Article (refereed) - postprint

Harris, Catherine A.; Scott, Alexander P.; Johnson, Andrew C.; Panter, Grace H.; Sheahan, Dave; Roberts, Mike; Sumpter, John P.. 2014. **Principles of sound ecotoxicology**. *Environmental Science & Technology*, 48 (6). 3100-3111. 10.1021/es4047507

Principles of Sound Ecotoxicology

Catherine A. Harris [*1], Alexander P. Scott [2], Andrew C. Johnson [3], Grace H. Panter [4], Dave Sheahan [5], Mike Roberts [6], John P. Sumpter [1]

[1] Institute for the Environment, Brunel University; [2] Cefas (Weymouth Laboratory); [3] CEH, Wallingford; [4] Astra Zeneca (Brixham Environmental Laboratory); [5] Cefas (Lowestoft Laboratory); [6] Defra (Chemicals and Emerging Technologies Division)

[*] Corresponding author:
Institute for the Environment, Brunel University, Uxbridge, Middlesex, UB8 3PH, UK.
Tel: +44 1895 266267
Fax: +44 1895 269761
E-mail: Catherine.Harris@brunel.ac.uk

Abstract

We have become progressively more concerned about the quality of some published ecotoxicology research. Others have also expressed concern. It is not uncommon for basic, but extremely important, factors to apparently be ignored. For example, exposure concentrations in laboratory experiments are sometimes not measured, and hence there is no evidence that the test organisms were actually exposed to the test substance, let alone at the stated concentrations. To try to improve the quality of ecotoxicology research, we suggest twelve basic principles that should be considered, not at the point of publication of the results, but during the experimental design. These principles range from carefully considering essential aspects of experimental design through to accurately defining the exposure, as well as unbiased analysis and reporting of the results. Although not all principles will apply to all studies, we offer these principles in the hope that they will improve the quality of the science that is available to regulators. Science is an evidence-based discipline and it is important that we and the regulators can trust the evidence presented to us. Significant resources often have to be devoted to refuting the results of poor research when those resources could be utilised more effectively.

Introduction

We, and others, have become increasingly concerned that the quality of a significant proportion of ecotoxicological research is not as high as it could, and probably should, be.[1-4] It is very common nowadays for us to read a scientific article published in a reputable journal and end up thinking "this effect of substance X is surprising", or even "I find it very difficult to believe that substance X really does cause those effects at those concentrations". Other scientists have also indicated that they have difficulty deciding what ecotoxicological research is sound and what is not.[5,6] Indeed, some have already published papers suggesting improvements that could be made to ecotoxicological research.[7]

We are not the first people to express concern about the quality of published research, either in our field (ecotoxicology) or any other field. For decades (and possibly hundreds of years), scientists have questioned the merits or demerits of particular pieces of research. Nearly half a century ago, an eminent physician, who was interested in possible links between the incidence of various diseases in people and their exposure to industrial chemicals in their working environments, published a set of criteria that he suggested should be used to support, or refute, reported associations between conditions in the workplace (e.g. exposure to industrial chemicals) and particular diseases.[8] In other words, he was interested in assessing the quality of research that purported to link substance X with adverse effect Y. More recently, various toxicologists and ecotoxicologists have published updated sets of criteria for quality assessment of published (eco)toxicological studies,[9-12] mainly as a prerequisite to determining what weight can be placed on a study before it can be used for environmental risk assessment.[5] The outcomes of these assessments do not inspire much confidence in the existing literature: many influential studies are rated as 'not reliable' or 'unacceptable';[6] at least one scientist has gone so far as to suggest that 'most published research findings are false'.[13]

We have no desire to undermine ecotoxicology; on the contrary, our desire is to improve it. We accept that some ecotoxicology, especially fieldwork, can be extremely difficult, if not impossible, to conduct in an ideal way. How, for example, does one obtain a clear-cut answer to a question such as 'are perfluorochemicals adversely affecting albatrosses?' in order to determine whether their documented exposure to these extremely persistent pollutants [14] is, or is not, of concern? Nevertheless we still consider that much ecotoxicology, including laboratory-based studies, is not being conducted (or interpreted) as well as it could be. In order to try and improve the situation in the future, we list, then briefly expand upon, the factors we consider

most important to defining the quality and usefulness of ecotoxicological research studies. We present a set of principles which, if adhered to, would improve considerably the quality of such research (see Table 1). Our approach is based on the very successful establishment of the principles of Green Chemistry.[15,16] However, whereas those principles were intended to accomplish the goals of green design and sustainability, ours are perhaps somewhat less ambitious and more practical, and specifically aimed at improving the quality of ecotoxicological research. Our principles also address the issue of reporting results in a balanced manner that reflects the results obtained.

Discussing the principles of sound ecotoxicology has necessitated mentioning some examples of what we consider poor ecotoxicology. We have attempted not to be unfair to any individual, or to any particular issue in ecotoxicology, and have tried to provide balance in this article by also mentioning examples of what we consider are good ecotoxicological studies. Most of our examples are in the field of aquatic ecotoxicology and, in particular, endocrine disruption, because this is our area of expertise, but we believe the principles outlined here are relevant to ecotoxicology as a whole.

The Principles of Sound Ecotoxicology

1. ***Adequate planning and good design of a study are essential***
   If the planning stages are not thought through adequately, an entire study could be wasted.

2. ***Define the baseline***
   When any endpoint is assessed, the 'normal' level of that endpoint in an unexposed organism should be established.

3. ***Include appropriate controls***
   Solvent controls and positive controls should be used where possible/appropriate. The number of controls should also be considered.

4. ***Use appropriate exposure routes and concentrations***
   Ensure that the route of exposure is appropriate (e.g. via water or via food) and that the concentrations applied are discussed within the context of concentrations measured in the environment.

5. ***Define the exposure***
   It is important to measure actual concentrations of the substance/s used, so that the real exposure scenario can be described, rather than a hypothetical one. Further, exposure media should be assessed for common contaminants.

6. ***Understand your tools***
   Knowledge of the particular test organism and test substance used are vital to generating reproducible results.

7. ***Think about statistical analysis of the results when designing an experiment***
   The number of exposure concentrations, as well as of target organisms, needs to be carefully considered prior to starting the experiment, in order that the results have sufficient statistical power to provide an answer to the hypothesis being tested.

8. ***Consider the dose-response***
   Consider the dose-response; any 'unusual' pattern of response needs further analysis and justification.

9.	***Repeat the experiment*** This may not be necessary where results are striking and statistical power is strong. However, in general, and particularly where results are unexpected and/or borderline, results must be shown to be repeatable.
10.	***Consider confounding factors*** Factors such as temperature, disease, and exposure to multiple substances should be taken into consideration; these may be especially relevant in fieldwork.
11.	***Consider the weight of evidence*** Results should be compared with previous studies, e.g. do fieldwork and laboratory studies support each other? Do the effects fit with known mechanism of action of the respective substance/s? Consider the plausibility of the results.
12.	***Report findings in an unbiased manner*** Do not over-extrapolate (e.g. from *in vitro* to *in vivo*); be aware of the limitations of the study; don't over-hype a result with very low significance; report negative (i.e. no effect) as well as positive findings.

Table 1. A summary of the Principles of Sound Ecotoxicology

Principle 1: Adequate planning and good design of a study are essential

It can often be the case that studies are undertaken in a hurry without sufficient forethought of the several critical factors involved. The first stage is to define the aim of the experiment. For example, is the aim to define the Lowest Observable Effect Concentration (LOEC) of a particular substance, or is it to establish whether effects might be seen at very low concentrations (equivalent to those seen in the natural environment)? Once the aim is agreed, a great deal of effort needs to go into planning the details of the study. Factors to be considered include: how many substances to investigate in any one study; how many exposure concentrations to use (and how far apart these should be spaced); how many replicates (e.g. tanks in the case of fish) of each concentration; how many subjects (e.g. fish per tank); the physico-chemical properties of the substance to be tested; whether the use of an organic solvent can be avoided; when to sample for chemical analysis; how many endpoints to assess (and which are the most relevant for the substance concerned). In addition, experimental planning needs to incorporate steps that can be taken to avoid operator bias, such as random allocation of animals between treatments and blinded analysis of samples where possible. Trying to achieve too much from a study can be as detrimental to the quality of the results as trying to do too little; a balance must be struck. The planning of a good ecotoxicological study can in some circumstances take longer than the exposure study itself.

Another factor which should be considered at this point is that adequate recording and documentation, not just of the outcome but of all the procedures undertaken along the way, are essential. If any queries arise during or after an experiment, researchers must be able to back up every step of their working, in order to be able to defend and, if necessary, correct what they have done. Furthermore, adequate information should be provided to enable others to repeat the study in full. We would not go so far as to say that all laboratories should follow 'Good Laboratory Practice' (GLP) guidelines, although we could learn much from these principles. Instead, we consider it sufficient to work to the spirit of GLP. For example, researchers should be prepared to share their raw data, (perhaps through a link to an appropriate database if the files are too large to be included as Supplementary Information), in addition to retaining a full report of how the study was designed, conducted and analysed in order to allow adequate interpretation of the results. Such steps were amongst those described in a recent editorial announcement in Nature,[3] a journal which is also recognising the problems faced by the recent spate of publications of unreliable data, and now, along with many other journals, stipulates that

scientists should deposit large datasets in an approved database prior to publication of the manuscript.

We cannot stress enough that good planning and management of an ecotoxicological study is vital for a successful outcome.

Principle 2: Define the baseline

To discriminate between exposed and unexposed test organisms, toxicological studies usually measure one or more biomarker or sublethal effect that occurs in response to substance exposure. Studies may include organisms sampled from wild populations or, in a laboratory context, the use of standard test species. Whatever the origin of the animals, it is important to characterise the natural variability in parameters or endpoints that form the basis of the investigation (in order to be able to design experiments that are sufficiently sensitive to discriminate real effects). In mammalian multi-generation studies, inter-laboratory variability in negative control data has been intensively studied to improve the sensitivity of the test methods.[17] Several fish species have also been the subject of detailed study to ensure that the experimental design is matched to the reproductive biology of the species used (e.g. the fathead minnow [*Pimephales promelas*] and zebrafish [*Danio rerio*]).[18,19] Essentially, if one of the endpoints in an exposure study is, for example, a plasma hormone concentration, it is important to know what are the 'normal' changes that occur within and between individuals over time. Plasma concentrations of sex steroid hormones in particular are highly dependent on the state of maturity of the gonads and thus show strong seasonal fluctuations. These need to be taken into account when planning and subsequently interpreting studies.

Another problematic area in the study of chemical effects on reproductive biology is sex differentiation. It is essential that there is a good understanding of the most sensitive period for this parameter for the species being studied, so that this can be taken account of in the experimental design, and exposure can be focussed on key windows – although it is acknowledged that there are in the zebrafish, in particular, contrasting views on the exact timing of sex differentiation, as discussed by Segner.[20]

Studies on the reproduction of molluscs have raised a number of disagreements with, for example, respect to whether or not substances such as Bisphenol-A (BPA) at very low concentrations increase fecundity in the ramshorn snail (*Marisa cornuarietis*).[21] Benstead *et al*

demonstrated the importance of establishing baseline seasonal fecundity patterns before investigating the effects of endocrine disrupting compounds on gastropod molluscs. [22] In that study, a clear correlation between number of eggs laid and photoperiod was established in the reference group, with a subsequent steep decline in egg production following the summer solstice. Although the effect reported in that paper (an extended reproductive season in snails exposed to 17β-estradiol [E2]) was observed to be a trend rather than being significantly different from the reference group at any one timepoint, the establishment of the baseline reproductive performance pattern of these snails is clearly important in determining whether or not estrogenic substances can impact on snail reproduction, and will provide useful background information for future research in this field.

Ultimately, the environmental significance of the results of a study can be better interpreted when there is a good understanding of baseline conditions.

Principle 3: Include appropriate controls

In theory, this is a relatively easy objective to achieve, at least in terms of laboratory exposure studies (perhaps less so in fieldwork situations). There are four main points that need to be considered:

a) Use appropriate 'negative' controls

Negative controls are those where no treatment is administered, and hence no response is expected. A scenario where particular thought should be given to the nature of the negative control is that where solvents are used to dissolve substances that are relatively insoluble in water, and thus the concentrated stocks require an organic solvent (such as ethanol, methanol, dimethylformamide, acetone) to deliver the substance to the exposure medium. It has been shown that such solvents can affect various endpoints in exposed organisms, even when used at low concentrations.[23] Hence it is imperative to minimise the use of solvents and also to include a control in which organisms are exposed to the same concentration of solvent as in the substance treatments. Crucially, these 'solvent controls' (as opposed to the 'dilution water controls') must also be used for comparison with the substance treatments when it comes to the statistical analysis of the results.

It has to be acknowledged that negative controls can be more difficult to implement properly in fieldwork, since there may simply not be any pristine sites available with which to compare

organisms from exposed sites. An example of this would be the work undertaken by Jobling *et al* (see also Iwanowicz et al),[24,25] where a small proportion of male roach (*Rutilus rutilus*) at supposedly clean sites (i.e. not exposed to Waste Water Treatment Plant (WWTP) effluents) were found to be intersex (albeit mildly so). The likely reason is that these sites are not as clean as we think (or hope) they are, and are likely often subject to diffuse pollution sources. These sites are nonetheless a useful source of reference values and scientists can use a measure of the relative contamination between sites (even if this is as simple as whether the site is upstream or downstream of an effluent outfall) to judge the influence of such contamination on the level of intersex in the fish under investigation.

b) Use a positive control where appropriate and/or available.

The use of a positive control with known levels of activity may not always be possible, but can be incredibly helpful in the interpretation of ecotoxicological data when implemented. One example of this is in endocrine disruption work, where the apparent hormonal activity of a substance is being investigated. For example, some synthetic estrogen mimics are very weak in comparison to the natural steroid hormone, E2, or the synthetic steroid, ethinylestradiol (EE2). If we compare the estrogenic potency of parabens *in vivo* with a control group, they are certainly estrogenically active. However, in comparison with a positive control, such as E2, these substances have been shown to be only weakly estrogenic both in mammals and in fish.[26,27] Thus a positive control allows us to put the results into perspective, as well as verifying that the bioassay (i.e. test procedure) is actually working properly.

c) Consider the number of controls.

Part of the reason for including controls in experiments is to establish the degree of variability in the responses of the test animals. Hence if an insufficient number of control subjects is used, then an inaccurate assessment of variability may be made and consequently the comparison with the treated subjects will be made on false assumptions. One example of a study which has demonstrated the importance of a robust experimental design including sufficient numbers of controls has been reported by Owen at al.[28] The authors initially found an effect of clofibric acid on the growth rate and condition of juvenile rainbow trout, but an expanded version of the study (using an increased number of control animals, spread over four tanks) did not repeat their original findings. Specifically, the effect observed in the original study was because the relatively small number of control fish (n=8) were exceptional and outperformed normal expectations, further highlighting the need for appropriate controls.

d) Use the appropriate type of control.

This advice refers to the fact that researchers must be aware that any bias introduced into the 'selection' or handling of control subjects is unacceptable. That is, the control organisms should be the same sex, age, of a similar size, from the same population as those in the treated groups, and definitively not pre-selected for desirable features which make them reliable controls. Further, all controls must be handled in the same way as treatment groups with respect to factors such as disturbance, food, and experimental conditions (such as light and temperature). If one tank requires cleaning, for example, all tanks should be cleaned.

Principle 4: Use appropriate exposure routes and concentrations

It is probably true to say that the weakest aspect of many ecotoxicological papers concerns exposure to the test substance(s). Most ecotoxicologists have their main training in biology, not chemistry. Ideally, ecotoxicologists should confer with environmental chemists and modellers before any experiments are designed. Some of the main issues to consider regarding exposure are examined in this and the next principle.

a) What is the most environmentally-relevant route of exposure?

Before exposing an organism to a substance in a laboratory experiment, it is wise to consider the most appropriate route of exposure. For aquatic organisms this is likely to be either via the water or the diet, depending in part on the hydrophobicity of the test substance as well as the behaviour and feeding characteristics of the organism concerned. In the wild, exposure to strongly hydrophobic chemicals may occur primarily via the diet; in which case, this route of exposure should be used, if at all practical. The toxicity of a substance can vary depending on the route of exposure.[29,30,31] In the terrestrial environment, exposure via diet is very common, and is often the main route of exposure to substances: recall the devastating effects that pesticides had on birds of prey in the 1950s and beyond.[32] In contrast, injecting any organism with a test substance (in the context of ecotoxicological studies) is wholly unrealistic and should be avoided, as it cannot shed any light on the real environmental exposure of wild animals.

b) What is meant by an 'environmentally relevant concentration'?

The concept of an 'environmentally relevant concentration' is clearly important in ecotoxicology, as it allows us to judge whether a substance is not merely a hazard but actually poses a risk. Since 1991, the phrase 'environmentally relevant concentrations' has appeared in

234 the title, or abstract, of 1675 papers according to the Web of Science (accessed January 2013).
235 Unfortunately, because there is no clear definition of the phrase 'environmentally relevant
236 concentration', the whole issue can be dangerously misleading. This problem can be illustrated
237 by considering the following issues.

238 *What is meant by 'environment': the sewage effluent, a sewage ditch, or a river?*
239 It is not unknown for scientists to use a value reported in sewage effluent when justifying their
240 experimental concentration as being environmentally relevant to aquatic organisms.[33,34,35] Some
241 WWTPs discharge into very small streams which are essentially formed from sewage effluent,
242 but could be classed as a water course. However, the vast majority of freshwater aquatic wildlife
243 live in rivers where considerable dilution of the sewage effluent is the norm.

244 *Could the quoted environmental concentration result from an unreliable measurement?*
245 Trying to detect a substance of interest at low and sub ng L^{-1} concentrations in complex matrices
246 is fraught with difficulties.[36] Hence it is also possible that reported exposure concentrations
247 (particularly in the environment) are in error, and require independent verification before they
248 are accepted. For example, the very high (hundreds of ng L^{-1}) concentrations of many sex steroid
249 hormones, particularly EE2, in UK and US streams reported by Aherne and Briggs [37]and Kolpin
250 et al [38]have proved not to be repeatable.[39] Such erroneous reports can have enormous influence
251 on what are, and are not, considered to be 'environmentally relevant' concentrations of
252 substances of concern. Hence the need for a broad review of the literature and/or the
253 collaboration with an analytical chemist in studies where necessary.

254 *Might the quoted environmental concentration be accurate but be entirely unrepresentative of*
255 *the majority of situations encountered by wildlife in time and space?*
256 Occasional very high concentrations can occur in the environment but in terms of probability
257 they are likely to be rare. A good example is the modelling of 11 large US catchments where the
258 50%ile cumulative probability for EE2 was between 0.0008 and 0.01 ng L^{-1} at mean and low
259 flow, respectively, but there remained in the 99%ile probability a potential for 0.3-1.0 ng L^{-1}
260 being detected. Thus, the vast majority of American aquatic wildlife would be most likely to be
261 exposed to concentrations in the 0.0008-0.01 ng L^{-1} EE2 range and only a tiny minority to
262 concentrations of $\geq$0.3 ng L^{-1}.[39] So whilst some authors might imply that 5 ng L^{-1} EE2 is
263 environmentally relevant,[33,35] the overwhelming evidence is that it would be atypical.

We should also make it clear that we are not asserting that only environmentally relevant concentrations should be used in ecotoxicological experiments. Indeed, there will be occasions where researchers have to use significantly higher concentrations in order to properly define a LOEC for a substance. The LOEC of a substance is, in fact, far more useful in the regulatory sphere than is a conclusion that no effect occurs at environmentally relevant concentrations, because a LOEC enables the regulators to impose more accurate and meaningful safety limits. Our primary message here is that the explanation of the concentrations selected for a particular study should be comprehensive, and the authors should be open and honest about the context of their results in relation to those concentrations which have been measured (or predicted) in real environmental samples. Thus the derivation of the measured environmental concentration (MEC) and the predicted environmental concentration (PEC) are also key factors here.

Principle 5: Define the exposure

A useful exercise to undertake when considering this principle is to remind ourselves of why we undertake ecotoxicological studies in the first place. The major reason is that we are concerned about the occurrence of certain substances in the environment, and we need to determine whether they are present at concentrations which can be harmful to living organisms. Thus, in conducting such studies, we hope to supplement the database which is used to risk assess environmental contaminants. Such risk assessments will clearly be inaccurate if the concentrations on which they are based are also inaccurate. The two main points to consider within the scope of this principle are outlined below.

a) The actual amount of exposure substance in the system must be measured

It is paramount that an attempt is made to determine the actual concentrations of substance/s present in the test media to which organisms are exposed. This can be done using either analytical chemistry or biological methods of analysis such as immunoassays or receptor binding assays. Which of these methods is more suitable is debatable; however, there is no doubt that without any attempt to measure concentrations of the test substance, the results of the study cannot be fully interpreted. We should also add at this juncture that it is important there is good quality control of analytical chemistry procedures employed, as the data obtained using such methods are of little use if it the associated methods have not been properly validated.

There are many examples in the literature where no analytical analyses have been performed. In these cases the researchers have no idea whether the concentration to which the organisms are exposed is (for example) 100 ng L^{-1} or 1 ng L^{-1}, leaving the results wide open to misinterpretation. Many of these studies also involved the use of a static-renewal system, which further increases the risk of unreliable results compared with a flow-through exposure system, hence rendering the measurement of the test substance even more important. Examples of such studies include those undertaken by Oehlmann *et al*,[21] whereby prosobranch snails were exposed to octylphenol (OP) and BPA at nominal concentrations ranging from 1 to 100 µg L^{-1}. The authors describe effects being observed 'at the lowest concentrations' but it is unclear as to what those concentrations actually were. This is critical information from a risk assessment point of view. Similarly equivocal information has been generated by Lister *et al*,[40] Di Poi *et al*,[41] Franzelletti *et al*[42] and Guler and Ford;[43] these studies tested pharmaceutical products, including fluoxetine, at nominal concentrations as low as 0.3 ng L^{-1} in static renewal systems; however no measurements of the actual concentrations of the substances that they were testing were performed. We accept that if a significant biological effect is observed in a dose-related manner it might be difficult to argue that something is not present in the water that is causing that response. But this information is of no use to the regulators if it is not known how much of the substance causes that response. In addition, if a response is observed which is unrelated to the concentration of the substance used, or if there is no response at all, it is impossible to provide an accurate interpretation of the data when the exposure concentrations are unknown. There may actually be no effect of the substance concerned at the (nominal) concentration, but it may be that no effect was observed because the chemical was not present in the tanks at anything like the concentrations that were expected. Finally, although technically more problematic, it is particularly important that verification of the actual exposure concentrations is provided when concentrations that are reported to be causing effects are extremely low (i.e. at concentrations similar to those found in the environment).

b) Potential contaminants in the system should also be monitored, thus providing an accurate profile of all major substances in the test media.

It is useful to have some knowledge of potential contaminants in the system. Clearly, not all eventualities can be accounted for, but what is looked for should include the more commonly occurring contaminants, to assess whether they are present at high enough concentrations to be of concern, or whether their presence can be ignored. There may be occasions where contaminants are found in sufficiently high concentrations that they are likely to act as

confounding factors in the toxicological assessment. Such a case was reported by Hala *et al*,[44] who discovered butyltin leaching from airline tubing in a flow-through exposure system at concentrations high enough to confer toxic effects on organisms. Conversely, Aoki *et al* reported intermittent detection of diethylhexyl phthalate (DEHP) in a study undertaken to assess the anti-androgenic nature of dibutyl phthalate (DBP) in fish;[45] however, in this case it was concluded that the DEHP originated from contamination during the extraction/analysis procedure (i.e. not from the tank water itself), and in any case it was present at such low levels as to be negligible in terms of its effect on the fish in this study. It is unlikely that any study which monitors concentrations of DEHP as a contaminant in water would not contain a trace of this chemical, but it is nonetheless wise to determine the concentrations of DEHP present (particularly in phthalate exposure studies) in order that their significance can be assessed. Another potentially problematic situation is where a test substance might be found to be present in the control tank (for example, via cross-contamination, or even due to inadequate cleaning of equipment between studies). Such information could be critical to understanding the results.

Principle 6: Understand your tools

When using live organisms to try to understand what are often dynamic processes, it is important to try to minimise the variability encountered by having a good understanding of the background of these organisms. For example, the quality of data obtained can be influenced by the age of animals, as well as by the conditions in which they were reared and/or maintained prior to the study. In addition, some species are very difficult to rear in the laboratory (or it is sometimes inappropriate for the particular assay in use) and if wild-caught organisms are used instead, it is vital that the conditions in the environment in which they have been living are well understood. The presence/absence of parasites should also be established. The presence of parasites can affect physiological parameters in animals,[22,46,47] and if those parameters overlap at all with those being used in a controlled exposure study, the interpretation of data obtained from infected animals can be problematic, to say the least.[48,49] Parasite infections such as microsporidians can cause gonadal disruption, produce intersex and female-biased populations, as well as affecting secondary sexual characteristics.[50] Such combinations of changes can be mistaken for changes that result from chemical exposure. Therefore baseline information on the prevalence of parasitism in different species and an awareness of the potential effects ensuing from this are essential considerations in studies undertaken with wild-caught animals.

In some mammalian studies, an understanding of the particular strain used in toxicological studies is necessary, as it is well known that some strains are more sensitive than others.[51] Likewise, with commonly used fish species, differences in sensitivities occur in the responses to stressors observed between different strains of the same species;[52,53,54] and Brown et al reported differences in growth and sexual development between inbred and outbred zebrafish,[55] which can impact on interpretation of data obtained from substance-exposure trials. It is also important to consider the relevance of the species selected in relation to the overall aim of the study.[7]

In vitro studies may appear to be more reproducible, but they are certainly not immune from variability. For example, the response of different cell lines to the same genotoxic agent can vary widely within and between laboratories. Therefore, the selection of the cell line to be used needs careful consideration;[56] also, even within the scope of analysing a single protein, different antibody preparations can elicit very different responses.[57] It is important that researchers are aware of these factors and are able to adequately define the reagents used.

Knowledge of the test substance is equally important. A confirmation of this knowledge should be communicated to the reader by simple means such as stating its purity and CAS number. A discussion of the impact of impurities on the interpretation of data obtained in an *in vitro* estrogen assay was presented by Beresford *et al*,[58] and has also been recognised by Harris *et al*,[59] who found that two different preparations of a phthalate presented very different estrogenic profiles as a result of one of these preparations having been supplemented with BPA. Some substances consist of different isomers which can have very different biological activities. For example, branched chain isomers of alkylphenolic compounds (such as 4-NP and 4-OP) induce estrogenic effects in fish, mammals and *in vitro* assays, in contrast to the straight chain isomer of the corresponding compound (4-n-NP and 4-n-OP) which are not estrogenic.[60,61] Hence the inadvertent use of the linear isomer of this substance in an ecotoxicology study could lead to erroneous conclusions of inactivity (as was the case, for example in Moore *et al*).[62]

Principle 7: Think about statistical analysis of the results when designing an experiment

The importance of appropriate statistical analysis cannot be overemphasised.[4] It is crucial that we are able to draw robust conclusions, and that we are able to justify them. In the case of an inappropriate statistical approach being used, an entire study can be undermined and, at worst,

misleading conclusions can be drawn. It may be necessary, particularly in some of the more complex analyses required, to enlist the help of professional statisticians. Different statistical approaches exist, the use of which are dependent on the aims of the study in question. These approaches range from testing methods to identify significant effect responses (e.g. to establish a no observable effect concentration [NOEC]); through empirical regression modelling (e.g. to estimate effect or benchmark concentrations); to complex biological modelling (e.g. DEBTOX). Although criticised by many statisticians,[63] the NOEC (i.e. the tested concentration just below the LOEC (lowest concentration that produced a significant response)) is still the most commonly used toxicity descriptor. This is derived by statistical testing approaches which assume 'no effect' (null hypothesis) and estimate the likelihood that an observed effect happened by chance alone (i.e. not statistically significant) or that it was unlikely to be due to chance alone (statistically significant).

Power analysis can be conducted to determine the size of a sample needed to reject a null hypothesis at given error rates, or it can be used to estimate, at given data variation and sample size, the minimal effect size that can be detected as statistically significant. This effect size defines the statistical detection limit which is always present in the data (also called 'minimal detectable significant difference'). Thus, an *a priori* power analysis can enable the scientist to design a study such that the sample size is high enough to provide reliable answers to the question posed, whilst not being so high that valuable resources are wasted. Nowadays, software packages exist which allow power and sample size calculation without the need to contact a professional statistician, at least for simple study designs. Recommended maximal error rates are usually α=5% and β=20%,[64] meaning that the minimal power is 80%, i.e. we would identify an effect above the detection limit in 4 out of 5 studies. Another parameter needed for the power calculation is an estimate about the most likely data variation, which can be derived either from previous studies or other historical data sources that are considered comparable to the researchers' own testing environment. So called 'range-finding' studies are often key to providing initial basic information.

An example of power analysis is given in table 2. This illustrates the issues involved with assessing the number of individuals required to produce an experiment which will offer a reasonable degree of power in the analysis. Two types of data have been assessed (the data used here are not real, but are derived from real exposure scenarios). The first is where the endpoint assessed is plasma E2 concentration in fish. The response of this parameter can be extremely low

(the maximum difference in mean plasma E2 concentration shown here was 2.2 ng ml^{-1}). The second scenario is where the response can be in several orders of magnitude (e.g. plasma vitellogenin concentration). In both cases a high and an intermediate effect detection limit are shown; in the case of plasma vitellogenin a 'low' response is also shown. The standard deviation (relative to the mean) is usually lower across individuals exposed to a high level of treatment than it is in the intermediate treatment group. What the information provided in Table 2 illustrates is that where the effect size is (or is expected to be) lower, more individuals are required to detect this size as significant at given error rates. Consequently, if the degree of change in a given endpoint is very small, providing robust evidence of any change can be challenging; where the degree of change is far greater, detecting a change in response to a stressor is much easier. In addition, the higher the variability observed within any treatment group, the more individuals are required.

Where good baseline (control) data are available, scientists will be able to determine the variability within control groups and use this to aid the experimental design. For example, extensive data sets have been published on the variability of a variety of reproductive and endocrinological parameters in fathead minnows,[18,65] which are extremely useful to researchers designing experiments using reproductive endpoints in these fish. Furthermore, Paull and colleagues considered that the level of inconsistency in reproductive success between breeding colonies of zebrafish maintained in the laboratory was so high that a minimum of six replicates per chemical treatment is necessary to discriminate a 40% change in egg output of females and sperm quality (in terms of motility) in male zebrafish (at α=5%).[19]

To conclude, it is important to remember that (i) error rates (and therefore a (controlled) uncertainty) are always present in our conclusions; (ii) statistical significance should not be confused with biological significance; (iii) "no effects" cannot be identified by statistics; and (iv) if one reaches the conclusion to accept a hypothesis, it does not mean that it is proven, it means that the hypothesis is supported given current data.

More detailed guidance on statistical approaches used in standard ecotoxicology studies can be found in the OECD Testing and Assessment guidelines.[63,64]

Endpoint	Treatment	Mean	Average or worst-case scenario standard deviation	Number of individuals required to give 80% power
Plasma E2 concentration (ng ml^{-1})	Control	3.84		
	Low	2.7	average	17
			worst-case	53
	High	1.62	average	4
			worst-case	7
Plasma vitellogenin concentration (ng ml^{-1})	Control	54		
	Low	85	average	16
			worst-case	30
	Medium	40000	average	3
			worst-case	5
	High	350000	average	2
			worst-case	2

*Table 2. The number of individuals required to provide data with a power of 0.8 and an α (probability of error) value of 0.05 in particular exposure scenarios. These a priori analyses, using log_{10} values of hypothetical data, were conducted using the statistical package 'G*Power'. A 'low' response example is not given for the endpoint of plasma E2 because the overall range of response is far smaller here than it is for the vitellogenin response.*

Principle 8: Consider the dose-response

In order to be able to deduce the dose-response of a substance (and hence put the results into any kind of environmental context), at least three concentrations need to be tested. A recent example of a study which does not report a full dose-response was published in Science,[66] where only two concentrations of the drug (oxazepam) were tested. Data from just one or two concentrations alone will be of little use in the regulatory field.

Secondly, we think that, in almost all cases, the relationship between dose and response should be regularly incremental (or decremental) – i.e. for each increase in dose, there should be a graded increase (or decrease) in response. This produces a 'monotonic' dose-response curve. Good examples of monotonic curves are those involving estrogen stimulation of vitellogenin production in fish and androgen stimulation of spiggin production in the stickleback (*Gasterosteus aculeatus*).[67,68] A key outcome of bioassays with monotonic curves (providing they can be consistently repeated) is that it is possible to accurately calculate the LOEC and the NOEC (or NOAEL) of compounds. These are very important for accurate ecological risk assessments.

There are numerous examples (many hundreds) of published dose-response curves in the field of ecotoxicology that are 'non-monotonic'.[69] These cover a whole range of shapes such as flat, U-shaped, J-shaped and inverted U, as well as many that are irregular (or 'multinodal'). When it comes to the interpretation of non-monotonic dose-response curves, a rift has developed between ecotoxicologists. In the view of Vandenberg and co-workers, non-monotonic curves form compelling evidence that low doses of compounds (in many cases well below the current NOAEL) are able to trigger effects that regulators do not currently take into account.[69] However, there are others who, while conceding that non-monotonic (especially inverted-U-shaped) curves are not unlikely to occur in some circumstances, are of the opinion that many of the non-monotonic relationships that have been reported can equally be ascribed to either poor experimental design and/or technique, or to the action of confounding factors. The gold test of whether a non-monotonic dose-relationship is a real phenomenon (as with other scientific endeavours) should be whether it can be reproduced consistently. Vandenberg *et al* appear, surprisingly, to argue that this is an unfair requirement in the field of low-dose effects, due to such effects tending to be more dependent on factors such as place, time, operators, strain of animal etc. than high dose effects. This view is obviously one that is open to debate.

We do accept that a dose-response relationship may, after further research, turn out to be genuinely non-sigmoidal (especially one that has a regular U or inverted-U shape). In such cases the burden of proof is on the researchers who report such data to, firstly, show that the phenomenon is repeatable and secondly, at some stage in the research process, to explain and, if possible, prove the underlying mechanism that causes the effect. Even if these two objectives can be achieved, there is still a major problem with using results from bioassays that have generated non-sigmoidal dose-response curves to guide environmental safety thresholds.

Principle 9: Repeat the experiment

a) Repeat the experiment in own laboratory in the first instance

With budgets tight and with scientists who undertake *in vivo* studies always looking to reduce the numbers of animals used, it is understandable that on many occasions a single experiment is cited as producing a particular and significant response pattern. This is especially true the closer the research is to fieldwork (for example, full life-cycle studies and/or mesocosm studies are, for some researchers, too expensive to undertake once, let alone twice). It is also a result of necessary legislation that exists to protect vertebrates used in experimental procedures, which means that researchers have to keep the number of animals used to a minimum. Hence *a priori* power analysis (see Principle 7) is an important tool to inform researchers of the minimum number of animals required to give a sound result in a given study. Furthermore, repeat studies must be justifiable to legislative bodies, and in some cases should include refinements (which aim to improve the robustness of the results obtained). However, all researchers must be aware that it is imperative that where the results are surprising, or especially hard-hitting (for example, a significant response to a very low dose of substance, or a response which contradicts previous studies), the onus is on the researchers concerned to repeat the experiment, in order to verify their conclusions. As is often quoted in the literature, 'extraordinary claims require extraordinary evidence'.

b) The importance of independent validation

Politicians or risk assessors must take great care when making decisions on the basis of observations that have not been independently confirmed. Unfortunately science funding is usually limited and, also, most scientists and funding bodies prefer to do 'original research' rather than confirm someone else's findings. Because of the consequent lack of independently validated studies, people who seek to make decisions on the basis of the scientific literature (such as risk assessors) instead rely heavily on the 'weight of evidence' (WoE) approach (i.e. where plausible evidence is built up from fragmented observations from a diverse range of species and approaches); see Principle 11. For example, the majority of us are agreed that in an ideal world we would like to be able to use invertebrates instead of vertebrate organisms in ecotoxicology. In the field of endocrine disruption, for example, molluscs might appear to be the ideal solution. There are at least 200 papers that suggest that the reproductive hormones of molluscs are the same as those of humans. However, as pointed out by Scott,[70] very few of these

studies have ever been properly independently validated (i.e. they have been on different species, with different endpoints, and different experimental designs).

Another important reason why one should wait for findings to be independently validated is that 'to err is human'. It should be safe to assume that any trained scientist (especially one with a good track record in research) should not make mistakes when, for example, working out dilutions and concentrations, making up solutions with defined molarities, or analysing data. However, it is not safe to assume this at all. In fact, the propensity of scientists to make errors appears to be rather high. It was the recognition that mistakes are easily made that was behind the issuance in 2003 of the Joint Code of Practice for Research by the main UK biological research funding bodies.[71] Its major requirement is that scientists should keep accurate and detailed records of all their actions in order that any such errors, if they occur, can be traced and corrected (even post-publication). It is also good practice to have other colleagues cross-checking calculations and/or data analysis, as a form of quality control.

The importance of reproducibility was discussed in a recent Nature World View article, in which the author asserts that "reproducibility separates science from mere anecdote".[72]

Principle 10: Consider confounding factors

Confounding factors are those 'conditions' present in the test environment which may influence the experimental result in addition to the specific parameter that is being assessed. These may include factors such as variations in temperature, disease and the presence of unexpected substances, amongst others. Although it is not always straightforward, or even possible, to actually quantify the confounding factors present, we must always be aware of their potential influence and be cautious in our interpretation of the results, especially when such factors are known to be present. Fieldwork scenarios, in particular, present a challenging and complex array of confounding factors which may enhance or mask the adverse effects of a chemical or mixture of chemicals. At the very least these must be acknowledged by the authors, and when known, accounted for in the analysis and interpretation of data arising from such studies.

As an example of good practice in relation to interpretation of field trials, we point to a study by Burkhardt-Holm et al that dealt with the issue of why fish catches (mainly of trout) have declined very significantly in Switzerland in the last few decades.[73] Instead of automatically linking the decline to the existence of estrogens in the aquatic environment (the fashionable

explanation at the time), the authors offered eight potential causes, ranging from poor water quality, increased predation (by birds), insufficient food, as well as changes in fisheries management. Each potential cause was discussed, in a very balanced manner, in order to rule them in or out. In the end, the researchers concluded that it is unlikely that the decline in fish stocks has a single cause; instead it is most likely due to a combination of factors (stressors).

As an example of bad practice in relation to interpretation of field trial data, we point to a study by Ginebrada et al that implies, in both its title "Environmental risk assessment of pharmaceuticals in rivers: relationships between hazard indexes and aquatic macroinvertebrate diversity indexes in the Llobregat River (NE Spain)" and abstract, that the reason for reduced macroinvertebrate diversity in the studied locations (namely, rivers receiving effluent inputs), is the presence of pharmaceuticals in the effluent discharge.[74] However, although the concentrations of the selected drugs were found to be correlated to both the density and biomass of macroinvertebrates, it seems inevitable that other properties of the effluents (such as other chemicals that are present in effluents, or the physico-chemical characteristics of the effluents concerned) would also have contributed to this reduction in diversity, and would probably also have shown a correlation. The authors did actually raise this point in the discussion section of the paper, but it should not (in our opinion) have been omitted from the title and the abstract.

One final example of a significant confounding factor is parasitic infection (see Principle 6 for further discussion on the impact of parasites on endpoints associated with endocrine disruption). As mentioned above, it is important to acknowledge the potential impact of such phenomena on the outcome of a study, even if the precise relationships are not clear-cut.

Principle 11: Consider the weight of evidence

The general principle behind assessing the weight of evidence (WoE) concerning the environmental risk posed by a particular substance involves taking all the available information, from whatever source (e.g. field and laboratory; *in vitro* and *in vivo*; ecological and physiological), and judging how well it does, or does not, tell a consistent story.

Many papers, especially reviews, refer to the 'WoE' for a particular theory, and this is what is used by regulators to determine the risk posed by a particular substance. However, according to Weed,[75] this term has not been scientifically defined and has been used in the majority of cases in a metaphorical sense (e.g. 'nine out of ten papers report a positive effect of compound X,

583 therefore surely, reader, you have to accept that compound X is an endocrine disruptor').
584 However, realising that this usage takes no account of the quality of the papers – and is, in all
585 probability, just a reflection of the prevailing bias in that particular field,[76] several people in
586 recent years have attempted to develop more focussed methods for quantifying WoE.[9,77]
587 However, whether this entails 'weighting' the studies on the basis of dataset size, or even simply
588 tabulating all the data points (where known) in the literature in an unbiased manner and allowing
589 the reader to make his/her own judgement,[39,78] all approaches suffer from the same inherent
590 weakness – namely that studies where no effects were observed are very often not published (see
591 Principle 12) and such studies cannot therefore be taken into account.

592 With regards to whether or not the research fits with existing literature, pharmaceuticals provide
593 an excellent example. They have an extremely well defined mechanism of action (at least as far
594 as their activity in humans is concerned). This information can be of immense value both in the
595 design of studies to assess ecotoxicity of pharmaceutical substances, and also in the
596 interpretation of results obtained from such studies, and should be taken into account when
597 assessing the weight of evidence for pharmaceutical substances for which the MOA (and also, in
598 many cases, their potential side effects) is well defined.

599 Despite some potentially difficult areas to negotiate, the WoE approach remains the only way
600 that scientists and policy makers can move forward in the uncertain world of science. A major
601 argument in the philosophy of science is that we can never prove a hypothesis, no matter how
602 many examples are provided, but only falsify it.[79] At first sight this would appear to keep science
603 in a prison of uncertainty, with nothing able to be proved. However, both Popper[80] and Hill[8]
604 allowed that where sufficient independently validated supporting evidence existed, the
605 hypothesis could be considered a working hypothesis and a basis for action.

606 ## *Principle 12: Report findings in an unbiased manner*

607 Researchers these days are under a great deal of pressure to attract research funding, to deliver a
608 positive outcome to their paymasters and to publish as many papers as possible in high impact
609 journals. We believe that these pressures are behind the increase in papers in which the title and
610 abstract tell one story (often with dramatic claims), while the methods and results tell another
611 (often containing weaknesses in design and/or mundane findings). Aside from the fact that the
612 publication of such papers is an indictment of the peer-review process, we believe that such use

613 of ‘spin’ is confusing for policy makers, a bad example for young researchers and ultimately
614 gives the profession a bad name. The problems occur when the researchers fail to acknowledge
615 or discuss the weaknesses and/or when they employ hyperbole (‘hype’) to exaggerate the
616 significance of their findings. A recent example of hype is the paper entitled ‘Antidepressants
617 make amphipods see the light’ in which the data purporting to show that the organisms
618 concerned move towards the light in response to exposure to fluoxetine is, in our opinion,
619 inconclusive (because although the data from one study show a significant effect, data from the
620 other study reported in the same paper show no such effect).[43]

621 One of the many reasons why such controversies arise, as proposed by Goldacre,[81] is the
622 ‘suppression of negative results’, a topic also addressed by Knight.[82] This is the (mostly passive)
623 tendency of researchers to publish only positive results (as negative results do not, except in a
624 few cases, attract research funding or ensure career progression). Goldacre argues, however, that
625 many scientists do not just tend to shy away from negative results, but actually have a bias
626 towards positive evidence, and points to a study that examined the outcome of FDA (Federal
627 Drug Administration) registered clinical trials on a class of antidepressant drugs.[83] Thirty seven
628 studies showed a positive effect, of which thirty six were published in peer-reviewed literature.
629 However, there were a nearly equal number of studies (thirty three) that gave negative results; of
630 these, twenty two were not published at all and another eleven were written up and published in
631 a way that implied they had a positive outcome. In the context of endocrine disruption, this
632 tendency for bias towards positive evidence probably explains why scientists, when including
633 negative (i.e. no effect) as well as positive data in their papers, tend to assume that the
634 experiments with the positive results are the ‘correct’ result, and any negative outcomes are due
635 to unforeseen circumstances – e.g. the experiments with negative outcomes have been variously
636 explained away on the basis that: ‘the experiment was not carried out at the right time of year’;
637 ‘the animals were not at the right stage of maturation’; ‘the experiment was done at the wrong
638 temperature’; or ‘the animals were not of the correct origin’. Although it cannot be denied that
639 there may be a valid explanation for a negative result, we suggest that, without actual hard
640 evidence, there is no *a priori* reason, in any study, to reject the experiments that give negative
641 results and only accept the ones that give positive results. Another reason that controversies
642 often arise in the reporting of ecotoxicological data is that there is no clear definition of what
643 constitutes an ‘adverse’ effect. Although it is not within the scope of this manuscript to address
644 this issue fully here, the authors recognise that this lack of definition can lead to subjective
645 presentation of data, depending on the personal opinion of the scientist concerned. For example,

some think that any alteration in the physiology of organisms, which has been induced by a substance to which that organism would not naturally be exposed, could be considered an adverse effect. On the other hand, others consider that it only becomes an adverse effect once there is an effect on population- or health-related endpoints. Still more may even believe that a reduction in the numbers of an over-crowded population would not necessarily be considered 'adverse'. This is perhaps an ethical issue that would be best discussed in another forum.

Causes and consequences of poor ecotoxicological research

Undoubtedly the most compelling reason for the rush to publish (and never mind the quality), is the fact that scientific research has become increasingly competitive over recent years. This has led to the need for scientists to publish prolifically in order to be able to secure both jobs and further funding. In many cases, quantity appears to rule over quality. The issue of the tendency not to publish 'negative' (no-effect) results may also be a factor here (scientists think that funders and future employers will be less interested in their work if they have not shown a newly discovered sensational effect of substance x on species y); although journal editors also have a duty to encourage the publication of no-effect data arising from well designed and executed studies. There is also evidence that there has been a proliferation of journal output over recent decades,[84] which may well have led to a dilution of good science with poor (although there are no studies that we know of that have investigated the change in number of ecotoxicological journals in particular over this time). We do agree with the sentiments expressed in a recent 'Nature' editorial that the frequently irreproducible data that are published these days are not usually a result of fraud, but of insufficient thoroughness in the analysis and presentation of data.[2]

The potential consequences of unsound ecotoxicology research can be profound. Ecotoxicologists presumably conduct their research because they want to protect wildlife from adverse effects of chemicals that already are, or could in the future be, present in the environment. In other words, they want to improve the environment (or prevent it deteriorating), by researching potentially hazardous chemicals and subsequently reducing chemical pollution in the environment. However, many ecotoxicologists have little or no contact with the people (regulators) who have to act on the results that they publish. Regulators have to assess the degree of risk posed by a substance (based primarily on the published research of ecotoxicologists) and, if necessary, take steps to reduce that risk to an acceptable level. The process of assessing the

677 degree of risk and taking any necessary risk reduction steps (such as setting environmental
678 quality standards, or restricting or even banning the use of a chemical) can often be a very
679 detailed and lengthy one. It often takes a decade or more and, these days, usually occurs at both
680 national and international levels. Hence it costs a great deal of money! Moreover, the funding
681 available for fundamental science to support the data produced by (eco)toxicologists is limited,
682 further hindering progress made by the regulators. In cases where data are published indicating
683 that a particular substance is likely to cause adverse effects to wildlife, it is naturally difficult to
684 change the negative public opinion towards this substance, even when the data concerned
685 emanated from just a single study. The cost of confirming or refuting the results of a poorly
686 designed study can be extremely high; and for fish chronic studies could amount to several
687 hundred thousand US dollars. The cause of protecting the environment itself may suffer as funds
688 are drawn away from studying other more harmful chemicals. In addition, the calculation of
689 Environmental Quality Standards (EQS) involves the evaluation of all studies published on the
690 particular chemical concerned. However, the existence of even one study that shows, for
691 example, that a 100-fold lower EQS should be applied, must be acknowledged by regulators
692 even if the vast majority of studies suggest otherwise. Any inadequacies in study design or
693 inaccuracies in the measurements made could have profound implications for regulators, for the
694 water industry, and ultimately for us as taxpayers, if they lead to a significantly lower acceptable
695 environmental concentration. Furthermore, there are undoubtedly environmental contaminants
696 upon which the regulator should be focussing their attention, and inaccurate data on other (less
697 harmful) substances may mean that their attention is not focussed on the chemicals that really
698 are of environmental concern.

699 In conclusion, ecotoxicologists need to think about the consequences of their research *before*
700 they publish it, and they need to take responsibility for it. This does not mean that results
701 suggesting a substance is of concern should be suppressed, or their publication significantly
702 delayed. Indeed, we embrace the process of publication as a major part of scientific discourse,
703 and its role in facilitating discussion around the subject in hand. But it does mean that scientists
704 have a duty to ensure that their research is sound, and therefore likely to be repeatable, before
705 publishing it. Likewise, readers should be aware that they should always critically appraise the
706 work contained therein, and not take it simply on trust. Scientists also need to give serious
707 consideration to making their raw data publically available, the benefits of which cannot be
708 overstated. Many high quality journals require that such data are deposited in a database prior to
709 publication; those that do not specifically require this do at least encourage authors to share their

data on request. Adhering to these guidelines will greatly enhance the trust afforded to individual scientists, and between scientists and policy makers. Transparency and robustness are key elements to a successful scientific outcome.

Acknowledgements

Many thanks to Luigi Margiotta-Casaluci, Martin Scholze and David Taylor for their comments and insight into the manuscript and the issues surrounding this topic. Andrew Johnson is grateful to CEH for supporting his contribution through science budget funding from NERC.

References

1. Agerstrand, M.; Edvardsson, L.; Ruden, C. Bad reporting or bad science? Systematic data evaluation as a means to improve the use of peer-reviewed studies in risk assessments of chemicals. *Hum. Ecol. Risk Assess*. **2013**, DOI: 10.1080/10807039.2013.854139

2. Nature Editorial. Must try harder. *Nature* **2012**, 483, 509.

3. Nature Editorial. Reducing our irreproducibility. *Nature* **2013**, 496, 398.

4. Vaux, D.L. Know when your numbers are significant. *Nature* **2012**, 492, 180-181.

5. Küster, A.; Bachmann, J.; Brandt, U.; Ebert, I.; Hickmann, S.; Klein-Goedicke, J.; Maack, G.; Schmitz, S.; Thumm, E.; Rechenberg B. Regulatory demands on data quality for the environmental risk assessment of pharmaceuticals. *Regul.. Toxicol. Pharmacol*. **2009**, 55(3), 276-280.

6. Agerstrand, M.; Küster, A.; Bachmann, J.; Breitholtz, M.; Ebert, I.; Rechenberg B.; Ruden, C. Reporting and evaluation criteria as means towards a transparent use of ecotoxicity data for environmental risk assessment of pharmaceuticals. *Environ. Pollut*. **2011**, 159(10), 2487-2492.

7 Breitholtz, M.; Ruden, C.; Hansson, S.O.; Bengtsson, B-E. Ten challenges for improved ecotoxicological testing in environmental risk assessment. *Ecotox. Environ. Safe*. **2006**, 63, 324-335.

8. Hill A.B. The environment and disease: association or causation? *Proc. Royal Soc. Med*. **1965**, 58, 295-300.

9. Klimisch, H.J.; Andreae, M.; Tillman, U. A systematic approach for evaluating the quality of experimental toxicological and ecotoxicological data. *Reg. Tox. Pharm*. **1997**, 25, 1-5.

10. Durda, J.L.; Preziosi, D.V. Data quality evaluation of toxicological studies used to derive ecotoxicological benchmarks. *Hum. Ecol. Risk Assess*. **2000**, 6(5), 747-765.

11. Hobbs, D.A.; Warne, M.,St.J.; Markich, S.J. Evaluation of criteria used to assess the quality of aquatic toxicity data. *Integr. Environ. Assess. Manag*. **2005**, 1(3), 174-180.

12. Schneider, K.; Schwarz, M.; Burkholder, I.; Kopp-Schneider, A.; Edler, L.; Kinsner-Ovaskainen, A.; Hartung, T.; Hoffmann, S. "ToxRTool", an new tool to assess the reliability of toxicological data. *Toxicol. Lett*. **2009**, 189(2), 138-144.

13. Ioannidis, J.P.A. Why most published research findings are false. *PLOS Med*. **2005**, 2(8), 696-701.

14. Giesy, J.P; Kannan, K. Global distribution of perfluorooctane sulfonate in wildlife. *Environ. Sci. Technol*. **2001**, 35(7), 1339-1342.

15. Anastas, P.T.; Zimmerman, J.B. Design through the 12 principles of green engineering. *Environ. Sci. Technol*. **2003**, 37(5), 94A-101A.

16. Anastas, P.T.; Eghbali, N. Green Chemistry: Principles and Practice. *Chem. Soc. Rev*. **2010**, 39(1), 301-312.

17. Marty, M.S.; Allen, B.; Chapin, R.E.; Cooper, R.; Daston, G.P.; Flaws, J.A.; Foster, P.M.D.; Makris, S.L.; Myclchreest, E.; Sandler, D.; and Tyl, R.W. Inter-Laboratory Control Data for Reproductive Endpoints Required in the OPPTS 870.3800/OECD 416 Reproduction and Fertility Test. Birth Defects Res. B **2009**, 86, 470–489.

18. Jensen, K.M.; Korte, J.J.; Kahl, M.D.; Pasha, M.S.; Ankley, G.T. Aspects of basic reproductive biology and endocrinology in the fathead minnow (*Pimephales promelas*). *Comp. Biochem. Physiol. C: Toxicol. Pharmacol*. **2001**, 128, 127-141.

19. Paull, G.C.; Van Look, K.J.W.; Santos, E.M.; Filby, A.L.; Gray, D.M.; Nash, J.P.; Tyler, C.R.; Variability in measures of reproductive success in laboratory-kept colonies of zebrafish and implications for studies addressing population-level effects of environmental chemicals. *Aquat. Toxicol*. **2008**, 87, 115–126.

765 20. Segner, H. Zebrafish (*Danio rerio*) as a model organism for investigating endocrine
766 disruption. *Comp. Biochem. Physiol. C: Toxicol. Pharmacol.* **2009**, 149, 187–195.

767 21. Oehlmann, J.; Schulte-Oehlmann, U.; Tillmann, M.; Markert, B. Effects of endocrine
768 disruptors on prosobranch snails (Mollusca: Gastropoda) in the laboratory. Part 1.
769 Bisphenol A and octylphenol as xeno-estrogens. *Ecotoxicology* **2000**, 9, 383–397.

770 22. Benstead, R.S.; Baynes, A.; Casey, D.; Routledge, E.J.; Jobling, S. 17β-Oestradiol may
771 prolong reproduction in seasonally breeding freshwater gastropod molluscs. *Aquat.*
772 *Toxicol*. **2011,** 101, 326–334.

773 23. Hutchinson, T.H.; Shillabeer, N.; Winter, M.J.; Pickford, D.B. Acute and chronic effects of
774 carrier solvents in aquatic organisms: A critical review. *Aquat. Toxicol*. **2006**, 76, 69-92.

775 24. Jobling S.; Nolan M.; Tyler, C.R.; Brighty, G.; Sumpter, J.P. Widespread sexual disruption
776 in wild fish. *Environ. Sci.Technol*. **1998**, 32(17), 2498-2506.

777 25. Iwanowicz, L.R.; Blazer, V.S.; Guy, C.P.; Pinkney, A.E.; Mullican, J.E.; Alvarez, D.A.
778 Reproductive health of Bass in the Potomac, USA, drainage: Part 1. Exploring the effects
779 of proximity to wastewater treatment plant discharge. *Environ. Toxicol. Chem*. **2009**,
780 28(5), 1072-1083.

781 26. Routledge, E.J.; Parker, J.; Odum, J.; Ashby, J.; Sumpter, J.P. Some alkyl hydroxy
782 benzoate preservatives (parabens) are estrogenic. *Toxicol. Appl. Pharmacol*. **1998**, 153,
783 12-19.

784 27. Pedersen, K.L.; Pedersen, S.N.; Christiansen, L.B.; Korsgaard, B.; Bjerregaard, P. The
785 preservatives ethyl-, propyl- and butylparaben are oestrogenic in an *in vivo* fish assay.
786 *Pharmacol. Toxicol.* **2000**, 86, 110-113.

787 28. Owen, S.F.; Huggett, D.B.; Hutchinson, T.H.; Hetheridge, M.J.; McCormack, P.; Kinter,
788 L.B.; Ericson, J.F.; Constantine, L.A.; Sumpter, J.P. The value of repeating studies and
789 multiple controls: replicated 28-day growth studies of rainbow trout exposed to clofibric
790 acid. *Environ. Toxicol. Chem*. **2010**, 29(12), 2831-2839.

29 Klimisch, H-J.; Deckardt, K.; Gembardt, Chr.; Hildebrand, B.; Küttler, K.; Roe, F.J.C. Subchronic inhalation and oral toxicity of n-vinylpyrrolidone-2. Studies in rodents. *Food Chem. Toxicol*. **1997**, 35, 1061-1074.

30 Irving, E.C.; Baird, D.J.; Culp, J.M. Ecotoxicological responses of the mayfly Baetis tricaudatus to dietary and waterborne cadmium: Implications for toxicity testing. *Environ. Toxicol. Chem*. **2003,** 22(5), 1058-1064

31 Gerhardt, A. Importance of exposure route for behavioural responses in *Lumbriculus variegatus* Müller (Oligochaeta: Lumbriculida) in short-term exposures to Pb. *Env. Sci. Pollut. Res.* **2007**, 14(6), 430-434.

32. Ratcliff, D.A. Decrease in eggshell weight in certain birds of prey. *Nature* **1967**, 215, 208-210.

33. Andrew, M.N.; O'Connor, W.A.; Dunstan, R.H.; MacFarlane, G.R. Exposure to 17α - ethynylestradiol causes dose and temporally dependent changes in intersex, females and vitellogenin production in the Sydney rock oyster. *Ecotoxicology*. **2010**, 19(8), 1440-1451.

34. Brion, F.; Tyler, C.R.; Palazzi, X.; Laillet, B.; Porcher, J.M.; Garric, J.; Flammarion, P. Impacts of 17β-estradiol, including environmentally relevant concentrations, on reproduction after exposure during embryo-larval-, juvenile- and adult-life stages in zebrafish (*Danio rerio*). *Aquat. Toxicol*. **2004**, 68(3), 193-217.

35. Nash, J.P.; Kime, D.E.; Van der Ven, L.T.M.; Wester, P.W.; Brion, F.; Maack, G.; Stahlschmidt-Allner, P.; Tyler, C.R. Long-term exposure to environmental concentrations of the pharmaceutical ethynylestradiol causes reproductive failure in fish. *Environ. Health Perspect*. **2004**, 112(17), 1725-1733.

36. Hummel, D.; Loffler, D.; Fink, G.; Ternes, T.A. Simultaneous determination of psychoactive drugs and their metabolites in aqueous matrices by liquid chromatography mass spectrometry. *Environ. Sci. Technol*. **2006**, 40, 7321-7328.

37. Aherne, G.W.; Briggs, R. The relevance of the presence of certain synthetic steroids in the aquatic environment. *J. Pharm. Pharmacol*. **1989**, 41(10), 735-736.

818 38. Kolpin, D.W.; Furlong, E.T.; Meyer, M.T.; Thurman, E.M.; Zaugg, S.D.; Barber, L.B.;
819 Buxton, H.T. Pharmaceuticals, hormones, and other organic wastewater contaminants in
820 U.S. streams, 1999-2000: a national reconnaissance. *Environ. Sci. Technol*. **2002**, 36,
821 1202-1211.

822 39. Hannah, R.; D'Aco, V.J.; Anderson, P.D.; Buzby, M.E.; Caldwell, D.J.; Cunningham,
823 V.L.; Ericson, J.F.; Johnson, A.C.; Parke, N.J.; Samuelian, J.H.; Sumpter, J.P. Exposure
824 assessment of 17α-ethynylestradiol in surface waters of the United States and Europe.
825 *Environ. Toxicol. Chem.* **2009**, 28(12), 2725-2732.

826 40. Lister, A.; Regan, C.; Van Zwol, J.; Van der Kraak, G. Inhibition of egg production in
827 zebrafish by fluoxetine and municipal effluents: A mechanistic evaluation. *Aquat. Toxicol*.
828 **2009,** 95, 320-329.

829 41. Di Poi, C.; Darmaillacq, A-S.; Dickel, L.; Boulouard, M.; Bellanger, C. Effects of perinatal
830 exposure to waterborne fluoxetine on memory processing in the cuttlefish *Sepia officinalis*.
831 *Aquat. Toxicol*. **2013,** 132-133, 84-91.

832 42. Franzellitti, S.; Buratti, S.; Valbonesi, P.; Fabbri, E. The mode of action (MOA) approach
833 reveals interactive effects of environmental pharmaceuticals on *Mytilus galloprovincialis*.
834 *Aquat. Toxicol*. **2013**, 140-141, 249-256.

835 43. Guler, Y.; Ford, A.T. Anti-depressants make amphipods see the light. *Aquat. Toxicol*.
836 **2010**, 99, 397-404.

837 44. Hala, D.; Bristeau, S.; Dagnac, T.; Jobling, S. The unexpected sources of organotin
838 contamination in aquatic toxicological laboratory studies. *Aquat. Toxicol*. **2010**, 96, 314-
839 318.

840 45. Aoki, K.A.A.; Harris, C.A.; Katsiadaki, I.; Sumpter, J.P. Evidence suggesting that di-n-
841 butyl phthalate has antiandrogenic effects in fish. *Environ. Toxicol. Chem*. **2011**, 30(6),
842 1338-1345.

843 46. Geraudie, P.; Boulange-Lecomte, C.; Gerbron, M.; Hinfray, N.; Brion, F.; Minier, C.
844 Endocrine effects of the tapeworm *Ligula intestinalis* in its teleost host, the roach (*Rutilus*
845 *rutilus*). *Parasitology* **2010**, 137, 697-704.

846 47. Trubiroha, A.; Kroupova, H.; Wuertz, S.; Frank, S.N.; Sures, B.; Kloas, W. Naturally-
847 induced endocrine disruption by the parasite *Ligula intestinalis* (Cestoda) in roach (*Rutilus*
848 *rutilus*). *Gen. Comp. Endocrinol*. **2010**, 166, 234-240.

849 48. Jobling, S.; Tyler, C.R. Endocrine disruption, parasites and pollutants in wild freshwater
850 fish. *Parasitology* **2003**, 126, S103-S108.

851 49. Sures, B. How parasitism and pollution affect the physiological homeostasis of aquatic
852 hosts. *J. Helminthol*. **2006**, 80, 151-157.

853 50. Ford, A.T.; Fernandes, T.F. Letter to the Editor: Better the devil you know? A
854 precautionary approach to using amphipods and daphnids in endocrine disruptor studies.
855 *Environ. Toxicol. Chem*. **2005**, 24(5), 1019–1021.

856 51. National Toxicology Program (NTP). (2001). Final report of the endocrine disruptors low
857 dose peer review panel. In Endocrine Disruptors Low Dose Peer Review. Raleigh, NC.
858 Available at: http://ntp.niehs.nih.gov/ntp/htdocs/liason/LowDosePeerFinalRpt.pdf.
859 Accessed May 23, 2013.

860 52. Loucks, E.; Carven, M.J. Strain-dependent effects of developmental ethanol exposure in
861 zebrafish. *Neurotoxicol. Teratol*. **2004**, 26(6), 745-755.

862 53. Soeffker, M.; Stevens, J.R.; Tyler, C.R.. Comparative breeding and behavioural responses
863 to ethinylestradiol exposure in wild and laboratory maintained zebrafish (*Danio rerio*)
864 populations. *Environ. Sci. Technol.* **2012**, 46(20), 11377-11383.

865 54. Vignet, C.; Begout, M-L.; Pean, S.; Lyphout, L.; Leguay, D.; Cousin, X. Systematic
866 screening of behavioural responses in two zebrafish strains. *Zebrafish* **2013**, 10(3), 365-
867 375.

868 55. Brown, A.R.; Bickley, L.K.; Ryan, T.A.; Paull, G.C.; Hamilton, P.B.; Owen, S.F.; Sharpe,
869 A.D.; Tyler, C.R. Differences in sexual development in inbred and outbred zebrafish
870 (*Danio rerio*) and implications for chemical testing. *Aquat. Toxicol.* **2012**, 112-113, 27-38.

871 56. Nehls, S.; Segner, H. Detection of DNA damage in two cell lines from rainbow trout,
872 RTG-2 and RTL-W1, using the comet assay. *Environ. Toxicol*. **2001**, 16, 321-329.

873 57. Fenske, M.; van Aerle, R.; Brack, S.; Tyler, C.R.; Segner, H. Development and validation
874 of a homologous zebrafish (*Danio rerio* Hamilton-Buchanan) vitellogenin enzyme-linked
875 immunosorbent assay (ELISA) and its application for studies on estrogenic chemicals.
876 *Comp. Biochem. Physiol. C. Toxicol. Pharmacol.* **2001**, 129, 217-232.

877 58. Beresford, N.; Routledge, E.J.; Harris, C.A.; Sumpter, J.P. Issues arising when interpreting
878 results from an in vitro assay for estrogenic activity. *Toxicol. Appl. Pharmacol*. **2000**,
879 162(1), 22-33.

880 59. Harris, C.A.; Henttu, P.; Parker, M.G.; Sumpter, J.P. The estrogenic activity of phthalate
881 esters in vitro. *Environ. Health Perspect*. **1997**, 105(8), 802-811.

882 60. Pedersen, S.N.; Christiansen, L.B.; Pedersen, K.L.; Korsgaard, B.; Bjerregaard, P. *In vivo*
883 estrogenic activity of branched and linear alkylphenols in rainbow trout (*Oncorhynchus*
884 *mykiss*). *Sci. Total Environ*. **1999**, 233(1-3), 89-96.

885 61. Odum, J.; Lefevre, P.A.; Tittensor, S.; Paton, D.; Routledge, E.J.; Beresford, N.A.;
886 Sumpter, J.P.; Ashby, J. The rodent uterotrophic assay: Critical protocol features, studies
887 with nonyl phenols, and comparison with a yeast estrogenicity assay. *Regul. Toxicol.*
888 *Pharmacol*. **1997**, 25(2), 176-188.

889 62. Moore, A.; Scott, A.P.; Lower, N.; Katsiadaki, I.; Greenwood, L. The effects of 4-
890 nonylphenol and atrazine on Atlantic salmon (*Salmo salar* L) smolts. *Aquaculture* **2003**,
891 222, 253-263.

892 63. OECD 1998. OECD Series on Testing and Assessment Number 10. Report of the OECD
893 workshop on statistical analysis of aquatic toxicity data. [Accessed via:
894 http://www.oecd.org/env/ehs/testing/seriesontestingandassessmentpublicationsbynumber.h
895 tm ;15-10-2013]

896 64. OECD 2006. OECD Series on Testing and Assessment Number 54. Current approaches in
897 the statistical analysis of ecotoxicity data: a guidance to application. [Accessed via:
898 http://www.oecd.org/env/ehs/testing/seriesontestingandassessmentpublicationsbynumber.h
899 tm ;15-10-2013]

900 65. Watanabe, K.H.; Jensen, K.M.; Orlando, E.F.; Ankley, G.T. What is normal? A
901 characterization of the values and variability in reproductive endpoints of the fathead

902 minnow, *Pimephales promelas*. *Comp. Biochem. Physiol. C: Toxicol. Pharmacol.* **2007,**
903 146, 348–356

904 66. Brodin, T.; Fick, J.; Jonsson, M.; Klaminder, J. Dilute concentrations of a psychiatric drug
905 alter behavior of fish from natural populations. *Science* **2013**, 339(6121), 814-815.

906 67. Sumpter, J. P.; Jobling, S. (1995). Vitellogenesis as a biomarker for oestrogenic
907 contamination of the aquatic environment. *Environ. Health Perspect*. **1995**, 103 (Suppl. 7),
908 173-178.

909 68. Katsiadaki, I.; Morris, S.; Squires, C.; Hurst, M. R.; James, J. D.; Scott, A. P. A sensitive,
910 in vivo test for the detection of environmental antiandrogens, using the three-spined
911 stickleback (*Gasterosteus aculeatus*). *Environ. Health Perspect*. **2006**, 114(Suppl. 1), 115-
912 121.

913 69. Vandenberg, L.N.; Colborn, T.; Hayes, T.B.; Heindel, J.J.; Jacobs, D.R., Jr.; Lee, D-H.;
914 Shioda, T.; Soto, A.M.; vom Saal, F.S.; Welshons, W.V.; Zoeller, R.T.; Myers, J.P.
915 Hormones and endocrine-disrupting chemicals: low-dose effects and nonmonotonic dose
916 responses. *Endocr. Rev*. **2012**, 33(3), 378-455.

917 70. Scott, A.P. Do mollusks use vertebrate sex steroids as reproductive hormones? II. Critical
918 review of the evidence that steroids have biological effects. *Steroids* **2013**, 78(2), 268-281.

919 71. Joint Code of Practice for Research. Issued by BBSRC; DEFRA; FSA; NERC (UK). **2003**.
920 [Accessed via:http://www.bbsrc.ac.uk/organisation/policies/position/policy/joint-code-of-
921 practice-for-research.aspx ; 15-10-13]

922 72. Russell, J. F. If a job is worth doing, it is worth doing twice. *Nature* **2013**, 496, 7.

923 73. Burkhardt-Holm, P.; Giger, W.; Guttinger, H.; Ochsenbein, U.; Peter, A.; Scheurer, K.;
924 Segner, H.; Staub, E.; Suter, M.J.F. Where have all the fish gone? *Environ. Sci. Technol*.
925 **2005**, 39(21), 441A-447A.

926 74. Ginebrada, A.; Muñoz, I.; López de Alda, M.; Brix, R.; López-Doval, J.; Barceló, D.
927 Environmental risk assessment of pharmaceuticals in rivers: Relationships between hazard
928 indexes and aquatic macroinvertebrate diversity indexes in the Llobregat River (NE
929 Spain). *Environ. Int*. **2010**, 36, 153-162.

930 75. Weed, D.L. Weight of evidence: A review of concept and methods. *Risk Analysis* **2005**,
931 25(6), 1545-1557.

932 76. Ioannidis, J.P.A. Contradicted and initially stronger effects in highly cited clinical
933 research. *JAMA-J. Am. Med. Assoc.* **2005**, 294(2), 218-228.

934 77. Brown, R. P. Greer, R.D. Mihaich, E.M. and Guiney. P.D. A Critical Review of the
935 Scientific Literature on Potential Endocrine-Mediated Effects in Fish and Wildlife.
936 *Ecotox. Environ. Safe*. **2001,** 49, 17-25

937 78. Carlsen, E.; Giwercman, A.; Keiding, N.; Skakkebaek, N.E. Evidence for decreasing
938 quality of semen during past 50 years. *BMJ* **1992**, 305, 609-613

939 79. Popper, K.R. The Logic of Scientific Discovery (translation of *Logik der Forschung*, first
940 published in 1935). Taylor & Francis e-Library, **2005**, ISBN 0-203-99462-0 Master e-
941 book. 545 pp.

942 80. Popper, K.R. Conjectures and Refutations: The Growth of Scientific Knowledge.
943 Routledge, London, **1963,** ISBN 0-415-28594-1. 431 pp.

944 81. Goldacre, B. Bad Science. Publ. Fourth Estate (London); **2009,** ISBN: 978-0-00-728487-0.
945 370 pp.

946 82. Knight, J. Null and void. *Nature* **2003**, 422, 554-555.

947 83. Turner, E. H.; Matthews, A. M.; Linardatos, E.; Tell, R. A.; Rosenthal, R. Selective
948 publication of antidepressant trials and its influence on apparent efficacy. *N. Engl. J. Med.*
949 **2008**, 358, 252-260.

950 84. Larsen, P-O.; von Ins, M. The rate of growth in scientific publication and the decline in
951 coverage provided by Science Citation Index.*Scientometrics* **2010,** 84(3), 575–603.