. doi: 10.3389/fevo.2022.844251

Duchateau MJ and Velthuis HHW, 1988. Development and Reproductive Strategies in Bombus terrestris Colonies. Behaviour, 107:186-207

Eeraerts M, Pisman M, Vanderhaegen R, Meeus I and Smagghe G, 2020. Recommendations for standardized oral toxicity test protocols for larvae of solitary bees, Osmia spp. Apidologie, 51:48-60. doi: 10.1007/s13592-019-00704-w

EFSA (European Food Safety Authority), 2013. Guidance Document on the risk assessment of plant protection products on bees (*Apis mellifera*, *Bombus* spp. and solitary bees). Efsa Journal, 11:3295

EFSA (European Food Safety Authority), 2015. Scientific Opinion addressing the state of the science on risk assessment of plant protection products for non-target arthropods. Efsa Journal, 13:3996

EFSA (European Food Safety Authority), 2018a. Peer review of the pesticide risk assessment for bees for the active substance clothianidin considering the uses as seed treatments and granules. Efsa Journal, 16:e05177. doi: https://doi.org/10.2903/j.efsa.2018.5177

EFSA (European Food Safety Authority), 2018b. Peer review of the pesticide risk assessment for bees for the active substance imidacloprid considering the uses as seed treatments and granules. Efsa Journal, 16:e05178. doi: https://doi.org/10.2903/j.efsa.2018.5178

EFSA (European Food Safety Authority), 2018c. Peer review of the pesticide risk assessment for bees for the active substance thiamethoxam considering the uses as seed treatments and granules. Efsa Journal, 16:e05179. doi: https://doi.org/10.2903/j.efsa.2018.5179

EFSA (European Food Safety Authority), Auteri D, Arce A, Ingels B, Marchesi M, Neri FM, Rundlöf M and Wassenberg J, 2022. Analysis of the evidence to support the definition of Specific Protection Goals for bumble bees and solitary bees. EFSA Supporting Publications, 19:7125E. doi: https://doi.org/10.2903/sp.efsa.2022.EN-7125

EFSA (European Food Safety Authority), Ippolito A, Focks A, Rundlöf M, Arce A, Marchesi M, Neri FM, Rortais A, Szentes C and Auteri D, 2021. Analysis of background variability of honey bee colony size. EFSA Supporting Publications. doi: https://doi.org/10.2903/sp.efsa.2021.EN-6518

EFSA FEEDAP Panel (EFSA Panel on Additives Products or Substances used in Animal Feed), Rychen G, Aquilina G, Azimonti G, Bampidis V, Bastos MdL, Bories G, Chesson A, Cocconcelli PS, Flachowsky G, Gropp J, Kolar B, Kouba M, López-Alonso M, López Puente S, Mantovani A, Mayo B, Ramos F, Saarela M, Villa RE, Wallace RJ, Wester P, Anguita M, Galobart J, Innocenti ML and Martino L, 2017. Guidance on the assessment of the safety of feed additives for the target species. Efsa Journal, 15:e05021. doi: https://doi.org/10.2903/j.efsa.2017.5021

EFSA GMO Panel (EFSA Panel on Genetically Modified Organisms), 2010. Statistical considerations for the safety evaluation of GMOs. Efsa Journal, 8:1250. doi: https://doi.org/10.2903/j.efsa.2010.1250

EFSA GMO Panel (EFSA Panel on Genetically Modified Organisms), 2011. Guidance for risk assessment of food and feed from genetically modified plants. Efsa Journal, 9:2150. doi: https://doi.org/10.2903/j.efsa.2011.2150

EFSA PPR Panel (EFSA Panel on Plant Protection Products and their Residues), 2012. Scientific Opinion on the science behind the development of a risk assessment of Plant Protection Products on bees (Apis mellifera, Bombus spp. and solitary bees). Efsa Journal, 10:2668

EFSA PPR Panel (EFSA Panel on Plant Protection Products their Residues), 2013. Guidance on tiered risk assessment for plant protection products for aquatic organisms in edge-of-field surface waters. Efsa Journal, 11:3290. doi: https://doi.org/10.2903/j.efsa.2013.3290

EFSA Scientific Committee, 2011. Statistical Significance and Biological Relevance. Efsa Journal, 9:2372. doi: https://doi.org/10.2903/j.efsa.2011.2372

EFSA Scientific Committee, Hardy A, Benford D, Halldorsson T, Jeger MJ, Knutsen HK, More S, Naegeli H, Noteborn H, Ockleford C, Ricci A, Rychen G, Schlatter JR, Silano V, Solecki R, Turck D, Younes M, Bresson J-L, Griffin J, Hougaard Benekou S, van Loveren H, Luttik R, Messean A, Penninks A, Ru G, Stegeman JA, van der Werf W, Westendorf J, Woutersen RA, Barizzone F, Bottex B, Lanzoni A, Georgiadis N and Alexander J, 2017. Guidance on the assessment of the biological relevance of data in scientific assessments. Efsa Journal, 15:e04970. doi: https://doi.org/10.2903/j.efsa.2017.4970

EMA (European Medicines Agency), 2010. Guideline on the investigation of bioequivalence. Available online: https://www.ema.europa.eu/en/documents/scientific-guideline/guideline-investigation-bioequivalence-rev1_en.pdf

Engel J and van der Voet H, 2021. Equivalence tests for safety assessment of genetically modified crops using plant composition data. Food and Chemical Toxicology, 156:112517. doi: https://doi.org/10.1016/j.fct.2021.112517

FAO (Food and Agriculture Organisation), IZSLT (Istituto Zooprofilattico Sperimentale del Lazio e della Toscana), Apimondia and CAAS (Chinese Academy of Agricultural Sciences), 2021. Good beekeeping practices for sustainable apiculture. Rome Available online: https://doi.org/10.4060/cb5353en

Faucon J-P, Aurières C, Drajnudel P, Mathieu L, Ribière M, Martel A-C, Zeggane S, Chauzat M-P and Aubert MF, 2005. Experimental study on the toxicity of imidacloprid given in syrup to honey bee (Apis mellifera) colonies. Pest Management Science, 61:111-125. doi: https://doi.org/10.1002/ps.957

Flores JM, Gámiz V, Gil-Lebrero S, Rodríguez I, Navas FJ, García-Valcárcel AI, Cutillas V, Fernández-Alba AR and Hernando MD, 2021. A three-year large scale study on the risk of honey bee colony exposure to blooming sunflowers grown from seeds treated with thiamethoxam and clothianidin neonicotinoids. Chemosphere, 262:127735. doi: https://doi.org/10.1016/j.chemosphere.2020.127735

Franke L, Elston C, Jütte T, Klein O, Knäbe S, Lückmann J, Roessink I, Persigehl M, Cornement M, Exeler N, Giffard H, Hodapp B, Kimmel S, Kullmann B, Schneider C and Schnurr A, 2021. Results of 2-Year Ring Testing of a Semifield Study Design to Investigate Potential Impacts of Plant Protection Products on the Solitary Bees Osmia Bicornis and Osmia Cornuta and a Proposal for a Suitable Test Design. Environ Toxicol Chem, 40:236-250. doi: https://doi.org/10.1002/etc.4874

Franklin EL and Raine NE, 2019. Moving beyond honeybee-centric pesticide risk assessments to protect all pollinators. Nat Ecol Evol, 3:1373-1375. doi: 10.1038/s41559-019-0987-y

Garbuzov M, Schürch R and Ratnieks FLW, 2015. Eating locally: dance decoding demonstrates that urban honey bees in Brighton, UK, forage mainly in the surrounding urban area. Urban Ecosystems, 18:411-418. doi: 10.1007/s11252-014-0403-y

Gill RJ, Ramos-Rodriguez O and Raine NE, 2012. Combined pesticide exposure severely affects individual- and colony-level traits in bees. Nature, 491:105-108. doi: 10.1038/nature11585

Haider M, Dorn S, Sedivy C and Müller A, 2014. Phylogeny and floral hosts of a predominantly pollen generalist group of mason bees (Megachilidae: Osmiini). Biological Journal of the Linnean Society, 111:78-91. doi: https://doi.org/10.1111/bij.12186

Hodge S, Stout, J., 2019. Protocols for methods of field sampling. Deliverable D1.1 PoshBee Project

Hoenig JM and Heisey DM, 2001. The Abuse of Power. The American Statistician, 55:19-24. doi: 10.1198/000313001300339897

Holzschuh A, Dormann CF, Tscharntke T and Steffan-Dewenter I, 2013. Mass-flowering crops enhance wild bee abundance. Oecologia, 172:477-484. doi: 10.1007/s00442-012-2515-5

Imdorf A, Buehlmann G, Gerig L, Kilchenmann V and Wille H, 1987. A test of the method of estimation of brood areas and number of worker bees in free-flying colonies. Apidologie, 18:137-146

Julious SA, 2004. Sample sizes for clinical trials with normal data. Stat Med, 23:1921-1986. doi: 10.1002/sim.1783

Kendall LK, Mola JM, Portman ZM, Cariveau DP, Smith HG and Bartomeus I, 2022. The potential and realized foraging movements of bees are differentially determined by body size and sociality. Ecology, 103:e3809. doi: https://doi.org/10.1002/ecy.3809

Klaus F, Tscharntke T, Bischoff G and Grass I, 2021. Floral resource diversification promotes solitary bee reproduction and may offset insecticide effects – evidence from a semi-field experiment. Ecology letters, 24:668-675. doi: https://doi.org/10.1111/ele.13683

Klein O, Roessink I, Elston C, Franke L, Jütte T, Knäbe S, Lückmann J, van der Steen J, Allan MJ, Alscher A, Amsel K, Cornement M, Exeler N, Guerola JS, Hodapp B, Jenkins C, Kimmel S and Tänzler V, 2022. Results of Ring-Testing of a Semi-Field Study Design to Investigate Potential Impacts of Crop Protection Products on Bumblebees (Hymenoptera, Apidae) and a Proposal of a Potential Test Design. Environ Toxicol Chem. doi: 10.1002/etc.5430

Kopit AM and Pitts-Singer TL, 2018. Routes of Pesticide Exposure in Solitary, Cavity-Nesting Bees. Environmental Entomology, 47:499-510. doi: 10.1093/ee/nvy034

Ladurner E, Bosch J, Kemp WP and Maini S, 2008. Foraging and Nesting Behavior of Osmia lignaria (Hymenoptera: Megachilidae) in the Presence of Fungicides: Cage Studies. Journal of Economic Entomology, 101:647-653. doi: 10.1093/jee/101.3.647

Laster LL and Johnson MF, 2003. Non-inferiority trials: the 'at least as good as' criterion. Stat Med, 22:187-200. doi: 10.1002/sim.1137

Lebreton J-D, Burnham KP, Clobert J and Anderson DR, 1992. Modeling Survival and Testing Biological Hypotheses Using Marked Animals: A Unified Approach with Case Studies. Ecological Monographs, 62:67-118. doi: 10.2307/2937171

Lehmann DM and Camp AA, 2021. A systematic scoping review of the methodological approaches and effects of pesticide exposure on solitary bees. PLOS ONE, 16:e0251197. doi: 10.1371/journal.pone.0251197

LOPEZ-VAAMONDE C, RAINE NE, KONING JW, BROWN RM, PEREBOOM JJM, INGS TC, RAMOS-RODRIGUEZ O, JORDAN WC and BOURKE AFG, 2009. Lifetime reproductive success and longevity of queens in an annual social insect. Journal of Evolutionary Biology, 22:983-996. doi: https://doi.org/10.1111/j.1420-9101.2009.01706.x

Lückmann J and Schmitzer S, 2019. The Oomen bee brood feeding test–revision of the method to current needs and developments. EPPO Bulletin, 49:137-146

Lückmann J and Tänzler V, 2020. Honeybee brood testing under semi-field and field conditions according to Oomen and OECD GD 75: is there a difference of the brood termination rate? Julius-Kühn-Archiv:91-95

Mair MM, Kattwinkel M, Jakoby O and Hartig F, 2020. The Minimum Detectable Difference (MDD) Concept for Establishing Trust in Nonsignificant Results: A Critical Review. Environ Toxicol Chem, 39:2109-2123. doi: https://doi.org/10.1002/etc.4847

Marchal P, Buatois A, Kraus S, Klein S, Gomez-Moracho T and Lihoreau M, 2020. Automated monitoring of bee behaviour using connected hives: Towards a computational apidology. Apidologie, 51:356-368. doi: 10.1007/s13592-019-00714-8

Medrzycki P, Giffard H, Aupinel P, Belzunces LP, Chauzat M-P, Claßen C, Colin ME, Dupont T, Girolami V, Johnson R, Le Conte Y, Lückmann J, Marzaro M, Pistorius J, Porrini C, Schur A, Sgolastra F, Delso NS, van der Steen JJM, Wallner K, Alaux C, Biron DG, Blot N, Bogo G, Brunet J-L, Delbac F, Diogon M, El Alaoui H, Provost B, Tosi S and Vidau C, 2013. Standard methods for toxicology research in Apis mellifera. Journal of Apicultural Research, 52:1-60. doi: 10.3896/IBRA.1.52.4.14

OECD (Organisation for Economic Co-operation and Development), 2014. Guidance Document on the Honey Bee (Apis Mellifera L.) Brood test Under Semi-field Conditions.

OEPP/EPPO, 2010. EPPO Standards PP1/170 (4) Efficacy evaluation of plant protection products. Side-effects on honeybees. Bulletin OEPP/EPPO, 40:313-319

Oomen P, De Ruijter A and Van Der Steen J, 1992. Method for honeybee brood feeding tests with insect growth-regulating insecticides. EPPO Bulletin, 22:613-616

Perry JN, Ter Braak CJ, Dixon PM, Duan JJ, Hails RS, Huesken A, Lavielle M, Marvier M, Scardi M, Schmidt K, Tothmeresz B, Schaarschmidt F and van der Voet H, 2009. Statistical aspects of environmental risk assessment of GM plants for effects on non-target organisms. Environ Biosafety Res, 8:65-78. doi: 10.1051/ebr/2009009

Peters B, Gao Z and Zumkier U, 2016. Large-scale monitoring of effects of clothianidin-dressed oilseed rape seeds on pollinating insects in Northern Germany: effects on red mason bees (Osmia bicornis). Ecotoxicology, 25:1679-1690. doi: 10.1007/s10646-016-1729-4

Pettis J, Tornier I, Clook M, Wallner K, Vaissiere B, Stadler T, Hou W, Maynard G, Becker R, Coulson M, Jourdan P, Vaughan M, Nocelli RCF, Scott-Dupree C, Johansen E, Brittain C, Dinter A and Kasina M, 2014. Assessing effects through semi-field and field toxicity testing. Pesticide risk assessment for pollinators. pp. 95-119.

Pettis JS, van Engelsdorp D, Johnson J and Dively G, 2012. Pesticide exposure in honey bees results in increased levels of the gut pathogen Nosema. Naturwissenschaften, 99:153-158. doi: 10.1007/s00114-011-0881-1

Pflugmacher D, Rabe A, Peters M and Hostert P, 2019. Mapping pan-European land cover using Landsat spectral-temporal metrics and the European LUCAS survey. Remote sensing of environment, 221:583-595

Pistorius J, Becker R, Lückmann J, Schur A, Barth M, Jeker L, Schmitzer S and Von der Ohe W, 2011. Effectiveness of method improvements to reduce variability of brood termination rate in honey bee brood studies under semi-field conditions. Proceedings of the Hazards of pesticides to bees-11 th International Symposium of the ICP-BR Bee Protection Group, 115-120 pp.

Roessink I HN, Schneider C, Quambusch A, Exeler N, Cabrera AR, Molitor A-M, Tanzler V, Hodapp B, Albrecht M, Brandt A, Vinall S, Rathke A-K, Giffard H, Soler E, Schnurr A, Patnaude M, Couture A, Lehman D., 2019. Progress on the Osmia acute oral test - findings of the ICPPR Non-Apis subgroup solitary bee laboratory testing. Proceedings of the Hazards of pesticides to bees - 14th international symposium of the ICP-PR Bee protection group, Bern, Switzerland

Roessink I and Van der Steen JJM, 2021. Beebread consumption by honey bees is fast: results of a six-week field study. Journal of Apicultural Research, 60:659-664. doi: 10.1080/00218839.2021.1915612

Ruddle N, Elston C, Klein O, Hamberger A and Thompson H, 2018. Effects of exposure to winter oilseed rape grown from thiamethoxam-treated seed on the red mason bee Osmia bicornis. Environ Toxicol Chem, 37:1071-1083. doi: https://doi.org/10.1002/etc.4034

Rundlöf M, Andersson GK, Bommarco R, Fries I, Hederström V, Herbertsson L, Jonsson O, Klatt BK, Pedersen TR and Yourstone J, 2015. Seed coating with a neonicotinoid insecticide negatively affects wild bees. Nature, 521:77-80

Rundlöf M, Stuligross C, Lindh A, Malfi RL, Burns K, Mola JM, Cibotti S and Williams NM, 2022. Flower plantings support wild bee reproduction and may also mitigate pesticide exposure effects. Journal of Applied Ecology, 59. doi: https://doi.org/10.1111/1365-2664.14223

Samuelson AE, Schürch R and Leadbeater E, 2022. Dancing bees evaluate central urban forage resources as superior to agricultural land. Journal of Applied Ecology, 59:79-88. doi: https://doi.org/10.1111/1365-2664.14011

Sandrock C, Tanadini LG, Pettis JS, Biesmeijer JC, Potts SG and Neumann P, 2014. Sublethal neonicotinoid insecticide exposure reduces solitary bee reproductive success. Agricultural and Forest Entomology, 16:119-128. doi: https://doi.org/10.1111/afe.12041

Schmitz OJ, 2008. From mesocosms to the field: the role and value of cage experiments in understanding top-down effects in ecosystems. Insects and ecosystem function, Springer. pp. 277-302.

Schuirmann DJ, 1987. A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. J Pharmacokinet Biopharm, 15:657-680. doi: 10.1007/bf01068419

Schumi J and Wittes JT, 2011. Through the looking glass: understanding non-inferiority. Trials, 12:106. doi: 10.1186/1745-6215-12-106

Schwarz JM, Knauer AC, Allan MJ, Dean RR, Ghazoul J, Tamburini G, Wintermantel D, Klein A-M and Albrecht M, 2022. No evidence for impaired solitary bee fitness following pre-flowering sulfoxaflor application alone or in combination with a common fungicide in a semi-field experiment. Environment International, 164:107252. doi: https://doi.org/10.1016/j.envint.2022.107252

Sgolastra F, Hinarejos S, Pitts-Singer TL, Boyle NK, Joseph T, Lūckmann J, Raine NE, Singh R, Williams NM and Bosch J, 2019. Pesticide Exposure Assessment Paradigm for Solitary Bees. Environmental Entomology, 48:22-35. doi: 10.1093/ee/nvy105 %J Environmental Entomology

Shao J, Chow S-C and Wang B, 2000. The bootstrap procedure in individual bioequivalence. Statistics in Medicine, 19:2741-2754. doi: https://doi.org/10.1002/1097-0258(20001030)19:20<2741::AID-SIM542>3.0.CO;2-3

Sponsler DB, Grozinger CM, Hitaj C, Rundlöf M, Botías C, Code A, Lonsdorf EV, Melathopoulos AP, Smith DJ, Suryanarayanan S, Thogmartin WE, Williams NM, Zhang M and Douglas MR, 2019. Pesticides and pollinators: A socioecological synthesis. Science of The Total Environment, 662:1012-1027. doi: https://doi.org/10.1016/j.scitotenv.2019.01.016

Sponsler DB, Matcham EG, Lin C-H, Lanterman JL and Johnson RM, 2017. Spatial and taxonomic patterns of honey bee foraging: A choice test between urban and agricultural landscapes. Journal of Urban Ecology, 3. doi: 10.1093/jue/juw008

Steffan-Dewenter I and Kuhn A, 2003. Honeybee foraging in differentially structured landscapes. Proceedings of the Royal Society of London. Series B: Biological Sciences, 270:569-575

Steffan-Dewenter I and Schiele S, 2004. Nest-site fidelity, body weight and population size of the red mason bee, Osmia rufa (Hymenoptera: Megachilidae), evaluated by mark-recapture experiments. Entomologia Generalis, 27:123-131

Steffan-Dewenter I and Schiele S, 2008. DO RESOURCES OR NATURAL ENEMIES DRIVE BEE POPULATION DYNAMICS IN FRAGMENTED HABITATS. Ecology, 89:1375-1387. doi: https://doi.org/10.1890/06-1323.1

Stuligross C and Williams NM, 2020. Pesticide and resource stressors additively impair wild bee reproduction. Proceedings of the Royal Society B: Biological Sciences, 287:20201390. doi: doi:10.1098/rspb.2020.1390

Tamburini G, Pereira-Peixoto M-H, Borth J, Lotz S, Wintermantel D, Allan MJ, Dean R, Schwarz JM, Knauer A, Albrecht M and Klein AM, 2021. Fungicide and insecticide exposure adversely impacts bumblebees and pollination services under semi-field conditions. Environment International, 157:106813. doi: https://doi.org/10.1016/j.envint.2021.106813

Tepedino VJ and Torchio PF, 1982. Phenotypic variability in nesting success among Osmia lignaria propinqua females in a glasshouse environment: (Hymenoptera: Megachilidae). Ecological Entomology, 7:453-462. doi: https://doi.org/10.1111/j.1365-2311.1982.tb00688.x

Tsvetkov N, Samson-Robert O, Sood K, Patel HS, Malena DA, Gajiwala PH, Maciukiewicz P, Fournier V and Zayed A, 2017. Chronic exposure to neonicotinoids reduces honey bee health near corn crops. Science, 356:1395-1397. doi: doi:10.1126/science.aam7470

US-EPA (Environmental Protection Agency) (U.S. Environmental Protection Agency), 2016. Guidance on Exposure and Effects Testing for Assessing Risks to Bees.

US FDA (US Food and Drug Administration), 2001. Guidance for Industry: Statistical Approaches to Establishing Bioequivalence. Available online: http://www.fda.gov/downloads/Drugs/Guidances/ucm070244.pdf

Van der Steen J and Cornelissen B (Plant Research International, Wageningen UR), 2015. Factoren die het foerageergedrag van honingbijen bepalen (deel I); Dracht in Nederland (cultuurgewassen en wilde planten)(deel II).

Visscher PK and Seeley TD, 1982. Foraging Strategy of Honeybee Colonies in a Temperate Deciduous Forest. Ecology, 63:1790-1801. doi: https://doi.org/10.2307/1940121

Walker E and Nowacki AS, 2011. Understanding equivalence and noninferiority testing. J Gen Intern Med, 26:192-196. doi: 10.1007/s11606-010-1513-8

Wang M, Braasch T and Dietrich C, 2020. Reduction of variability for the assessment of side effects of toxicants on honeybees and understanding drivers for colony development. PLOS ONE, 15:e0229295. doi: 10.1371/journal.pone.0229295

Wellek S (Hall/CRC Ca), 2010. Testing Statistical Hypotheses of Equivalence and Noninferiority (2nd ed.).

Whitehorn PR, O'Connor S, Wackers FL and Goulson D, 2012. Neonicotinoid pesticide reduces bumble bee colony growth and queen production. Science, 336:351-352. doi: 10.1126/science.1215025

Willis Chan DS and Raine NE, 2021. Population decline in a ground-nesting solitary squash bee (Eucera pruinosa) following exposure to a neonicotinoid insecticide treated crop (Cucurbita pepo). Scientific Reports, 11:4241. doi: 10.1038/s41598-021-83341-7

Wintermantel D, Pereira-Peixoto M-H, Warth N, Melcher K, Faller M, Feurer J, Allan MJ, Dean R, Tamburini G, Knauer AC, Schwarz JM, Albrecht M and Klein A-M, 2022. Flowering resources modulate the sensitivity of bumblebees to a common fungicide. Science of The Total Environment, 829:154450. doi: https://doi.org/10.1016/j.scitotenv.2022.154450

With KA, 2019. Essentials of Landscape Ecology. Oxford, Oxford University Press.

Woodcock BA, Bullock JM, Shore RF, Heard MS, Pereira MG, Redhead J, Ridding L, Dean H, Sleep D, Henrys P, Peyton J, Hulmes S, Hulmes L, Sárospataki M, Saure C, Edwards M, Genersch E, Knäbe S and Pywell RF, 2017. Country-specific effects of neonicotinoid pesticides on honey bees and wild bees. Science, 356:1393-1395. doi: doi:10.1126/science.aaa1190

ZANETTE LRS, MILLER SDL, FARIA CMA, LOPEZ-VAAMONDE C and BOURKE AFG, 2014. Bumble bee workers drift to conspecific nests at field scales. Ecological Entomology, 39:347-354. doi: https://doi.org/10.1111/een.12109

Zurbuchen A, Landert L, Klaiber J, Müller A, Hein S and Dorn S, 2010. Maximum foraging ranges in solitary bees: only few individuals have the capability to cover long foraging distances. Journal of Biological conservation, 143:669-676

1

Appendix A – Foraging range of honey bees

A.1. Introduction

The foraging range of honey bees is important to consider when setting the minimum required distance between treated and control fields in PPP field effect studies. It is also considered in the requirement for the size of the area around the test fields which needs to be free of alternative forage in both field exposure and effect studies.

A.2. Foraging range in current guidance

Current guidelines and guidance for honey bee field studies to test the effect or the exposure of a PPP report varying ranges between and around fields:

- The OEPP/EPPO (2010) protocol states that for field effect studies "*the plots should not be close to other flowering crops or non-cultivated areas which are significantly attractive to bees. As a guide, the same separation distance as for the test plots should be considered* [2-3 km]*, taking into account the size and attractiveness of the other crops or noncultivated areas."*
- The Scientific Opinion EFSA PPR Panel (EFSA Panel on Plant Protection Products and their Residues) (2012) recommended a distance between control and test fields in effect studies of 4-6 km.
- The previous guidance document (EFSA (European Food Safety Authority), 2013) requires for effect studies that '*sources of nearby alternative forage should be sparse*' and that test fields are separated by at least 4 km to account for the foraging distance of honeybees. However, for exposure field studies (determining the level of residues taken back to the hive by honey bees) it is stated that "no bee attractive crop and no flowering hedges or trees should be present within 2 km around the colonies. No flowering weeds in the treated fields and no flowering plants in field margins of the treated crop and adjacent crops."
- The US-EPA (Environmental Protection Agency) (2016) requests that '*crops/alternative forage within 3 km of colonies are accounted for*' and reports a mean foraging range of honeybees of 1.5-3 km. In addition, it is acknowledged that in any landscape there will be at least some alternative forage available and that full exclusion can never be reached.
- In their recommendations for data collection to assess the health status of managed honey bee colonies, the EFSA Panel on Animal Health and Welfare suggests using an average foraging distance of 3 km around the colony and states that "*the surfaces of the habitat need to be estimated in order to define the relative contribution of different source of feed and other resources*.' and "*it is important to define the relative contribution of different sources of food and other resources"* (EFSA, 2016).
- Medrzycki et al. (2013), in the COLOSS BEEBOOK recommendations for testing toxicity on bee colonies in field conditions, Section 4.3.1.3, state that the negative reference field should be located at least 4 km from the treatment field.

In addition, the hearing experts that were invited by the working group considered 4 km a reasonable distance to consider as foraging range in the design of honey bee exposure refinement studies.

A.3. Summary of open literature findings

Factors to consider with regard to the foraging range of honey bees are, among others, honey bee colony size, season and landscape context (agriculture, urban, forest etc.). No systematic literature search could be done within the available timeframe. However, relevant literature known to the working group is summarized here.

With one exception of a study in a forest in the USA, all studies were performed in either the UK or Germany. All available studies used waggle dance decoding to determine the foraging range of honey bees. The studies include a variable number of hives (1-16), variable observation length (8-480 h), a variable number of analysed dances (hundreds to thousands) and span variable times of year. Some

studies focussed on pollen foragers only while others also included nectar foragers. The foraging range is reported in varying percentiles. The mean or the median is almost always available, and the 90th and/or 95th percentile and/or maximum are reported in several studies. Hives were small in most studies (4,000 workers), some hives were kept indoors, and several were controlled for size during the observation period and fed when needed. All of these factors may have influenced their foraging range relative to that of a honey bee hive under more normal beekeeping conditions. Especially in more recent studies, foraging ranges were compared between different landscapes, e.g. containing a large or a small area with maize or oilseed rape, or in agricultural versus urban areas, or between parts of the year, especially spring and summer.

Noting these differences, and noting also the great variability observed within many of the studies between different observation days and/or hives, nevertheless a general picture emerges. Several studies show larger foraging distances in summer compared to spring. While honey bees have been recorded as foraging as far as 10 km from their hives, mean and median foraging trips rarely exceeded 2 km (Table). When reported, 90th – and 95th percentiles are <2,500 m in datasets that were based on more than one hive (4, 10, 10 or 16 hives in the four datasets). Larger 90th and 95th percentiles (foraging range 3, 5, 6 or >9.5 km) come from studies on single hives.

It should be considered that the scientific literature on foraging range studies is different from studies that test the effect of PPPs with respect to the selection of locations. In the latter, hives have to be placed at the edge of the treated and control fields, which will usually contain an attractive and flowering crop. The size of the test field should meet the nutritional requirements of the bee hives placed at the edge. Under such conditions, there may assumed to be less need for the bees to search for forage further from the hive. Nevertheless, even under such conditions, honey bees have been shown to forage elsewhere than the field they are located next to (see supplementary document, Section 5.3.7 on the Landscape Factor). Additionally, some of the studies on foraging range were done in urban areas. These are different from the agricultural areas in which effect studies for PPPs are done.

Also, hives in honey bee dossier studies are normally larger than the hives tested here. The role of hive size in foraging distance is unclear: while Beekman et al. (2004) found larger distances for larger hives, they based this finding on relatively few observations. Samuelson et al. (2022) found no relation between colony strength and foraging distance.

The working group noted the differences between the studies. As setting clear exclusion criteria is difficult, it was concluded to keep all studies to derive relevant foraging ranges for honey bee field studies. It was however acknowledged that if additional data become available that are more fit for purpose, this parameter can be changed in the future. Note that all the data are gathered in studies where hives were not purposefully placed close to a mass flowering crop, and therefore they can be considered conservative.

In the table below, all studies are summarized. Where a study reports different findings for different study setups, these different setups are included in the table as different datasets. The exception is one study where all subsets yielded very similar results (Steffan-Dewenter and Kuhn, 2003); this study is included as one dataset only. This leads to a total of 14 datasets. It is noted that these are of varying size and quality, with some of the datasets lacking detail or being based on relatively few observations and/or hives. As the latter generally show larger foraging ranges, they are all kept in, in a conservative approach.

Table 4: summarizes the findings from the scientific literature on foraging range of honey bees. Where publications clearly report different findings for different study setups, these are given separate rows in the table

Reference	Number of colonies	Colony size	Timing of measurement and	Type of measurement and number	Area type and location	Foraging range (m)	Comment

			total observation period								
						Min-max	Mean	Median	90th p.ile	95th p.ile	
Beekman and Ratnieks (2000)											
(a)	1	4,000	15, 16 and 19 August 1996 for 8, 8 and 4 hours = 20 h	Dance decoding (444 dances)	Sheffield, UK, east of the heather moors		5,500	6,100	>9,500		Only 1 colony. Extensive patches of heather were in bloom far from the hive. 10th% <500 m; 75th% >7,500 m
(b)	1	4,000?	1-3 May 1997, observation period unknown	Dance decoding (456 dances)	Sheffield, UK, east of the heather moors		1,000		3,000-3,500		Only 1 colony. Heather was not in bloom. Few details on methods.
Beekman et al. (2004)											
(a)	2	6,000	23,24,26, 29,30 July 1999, 4 h/day: 20 h/colony = 40 h	Dance decoding (ca. 1285 dances)	University of Sheffield, Yorkshire, UK		670-680	350-530			
(b)	2	6,000	2 August 1999, 4 h/colony = 8 h	Dance decoding (204 dances)	University of Sheffield, Yorkshire, UK		1,020-1,970	230-380			One day only, few data
(c)	2	20,000	23,24,26, 29,30 July 1999, 4 h/day: 20 h/colony = 40 h	Dance decoding (1185 dances)	University of Sheffield, Yorkshire, UK		620	480-500			

(d)	2	20,000	2 August 1999, 4 h/colony = 8 h	Dance decoding (211 dances)	University of Sheffield, Yorkshire, UK		2,630 - 3,010	2,750- 2,810			One day only, few data
Couvillon et al. (2015)											
	3	5,000	July-August + March + Sept-Oct	Dance decoding (5484 dances of which 4562 nectar dances and 922 pollen dances)	University of Sussex campus in East Sussex, UK	Max <6000 m	1,408 (nectar) / 1,074 (pollen)				Hives kept indoor and were managed to keep the number of workers consistent between the hives and over time; also they were sometimes fed.
Danner et al. (2014)											
	4	4,000	22 July to 7 August, 90 min./hive/d: 17 days*4 hives *90min = 102 h	Dance decoding for pollen (614 dances)	northern Bavaria, Germany	14-4,439	820 ± 582 m		1,538		Hives were rotated along landscapes with different maize acreage proportions
Danner et al. (2016)											
	16	4,000	18 April-20 August, 7*90min. /hive = 168 h	Dance decoding for pollen (1347 dances)	Würzburg, Germany	35-9,510	1,015 ± 26 m		2,193		Hives were rotated along landscapes with different OSR and semi-natural areas acreag

											e proportions. 75th% 1,355 m
Garbuzov et al. (2015)											
	3	2,000-5,000	April – October 2001, 1 h/~week/hive, 25,5 weeks: 77 h	Dance decoding (931 dances)	Brighton, UK		4–1 - 1229				Urban area – range of means per month shown
Samuelson et al. (2021)											
a)	10	6,000 - 18,000?	24 weeks in April-Sept 2017, 2 h/hive every 2 weeks: 480 h	Dance decoding (1,428 dances)	SE UK, urban areas	Max 9,375		492		<2,500	Hives located in urban area
b)	10	6,000 - 18,000?	24 weeks in April-Sept 2017, 2 h/hive every 2 weeks: 480 h	Dance decoding (2,827	SE UK, agricultural areas	Max 8,158		743		<2,500	Hives located in agricultural area
Steffan-Dewenter and Kuhn (2003)											
	4	4,000	12 May-31 July 2001, 30-40 minutes per colony = 80days * 30-40 min = –0 - 53 h	Dance decoding (1137 dances)	Lower Saxony, Germany, 3 simple + 3 complex landscapes	62 – 10,037	1,526 ±37	1182			Colonies were moved between the landscapes over the study period
Visscher and Seeley (1982)											
	1	20,000 bees	Four 9-d periods spanning the 1980 summer (June-August), i.e. dances from 36x9=324 h	Dance decoding from observation hive	forest in New York, USA	50-10100 m	2260 ± 1890 m	1650	5000	6000	Only 1 colony. Great variability in patterns of foraging distance at different times during summer. 99th%

											7,700 m
							Mean of all means[1]: 1,591 m		**Mean of the 90th percentiles[2]: 3,819 m**		

1) When two values are reported for a dataset, the highest value is entered. When no mean is reported, the median is entered.

2) When a range is reported, the highest value is entered. When no 90th percentile is reported, the 95th percentile is entered when available. When only an unbound value is reported, it is entered as a bound value.

A.4. Conclusion

From all datasets in Table 4, the mean of all means is 1.6 km. The mean of the 90-95%tile values is 3.8 km. The max is around 10 km. It is acknowledged that in view of the complexity of the data, such calculations are an oversimplification. Nevertheless, the working group considers the data useful to conclude on the foraging range of honeybees for the purpose of field study design.

Based on the above, the working group considers that in **field exposure studies**, the area around the hive that needs to be characterized **in detail** with regard to alternative forage is **1.5 km**, and that **the area around the test hive that needs to be free of mass-flowering alternative forage is 4 km**.

For **field effect studies,** the same is recommended (though not required), to increase the chances of meeting the exposure assessment goal. Additionally, in field effect studies, the **distance between test item and control fields within a block should be at least 4 km** to exclude systematic exchange of bees and the **distance between blocks should be at least 10 km** to exclude exchange of bees.

A.5. Study details

Beekman and Ratnieks (2000) decoded Waggle dances of honey-bees during the blooming of heather (*Calluna vulgaris* L.) in August 1996 using a hive located in Sheffield, UK, east of the heather moors. The hive contained ca. 4000 workers and one frame of brood. Dances of returning foragers were videotaped on 15, 16 and 19 August for 8, 8 and 4 hours, respectively. At this time, extensive patches of heather were in bloom. 444 bee dances were decoded. The median distance foraged was 6.1 km, the mean 5.5 km. Only 10% of the bees foraged within 0.5 km of the hive whereas 50% went more than 6 km, 25% more than 7·5 km and 10% more than 9·5 km from the hive.

On 1-3 May 1997, the experiment was repeated with a different hive. 456 dances were decoded. The mean patch distance in May was 1 km. No further details or calculations are given for this month but based on figure 3 in the paper, a 90th percentile foraging distance of 3-3.5 km is estimated.

Foraging may also depend on season, as shown by Beekman et al. (2004), who studied foraging distance of two small and two large hives based on dance decoding at two timepoints (July and August). The hives were located at the Laboratory of Apiculture and Social Insects, University of Sheffield, Yorkshire, UK. Dances of returning foragers were videotaped simultaneously on 23, 24, 26, 29, 30 July and 2 August 1999, for 4 hours each day per colony. Ca. 30 dances were decoded for each hour of videotape. Total number of dances decoded for L1, L2, S1 and S2 are, respectively: 120, 124, 119, 120 (23 July); 124, 122, 120, 120 (24 July); 138, 76, 181, 131 (26 July); 120, 120, 120, 132 (29 July); 121, 120, 120, 122 (30 July) and 91, 120, 115, 89 (2 August). It is noted that the data in the Table below are presented for July and August separately, suggesting an influence of month, while the measuring days were actually very close together and the datasets are imbalanced (5 measuring days for July, only 1 for August). The authors presented their data in this way since they state that on 2 August, bees started foraging on distant heather moors. They present mean and median distances. S refers to small colonies

(approximately 6,000 bees) and L to large colonies (approximately 20,000 bees), see **Table** (data copied from article). A figure in the publication presents foraging locations around each hive at each day, but it is difficult to read exact values from this.

Table 5: Mean and median foraging distances in July and August for all four colonies (S1, S2, L1, L2), as presented in Beekman et al. (2004).

	S1	S2	L1	L2
Median foraging distance July (km)	0.35	0.53	0.50	0.48
Mean foraging distance July (km)	0.67	0.68	0.62	0.62
Median foraging distance August (km)	0.23	0.38	2.81	2.75
Mean foraging distance August (km)	1.02	1.97	2.63	3.01

Couvillon et al. (2014) found results within the ranges reported in the Scientific Opinion: mean foraging distance for nectar 1408 m (based on 4,562 decoded dances), mean pollen 1074 m (922 dances), max.<6000 m (5484 decoded dances). They found that median foraging distance tended to be higher in summer (July-August) than in spring (March) and autumn (September-October), see figure 2 in the paper. Note that seasons were not defined by calendar months per se but by presence or absence of flowering plants. For their study, the authors used three indoor observation hives of 5,000 workers which were located at the laboratory on the University of Sussex campus in East Sussex, UK. Videos were made from 11 August 2009 to 31 August 2011 on most days when bees were foraging, March-October. Each hive was filmed for 1 h per study day. Dancing bees were classified as pollen foragers when carrying pollen, otherwise as nectar foragers. Based on the same data, **Couvillon et al. (2014)** conclude that honey bees can be used as bio-indicators to monitor large areas and provide information relevant to better environmental management. The hives may have been restricted to this size throughout the season (the method Section states: "All three colonies were queen-right and managed using standard beekeeping techniques to keep the numbers of workers consistent between the hives and over time, and to prevent overcrowding, which leads to swarming."). Hives were given supplemental feed when needed, which was rare. The fact that they were kept indoors may influence thermoregulation and labour devision, hive size restriction and feeding may have consequences for their foraging ranges.

Danner et al. (2014) used four hives of 4,000 workers each which were rotated along different landscape types in northern Bavaria, Germany, containing varying proportions of maize acreage in a radius of 1.5 km around the hive. All hives were observed for normally 90 minutes each day from 22 July to 7 August, i.e. 17 days, when the maize in the area was flowering. Waggle dances were decoded to determine foraging distance. During maize flowering a total of 614 bee dances for pollen sources in 11 landscapes with varying maize acreage were recorded and decoded. 125 dances (19%) advertised for maize pollen (determined by colour/pollen analysis) and 489 for other pollen species. The mean pollen foraging distance was 820 ± 582 m with a range from 14 to 4439. 90% of all foraging locations were within 1538 m around the hive. The mean foraging distances of bees that collected maize pollen were significantly lower (589 ± 41 m, range 27–3040 m) than for other pollen origins (879 ± 27 m, range 14–4439 m. 5% of all maize pollen foraging locations were beyond 1456 m.

In a comparable setting, Danner et al. (2016); Danner et al. (2017) used sixteen hives of 4,000 workers each which were rotated along different landscape types around Würzburg, Germany, containing as flower sources different proportions of oilseed rape and semi-natural habitats in a radius of 2 km (2016, 2017) around the hive. All hives were observed for 90 minutes in seven observation rounds (four rounds during spring and OSR bloom, 18 April-24 May, and three blocks in summer after OSR bloom, 16 July to 20 August). Waggle dances were decoded to determine foraging distance. Pollen diversity in pollen collected with bee traps was determined using DNA sequencing.

Pollen foraging distances of 1347 observed and decoded bee dances ranged between 35 and 9510 m with a mean of 1015 m (± 26 m; standard error of the mean; SEM). Ninety percent of all dances were within 2193 m and 75% were within 1355 m around the hives. In spring, increasing area of flowering OSR within 2 km reduced mean pollen foraging distances from 1324 m to only 435 m. In summer, increasing cover of SNH areas close to the colonies (within 200 m radius) reduced mean pollen foraging distances from 846 to 469 m.

The authors demonstrated that foraging distances increase with decreasing landscape diversity. At the same time, neither pollen amount nor diversity were influenced by landscape diversity, suggesting that honey bees compensate for lower resource availability by increasing their foraging range to maintain pollen amount and diversity. During the mass-flowering of oilseed rape, this crop was the dominant pollen source. Nevertheless, decoding waggle dances showed that during this time, foragers danced more to alert for diverse pollen sources located in habitats with patchy plant distribution, possibly to ensure that a diverse pollen diet is collected despite the mass abundance of oilseed rape.

Garbuzov et al. (2015) demonstrated that urban honey bees from three colonies held in observation hives in Brighton, UK, foraged mainly in the surrounding urban area (up to 1 km) during the main bloom period of oilseed rape in April-May 2011, using three hives of 2,000-5,000 workers. The bees were kept ca. 2.2 km from countryside. Worker bees and brood were removed as necessary to prevent swarming, which is triggered by overcrowding. To prevent possible starvation, colonies were fed 500 ml of 2 M sugar solution most weeks after videoing for data collection, so that the syrup had been consumed several days before data collection resumed. Colonies were monitored from 20 April to 16 October (i.e. 25.5 weeks) in 2011, which encompasses most of the foraging season (March/April–October/November) in the UK. The dance area of each hive was video-recorded for 1 h at approximately weekly intervals between 10:00 and 16:00 BST during favourable foraging weather. Details about the findings in the different months are shown in Figure 2 in the paper and in the Table below, in which data from

Table in the paper were copied. Generally, the foraging range was very small compared to findings in other studies.

Table 6: Estimated foraging distances with confidence interval and number of waggle dances, in April-October, as presented in Garbuzov et al. (2015)

Month	**Estimated foraging distance (m)**		**Number of waggle dances**
	Mean	**Confidence interval**	
April	518	86	61
May	461	41	234
June	670	110	116
July	1,229	175	209
August	589	64	166
September	685	96	95
October	846	235	50

Samuelson et al. (2022) analysed 2,827 waggle dances of 20 honey bee colonies placed at either the urban or the agricultural extremes of an urbanization gradient in SE England. Each hive was recorded every 2 weeks for 2 h a day, over 24 weeks from April to September 2017. They also analysed 551 dances from a subset of these hives in 2016, to investigate consistency across years. Hives contained three to eight frames of workers, brood and a queen. Number of bees not given and are estimated by the working group at 6,000-18,000 ca.2,200 per frame based on various numbers for European hives in Delaplane et al. 2013). Colonies were only fed when at risk of starvation. Swarm control (removal of a brood frame and/or queen cells) was carried out between April and July. At each session, they also recorded nectar sugar concentration. Foraging trips were generally longer in agricultural areas than in urban areas, which was not compensated for by collection of nectar with higher sugar content. Their model found no effect of a.o, colony strength on foraging distance.

In urban areas (1,428 dances), median foraging distance was 492 m, max was 9,375 m. In agricultural areas (1,399 dances), median was 743 m, max was 8,158 m. While no other percentiles are quantified, it is stated that a 2,5 km radius around each site incorporated the 95th percentile of recorded dances, and Figure 3a in the paper shows the distribution of foraging distances.

Sponsler et al. (2017) analysed dance behaviour from 7 August to 26 September 2014 one day per week from three three-frame observation hives to determine whether bees from hives located on the

border of urban and agricultural landscapes. Waggle dances were analysed from a morning and an afternoon recording session of 45 minutes each, for seven days in total, resulting in a total observation time of 10.5 h per hive or 31.5 h in total. The number of analysed dances ranges between 17 and 347 on different dates and is 694 in total. The spatial foraging patterns are depicted in a figure, showing results for eight different dates (Figure 3 in the paper). The authors state that '*Foraging activity was most concentrated near the apiary […]. When foraging activity ranged >1 km from the apiary, it was consistently concentrated in the agricultural landscape to the south and west, though occasional foraging occurred along the urban–agricultural interface to the north (7 August, 4 September) and in the urban landscape to the east (12 September) (Fig. 3). The most distant foraging occurred 4 September, when a small amount of activity occurred in the agricultural landscape ~4 km southwest of the apiary*'.

The results suggest that the 90th percentile of the data over the whole period is < 3 km, but without actual foraging range values given in the article, this cannot be stated with certainty, nor can mean/median or any other percentiles be calculated. The study is therefore not included in the summary table, but the working group notes that the results do not indicate that our overall conclusions on foraging range are under conservative.

Steffan-Dewenter and Kuhn (2003) compared three structurally simple landscapes characterized by a high proportion of arable land and large patches, with three complex landscapes with a high proportion of semi-natural perennial habitats and low mean patch size. Four glass-sided observation hives were placed in the centre of the landscapes and switched at regular intervals between the six landscapes from the beginning of May to the end of July. The study was conducted in 2001 in southern Lower Saxony, Germany. The observation hives were built in the spring from artificial swarms and sister queens and started with approximately 4000 individuals. Part of the brood was removed now and then to prevent overcrowding and swarming. Honeybee dances were decoded in each colony at least once each foraging day (i.e. when nectar and pollen foraging took place) for 30-40 minutes per colony, from 12 May until 31 July. The sequence of the observations was randomly changed each day. A total of 1137 bee dances were observed and decoded, 376 of pollen foragers, 688 of nectar foragers and 73 unknowns (because a pollen trap was active during those observations). Also, dance activity was measured.

The following foraging distances can be read from the paper:

Table 7: Foraging distances as presented in Steffan-Dewenter and Kuhn (2003)

	Overall (1137 dances)	**Simple landscapes (527 dances)**	**Complex landscapes (610 dances)**	**May**	**June**	**July**
Mean	1526 ±37	1569 ±56	1489±50	1319±53	1787±97	1518.2±51
Median	1182	1265	1145	1076	1329	1184
range	62 – 10037					
95% confidence interval (corrected for colony effects) (values estimated from Figures 2 and 3 in the paper)				1150-1320 m	1410-2040 m	1120-1580 m

Statistical analysis (three-way ANOVA) showed that foraging distances differed significantly between simple and complex landscapes for pollen- but not for nectar-collecting bees. Furthermore, landscape effects were subjected to seasonal changes and foraging distances varied between individual colonies.

Dancing activity was significantly higher in complex than in simple landscapes but did not depend on month or colony.

Visscher and Seeley (1982) studied one colony under primarily non-agricultural conditions: a forest in New York, USA. The colony contained ~20,000 bees and was placed in an observation hive. They recorded the dances of bees continually from 0800 to 1700 each day for four 9-d periods spanning the 1980 summer: 12-20 June, 9-17 July, 28 July-5 August, and 19-27 August. Pooling data from all 36 d of observation yields the following statistics on foraging distance: 2260 ± 1890 m (mean + 1 SD); range 50-10 100 m; 50th percentile (median), 1650 m; 90th percentile, 5000 m; 95th percentile, 6000 m; 99th percentile, 7700 m. There was a great deal of variability in the patterns of foraging distance at different times during the summer.

In addition to the measurements of foraging distance presented above, there are two more references of a more theoretical nature that are not included in the overview table: **Van der Steen and Cornelissen (2015)**, in a review which discusses the balance between dietary needs and foraging costs, concluded that honey bee foraging flights are usually restricted to less than one kilometre. **Baveco et al. (2016)** assumed a maximum foraging distance of 2 km in their energetics-based honey bee nectar-foraging model, arguing that in their model, by definition, sites were next to a large resource patch.

Appendix B – Exposure via pollen patties in colony feeder studies in honey bees

Most feeder studies use sugar solution as an exposure matrix. To investigate the possibilities of an alternative or additional exposure matrix, this document summarizes the papers known to the Working Group that used pollen patties as exposure matrices to honey bees in pesticide effect studies.

B.1. Summary of experiments

Five experiments were found that used pollen patties inserted directly into the colony to ensure exposure of honey bees. The colonies were kept outside under normal conditions. Pollen patties were composed of a mixture of pollen and sugar solution or honey, of varying origin. Some authors report that they avoid using pollen taken directly from bees since according to them it contains many pesticides. In all but one case, it was reported that pollen traps were placed in front of the hive to limit pollen entry from foraging and stimulate consumption of the offered patties. Frequency of feeding ranged between every 2-3 days to once per week. Details on hive size at the start of the exposure are not given in most studies. 360-1200 g pollen patties were given per week. Verification of consumption was done in some cases; consumption was almost complete in one study but incomplete in other studies.

In addition, one paper calculates the theoretical amount of beebread consumption. This paper is also included below.

Table 8: Honey bee feeding studies using pollen patties - summary table

Reference	Colony size	Patty composition	Amount and frequency of feeding	Remarks on consumption
Pettis et al. (2012)	30-40.000 adult bees	Megabee® protein patties (100 g each) or patties containing 5 and 20 ppb of imidacloprid made by mixing neat material with the sucrose solution used to make the protein patties.	Each colony received four 80-g patties per week = 360 g/week, for 10 weeks	Consumption ca. 30g patty/colony/d, i.e. 210 g/week.
Tsvetkov et al. (2017)	Two deep chambers separated by an 1-inch spacer; the bottom chamber containing brood and food stores and the upper brood chamber containing empty foundation frames to allow colonies to grow	56% FeedBee pollen supplement (Bee Processing Enterprises LTD.), 33% sugar syrup and 11% water	Each colony received a 200g patty 3x/week = 600 g/week, for 12 weeks.	Pollen traps in place. Consumption not reported.
Williams et al. 2015	Not reported ('typical for the season': early May in Switzerland)	Bee-collected pollen and honey (3:1 by mass, respectively) obtained from non-intensive agricultural areas of Switzerland	Each colony received 100g patty 7x/week = 700 g/week, for 36 days.	Pollen traps in place. Pollen patties 'well received, but never completely consumed'.

Sandrock et al. 2014	Unclear. Colonies started with 1.5 kg of bees one year earlier and had three hive bodies with 11 frames each when exposure started.	55% honeybee pollen (common stock of commercial pollen with mixed floral content of at least 19 plants; Sonnentracht Imkerei, Bremen, Germany), 5% brewer's yeast and approximately 40% sucrose (two thirds 73% sugar syrup and one third powder sugar)	2x200 g 3x/week = 1200 g/week for 46 days	Pollen traps in place. The two patties were generally consumed within 48 hours.
Dively et al. 2015	Size unclear at start of experiment, hives had started with 900 bees 4 weeks earlier	Pollen diet substitute powder from MegaBee, Dadent &Sons, Inc., Hamilton, IL and sugar solution, at 1.7:1 diet to sucrose solution ratio	Each colony received four 80-g patties per week = 360 g/week, for 12 weeks	Pollen traps in place. Measured consumption 265-277 g/colony/week.
Study on pollen turnover in colony with estimations of pollen intake				
Roessink and van der Steen	2 colonies with ca.5,600 bees	Not relevant	Almost 75% of collected pollen consumed within ~one week. 95% consumed within two weeks and only a small remainder was stored for a prolonged period.	Theoretical consumption: 21 g/d/10.000 bees, fits with what is estimated in their field study.

B.2. Honey bee feeding studies using pollen patties – summaries of individual studies

(Pettis et al., 2012) continually exposed full sized colonies of bees (30–40,000 adults) for ten weeks to 5 and 20 ppb imidacloprid by provisioning colonies with protein supplement patties spiked with the pesticide. Thirty honey bee colonies were used and divided into three treatment groups of ten colonies each. Colonies were established in April 2008 in five apiaries, approximately 0.5 km apart containing two colonies from each treatment group (total colonies per apiary 0 6). Packages of bees (1.8 kg) were installed in new hive equipment including frames with waxcoated plastic foundation. All queens established in study colonies came from the same genetic source. All colonies were managed to limit the levels of other pests and pathogens. All colonies were fed equal amounts of sucrose syrup until natural forage was abundant in May. It is not reported if pollen traps were in place during exposure.

Treatments consisted of untreated Megabee® protein patties (100 g each) or patties containing 5 and 20 ppb of imidacloprid made by mixing neat material with the sucrose solution used to make the protein patties. Samples were taken of fresh treated protein patties and analyzed for imidacloprid content to insure proper delivery of the target dose. Beginning in May 2008, each colony received four 80-g patties per week for 10 weeks. Unconsumed patties were removed after 7 days, weighed to measure consumption, and replaced with new treatment protein patties.

Pesticide exposure to colonies was verified by measuring the weekly consumption of the treated protein patties and by analyzing the level of imidacloprid in stored bee bread and random-aged bees removed from colonies 1 week after the exposure period (Table). Daily protein patty consumption averaged 29.0±0.84, 29.3±0.78, and 31.1±0.85 g for the control, 5 and 20 ppb colonies, respectively, and was not significantly different among treatments (ANOVA, F00.83, d.f.02, 7, P00.39).

Tsvetkov et al. (2017) treated honey bees with clothianidin in an artificial pollen supplement over a 12 week period in 2015, to investigate potential effects of this "season long" exposure to clothianidin. 10 standard honey bee colonies, which contained two deep chambers separated by an 1-inch spacer;

the bottom chamber containing brood and food stores and the upper brood chamber containing empty foundation frames to allow colonies to grow. Pollen traps were placed on all colonies and activated at the beginning of the experiment. Every 2-3 days each colony received a fresh 200 g artificial pollen patty, which was placed between the two chambers. The researchers used artificial pollen patties in lieu of real pollen because real pollen often contains a number of agrochemicals, even when collected from natural areas, as shown in the current and other articles. The artificial pollen patties were mixed using 56% FeedBee pollen supplement (Bee Processing Enterprises LTD.), 33% sugar syrup and 11% water.

Multi-residue testing on 5 treated and 5 control pollen patties was done to confirm dosing and check for contamination. All patties contained trace levels of two fungicides likely used as preservatives.

The study of **Williams et al., 2015,** was considered in the EFSA 2018 reassessment of clothianidin and thiamethoxam, (EFSA (European Food Safety Authority), 2018c, a), see C_T.Colony_feeder, study C*T.1515. The experiment was performed in Switzerland, during May-September 2013 using A. mellifera carnica honey bees. Six sister queen experimental colonies were established in early May; each contained typical quantities of adults, immatures and food for the season. Colonies were randomly assigned to either neonicotinoid or control treatments, with each group represented equally.

Treatments were administered via pollen supplements that were prepared from bee-collected pollen and honey (3:1 by mass, respectively) obtained from non-intensive agricultural areas of Switzerland. Supplements for the neonicotinoid treatment were additionally spiked with 4 ppb thiamethoxam and 1ppb clothiandin. These amounts were confirmed (4.16 and 0.96 ppb for thiamethoxam and clothianidin, respectively) by chemical analysis. All colonies were equipped with pollen trap for limiting pollen inflow. Each colony received 100 g pollen supplement every day for 36 days. Authors reported that supplements were "well-received", but never completely consumed.

The study of (Sandrock et al., 2014) was considered in the EFSA 2018 reassessment of clothianidin and thiamethoxam, (EFSA (European Food Safety Authority), 2018c, a), see C_T.Colony_feeder, study C*T.1212. An experiment involving 24 honeybee colonies was carried out. Such colonies were established using artificial swarms (1.5 kg of bees) in summer 2010. 14 sister queens originated from an intensely cultivated area in Germany (strain A), while 10 sister queens originated from an alpine area in Switzerland (strain B). Two groups of 12 hives (containing each 7 queens from strain A and 5 queens from strain B) were put in a single row, separated by 20 meters.

Colonies were treated against Varroa with oxalic acid several times, the first being five days after the establishment of the colonies (no capped brood present). During 2010, colonies were fed with sugar syrup, while pollen was naturally collected by the bees.

During spring and summer 2011 colonies were not fed but left to freely collect nectar and pollen. All colonies were simultaneously provided with a second and third hive body containing 11 frames with wax-foundations for comb building in early April 2011 and in mid-May 2011, respectively. The upper hive body provided last was separated by a queen excluder to ensure honey storage only, and on the same day it was provided the experimental treatment was initiated and lasted until end of June 2011 The exposure started in mid-May 2011. Nevertheless, pollen inflow was limited by installing pollen traps at the hives' entrances. Exposure was achieved by directly inserting pollen patties (2x200 g, 55% honeybee pollen (common stock of commercial pollen with mixed floral content of at least 19 plants; Sonnentracht Imkerei, Bremen, Germany), 5% brewer's yeast and approximately 40% sucrose (two thirds 73% sugar syrup and one third powder sugar)) in the hives. Control hives received plain pollen, while the other received patties containing 5 ppb TMX (analysed 5.31±31 ppb) and 2 ppb CTD (analysed 2.05±1.18 ppb). Pollen patties were provided three times per week, over 46 days. Consumption of the spiked pollen was checked. The authors reported that the two patties were generally consumed within 48 hours.

The study of **Dively et al., 2015,** was considered in the EFSA 2018 reassessment of imidacloprid,(EFSA (European Food Safety Authority), 2018b), see IMI.Colony_feeder, studies I.362 and I.2017. In early April, hives started with ca. 900 g of bees from a commercial supplier and sister queens in new hive boxes and new plastic foundations. For the first four weeks, colonies were fed sucrose syrup to allow colonies to build up before they were assigned to treatment groups. The colonies were equalized for colony strength in early May (brood frames with workers were exchanged among hives) before the start of the exposure, then randomly allocated to groups.

The hive inspections were done biweekly.

Once a week over a 12-week period, four 80 g patties spiked or non-spiked diet patties (pollen diet substitute powder from MegaBee, Dadent &Sons, Inc., Hamilton, IL and sugar solution, at 1.7:1 diet to sucrose solution ratio) were placed within the hives for ad libitum consumption. At each diet placement, remaining portions of the old patties were removed and weighed to keep track of the cumulative weight of diet consumed per colony.

To verify the exposure dose, samples of freshpatties of each treatment and portions opf paties removed after 7 days were collected and analyzed for imidacloprid residues.

The hives were located in Beltsville, MD, USA, in an area massively dominated (3 kms) by corn, soybean and small grains and reported as relatively free from insecticide exposure. None of these crops were treated with imidacloprid, although a portion of the corn acreage was seed-treated at a low rate with other neonicotinoids.

Queen cells when seen were destroyed, missing queens were replaced in the first part of the exposure, but thereafter colonies were left to replace queens naturally. Additional boxes/supers were added in a later stage (as the colony was growing): a second full box in mid-May and a super in mid-June. Pollen trap was fitted to the hives for the exposure period to induce consumption from the in-hive patties. After the exposure period (mid August) each colony was given sucrose syrup to prepare to the winter.

Weekly consumption of diet patties varied significantly over the exposure period but was not different among treatment groups, which ranged from 265.3 to 277.2 g per colony.

In a separate experiment, the fate of imidacloprid in the colony was tracked.

Note: The study was conducted for 2 years (2009 and 2010), with practically the same method, but in both years the experiment started with new colonies (therefore separately assessed by EFSA).

To investigate the turnover of beebread (stored pollen in cells),Roessink and Van der Steen (2021) tracked the presence of beebread cells over 25 days in two honey bee colonies under field conditions in Renkum, The Netherlands. Colonies contained ca. 5,600 bees in a ten-frame hive. Beebread monitoring was performed from 7 June to 28 July 2018 in colony A and from 21 June to 2 August in colony B. Beebread cells in all frames of the colonies were marked on transparent sheets twice per week so that their continuation, disappearance and initiation could be followed. The content of empty but previously filled cells was considered to be consumed in the period between recordings. Almost 75% of the collected pollen was consumed within approximately one week. Almost all pollen (95%) was consumed within two weeks and only a small remainder was stored for a prolonged period. In addition, the number of bees and the number of capped brood cells were recorded over the 25 days. The colonies were either still increasing or stabilizing their total number of bees. The number of capped brood cells varied and was always higher than the number of recorded bees. The authors calculate a pollen consumption of between 8,325 and 11,100 mg pollen/day, which, they state, corresponds with a consumption of 10,375 mg pollen/day for a colony of 5,000 worker bees calculated using the intake values from Rortais et al. (2005).