2.6 Design and analysis of field studies with bees: a critical review of the draft EFSA guidance

Frank Bakker[6]

Independent Ecotoxicologist, Dr Frank Bakker is Managing Director Eurofins/MITOX; Lieu dit Pichoy; 32250 Fourcès, France. E-mail: frank.bakker@mitox.eu

Abstract

The specific protection goal, primary assessment endpoints, acceptable effect thresholds and experimental design proposed in the EFSA update of the bee guidance document are subjected to critical review. It is concluded that the negligible effect criteria were established without sufficient regulatory definition and without convincing scientific argumentation. For the assessment endpoints, effects on hive strength lack temporal definition and the reduction to numbers of bees is inappropriate to evaluate effects. Restricting mortality assessments to homing failure is not theoretically justified and specific criteria were incorrectly derived. The combination of acute effect estimates with models for chronic stressors is biased risk assessment and a temporal basis for the acceptability of effects is missing. Effects on overwintering success cannot be experimentally assessed using the proposed criteria. The experimental methodology proposed is inappropriate and the logistical consequences, in particular those related to replication and land use are such that field studies are no longer a feasible option for the risk assessment. It may be necessary to explore new lines of thought for the set-up of field studies and to clearly separate experimentation from monitoring.

Key-words: honeybee risk assessment, field study, regulatory guidance

1. Introduction

Growers use crop protection products to improve yields. However, these products may adversely affect arthropods, including honeybees. Because declining pollination services may induce food shortages, bee health issues are of public concern and consequently regulatory authorities request bee safety data. Supranational organizations such as EPPO and OECD have developed and adopted guidelines to provide this experimental evidence. In recent years the appearance of neonicotinoid insecticides has been associated with declining honeybee populations. As this implicitly means that current regulation was not sufficiently fit to prevent bee losses, there has been a call for a review of the current regulatory model.

Recently (4 July 2014) the EFSA published a restructured version of their draft guidance document on risk assessment of plant protection products on bees (*Apis mellifera*, *Bombus* spp. and solitary bees) (EFSA, 2013). The document proposes a scheme designed to ensure crop protection will only result in a negligible effect on the ecosystem service of in-field pollination. To meet this protection goal, explicit rules related to study design, test endpoints, data analysis and interpretation for different testing tiers are provided. With this contribution to the symposium, I present a personal biologist's perspective on the proposals for field study design and analysis made in the guidance document. The EFSA makes a clear distinction between the assessment of effects and the assessment of exposure. I restrict myself to discuss effect assessments only.

In relation to protection goals, the guidance document states "the viability of each colony, the pollination services it provides, and its yield of hive products all depend on the colony's strength and, in particular, on the number of individuals it contains". In relation to this, the primary assessment endpoints for field studies have been defined as follows: (1) the magnitude of effects on colonies should not exceed 7% reduction in colony size; (2) foragers mortality should not be increased compared with controls by a factor 1.5 for six days or a factor of 2 for three days or a

[6] This paper is presented on personal title and the views expressed herein are not necessarily the views held by Eurofins.

factor of 3 for two days. For the third primary assessment endpoint; (3) overwintering, no quantitative interpretation criteria are given although the experimental conditions are outlined.

2. A review of the three primary assessment endpoints

2.1 Colony strength

The explicit values for the assessment endpoints colony strength (=number of bees in the colony) and the associated values for forager mortality point at a thorough scientific underpinning. It has been suggested that the 7% cut-off for negligible effect comes from a modelling exercise. However, the draft guidance specifies that the valuation of reductions in colony size of different magnitude was based on 'expert judgment'. Four categories of "detrimental impact" (defined as reduction in colony size) are recognized: large (>35% reduction), medium (15% to 35%), small (7% to 15%) and negligible (3.5% to 7%). The document states that the experts in the EFSA working group unanimously agreed that a reduction in colony size greater than one third should be seen as a large effect compromising viability, pollination services and [honey] yield. The scientific basis of this judgment however, remains unclear. The negligible effect class was apparently derived in the same manner, but in addition by "reference to the potential for experimental detection" (EFSA, 2013). The 3.5% lower limit to negligible effect is puzzling. The intermediate classes were "defined arbitrarily". It is surprising that the setting of cut-off criteria for the single most important assessment endpoint is so poorly documented. Important questions remain to be answered.

What is the permissible time scale over which a reduction in colony size might exceed 7%? Is this daily, weekly, seasonally? Considering the manner in which criteria for forager mortality were derived (see next section), it follows that at no point in time the reduction may exceed 7%. This is an untenable position. Natural fluctuations in hive size exceed this figure by an order of magnitude and e.g. the production of a swarm would be seen as an unacceptable impact on hive strength. Apidologists have documented honeybee dynamics empirically for over a century and this shows that seasonal fluctuations typically range from 4000 to 40000 worker bees (see e.g. Imdorf *et al.* 1987, Harbo, 1986). In fact, the graphs presented by Imdorf *et al.* (1987) show that when measured in 3-weekly intervals most measurements of "colony strength" deviate with more than 7% from the previous one. Such variation is also evident from recent modelling work, such as by Russell *et al.* (2013). Because colonies do not develop in complete synchrony, any colony is therefore likely to differ from another, presumably by at least 7%.

What is meant with "the number of bees", taken as the equivalent of colony strength? When the assessment endpoint concerns adult bees only, as the document suggests, a proper evaluation of colony strength will be impossible. Hive strength is an important endpoint when it comes to evaluating the impact of plant protection products on honeybees. However, defining hive strength solely in terms of the number of adult bees is an unacceptable oversimplification. As such it cannot be used to assess hive strength in a proper manner. A colony with 10.000 worker bees and 30.000 brood cells is stronger than a colony with 10.701 worker bees and no brood cells, yet the difference in the number of adult bees exceeds 7%.

The strength of a hive also implies the potential to respond to adverse conditions or to anticipate these. It is clear that the 7% reduction criterion was established without appropriate scientific rigour and without sufficient regulatory definition as to when, how and for how long it should (not) be observed. The temporal aspect is important because healthy bee colonies exhibit a striking potential to recover from acute catastrophic events. This follows not only from abundant empirical evidence, but also from modelling. For example the model explored by Khoury et al. (2011) shows that under a wide range of conditions bee colonies will return to a stable equilibrium even when catastrophic reductions in numbers of forager or hive bees would occur. Cresswell and Thompson (2012) remark that at stake is not the harm per se but whether exposure is capable of causing colony collapse. From a regulatory point of view colony collapse may not be an appropriate endpoint, but a zero (or 7%) tolerance is equally inappropriate. A *temporal* reduction

in the number of adult bees should not be an issue as long as the hive is sufficiently strong to buffer against such a depression, even if this would imply a short-term reduction of pollination services.

The models underlying some of EFSA's recommendations for effect thresholds are based on chronic stressors. The authors of these models (e.g. Khoury *et al*. 2011, Russell *et al*, 2013), explicitly mention pathogens but not toxicants. The question is whether pesticides should be evaluated as chronic stressors or as acute ones. Whereas agricultural use of several plant protection products sequentially may arguably cause chronic stress, normal product use should rarely result in chronic exposure. This is both for agronomical (product use scenarios) and for biological (learning, information transfer) reasons. Consequently, regulatory and monitoring schemes in the past have focused on incidental rather than chronic exposure. Bee incidents have normally been linked to catastrophic exposure events only. The EFSA guidance proposal however, combines acute exposure events effect measurements with chronic exposure models and this results in unrealistically high estimates of the long term effects of short term exposure.

2.2 Forager bee mortality

EFSA has come to the conclusion that the second primary assessment endpoint should be the mortality of *forager* bees. The document explicitly earmarks e.g. dead bee traps as an inappropriate tool because these also measure other sources of mortality. The line of thought that forager mortality should be measured exclusively in order to protect honeybee colonies from being reduced by more than 7% is not straightforward, but the document provides some insight into the sources of this thinking. A model published by Khoury *et al.* (2011) and in particular the use of that model fed with empirical data on homing behavior (Henry *et al.* 2012a) have been the principal sources of information for this regulatory reasoning.

Indeed, modelling work that assumes a division of tasks (nursing vs foraging) by social inhibition, such as the analytical compartment model by Khoury *et al.* (2011, 2013) and the time-based simulation model by Russell et al. (2013) demonstrates how sustained increases in forager mortality, caused by *chronic* stressors, may cause lower equilibrium densities, reduced food supplies or eventually lead to colony collapse. The mechanism is fairly intuitive. Linked by social inhibition, forager death drains the nurse bee population, which results in a decline in brood rearing and, as forager bees typically have higher mortality rates increases in overall bee mortality. The process is reinforced by a reduction in food supply. With their model Khoury *et al.* (2011) sought to explore the effect of varying forager death rates on hive dynamics and for this purpose they studied this parameter in isolation of other possible factors influencing hive dynamics. Although the authors made an attempt to parameterize their model within realistic ranges, obviously the resulting output was not intended to be taken as representative for the dynamics of real hives. The important finding is that there may exist a threshold to chronic forager bee mortality, such as may result from pathogens, above which colonies will collapse. However, equally important is the finding that with chronic forager mortality rates below the critical value, the social inhibition mechanism provides a buffer that helps the hive to return to an equilibrium with a constant ratio of hive to forager bees even when catastrophic events occur. Khoury *et al.* (2011) show that as long as forager bee mortality remains chronically below the threshold, the trajectories will always lead back to the non-zero equilibrium, even when near to 100% of the forager bees would disappear as a consequence of catastrophe.

From their modelling Khoury *et al.* (2011) conclude that chronic stressors that reduce forager survival by approximately two thirds, i.e. a reduction in the life span from 6.5 to 2.8 days, will place a colony at risk if the colony does not respond, e.g. by adapting recruitment rates. However, whereas Khoury *et al.* (2011) define risk in terms of probable colony extinction, the EFSA work with the specific protection goal (SPG) of a maximum of 7% reduction in colony strength. Using Khoury *et al*'s (2011) model they calculate the time period during which the colony can sustain a certain forager mortality rate before reaching a size less than 93% of the pre-exposure situation. This

exercise resulted in very specific requirements, viz. an increase of forager mortality by a factor 1.5 can be tolerated for six days (average over six days), a factor 2 for three days, a factor 3 for 2 days.

There are several problems associated with this approach.

(1) The mortality increase is defined in terms of multiples of the background mortality, whereas there is no guidance as to what the background mortality can be. Obviously, doubling a background mortality of 0.355/day does not have the same consequences as doubling a background mortality of 0.035/day. In the first case the colony will go extinct, whereas in the second no effect will be observed.

(2) Khoury *et al.*'s (2011) differential equations model, and consequently EFSA's use of it, uses constant rates for the driving parameters and excludes variation in food supply (but see Khoury *et al.* 2013) and seasonally fluctuating parameter values. Russell *et al.*'s (2013) difference equations model does capture this variability and their exercise clearly shows the differential sensitivity of the various parameters recognized in the model for (seasonal) changes in value. The later the forager death rate starts to rise in spring, the better for colony survival. The EFSA does recognize that seasonal variation may have to be taken into account, but this recognition is restricted to stating that model parameters can be calibrated for spring colonies by using data from Henry *et al.* (2012a) and for autumn colonies by using data from Cresswell and Thompson (2012) and Henry *et al.* (2012b). Given the outcome of the various model analyses one would expect the tolerance for adverse effects to be related to hive developmental status. Effects in March will impact a colony in a different manner than effects incurred in September.

(3) The compartment models on which the arguments are based are extremely sensitive to changes in the inhibition factor (σ) that determines transition rate of hive bees to forager bees. Russell *et al.* (2013) show that any agent that alters this rate could have an enormous impact on the development of a colony. However, actual values of σ cannot be measured and potential effects of chemical stressors on σ remain unknown. It is clear from field studies with e.g. dimethoate that bees may completely stop foraging for prolonged periods until relocated to a site without exposure (pers. obs.). Under these circumstances an effect on σ could be expected.

(4) Stressors such as crop protection products will rarely have chronic and constant effects. With the suggested guidance it will be difficult to evaluate the acceptability of gradually decreasing mortality (e.g. 30% on day 1, 15% on day 2, 3% on day 3 etc.).

(5) The suggested cut-off values are based on the definition of negligible effect being equal to 7% reduction in colony size and a hive size of 5000 bees as the minimum size fit for overwintering. Both criteria require a more rigorous scientific evaluation before implementation into legal guidance.

The compartment models explored by Khoury et al. (2011, 2013) and Russell et al. (2013) underscore the importance of assessing forager bee mortality. In fact, Russell et al. (2013) show that hive vigour is much less affected by mortality of hive bees than by mortality of forager bees. Forager bees may be lost from the hive population due to mortality, but also due to sub-lethal causes. The work by Henry et al. (2012a) show that exposure to sub-lethal doses of pesticides may affect cognitive capacities to the extent that foragers fail to return to the hive. Although the study has been criticized for certain aspects of the experimental design (Cresswell and Thompson, 2012; Guez, 2013a,b), the potential for sub-lethal effects leading to a drain on the forager population was well demonstrated. Homing failure may result from cognitive dysfunctions at sublethal doses, but also from mortality. In a field study with honeybees it will be impossible to distinguish between these two sources of forager bee loss, even with RFID-techniques. A third cause for homing failure may be altruistic self-removal (Rueppel et al. 2010), i.e. a situation where foragers decide not to return to the hive to avoid contaminating their kin. This issue is, of course, highly academic. The important point is that a certain proportion of the forager bee population may disappear or die away from the hive and as long as this proportion is unknown it remains important to provide an accurate estimate. Homing failure cannot be assessed with classical tools

such as dead bee traps, but RFID-techniques (see Henry et al. 2012) seem to provide a good solution to this problem. As with any capture-mark-recapture method there are *caveats* associated with the method, certainly because the size of the forager population is dynamic in honeybees. Validation such as undertaken by the CEB (Biological Tests Commission (Commission des Essais Biologiques), of the French Plant Protection Association (AFPP - Association Française de Protection des Plantes) is therefore important.

However, forager bees not only die during foraging bouts, but also upon return to the hive. It is important to include this mortality in the evaluation. For this purpose dead bee traps are excellent tools. In this regard it is surprising that EFSA considers dead bee traps as "not totally appropriate, as they tend to measure dead bees at the colony (colony bees) and not foragers in the field". Moreover, in the absence of pathogens, the mortality of hive bees is known to be low (Harbo 1993b), which would imply that most dead bees in a dead bee trap will be forager bees. If, however, exposure to pesticides would lead to an increase in hive bee mortality, e.g. by cross-contamination it becomes important to also assess this source of mortality (Brown 2013, Russell et al., 2013) as it will contribute to colony failure. The recommendation should therefore be to assess bee mortality both inside the hive and away from it, rather than homing failure alone.

2.3 Overwintering success

The guidance document does not dwell in depth on the primary assessment endpoint of overwintering success. It is clear that the guidance does not imply the proportion of hives that survive the winter, but rather the relative condition of the hives in spring. A practical recommendation in the guidance is that hive monitoring should be maintained for a time after the wintering period and that the colonies own honey should be used to sustain the colony during winter, the idea being that the colony will then consume potentially contaminated honey and pollen during the initial start-up phase in spring. Hive strength may be assessed using the methods of Costa *et al.* 2012, but other than that there is not much guidance on experimental design.

The main statement is that overwintering success should be assessed by comparing the colony strength of the treatment colonies with the control colonies and that there should not be a significant difference between the control and the treatment. The 7% is not explicitly mentioned in this context, but as the assessment is said to be linked to the Specific Protection Goal, which is a negligible effect (= <7%) on colony strength, this may be assumed. How likely is it that a study may have sufficient experimental power to detect small effects on overwintering success with less than 5% risk on a false positive result? A great many factors determine the overwintering success of a colony and many interactions of these are not yet understood or even recognized. It is well known that among beekeepers enormous variability is observed, but the relative importance of the various factors involved in causing this variability remains unknown. Against such a noisy background it will not be possible to design a study where differences in overwintering success (measured as hive strength) as low as 7% can be attributed to treatment with an 80% confidence in the correct conclusion. In my opinion effects of exposure to a certain crop protection product on overwintering success cannot be experimentally assessed in a reliable manner. However, monitoring studies may, or rather should be invoked to assess overwintering success, which is after all, key to bee health.

3. A review of the proposals for the experimental design of bee field studies

Whereas the primary assessment endpoints, in particular hive strength and overwintering success, for field studies are in need of more specific definition, the guidance document is rather explicit about the appropriate study design to assess these endpoints. In the following section I review the design proposals.

3.1 Choice of crop

Two different recommendations are found in the EFSA guidance when it comes to the choice of test crop, a flexible one and a strict one. In Appendix O (Effects studies—protocols, guidance and guidelines for honey bee, bumble bee and solitary bee), the choice of crop that can be used is left open. It may be the proposed crop for the test item, but it may also be possible to use a highly attractive model plant (e.g. *Phacelia tanacetifolia* or oilseed rape) and to extrapolate the study findings to a range of crops. As EFSA specifies: the key issue in selecting a suitable crop is to ensure that it is attractive to honey bees and that the residues, and hence the exposure to honey bees, is environmentally relevant and at least as high as predicted in the exposure section. This flexible recommendation is in line with EPPO 170 guidance. According to EPPO 170 (4), for testing of effects on honey bees following spray applications, in the first instance, rape, mustard, *Phacelia* or another crop highly attractive to bees should be used as test plants, e.g. in the case of a standard semi-field or field trial based on acute toxicity.

However, in Appendix R (Test crops to be used), the EFSA Working Group cites this text from EPPO 170, but recommends that *Phacelia* be used in semi-field and field tests and a number of reasons are specified. Among these are biological reasons, in particular the attractiveness of *Phacelia* for honeybees and the open architecture of the flowers resulting in worst case exposure of nectaries and anthers. A comment here is that although *Phacelia* is indeed an attractive crop, it is certainly not by far the most attractive crop. Indeed rape and mustard may show similar densities of forager bees. In addition, the nectaries of *Phacelia* are not more exposed than e.g. the nectaries of mustard. In this respect there seems to be no real justification for strict crop recommendation. A number of practical advantages are also given, some of which relate to a presumed flexible sowing date. However, from an agronomical perspective *Phacelia* crop may not always be the best choice for all soils and there may be a substantial risk of crop failure when sown at the wrong time of the year.

My recommendation would be to abandon Appendix R and leave the text in Appendix O in as far as this concerns the choice of crop.

3.2 Field size and replication

According to the EFSA guidance the choice of field sites must ensure that no cross foraging will occur and bee attractive crops in the surroundings should be sparse to constrain the foraging to the test fields. Therefore, it is proposed to choose areas presenting similar environmental conditions, where possible at least four kilometers apart. To ensure appropriate exposure the recommendation is to have sites of at least 2 ha flowering crop. In practice this implies that each field must be surrounded by a radius of 2 km to ensure sites are 4 km apart. A circle with a diameter of 4 km is equivalent to a surface of 1256 ha and inside this area no other flowering crop may be found. For many landscapes this will be utopic for most of the year and certainly in periods when bees are active.

In accordance with standard experimental practice the EFSA guidance document specifies that the experiment should be such that a 7% detrimental effect should be detected with a power of 80% and a risk of accepting a false positive result of 5% or less. On this basis and with some assumptions concerning within and between field variability in hive strength, a calculation example is then provided. The result of the calculation is that it should theoretically be possible to detect a 7% effect on hive strength in a field study that has 28 field sites with 7 colonies in each field (14 control and 14 treated replicates), i.e. a total of 196 colonies. Alternatively, in a set-up with 1 colony per field a total of 120 fields would be needed. In terms of surface this implies that a field study must be performed over a total surface of 28*1256=35168 ha or with one colony per field 120x1256 =150720 ha (1507 km^2).

The variability assumptions are rather restrictive (Coefficient of Variation (CV) =15% between hives and CV=5% between fields) and will in reality often be exceeded. This implies there is high risk

that even after setting up a study with 196 hives in 28 locations one may end up with a study of insufficient power to detect an effect of 7%. With this replication, practical feasibility is an issue both for the assessments and the experimental treatment applications as the logistics will get overly cumbersome quickly. Thus, even under unrealistically low assumptions of variability the feasibility of performing a field study that can correctly identify an effect of 7% is minimal and the associated costs will be enormous. Is there a way to reduce the logistic effort to a practicable minimum? The most straightforward solution would be to abandon the 7% threshold. As argued before, there is no solid scientific basis underlying the 7% criterion and given the practical consequences it is stunning that the European Food and Safety Authority is requesting such amazing investments without convincing theoretical justification.

Intuitively more replicates means more precision and this also follows from the EFSA calculations. However, for field studies with bees this may not be necessarily true. Increasing the number of test fields will not necessarily reduce or cancel out noise, but may actually introduce bias. This is a consequence of the enormous surface over which the studies must be laid out. The example above with 28 test fields has a surface of >35000 ha. In this desert of non-flowering crop would be the 28 oases of 2 ha flowering *Phacelia*. Obviously there is an enormous risk for spatial inhomogeneity or gradients over such a vast area, such that these fields may differ in several important respects and it is not certain that these will average out by randomly assigning treatments to fields. Thus, from a practical point of view replication by fields is not an easy solution. The question is whether it is a prerequisite for statistically sound study design and this revolves around the issue of pseudo-replication.

According to most textbooks on statistics a *replication* of a treatment is an independent observation of the [effect of] treatment and thus *n* replications must involve *n* experimental units. The conceptual difference underlying different opinions on (pseudo-)replication in bee studies is related to the interpretation of what constitutes a unit. Although the test item is physically applied to a field, the field is not necessarily the unit of observation. In this context the field is the treatment unit and the hive is the independent observation unit to be replicated. (Note that to ensure this independence hives should either not be populated with sister queens, or it should be done only in treatment-control pairs). Obviously fields may differ in crop attributes, but as long as exposure is quantified (e.g. by assessing levels of foraging, nectar and pollen uptake) we can assert with a high level of confidence whether fields were sufficiently comparable for the purpose of the study. Increasing the number of fields will not necessarily reduce noise or 'cancel out' random factors.

Even if the debate on study design would lead to the conclusion that multiple hives in a field are pseudo-replicates, what would be the consequence; that we can no longer do field studies? The consequence of potential interdependence of the replicates may lead to an underestimation of the error variance, which may affect the risk of committing type I errors. However, this does not make the analysis faulty *per se* and indeed one may question the relevance of type I errors in a regulatory context. A consequence of replicating hives but not fields is that if fields have an important contribution on observed effect levels, conclusions drawn from the analysis of replicate hives set up in e.g. two fields only may be restricted to fields of these particular conditions. In other words the results may not be general. This is the true risk of pseudo-replication. But how bad is that in the general context of regulatory ecotoxicology? In a wide range of highly comparable study designs we are comfortable with the use of statistics to analyze data from (pseudo-)replicated designs because we believe it helps to have a formal analysis that quantifies the risk of finding a false positive result, while at the same time we accept the 'pseudo' aspect of the study and assume that the interdependence of treatment replicates most likely has no significant bearing on the toxicological effects we are interested in.

3.3 Duration of a study

Because overwintering is one of the primary assessment endpoints, the study duration will extend into the next season and, by recommendation of EFSA, it should extend at least into next spring to ensure that any possibly contaminated storage has been consumed. In itself that is a valid request, however, as discussed above, the question is whether an experiment can be designed that measures differences in overwintering success in terms of hive strength and with a precision of 7%. My personal opinion is that it is not and consequently that studies can be of shorter duration. As outlined in the guidance, post-exposure monitoring also comes with logistic requirements (hives at the same location in an area far from fields or potentially treated crops) and with the replication mentioned earlier this may be incompatible with good beekeeping. For measurement of the other endpoints a period of two brood cycles is recommended and this is in line with current practice.

3.4 Colony condition

The straightforward recommendation in the guidance is that at the start of an experiment colonies must be in the same 'state', specified as genetic origin, population size and health status and implying equal 'strength'. In addition to this come recommendations concerning season-specific colony composition and size and last but not least the use of sister queens to reduce genetic variation. The use of sister queens has a long tradition in apidological research (see e.g. Harbo, 1986) but the question is whether it should be recommended in regulatory studies. The purpose of regulatory work is to assess the risk for honeybees in general and not just for a specific genotype. In this respect sister queens are a perfect example of pseudo-replication (see above). In addition a large contribution of genetic background to experimental variability is often assumed, but in reality it remains largely unknown how much the queen's genotype contributes to the variability in the primary assessment endpoints in a variable environment.

4. Alternative ideas for field experimental design

Performing ecotoxicological field studies with honey bees using hives that resemble commercial hives comes with many pitfalls and *caveats* as will be clear already from this limited review of the proposed EFSA guidance. One of the main obstacles to solid experimental design is the inherent variability in hive development and the myriad of interdependent and mostly uncontrollable factors interacting with the colony. It is an illusion to think that longer term studies, such as those addressing overwintering success, can be designed such that the impact of a single stressor (the test item) can be singled out and tested for in an experimental study that involves commercial hives. The debate on CCD illustrates this point well. In this respect it is important to realize that the field study is the final step in a series of experiments designed to demonstrate the *potential* impact of a test item on honeybee health (or in EFSA terms, the ecosystem service of in-field pollination). It is therefore by nature an experimentation and not a monitoring exercise. Monitoring can be seen as an exercise to validate the predictions of the sequential testing scheme, including the higher tier studies, or rather the regulatory decisions that were made on basis of the results obtained therein. Thus, whereas monitoring must involve commercial hives, this does not necessarily apply to experimentation.

What distinguishes a field study from studies at lower testing tiers is the freedom of choice given to the bees when it comes to foraging decisions. As a consequence the colony will be better able to tune its development to available forage and to intra-colony conditions. Thus, the field study allows for an assessment of colony development parameters such as egg laying and brood development under conditions that are in principle less restrictive than conditions in e.g. tunnel studies. However, this does not imply that the test hive should also mimic realistic conditions. In fact hives can be prepared and tuned to specific experimental purposes. In this respect there is a lot to learn from John Harbo, who achieved a high degree of standardization of experimental hives

and with his set-up managed to assess basic biological parameters governing hive development with high precision.

It is beyond the scope of this paper and beyond my individual capacity to provide a full alternative to the proposals under review, but some initial ideas or starting points may be worth mentioning. My proposal for a field study design would involve the 'artificial or shook swarm' technique, following the recommendations as described by Harbo (1986), and an assessment of basic parameters under specific experimental conditions, using the methodology as described by Delaplane et al. (2013), which can be found in the excellent COLOSS Beebook. There is discussion whether the shook swarm method is appropriate for early spring conditions, but it seems definitely a good option for summer (Pistorius, pers. comm.). Previous work by Bakker and Calis (2003) that involved hive preparation by the shook swarm method in combination with age-controlled brood provisioning showed that at least under semi-field conditions mini-hives prepared in this manner provided for consistent and statistically powerful assessments of mortality and foraging parameters and, to a lesser extent, on hive weight even with only four mini-hives per treatment.

In designing a field study it should also be realized that an experiment does not necessarily have to combine assessments of all parameters of interest. Studies could be separated, e.g. a study for effects on egg laying and egg survival (*cf* Harbo 1982, 1985), a study for effects on brood development, a study for effects on general hive maintenance such as food storage, cell cleaning etc. The relative importance of effects on these parameters could then be assessed using a simulation model such as the one described by Russell et al. (2013). Obviously, a field study is also the ideal setting for an assessment of forager mortality. However, the emphasis on non-returning foragers is not justified. What will drive the dynamics in the hive is the total number of bees dying as a consequence of exposure. In addition to dead bee traps at the hive entrance, and in addition to the RFID-technique discussed above, several new sophisticated and sensitive methods are available to monitor numerical changes inside the hive continuously and precisely (see e.g. the presentation of Sandra Evans in this symposium). This should be the way forward.

5. Acknowledgements

The symposium was a stimulating event and I thank organizers and participants for their great input. Special thanks to Saskia Aldershof, Pieter Oomen and Jens Pistorius for their constructive comments on the manuscript.

6. References

Bakker, F. and Calis, J. 2003. A semi-field approach to testing effects of fresh or aged pesticide residues on bees in multiple-rate test designs. Bull. Insectol. 56 (1): 97-102

Brown, K.M. 2013. Mathematical models of honey bee populations: Rapid Population Decline. Thesis Univ. of Mary Washington, Fredericksburg, Virginia, 20 pp. kellybrown.umwblogs.org/files/2013/04/KellyBrownHonors.pdf

Costa C, Buchler R, Berg S, Bienkowska M, Bouga M, Bubalo D, Charistos L, Le Conte Y, Drazic M, Dyrba W, Fillipi J, Hatjina F, Ivanova E, Kezic N, Kiprijanovska H, Kokinis M, Korpela S, Kryger P, Lodesani M, Meixner M, Panasiuk B, Pechhacker H, Petrov P, Oliveri E, Ruottinen L, Uzunov A, Vaccari G and Wilde J, 2012. A Europe-Wide Experiment for Assessing the Impact of Genotype-Environment Interactions on the Vitality and Performance of Honey Bee Colonies: Experimental Design and Trait Evaluation. Journal of Apicultural Science, 56, 147-158.

Cresswell JE and Thompson HM, 2012. Comment on "A Common Pesticide Decreases Foraging Success and Survival in Honey Bees". Science, 337.

Delaplane, K.S., van der Steen, J., Guzman, E. 2013. Standard methods for estimating strength parameters of *Apis mellifera* colonies. *In V Dietemann; J D Ellis; P Neumann (Eds) The COLOSS BEEBOOK, Volume I: standard methods for* Apis mellifera *research.* Journal of Apicultural Research 52(1): http://dx.doi.org/10.3896/IBRA.1.52.1.03

European Food Safety Authority, 2013. EFSA Guidance Document on the risk assessment of plant protection products on bees (*Apis mellifera*, *Bombus* spp. and solitary bees). EFSA Journal 2013;11(7):3295, 268 pp., doi:10.2903/j.efsa.2013.3295 NOTE: here the update published 04 July 2014 was used.

EPPO (European and Mediterranean Plant Protection Organization), 2010. PP 1/170 (4): Side-effects on honeybees. 40, 313-319.

Guez, D. 2013a. A common pesticide decreases foraging success and survival in honey bees: questioning the ecological relevance. Front Physiol. 2013; 4: 37

Guez, D. 2013b. Henry et al. (2012) homing failure formula, assumptions, and basic mathematics: a comment. Front Physiol. 2013; 4: 142.

Harbo, J.R. 1986. Effect of population size on brood production, worker survival and honey gain in colonies of honeybees. J. Apic. Res. 25 (1): 22-29.

Harbo, J.R. 1988. Effect of Comb Size on Population Growth of Honey Bee (Hymenoptera: Apidae) Colonies. J. Econ. Entomol. 81(6): 1606-1610.

Harbo, J.R. 1993a. Worker-Bee Crowding Affects Brood Production, Honey Production, and Longevity of Honey Bees (Hymenoptera: Apidae). J. Econ. Entomol. 86(6): 1672-1678.

Harbo J.R. 1993b. Effect of brood rearing on honey consumption and the survival of worker honey bees. J. apic. Res. 32(1): 11-17.

Henry M, Beguin M, Requier F, Rollin O, Odoux JF, Aupinel P, Aptel J, Tchamitchian S and Decourtye A, 2012a. A Common Pesticide Decreases Foraging Success and Survival in Honey Bees. Science, 336: 348-350.

Henry M, Beguin M, Requier F, Rollin O, Odoux JF, Aupinel P, Aptel J, Tchamitchian S and Decourtye A, 2012b. Response to Comment on "A Common Pesticide Decreases Foraging Success and Survival in Honey Bees". Science, 337.

Henry, M., Decourtye, A. 2013. Ecological relevance in honeybee pesticide risk assessment: developing context-dependent scenarios to manage uncertainty. Front Physiol. 2013; 4: 62.

Henry, M. 2013. Assessing homing failure in honeybees exposed to pesticides: Guez's (2013) criticism illustrates pitfalls and challenges. Front Physiol. 2013; 4: 352.

Imdorf, A.; Buehlman, G.; Gerig, L.; Kilchenmann, V., Wille, H. 1987. Ueberpruefung der Schaetzmethode zur ermittlung der brutflaeche und der anzahl arbeiterinnen in Freifliegenden Bienenvoelkern. [*A Test of the Method of Estimation of Brood Areas and Number of Worker Bees in Free-Flying Colonies*]. Apidology **18**, p 137-146.

Khoury DS, Myerscough MR and Barron AB, 2011. A Quantitative Model of Honey Bee Colony Population Dynamics. PloS one, 6, e18491.

Khoury DS, Barron AB, Myerscough MR (2013) Modelling Food and Population Dynamics in Honey Bee Colonies. PLoS ONE 8(5): e59084. doi:10.1371/

journal.pone.0059084

Rueppell, O., Hayworth, M.K., Ross, N.P. 2010. Altruistic self-removal of health compromised honey bee workers from their hive. J. Evol. Biol. **23**: 1538-1546. Doi:10.1111/j. 1420-9101.2010.02022.x

Russell, S., Barron, A.B., Harris, D. 2013. Dynamic modelling of honey bee (*Apis mellifera*) colony growth and failure. Ecological Modelling 265: 158-169. http://dx.doi.org/10.1016/j.ecolmodel.2013.06.005Szabo, T.L., Lefkovitch, L.P. 1989. Effect of brood production and population size on honey production of honeybee colonies in Alberta, Canada. Apidology **20**: 157-163.