Considering aspects of the 3Rs principles within experimental animal biology

Lynne U. Sneddon[1]*, Lewis G. Halsey[2], and Nic R. Bury[3]

[1]Institute of Integrative Biology, University of Liverpool, The BioScience Building, Liverpool, L69 7ZB, UK

[2]Department of Life Sciences, University of Roehampton, London, SW15 4JD, UK

[3]University of Suffolk, Faculty of Health Sciences and Technology, James Hehir Building, Neptune Quay, Ipswich, IP4 1QJ, Suffolk, United Kingdom.

*Author for correspondence: lsneddon@liverpool.ac.uk

ABSTRACT

The 3Rs – Reduction, Replacement and Refinement – are embedded into the legislation and guidelines governing the ethics of animal use in experiments. Here, we consider the advantages of adopting key aspects of the 3Rs into experimental biology, represented mainly by the fields of animal behaviour, neurobiology, physiology, toxicology and biomechanics. Replacing protected animals with less sentient forms or species, cells, tissues or computer modelling approaches has been broadly successful. However, many studies investigate specific models that exhibit a particular adaptation, or a species that is a target for conservation, such that their replacement is inappropriate. Regardless of the species used, refining procedures to ensure the health and wellbeing of animals prior to and during experiments is crucial for the integrity of the results and legitimacy of the science. Although the concepts of health and welfare are developed for model organisms, relatively little is known regarding non-traditional species that may be more ecologically relevant. Studies should reduce the number of experimental animals by employing the minimum suitable sample size. This is often calculated using power analyses, which is associated with making statistical inferences based on the *P* value, yet *P* values often leave scientists on shaky ground. We endorse focussing on effect sizes accompanied by confidence intervals as a more appropriate means of interpreting data; in turn, sample size could be calculated based on effect size precision. Ultimately, the appropriate employment of 3Rs principles in experimental biology empowers scientists in justifying their research, and results in higher-quality science.

KEY WORDS: Animal welfare; Environmental enrichment; Replacement; Reduction; Refinement; Toxicology

INTRODUCTION

Animal research is essential for the advancement of new technologies and medicines crucial to improving human and animal health. It is also vital for our understanding of fundamental animal biology, as well as essential areas of applied animal science, such as how animals function in the face of climate change or anthropogenic disturbance. Further, studies exploring animal health and welfare enable us to manage captive animals more effectively, and prevent poor welfare that leads to disease. Against this backdrop of necessary animal research, scientists are increasingly asked to justify their experimental approaches when using protected animals. This is partly driven by demands from the general public that the use of animals in research is moral and ethically justifiable. A recent poll in the United States demonstrated that 50% of the public were opposed to the use of animals in research (Pew Research Center, 2015). In 2015, nine European countries presented a petition to the European Commission (EC) to ban animal research. However, the EC opposed this movement, but responded by stating that ethical justification and adoption of the 3Rs (Reduction, Replacement and Refinement) is a must for experimental studies (EC, 2015). Of course, it is in scientists' interest to adopt an ethical and humane approach to husbandry and experimental design, since healthy animals produce robust, reliable results, underlying valid scientific outputs. For example, improved husbandry and handling of rodents reduces stress, and this leads to less variable data and more meaningful results (Hurst & West, 2010; Singhal et al., 2014). Embedding the 3Rs principles into scientific planning and execution therefore directly benefits data quality.

The 3Rs concepts were first developed by Russell and Burch (1959) and have become rooted in legislation and guidelines concerning animal experimentation in many countries (Fig. 1). Refinement involves either reducing the invasiveness of a technique or improving animal welfare and health during scientific studies. This can be achieved through better assessment of the animal's state or improved husbandry and housing. Reduction concerns minimising the number of animals used to effectively achieve the goals of an experiment. Replacement involves the adoption of alternatives to protected animals – such alternatives may be non-protected species or immature forms; cell lines or cultured tissues; mathematical modelling of existing data sets or conceptual data; or the use of humans, their tissues or their cells (with permission). Many funding bodies in the UK and Europe now have dedicated application sections on each of the 3Rs that must be completed, thus requiring justification of the use of protected animals. In this Commentary, we discuss current knowledge and recent developments in the 3Rs relevant to the field of experimental animal biology. Our views are fuelled by a recent symposium funded by the Society for Experimental Biology (SEB) and co-funded by the Association for the Study of Animal Behaviour (ASAB), held in London in 2016 (Knight, 2016).

REFINEMENT

Refinement is an integral component of improving laboratory animal welfare, which is vital for healthy biological functioning and a normal behavioural repertoire. Therefore, refining procedures to reduce their invasiveness or the degree of stress they cause and perfecting housing and husbandry should be the goal of any scientist. However, some animal groups have received relatively little attention in this area, resulting in less-developed tools or knowledge to assess their health and welfare (e.g. pain assessment is highly developed for mammals compared with other animal groups, Sneddon et al., 2014; Sneddon, 2015). Additionally, good husbandry practices improve animal wellbeing and the reliability of experimental results; thus, it is important to know what different species require in their environment in order to maintain their health and welfare. The necessity to develop refinement recommendations and good laboratory practices for both traditional and non-traditional species has driven this vibrant research field.

Environmental enrichment

The EC Directive (2010) proposes that all protected animals should have enriched environments in which to live. Enrichment can involve physical objects that either make an environment more complex (e.g. plastic plants, gravel substrate and overhead cover in a fish tank; Pounder et al., 2016) or can be used by the animals (e.g. perches in bird enclosures; Kalmer et al., 2010). Alternatively, enrichment can involve appropriate social housing (e.g. gregarious species not kept in isolation or territorial species held in groups), apparatus to allow exercise (e.g. rodent running wheel), nutritional enrichment (e.g. diversity of feeding regimens) and sensory stimulation (visual, olfactory and aural; see Singhal et al., 2014). Understanding the appropriate type of enrichment can have tremendous benefits, reducing stress and the inter-individual variation in behavioural and physiological variables (Singhal et al., 2014). Preference testing can provide insight into what an animal would choose, although this depends on the resources tested and so caution should be applied. As an example of the effect that refinement can have, it is known that zebrafish have relatively smaller brains when reared in barren conditions compared with enriched tanks (DePasquale et al., 2016), which might indicate chronic sensory deprivation. This raises both ethical issues and concerns about the veracity of neurobiological and behavioural research conducted on such individuals. Indeed, zebrafish housed for seven months in barren tanks choose to interact with enrichment when given the option (Schroeder et al., 2014). In addition, rainbow trout housed in enriched tanks recover from stressors more quickly (Pounder et al., 2016; Fig. 2), and it is known that background colour influences growth rates, physiological stress and behaviour in *Xenopus* (Holmes et al., 2016; Fig. 2). These studies can have real impact upon husbandry protocols, which are essential for guaranteeing the health of experimental animals.

Refining experimental procedures

Refinements to reduce the invasiveness of a procedure can be as simple as improving the manner in which animals are handled. Hurst and West (2010) showed that handling mice by allowing them to voluntarily sit in a cupped hand or enter a plastic tunnel reduced anxiety and stress compared with the traditional method of picking up mice by the tail. Non-invasive imaging of molecular responses – using techniques such as magnetic resonance imaging (MRI), positon emission tomography (PET), single positron emission computed tomography, ultrasound and optical imaging (bioluminescence and fluorescence) – circumvents the need to humanely kill or biopsy animals for samples: imaging can be performed *in vivo* and in real time, negating the necessity for sampling groups of animals at various time points (O'Farrell et al., 2013). These imaging techniques can monitor molecular and cellular changes non-invasively in intact animals, although repeated anaesthesia may be necessary and is likely to be stressful. These approaches have facilitated significant advances in preclinical research and, consequently, fewer animals are required, individuals can be tracked over a longer time period and they are not subjected to invasive, potentially painful, procedures (reviewed in O'Farrell et al., 2013). Thus, there is scope for these non-invasive technologies to be applied to a wide variety of contexts in experimental animal biology, but there is a substantial economic cost to employing imaging techniques.

Assessing welfare is key to ensuring that animals are healthy before, during and after experiments where post-surgical care is vital. Laboratory rodents have been well studied, and key behavioural changes (Sneddon et al., 2014), as well as the more recent grimace scales for rats, mice and rabbits, can be used to gauge their pain levels (Langford et al., 2010; Sotocinal et al., 2011; Keating et al 2012 see NC3Rs, 2017 for scales). Extensions of the grimace scales have been applied to horses (Dalla Costa et al 2014), and are likely to be applicable to other non-model mammals. Although non-mammalian animals are less well studied, advances are being made. For example, fin clipping of zebrafish, a routine procedure for genomic screening, is normally conducted under anaesthesia, but analgesics are not routinely applied. However, Schroeder and Sneddon (2017) demonstrated substantial changes in behaviour after fin clipping that were ameliorated by pain-relieving drugs (Fig. 2). Rather than injecting these relatively small fish, this study showed that adding the drugs to the tank water effectively reduces pain, and this could be extrapolated to other aquatic species. Further research is required to develop robust indicators of welfare and health in a variety of common laboratory models, since species can differ in their expression of poor welfare. Automated monitoring of animal health through non-invasive use of behavioural recording equipment would be ideal (e.g. Deakin et al., 2017 MS submitted; Rushen et al., 2012; Noldus, 2016).

Refinement for non-traditional experimental species

Although much is known about refinement in model organisms, many experimental animal biologists use non-traditional species to answer important and ecologically relevant physiological questions. While refinements therefore need to be employed on a species-by-species basis, general principles from model organisms should make a good starting point from which welfare testing can begin. A further confounding issue is that many experiments take place in the field rather than a laboratory. General principles of refinement can be applied, with the capture, handling, tagging and sampling of animals done in the most humane way. If invasive methods are appropriate, ways to improve animal welfare and health can be considered. Obviously it can be difficult to assess health and welfare if the animals are returned to their natural environment. However, recapture studies (e.g. intraperitoneal tags, Gardner et al., 2015; radio collars, Hopkins & Milton, 2016) and assessment of subsequent breeding success (Phillips et al., 2003) can provide some measure of survivorship. This is pertinent to understanding how previous procedures may have affected the animals, given that survival and reproduction can be affected by vulnerability to predators, and the ability to harvest resources and to cope with intraspecific agonistic interactions.

REPLACEMENT

Replacement in a comparative physiology context

Studying physiological adaptation or the response of vulnerable species to environmental perturbations is at the core of comparative and conservation physiology. Krogh’s principle states that “for such a large number of problems there will be some animal of choice, or a few such animals, on which it can be most conveniently studied.” Thus, often in the comparative and conservation disciplines, animals cannot be easily replaced, and reduction and refinement are more realistic ethical strategies. However, the evolutionary conservation of physiological traits throughout the eukaryotes means that alternative non-vertebrate organisms can provide valuable information where processes are shared with sentinel organisms, enabling experimental biologists to embrace the replacement approach. For example, the cell behaviour of the soil-dwelling amoeba *Dichtyostelium* can be used as a rapid screen for the effects of medicinal products (Otto et al., 2016). As another example, the simplified neuronal network of the pond snail *Lymnaea stagnalis* can be used to study the neurobiological processes involved in decision making and motivational state (Crossley et al., 2016), as well as the effects of stressors on memory formation (Lukowiak et al., 2014). In addition, e*x vivo* systems, organoid cell cultures and immortalised cell lines are often utilised and, although they cannot replace the complex interactions between tissues in intact vertebrates, they can provide insight when investigating intra- and inter-cellular biological processes or tissue-level responses. The key is to find the right non-vertebrate model organism or *in vitro* system to answer the question of interest – a concept that will be very familiar to a comparative physiologist audience.

Factors driving replacement research

Recent advancements in replacement approaches within experimental biology have occurred in identifying alternatives to the use of vertebrates in regulatory tests; tests which are required by law as part of any chemical's risk assessment, such as OECD 305 (Bioaccumulation in Fish: Aqueous and Dietary Exposure) and OECD 203 (Fish, Acute Toxicity Test) (Lillicrap et al., 2016) for aquatic environmental risk assessment. For example, within Europe, the regulations concerning the Registration, Evaluation, Authorisation and restriction of Chemicals (REACH) have resulted in many thousands of chemicals requiring further animal testing. Though the European Union (EU) did not ban animal testing as part of REACH, animal welfare legislation requires the incorporation of the 3Rs principles. This has led to a strong impetus for regulatory authorities to accept replacement test systems as part of risk assessment evaluation (Burden et al., 2016). Acceptance requires a rigorous scientific understanding about whether such alternatives adequately reflect physiological processes observed in intact adult fish.

Suitable replacements

Embryonic and young forms

The young forms of many species are not considered to suffer. Thus, the United Kingdom Animals (Scientific Procedures) Act 1986 and European Directive 2010/63 specifies that fish become a protected animal once they are capable of independent feeding [e.g. zebrafish after 120 hours post fertilization (120 hpf) at 28°C; Strähle et al., 2012]. However, this is not case for all countries (Box 1). This threshold is based upon the concept that, before this stage, fish are not fully developed and are unable to experience external stimuli, meaning there is no obligation to report the number of fish embryos used. But recent studies show that 120 hpf larval zebrafish respond to noxious stimuli, and that this is ameliorated by administration of pain-relieving drugs (Lopez Luna et al., 2017a; 2017b). From a regulatory perspective, the fish embryo toxicity (FET) test, which lasts for 96 hpf for zebrafish (Henn and Braunbeck, 2011), correlates well with adult acute toxicity (Lammer et al., 2009, Scholz et al 2014), and the OECD have approved OECD 236 test FET guidelines (Busquets et al., 2014).

In basic research, embryos, including those from chickens, have been used extensively to study the development and functioning of organs within the context of a whole organism (e.g. Tazawa et al 2002). Zebrafish embryos are now used for many basic physiological and behavioural studies; for example, sophisticated video imaging packages can be used to record their movement in response to chemical exposure (e.g. Nüßer et al., 2016), translucent fish embryos provide an ideal model to study cardiovascular function (Incardona and Scholz,, 2016, Yozzo et al., 2013), and genetic manipulation has enabled a study of the functional regulation of ionoregulation (Cruz et al 2013, Guh et al 2015).

Cell lines and organoid cultures

The EU’s decision to ban animal testing for cosmetics ingredients (EU1223/2009) provided the momentum to develop alternative mammalian *in vitro* models to identify chemicals that pose a health risk. In addition, there is a long history of the development of fish cell lines from a variety of tissues and organisms (Bols et al., 2005). For example, the cell line derived from the gills of rainbow trout (RTgill-W1) (Bols et al., 1994) is promising as a replacement for OECD203 (Tanneberger et al., 2013; Lillicrap et al., 2016) and for chronic toxicity tests. But further basic mechanistic understanding of how cell growth in culture correlates with somatic growth in a whole fish is necessary for *in vitro* to *in vivo* extrapolation (Stadnicka-Michalak et al., 2015).

Extensive research has gone into mammalian tissue and stem cell-derived organoid cultures for disease and drug development research (Liu et al 2016; Muthuswamy, 2017). The time it takes to develop these types of *in vitro* model may make them unsuited to comparative physiological studies, but they are of interest for basic research because these systems better replicate in situ tissue physiology than do 2-D cell cultures.

A further development is the potential replacement of the OECD 305 test, which has led to technical advancements in fish *in vitro* organoid cultures (Baron et al., 2012, Schnell et al., 2016). Data on the basic characteristics of chemical uptake, metabolism and excretion by these organoid cultures provide the scientific rigor which supports their use in alternative testing procedures for bioconcentration studies. For example, a primary fish gill culture technique has been developed from which two fish (subject only to humane killing) can produce between 48 and 72 cell culture inserts: harvesting of cells for primary culture in the UK is not defined as a procedure, so this approach replaces the use of animals (Schnell et al., 2016). The system has been used to study branchial physiological processes, such as ammonia excretion and endocrine control of epithelial tight junction formation (see Bury et al., 2014). The liver is the main site of metabolism and excretion, and a number of *ex vivo* and *in vitro* methods (e.g. liver slices, primary hepatocytes, S9 fraction and cell cultures) have been deployed to estimate the ability of the liver to metabolise compounds (see Weisbrod et al., 2009). Recent advances in liver organoid cell culture techniques generate three-dimensional spheroidal hepatocytes (Uchea et al., 2013; Baron et al., 2012) that better represent the metabolic capabilities of the intact liver (Baron et al., 2017). Encouragingly, there are a number of studies that extrapolate the hepatocyte *in vitro* biotransformation data to *in vivo* scenarios (Nichols et al., 2006, 2007; Cowan-Ellsberry et al., 2008), allowing derivation of bioconcentration factors BCF (Nichols et al., 2013).

High-throughput FET or *in vitro* screens are being used as part of the Adverse Outcome Pathways (AOP) conceptual framework to identify molecular initiating events (MiE) induced by a compound (Ankley et al., 2010, Wittwehr, et al., 2017). AOPs aim to use empirical mechanistic data at lower levels of biological organisation (e.g. cells) to predict higher level effect (e.g. whole-organism toxicity). MiE identification can uncover chemicals of unknown toxic action or off-target effects (Villeneuve et al., 2014). Ultimately, it is envisaged that the AOP concept can lead to computer-based predictive models to assist environmental risk assessment (Wittwehr et al., 2017), replacing many, if not all, animals used in regulatory procedures. The AOP concept is a wonderful example of how toxicology and physiology are intertwined. The wealth of data on the downstream effects of stimulating a receptor within a cell, whether by a synthetic or natural chemical, will potentially aid the identification of regulatory mechanisms and feedback control of physiological processes.

REDUCTION

'Reduction' proposes that researchers reduce the number of experimental animals used such that just enough data and no more are obtained to give sufficiently informative results. Experimental designs that incorporate stronger perturbations or support greater measurement precision improve the signal-to-noise ratio of the data analysis (see Halsey, 2007), which enables the sample size to be reduced. Put simply, cleaner and clearer experiments require fewer experimental animals for the analysis to be robust. Authors such as McClelland (2000), Eng (2003) and de Boo and Hendriksen (2005) suggest various avenues for improving measurement precision, including: (1) using more reliable measures, repeating measurements, using experienced staff and well-honed experimental procedures; (2) including measures of concomitant variables (such as body mass) to account for measurable variability; (3) experimentally reducing variability, e.g. by working with one age group or sex [the latter pertains to both study animal and researcher (Sorge et al., 2014)]; however, this reduces the generalizability of the findings (Würbel, 2000), and thus has been disallowed by the National Institutes of Health in the US; (4) increasing the variance in the predictor variable(s); for example, including animals with a greater age range if studying correlates of senescence; (5) using subjects as their own controls (e.g. testing each animal after a saline injection as well as a hormone injection). However, we argue that there is an over-arching research problem that typically supersedes tweaks made to experimental designs – the focus on the ubiquitous *P* value when interpreting data analyses. Regardless of the experimental design, due to some intrinsic frailties of *P* value-based data analysis, such studies will usually have employed a sample size too small for robust conclusions to be made.

Reduction… in the use of the P value for data interpretation

Typically, the number of animals included in an experiment is determined using statistical power analysis to calculate the sample size required for an estimated probability of correctly rejecting the null hypothesis. Statistical power of 80% is the norm (Cohen, 1988), which means that when the null hypothesis being tested is false, a statistically significant result will be reported 80% of the time. The number of animals necessary to achieve 80% power in a well-designed experiment is deemed ‘required’ and is thus ethically acceptable according to the 3Rs philosophy. Power analysis is intimately tied to the *P* value, since the latter is used to decide whether the null hypothesis is rejected or not (and thus whether a finding is deemed ‘significant’).

Recently it has become evident that many scientific findings are not reproducible (Baker, 2016; Collaboration, 2015), shaking the pursuit of science to its core (Economist, 2013; Freedman et al., 2015; Mobley et al., 2013; Ioannidis, 2005). To conduct a study on animals that is not reproducible is fundamentally counter to the 3Rs principle; animals have been used in fruitless and even misleading experiments (Button et al., 2013). Many authors have discussed how to combat irreproducibility (Freedman et al., 2015; Ioannidis et al., 2015; McNutt, 2014; Nosek et al., 2015; Reproducibility-Initiative, 2014; Woolston, 2014). While only a few publications have targeted the *P* value as a potential culprit, these papers have compellingly argued that over-reliance on *P* values for data interpretation is helping drive irreproducibility (Colquhoun, 2014; Cumming, 2008; Halsey et al., 2015; Nuzzo, 2014; although other factors, such as lack of homogeneity in protocols, can contribute). Crucially, this is the case even when statistical power is 80%.

First, interpretation of data based on *P* values will often produce misleading conclusions owing to the false discovery rate, which is the probability of calculating a *P* value sufficiently low to claim 'significance' when in fact the null hypothesis is true (Colquhoun, 2014). Assuming *P* values <0.05 are those considered 'significant', and that the proportion of studies conducted where the null hypothesis is false is 10%, the false discovery rate is at least 36% according to Colquhoun (2014) and Sellke et al. (2001) (although it could be less in research fields where scientists conduct the experiments they anticipate are likely to return 'significant' results; Wacholder et al., 2004). Second, models have highlighted that *P* typically varies dramatically between replicates of a study, and this 'fickleness' in *P* is present even when statistical power is quite high (Cumming, 2008; Halsey et al., 2015).

In the biological disciplines, average statistical power, including in fields such as neuroscience (Button et al., 2013; Macleod et al., 2009) and behavioural ecology (Jennions and Møller, 2003), is consistently less than 50% and often considerably lower (Smith et al., 2011). Such low power exacerbates the problem of false discoveries and *P*'s inherent fickleness. Simply put, when a study reports a *P* value indicating strong evidence against the null hypothesis, there is every chance that a replication of that study would report a *P* value indicating much less evidence against the null hypothesis (and *vice versa*). Furthermore, studies that do yield significant results tend to exaggerate the true effect size, and this is exacerbated when statistical power is low (Button et al., 2013; Halsey et al., 2015). Consequently, the interpretation of one-off experiments based on the *P* value may explain why so many studies are irreproducible (Halsey et al., 2015).

There are further valid reasons to question the usefulness of *P* for data interpretation (Cohen, 1994; Tressoldi, 2013). Of particular relevance is that significance testing of the null hypothesis only allows us to ask a very limited question about our data, simply 'is there or isn't there?'. For example, 'is there a difference in metabolic rates between two mouse strains?' or 'is there a relationship between metabolic rate and risk-taking behaviour?'. Given a large enough study we can always find a difference, or a relationship, to some degree (Cohen, 1994; Loftus, 1993), and so answering these questions tells us very little about our data.

Once these sobering facts about the *P* value have sunk in, the only conclusion open to us is to greatly reduce, or even discard, our use of *P* in statistical analyses. Although *P* values are entrenched within the research culture of experimental biology, when animal health and welfare is at stake it is surely unethical to continue using an inadequate statistical index for data interpretation. In turn, the use of power analysis to calculate the necessary numbers of experimental animals becomes questionable.

What alternatives do we have?

There are several alternatives available, such as Bayesian analysis and the Akaike Information Criterion, although no method is perfect (Ellison et al., 2014). We suggest that instead of focussing on the standard approach of ‘is there or isn’t there?’, it is more illuminating to ask ‘how big is the difference?’ or ‘how strong is the relationship?’, coupled with the question ‘how precise is the estimate of the magnitude of the difference or relationship?’. The answers to these two questions not only tell us if there is a difference or a relationship, but much more by also informing us of its (estimated) magnitude coupled with how precise that estimate is likely to be; all in all – a much better use of experimental animals. The most straightforward way to analyse our data in order to answer these two questions is first to calculate the effect size – the size of the difference between conditions or the strength of the correlation between two variables. Second, because our experiment only estimates rather than measures the population effect size, we should also provide the confidence intervals for that estimate, to indicate how precisely the effect is known (Cumming, 2008; Halsey et al., 2015; Johnson, 1999; Nakagawa and Cuthill, 2007).

More is less

When basing data interpretation on effect size estimates and their precision, the number of experimental animals required should relate to how precisely we need our sample to represent the population. ‘Planning for precision’ calculates the sample size required for the effect size needed in order to provide a defined degree of precision, based on the predicted effect size and variance within the data (Maxwell et al., 2008). Currently, few studies take this approach – when presented, 95% confidence intervals are often large, showing poor precision; a fact that may explain the omission of confidence intervals from many figures. But it is important that we are aware of the level of precision (or otherwise) in our experimental results (rather than hiding it behind a *P* value; Cumming, 2008); if necessary we should adjust our sample size accordingly. Designing experiments around precision rather than power analysis is likely to increase experimental animal numbers. However, if the results are more meaningful then this should reduce the number of experiment repetitions needed, hence reducing experimental animal numbers in the long run.

Perhaps the strongest argument for analyses based on effect sizes combined with confidence intervals is that where multiple studies on a particular question have been published and this information included, it can then be combined in a meta-analysis, enabling us to home in on the statistical truth (e.g. Sena et al., 2010). Typically, the confidence intervals around an effect size calculated from meta-analysis are much smaller than those of the individual studies (Cohn and Becker, 2003), thus giving a much clearer picture about the true, population-level effect size (Fig. 3). Indeed, sample sizes required to detect effect sizes with suitable precision are often prohibitive or deemed unethical for individual researchers, necessitating future meta-analyses (Maxwell et al., 2008). And meta-analyses are efficient on experimental animal numbers. First, where a meta-analysis is undertaken solely on previously published data, it represents an experiment-free study; the ultimate in 3Rs Reduction. Second, where multiple studies of a similar nature are conducted on a relatively intractable research question (Nature Magazine, 2016), within as well as across publications (Harris et al., 2014), meta-analyses give good indication of when such replicate experiments are no longer necessary (Fig. 3). However, the Achilles heel of the meta-analysis is the ‘file drawer phenomenon’. Data on animal experiments are often filed away and not published if found to be ‘non-significant’ (Dwan et al., 2013) – another example of the need to remove the focus from the *P* value. Yet the results of all robust and relevant studies provide invaluable grist to the mill for a future meta-analysis, regardless of their supposed ‘interest’, and meta-analyses often highlight approximate agreements between multiple studies that appear contradictory when viewed as providing either ‘significant’ or ‘non-significant’ findings. Indeed, filing away uninteresting data skews the distribution of published data and distorts the truth, which in the long run will lead to a greater overall number of animals being subjected to experiments. It is therefore essential for 3Rs Reduction, and for the pursuit of science in general, that all valid experimental data are published. Fortunately, there are progressively more journals that explicitly judge whether a submission is suitable for publication on merit alone without consideration of impact. And for those researchers who insist on *P* value-based interpretations, the revised version of the European code of research integrity states that non-significant results should be treated as valid findings worthy of publication {Wissenschaftsstiftung, 2017 #4816; Box 2}; a standard that the EU’s Horizon 2020 programme now expects its recipients to abide by.

CONCLUSIONS

Here, we have highlighted the benefits of adopting the 3Rs into experimental biology: there are advantages for the quality of data obtained, the robustness of the experimental design – including statistical analyses – and the validity of the scientific outputs. Adopting an ethical approach allows researchers to justify their studies not only to legislators and ethics committees but also to funding bodies and the public.

Refinement of both husbandry practices and experimental design is an important aspect of the 3Rs. Developing optimal husbandry and housing to ensure animal health and welfare and a means of monitoring animal welfare before, during and after experiments is paramount. Additionally, experimental design should be carefully thought through and possibly logged in a database prior to the study commencing. NC3Rs have developed an online tool – the Experimental Design Assistant (EDA, 2017) – to assist researchers in developing their approach and to encourage randomisation and blinding where possible to prevent bias. Reproducibility and translatability of published studies has recently come under scrutiny, and where this is due to the lack of full reporting of methods, many journals are tackling this via adopting the ARRIVE guidelines, using a checklist to ensure that all experimental details are provided to allow researchers to fully replicate studies (ARRIVE, 2017). To encourage ethical thinking, we propose that all journals reporting animal research could ask authors to include a section on ethical justification of the study so that the 3Rs thought-process is clear (some journals already do).

In terms of Reduction, there is a conflict between minimising the number of animals used versus recent revelations that published results may not be robust. How can a balance be struck between keeping animal use as low as possible while including a large enough ‘N’ to ensure the study was worth doing? In debating this question it is counter-productive to couch it within the concept of power analysis and implicitly therefore the fickle *P* value. We need to put the health and welfare of animals ahead of our statistical traditions. In turn, when designing experiments we should plan for precision; we urge biology journals to encourage this analysis rather than requesting power analysis information as they do at present. For authors, we suggest some draft text that could form the basis of a statement included in the Methods section of a manuscript to highlight and justify the authors’ focus on statistical analyses other than the *P* value (Box 2).

The biggest Reduction sin of all is not publishing our data – animals have been used and zero knowledge accumulated. We must strive to publish all results, however interesting or otherwise we consider them to be, to make full use of the experimental animals and to maximise the accuracy of future meta-analyses. Journals publishing non-significant results and demanding high clarity are invaluable in supporting this endeavour, ensuring the lives of all animals used are respected.

Developments in the use of non-protected species and young forms alongside the validation of cell and tissue preparations in a variety of contexts leave much scope for considering Replacement. Other options, such as the use of human volunteers (e.g. Halsey et al. 2017), human samples or modelling of existing data sets, may avoid animal use. However, it is crucially important that when animals are used the species chosen is relevant to the question being addressed; the careful choice of model underpins the utility of the scientific outcomes from any study. Therefore, Relevance could be considered as a 4th R. The importance of Relevance is highlighted by scientists that, for example, interrogate questions at the species-specific level, particularly where adult forms cannot be replaced by juveniles. In this situation, Replacement is not an R that can be deployed. In turn, Refinement and Reduction become all the more important levers to pull in seeking to maximise the health and welfare of the experimental animals.

Acknowledgements

We are grateful to the organisations that sponsored and made contributions to the SEB Animal Section Symposium “Implementing the 3Rs: Improving experimental approaches in animal biology” including ASAB Workshop funding; contributions to speakers’ expenses and/or provision of speakers by National Centre for the 3Rs (NC3Rs UK, Nathalie Percie du Sert and Mark Prescott); Royal Society for the Prevention of Cruelty to Animals (RSPCA, Penny Hawkins); Laboratory Animal Science Association (LASA, Caroline Chadwick); Biotechnology and Biological Sciences Research Council (BBSRC, Lydia Darragh) and Understanding Animal Research (UAR, Wendy Jarrett); and to ASAB, RSPCA and Universities Federation for Animal Welfare (UFAW, Huw Golledge) for sponsoring presentation prizes. We also thank the participants of this symposium for their presentations and open sharing of ideas. We are grateful to Gordon Drummond for his feedback on the Reduction section of this article, and to Malcolm Macleod for help in sourcing a suitable example of a cumulative meta-analysis.

Competing interests

The authors declare no competing interests.

References

Ankley, G. T., Bennett, R. S., Erickson, R. J., Hoff, D. J., Hornung, M. W., Johnson, R. D., Mount, D. R., Nichols, J. W., Russom, C. L., Schmieder, P. K., Serrrano, J. A., Tietge, J. E. and Villeneuve, D. L. (2010). Adverse outcome pathways: a conceptual framework to support ecotoxicology research and risk assessment. Environ. Toxicol. Chem. **29**, 730-741.

ARRIVE (2017). http://www.nc3rs.org.uk/arrive-guidelines

Baker, M. (2016). 1,500 scientists lift the lid on reproducibility. *Nature* **533**, 452-454.

Baron, M. G., Mintram, K. S., Owen, S. F., Hetheridge, M. J., Moody, A. J., Purcell, W. M., Jackson, S. K. and Jha, A. N. (2017). Pharmaceutical Metabolism in Fish: Using a 3-D Hepatic In Vitro Model to Assess Clearance. *PLoS One*. **12**:e0168837.

Baron, M. G., Purcell, W. M., Jackson, S. K., Owen, S. F. and Jha, A. N. (2012).Towards a more representative in vitro method for fish ecotoxicology: morphological and biochemical characterisation of three-dimensional spheroidal hepatocytes. *Ecotoxicol*. **21**, 2419-2429.

Bols, N. C., Barlian, A., Chirinotrejo, M., Caldwell, S. J., Geogan, P. and Lee, L. E. J. (1994). Development of a cell line from the primary cultures of rainbow trout, Oncorhynchus mykiss (Walbaum), gills. *J. Fish Dis*. **17**, 601-611.

Bols, N. C., Dayeh, V. R., Lee, L. E. J. and Schirmer, K. (2005). Use of fish cell lines in the toxicology and ecotoxicology of fish. In *Biochemistry and Molecular Biology of Fishes* Vol 6, Eds T. M. Moon & Mommensen, T. P. Elsevier, Amsterdam, Netherlands.

Boos, D. D. and Stefanski, L. A. (2011). P-Value precision and reproducibility. *The American Statistician* **65**, 213-221.

Burden, N., Benstead, R., Clook, M., Doyle, I., Edwards, P., Maynard, S.K., Ryder, K., Sheahan,, D., Whale, G., van Egmond, R., Wheeler, J.R. and Hutchinson T.H. (2016).Advancing the 3Rs in regulatory ecotoxicology: A pragmatic cross-sector approach. *Integr. Environ. Assess. Manag*. **12**, 417-421.

Bury,N. R., Schnell, S. and Hogstrand, C. (2014). Gill cell culture systems as models for aquatic environmental monitoring. *J. exp. Biol*. **217**, 639-650.

Busquets, F., Strecker, R., Rawlings, J. M., Belanger, S. E., Braunbeck, T., Carr, G.J., Cenijn, P., Fochtman, P., Gourmelon, A., Hübeler, N., Kleensamg, A., Knöbel, M., Kussatz, C., Legler, J., Lillicrap, A., Martinez-Jeronimo, F., Plleichtner, C., Rzodeczko, H., Salinas, E., Schneider, K. E., Scholz, S., van den Brandhof, E-J., van der Ven, L. T. M., Walter-Rohde, S., Weight, S., Witters, H. and Halder, M. (2014). OECD validation study to assess intra- and inter-laboratory reproducibility of the zebrafish embryo toxicity test for acute aquatic toxicity testing. *Regul. Toxicol. Pharmacol*. **69**, 496-511.

Button, K., Ioannidis, J., Mokrysz, C., Nosek, B., Flint, J., Robinson, E. and Munafo, M. (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nat. Rev. Neurosci* **14**, 365-376.

Carlsson C. and Pärt P. (2001). 7-Ethoxyresorufin O-deethylase induction in rainbow trout gill epithelium cultured on permeable supports: asymmetrical distribution of substrate metabolites. *Aquat. Toxicol*. **54**, 29-38.

Cohen, J. (1988). *Statistical power analysis for the behavioural sciences*. Hillside. NJ: Lawrence Earlbaum Associates.

Cohen, J. (1994). The Earth is round (p < 0.05). *Amer. Psychologist* **49**, 997-1003.

Cohn, L. D. and Becker, B. J. (2003). How meta-analysis increases statistical power. *Psychological Meths*. **8**, 243.

Collaboration, O. S. (2015). Estimating the reproducibility of psychological science. *Science 349*.

Colquhoun, D. (2014). An investigation of the false discovery rate and the misinterpretation of p-values. *Roy. Soci. Open Science* **1**, 140216.

Cowan-Ellsberry, C., Dyer, S. D., Erhardt, S., Bernhard, M. J., Roe, A. L., Dowty, M. E. and Weisbrod, A. V. (2008). Approach for extrapolating in vitro metabolism data to refine bioconcentration factor estimates. *Chemosphere* **70**, 1804 – 1817.

Crossley, M., Staras, K., Kemenes, G. (2016). A two-neuron system for adaptive goal-directed decision-making in Lymnaea. *Nat. Commun*. 7:11793.

Cruz, S.A., Lin, C.H., Chao, P.L., Hwang, P-P. (2013). Glucocorticoid receptor, but not mineralocorticoid receptor, mediates cortisol regulation of epidermal ionocyte development and ion transport in zebrafish (*Danio rerio*). *PLoS One* **8**: e77997. **Cumming, G.** (2008). Replication and p intervals. p values predict the future only vaguely, but confidence intervals do much better. *Persp. Psycholog. Sci.* **3**, 286-300.

Cumming, G., Fidler, F. and Vaux, D. (2007). Error bars in experimental biology. *J. Cell Biol*. **177**, 7-11.

Dalla Costa, E., Minero, M., Lebelt, D., Stucke, D., Canali, E. and Leach, M. C. (2014) Development of the horse grimace scale (hgs) as a pain assessment tool in horses undergoing routine castration. *PLoS ONE* **9**, e92281.

de Boo, J. and Hendriksen, C. (2005). Reduction strategies in animal research: a review of scientific approaches at the intra-experimental, supra-experimental and extra-experimental levels. *ATLA* **33**, 369.

Deakin, A. G., Buckley, J., AlZu'bi, H. S., Cossins, A. R., Spencer, J. W., Al'Nuaimy, W., Young, I.S. and Sneddon, L. U. Automated fish behaviour monitoring: A novel tool to characterize the status of zebrafish. Manuscript submitted.

DePasquale, C., Neuberger, T., Hirrlinger, A. M., & Braithwaite, V. A. (2016). The influence of complex and threatening environments in early life on brain size and behaviour. *Proc. Roy. Soc. Lond. B,* **283**, 20152564

Dwan, K., Gamble, C., Williamson, P. R. and Kirkham, J. J. (2013). Systematic review of the empirical evidence of study publication bias and outcome reporting bias — An updated review. *PLoS ONE* **8**, e66844

Economist (2013). Unreliable Research: Trouble at the Lab. In *The Economist*, pp. 26-30 ['Trouble at the Lab']: The Economist Newspaper Limited.

EDA (2017). https://eda.nc3rs.org.uk/

Ellison, A., Gotelli, N., Inouye, B. and Strong, D. (2014). P values, hypothesis testing, and model selection: it's de´ja` vu all over again. *Ecol*. **95**, 609-610.

Eng, J. (2003). Sample Size Estimation: How Many Individuals Should Be Studied? 1. *Radiol*. **227**, 309-313.

European Commission (2015) Communication From The Commission on the European Citizens' Initiative "Stop Vivisection". http://ec.europa.eu/environment/chemicals/lab_animals/pdf/vivisection/en.pdf

Fisher, R. (1959). *Statistical Methods and Scientific Inference.* New York: Hafner Publishing.

Freedman, L. P., Cockburn, I. M. and Simcoe, T. S. (2015). The Economics of Reproducibility in Preclinical Research. *PLoS Biol* **13**, e1002165.

Gardner, C. J., Deeming, D. C., Wellby, I., Soulsbury, C. D. and Eady, P. E. (2015). Effects of surgically implanted tags and translocation on the movements of common bream *Abramis brama* (L.). *Fisheries Res.* **167**, 252-259.

Geppert, M., Sigg, L. and Schirmer, K. (2016). A novel two-compartment barrier model for investigating nanoparticle transport in fish intestinal epithelial cells. *Environ. Sci. Nano*. **3** , 388.

Guh, Y-J., Lin, L-H. and Hwang, P-P. (2015) Osmoregulation in zebrafish: ion transport mechanisms and functional regulation. *EXCLI J*. **14**: 627-659.

Halsey, L. G., Curran-Everett, D., Vowler, S. and Drummond, G. (2015). The fickle P value generates irreproducible results. *Nature Meths*. **12**, 179-185.

Halsey, L.G., Coward, S. R. L., Crompton, R. H. and, Thorpe, S. K. S. (2017) Practice makes perfect: performance optimisation in ‘arboreal’ parkour athletes illuminates the evolutionary ecology of great ape anatomy. *J. Human Evol*. **103**, 45-52.

Halsey, L. G. (2007). 'Traversties of justice': The noise to signal ratio in association football. *Soccer and Society* **8**, 68-74.

Handy, R. D., Musonda, M. M., Phillips, C. and Falla, S. J. (2000). Mechanisms of gastrointestinal copper absorption in the African walking catfish: Copper dose-effects and a novel anion-dependent pathway in the intestine. *J. exp. Biol*. **203**, 2365-2377.

Henn, K. and Braunbeck, T. (2011). Dechorionation as a tool to improve the fish embryo toxicity test (FET) with the zebrafish (*Danio rerio*). *Comp. Biochem. Physiol.* **153C**, 91-98.

Holmes, A. M., Emmans, C. J., Jones, N., Coleman, R., Smith, T. E. and Hosie, C. A. (2016). Impact of tank background on the welfare of the African clawed frog, *Xenopus laevis* (Daudin). *Appl. Anim. Behav. Sci*. **185**, 131-136.

Hopkins, M. E. and Milton, K. (2016). Adverse effects of ball-chain radio-collars on female mantled howlers (*Alouatta palliata*) in Panama. *Internat. J. Primatol*., **37**, 213-224.

Hurst, J. L. and West, R. S. (2010). Taming anxiety in laboratory mice. *Nat Methods*. **7**, 825-826.

Incardona, J. P. and Scholz, N. L. (2016). The influence of heart developmental anatomy on cardiotoxicity-based adverse outcome pathways in fish. *Aquat. Toxicol*. **177**, 515-525.

Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS Med* **2**, e124.

Ioannidis, J. P. A., Fanelli, D., Dunne, D. D. and Goodman, S. N. (2015). Meta-research: Evaluation and Improvement of Research Methods and Practices. *PLoS Biol* **13**, e1002264.

Jennions, M. D. and Møller, A. P. (2003). A survey of the statistical power of research in behavioral ecology and animal behavior. *Behav. Ecol*. **14**, 438-445.

Johnson, D. (1999). The insignificance of statistical significance testing. *J. Wildlife Manag*. **63**, 763-772.

Kalmar, I. D., Janssens, G. P. J. and Moons, C. P. H. (2010). Guidelines and ethical considerations for housing and management of psittacine birds used in research. *ILAR J*. **51**, 409-423.

Kawano, A., Haiduk, C., Schirmer, K., Hanner, R., Lee, L. E. J., Dixon, B. and Bols, N. C. (2011). Development of a rainbow trout intestinal epithelial cell line and its response to lipopolysaccharide. *Aquacul. Nut*. **17**, E241-E252.

Keating, S. C. J., Thomas, A. A., Flecknell, P. A. and Leach, M. C. (2012). Evaluation of EMLA cream for preventing pain during tattooing of rabbits: Changes in physiological, behavioural and facial expression responses. *PLoS One* **7**, e44437

Knight, K. (2016). Implementing the 3Rs: improving experimental approaches in animal biology. *J. Exp. Biol*. **219**, 2414-2415.

Lammer, E., Kamp, H., Hisgen, V., Koch, M.,Reinhard, D., Salinas, E. R., Wendler, K., Zok, S. and Braunbeck, T. (2009). Development of a flow-through system for the fish embryo toxicity test (FET) with the zebrafish. *Toxicol. In Vitro* **23**, 1436 – 1442.

Langford, D. J., Bailey, A. L., Chanda, M. L., Clarke, S. E., Drummond, T. E., Echols, S., Glick, S., Ingrao, J., Klassen-Ross, T., LaCroix-Fralish, M. L., Matsumiya, L., Sorge, R. E., Sotocinal, S. G., Tabaka, J. M., Wong, D., van den Maagdenberg, A. M. J. M., Ferrari, M. D., Craig, K. D. and Mogil, J. S. (2010). Coding of facial expressions of pain in the laboratory mouse. *Nat. Meths*. **7**, 447-449.

Lavine, M. (2014). Comment on Murtaugh. *Eco*l. **95**, 642-645.

Lillicrap, A., Belanger, S., Burden, N., Du Pasquier, D., Embry, M., Halder, M., Lampi, M. A., Lee, L., Norberg-King, T., Rattner, B. A., Schirmer, K. and Thomas, P. (2016). Alternative approaches to vertebrate ecotoxicity tests in the 21st Century: A review of developments over the last 2 decades and current status. *Environ. Toxicol. Chem*. **35**, 2637-2646.

Liu, F., Huang, J., Ning, B., Liu, Z., Chen, S. and Zhao, W. (2016) Drug discovery via human-derived stem cell organoids. *Front. Pharmacol*. **7**:334

Loftus, G. R. (1993). A picture is worth a thousand p values: On the irrelevance of hypothesis testing in the microcomputer age. *Behav. Res. Meths. Instruments Comps*. **25**, 250-256.

Lopez-Luna, J., Al-Jubouri, Q., Al-Nuaimy, W. and Sneddon, L. U. (2017a). Impact of analgesic drugs on the behavioural responses of larval zebrafish to potentially noxious temperatures. *Appl. Anim. Behav. Sci*. **188**, 97–105.

Lopez-Luna, J., Al-Jubouri, Q., Al-Nuaimy, W. and Sneddon, L.U. (2017b). Activity reduced by noxious chemical stimulation is ameliorated by immersion in analgesic drugs in zebrafish. *J. Exp. Biol*, **220**, 1451-1458.

Lukowiak, K., Sunada, H., Teskey, M., LukowiaK, K. And Dalesman, S. (2014). Environmentally relevant stressors alter memory formation in the pond snail *Lymnaea*. *J. Exp. Biol.* **217**, 76-83

Macleod, M. R., Fisher, M., O'Collins, V., Sena, E. S., Dirnagl, U., Bath, P. M., Buchan, A., van der Worp, H. B., Traystman, R. J. and Minematsu, K. (2009). Reprint: Good laboratory practice: preventing introduction of bias at the bench. *J. Cerebral Blood Flow Metab*. **29**, 221-223.

Maxwell, S., Kelley, K. and Rausch, J. (2008). Sample size planning for statistical power and accuracy in parameter estimation. *Ann. Rev. Psychol*. **59**, 537-563.

McClelland, G. H. (2000). Increasing statistical power without increasing sample size.

NC3Rs (2017). http://www.nc3rs.org.uk/grimacescales.

Maxwell, S., Kelley, K. and Rausch, J. (2008). Sample size planning for statistical power and accuracy in parameter estimation. *Ann. Rev. of Psychol*. **59**, 537-563.

McNutt, M. (2014). Journals unite for reproducibility. *Science* **346**, 679.

Minghetti, M. and Schirmer, K. (2016). Effect of media composition on bioavailabilty and toxicity of silver and silver nanoparticles in fish intestinal cells (RTgutGC). *Nanotoxicol.*, **10**, 1529 – 1534.

Mobley, A., Linder, S. K., Braeuer, R., Ellis, L. M. and Zwelling, L. (2013). A Survey on Data Reproducibility in Cancer Research Provides Insights into Our Limited Ability to Translate Findings from the Laboratory to the Clinic. *PLoS ONE* **8**, e63221.

Muthuswamy, S.K. (2017). Bringing together the organoid field: from early beginnings to the road ahead. *Development* **144**, 963-967.

Nakagawa, S. and Cuthill, I. (2007). Effect size, confidence interval and statistical significance: a practical guide for biologists. *Biological Rev*. **82**, 591-605.

Nature Magazine. (2016). Go forth and replicate! *Nature* **536**, 373.

Nichols, J. W., Fitzsimmons, P. N. and Burkhard, L. P. (2007). In vitro-in vivo extrapolation of quantitative hepatic biotransformation data for fish. II Modeled effects on chemical bioaccumulation. *Environ. Toxicol. Chem*. **26**, 1304-1319.

Nichols, J.W., Huggett, D.B., Arnot, J.A., Fitzsimmons, P.N. and Cowan-Ellsbery, C.E. (2013). Toward improved models for predicting bioconcentration of well-metabolized compounds by rainbow trout using measured rates of in vitro intrinsic clearance. *Environ. Toxicol. Chem*. **32**, 1611-1622.

Nichols, J. W., Schultz, I. R. and Fitzsimmons, P. N. (2006). In vivo-in vitro extrapolation of quantitative hepatic biotransformation data for fish. I A review of methods, and strategies for incorporating intrinsic clearance estimates into chemical kinetic models. *Aquat. Toxicol.* **78**, 74-90.

Noldus (2016). Sense Well, www.noldus.com/projects/sensewell

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., Buck, S., Chambers, C. D., Chin, G., Christensen, G. et al., (2015). Promoting an open research culture. *Science* **348**, 1422-1425.

Nüßer, L.K., Skulovich, O., Hartmann, S., Seiler, T-B., Cofalla, C., Schuttrumpf, H., Hollert, H., Salomons, E. and Ostfeld, A. (2016). A sensitive biomarker for the detection of aquatic contamination based on behavioural assays using zebrafish larvae. *Ecotoxicol. Environ. Saf*. **133**, 271-280.

Nuzzo, R. (2014). Statistical Errors. *Nature* **506**, 150-152.

OECD Guidelines for testing chemicals. Test No. 203: Fish, Acute Toxicity Test (1992). www.oecd-library.org

OECD Guidelines for testing chemicals. Test No. 305: Bioaccumulation in Fish: Aqueous and Dietary Exposure (2012). www.oecd-library.org

O'Farrell, A. C., Shnyder, S. D., Marston, G., Coletta, P. L. and Gill, J. H. (2013). Non-invasive molecular imaging for preclinical cancer therapeutic development. *Br J Pharmacol.* **169**, 719–735.

Otto, G. P., Cocorocchio, M., Munoz, L., Tyson, R. A., Bretschneider, T. and Williams, R. S. (2016). Employing Dictyostelium as an Advantageous 3Rs Model for Pharmacogenetic Research. *Methods Mol. Biol*. **1407**, 123-130.

Pew Research Center (2015). "Public and Scientists' Views on Science and Society" http://www.pewinternet.org/2015/01/29/public-and-scientists-views-on-science-and-society/

Phillips, R. A., Green, J. A., Phalan, B., Croxall, J. P. and Butler, P. J. (2003). Chick metabolic rate and growth in three species of albatross: a comparative study. *Comp. Biochem. Physiol. A*, **135**, 185-193.

Pounder, K. C., Mitchell, J. L., Thomson, J. S., Pottinger, T.G., Buckley, J. and Sneddon, L.U. (2016). Does environmental enrichment promote recovery from stress in rainbow trout? *Appl. Anim. Behav. Sci*. **176**, 136–142

Reproducibility Initiative. (2014). Reproducbility Initiative. In http://validation.scienceexchange.com/#/reproducibility-initiative, vol. 2014.

Rushen, J., Chapinal, N. and de Passillé, A. M. (2012). Automated monitoring of behavioural-based animal welfare indicators. *Animal Welf.* **21**: 339-350

Russell, W. M. S., and Burch, R. L. (1959). *The principles of humane experimental technique*. London: Methuen.

Russom, C. L., Bradbury, S. P., Broderius, S. J., Hammermeister, D. E. and Drummond, R. A. (1997). Predicting modes of toxic action from chemical structure: Acute toxicity in the fathead minnow (*Pimephales promelas*). *Environ. Toxicol. Chem*. **16**, 948-967.

Schnell, S., Stott, L. C., Hogstrand, C., Wood, C. M., Kelly, S. P., Pärt, P., Owen, S. F. and Bury N. R. (2016). Procedures for the reconstruction, primary culture and experimental use of rainbow trout epithelia. *Nature Prot*. **11**, 490-498.

Scholz, S., Ortmann, J., Klüver, N. and Léonard, M. (2014). Extensive review of fish embryo acute toxicities for the prediction of GHS acute systemic toxicity categories. *Regul. Toxicol. Pharmacol*. **69**, 572-579.

Schroeder, P., Jones, S., Young, I. S. and Sneddon, L. U. (2014). What do zebrafish want? Impact of social grouping, dominance and gender on preference for enrichment. *Lab. Anim*. **48**, 328–337.

Schroeder, P. and Sneddon, L. U. (2017). Exploring the efficacy of immersion analgesics in zebrafish using an integrative approach. *Appl. Anim. Behav. Sci*. **187**, 93–102.

Sellke, T., Bayarri, M. and Berger, J. O. (2001). Calibration of ρ values for testing precise null hypotheses. *The American Statistician* **55**, 62-71.

Sena, E. S., Briscoe, C. L., Howells, D. W., Donnan, G. A., Sandercock, P. A. and Macleod, M. R. (2010). Factors affecting the apparent efficacy and safety of tissue plasminogen activator in thrombotic occlusion models of stroke: systematic review and meta-analysis. *J. Cerebral Blood Flow Metab*. **30**, 1905-1913.

Singhal, G., Jaehne, E. J., Corrigan, F., and Baune, B. T. (2014). Cellular and molecular mechanisms of immunomodulation in the brain through environmental enrichment. *Front. Cell. Neurosci*. **8**, 97.

Smith, D. R., Hardy, I. C. and Gammell, M. P. (2011). Power rangers: no improvement in the statistical power of analyses published in Animal Behaviour. *Anim. Behav*. **81**, 347-352.

Sneddon, L. U. (2015). Pain in aquatic animals. *J. Exp. Biol*. **218**, 967-976

Sneddon, L. U., Elwood, R. W. Adamo S. and Leach M. C. (2014). Defining and assessing pain in animals. *Anim. Behav*. **97**, 201-212.

Sorge, R. E., Martin, L. J., Isbester, K. A., Sotocinal, S. G., Rosen, S., Tuttle, A. H., Wieskopf, J. S., Acland, E. L., Dokova, A. and Kadoura, B. (2014). Olfactory exposure to males, including men, causes stress and related analgesia in rodents. *Nature Meths*. **11**, 629-632.

Sotocinal, S.G., Sorge, R.E., Zaloum, A., Tuttle, A.H., Martin, L.J., Wieskopf, J. S., Mapplebeck, J. C. S., Wei, P., Zhan, S., Zhang, S., McDougall, J. J., King, O. D. and Mogil, J. S. (2011). The Rat Grimace Scale: a partially automated method for quantifying pain in the laboratory rat via facial expressions. *Mol. Pain* **7**, 55.

Stadnicka-Michalak, J., Schirmer, K. and Ashauer, R. (2015). Toxicology across scales: cell population growth in vitro predicts reduced fish growth. *Science Adv*. **1**:e1500302

Strähle, U., Scholz, S., Geisler, R., Greiner, P., Hollert, H., Rastegar, S., Schumacher, A., Selderslaghs, I., Weiss, C., Witters, H. and Braunbeck, T. (2012). Zebrafish embryos as an alternative to animal experiments--a commentary on the definition of the onset of protected life stages in animal welfare regulations. *Reprod Toxicol*. **33**,128-32.

Tanneberger, K., Knöbel, M., Busser, F. J. M., Sinnige, T.L., Hermens, J. L. M. and Schirmer, K. (2013). Predicting fish acute toxicity using a fish gill cell line-based toxicity assay. *Environ. Sci. Technol*. **47**, 1110-1119.

Tazawa, H., Aliyama, R., Moriya, K. (2002). Development of cardiac rhythms in birds. *Comp. Biochem, Physiol* **132A**, 675-689.

Tressoldi, P. E., Giofré, D., Sella, F. and Cumming, G. (2013). High Impact=High Statistical Standards? Not Necessarily So. *PLoS ONE* **8**, e56180.

Uchea, C., Owen, S. F. and Chipman, J. K. (2015). Functional xenobiotic metabolism and efflux transporters in trout hepatocyte spheroid cultures. *Toxicol. Res*. **4**, 494-507.

van Helden, J. (2016). Confidence intervals are no salvation from the alleged fickleness of the P value. *Nat. Meth.* **13**, 605-606.

Villeneuve, D., Volz, D. C., Embry, M.R., Ankley, G .T., Belanger, S.E., Léonard, M., Schirmer, K., Tanguay, R., Truong, L. and Wehmas, L. (2014). Investigating alternatives to the fish early-life stage test: a strategy for discovering and annotating adverse outcome pathways for early fish development. *Environ. Toxicol. Chem*. **33**, 158-169

Wacholder, S., Chanock, S., Garcia-Closas, M. and Rothman, N. (2004). Assessing the probability that a positive report is false: an approach for molecular epidemiology studies. *J. Nat. Cancer Institute* **96**, 434-442.

Weisbrod, A.V., Sahi, J., Segner, H., James, M. O., Nichols, J., Schultz, I., Erhardt, S., Cowan-Ellsberry, C., Bonnell, M. and Hoeger, B. (2009). The state of in vitro science for use in bioaccumulation assessments for fish. *Environ. Toxicol. Chem*. **28**, 86-96.

Wissenschaftsstiftung, E. (2017). The European Code of Conduct for Research Integrity. Revised Edition.

Wittwehr, C., Aladjov, H., Ankley, G., Byrne, H. J., de Knecht, J., Heinzle, E., Klambauer, G., Landesmann, B., Luijten, M., MacKay, C., Maxwell, G., Meek, M. E.,

Paini, A., Perkins, E., Sobanski, T., Villeneuve, D., Waters, K. M. and Whelan, M. (2017). How adverse outcome pathways can aid the development and use of computational prediction models for regulatory toxicology. *Toxicol. Sci.* **155**, 326-336.

Woolston, C. (2014). A blueprint to boost reproducibility of results: Online commenters show support for a call to shake up science. *Nature*. **513**, 283.

Würbel, H. (2000). Behaviour and the standardization fallacy. Nature Genetics 26, 263-263.

Yozzo, K.. L., Isales, G. M., Raftery, T. D. and Volz, D.C. (2013). High-content screening assay for identification of chemicals impacting cardiovascular function in zebrafish embryos. *Environ. Sci. Technol.* **47**, 11302-11310

Box 1: Which animals are protected under the legislation of selected countries?

Globally, legislation differs between countries and geographical regions. Either all animals used in research are protected (specific species or ages are not prescribed) or the legislation identifies which animals at what stage of development are included.

Country or region	**Protected animals**
Australia	Vertebrates of all developmental stages Cephalopods of all developmental stages
Brazil	All animals
China	All animals
Europe	Adult vertebrates Mammalian, bird and reptile foetuses in last third of development Amphibian and fish at the free-feeding stage Cephalopods at the free feeding stage
India	All animals
South Africa	All vertebrates including eggs, foetuses and embryos Cephalopods Decapods
USA	Warm-blooded vertebrates except farm animals used in food and fibre research, rats of the genus *Rattus* and mice of the genus *Mus*

Box 2: P is for Publication

Many journals, funding bodies and reviewers like to see *P* values and power analyses. For this reason, experimenters might be concerned about disadvantaging themselves if they become apostates of the *P* value doctrine. They might best be advised to continue reporting *P* values in their manuscripts but to shift the focus of interpretation onto effect sizes. For project proposals, perhaps providing both a power analysis and a plan for precision would be sensible. Below is a text template that can be used for inclusion in the Methods section of manuscripts to flag up that data interpretation will be based on effect sizes, and to justify why, while reassuring that *P* values will remain present:

In the current article, the *P* value is treated as a continuous variable (Fisher, 1959; Boos and Stefanski, 2011), and because it is typically highly imprecise it is considered to be only a tentative indication of the strength of evidence for observed patterns in the data (Fisher, 1959; Boos and Stefanski, 2011; Halsey et al., 2015). Primarily, patterns in the data are interpreted from graphs of sample effect sizes and their precision (quantified by 95% confidence intervals) (Lavine, 2014; Loftus, 1993).

Figure legends

Fig. 1. Ethical thinking when planning animal experiments from conceiving an experiment, applying the 3Rs and finally publication. The figure shows a diagrammatic representation of the major ethical concepts and key questions that scientists must address under the traditional view of the 3Rs – Reduction, Replacement, Refinement – to justify the use of animals in experimentation, from planning the programme of work through to publication. *Except cephalopods, which are protected animals in Australia, Europe and South Africa as listed in Box 1.

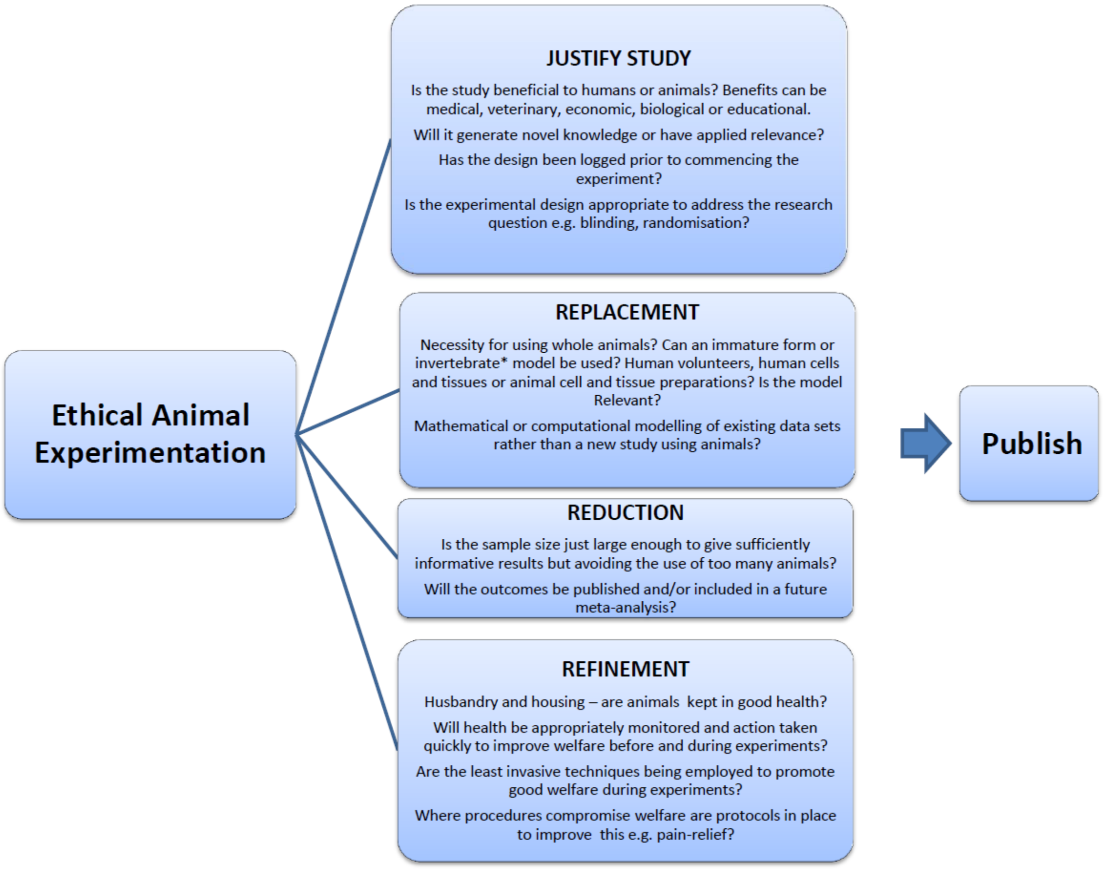
Fig. 2. Some examples of studies where refinement has proved to be beneficial to the welfare of the experimental animals. (A) Impact of enrichment (gravel, plastic plant and overhead cover) on improving recovery rates in rainbow trout: mean (±SE) opercular beat recovery rate (OBR; beats min−1) post treatment, in rainbow trout held in either enriched (dark bars) or barren (light bars) environments. Recovery OBR rate was estimated for each individual fish by subtracting OBR at time of recovery from OBR rate after either one minute of air emersion (Stress) or after deep-plane anaesthesia, and divided by the time between time points (Adapted from Pounder et al., 2016 with kind permission from Elsevier). (B) Impact of background colour in the tanks of *Xenopus laevis*, demonstrating that a white background results in greater body mass change (BMC, g) than a black background (Taken from Holmes et al., 2016 with kind permission from Elsevier). (C) The use of pain-relieving drugs during recovery from fin clipping in zebrafish ameliorates a reduction in activity. The graph shows the mean percentage change in activity level (number of swimming movements) 80 mins after tail fin clipping without analgesia (Fin clip) or in conjunction with immersion in lidocaine

(5mg/L) in zebrafish (adapted from Schroeder & Sneddon, 2017 with kind permission from Elsevier).

Fig. 3. Cumulative meta-analysis of the efficacy of lytic treatments (e.g. tissue plasminogen activator) in thrombotic animal models of stroke. The data have been adapted to illustrate key points explained and discussed in this article. Studies are in order of their publication date. The greater the value on the x axis, the greater the positive effect of the treatment. Treatment improves outcome; however, the estimate of efficacy (effect size) decreased as more data became available. This often happens, because studies are typically underpowered and therefore, when statistically significant, tend to overestimate the true effect size (Halsey et al. 2015). Note also the considerable size of the 95% confidence intervals (thin horizontal bars) for the first study and even once the first few studies are combined; this is common and demonstrates the lack of precision that individual studies often provide about the true (population) effect size, but is not apparent when focussing on the associated *P* value. Indeed, focussing on the *P* value of each study to synthesise the findings would return a confused conclusion, since while many of the studies report a statistically significant effect of the treatment (black data points and 95% confidence intervals), many of the studies indicate no treatment efficacy (blue). In contrast, focussing on the effect size and 95% confidence intervals of each study shows a relatively consistent pattern of evidence of treatment efficacy (as illustrated), and estimate accuracy of the degree of treatment efficacy steadily improves as mores studies are combined into the meta-analysis. The thick horizontal line shows a suggested approximate date at which the efficacy of the treatment was well known and further studies were unlikely to substantially refine this. Although studies published subsequent to 2001/2002 probably included other valuable experiments and/or analyses, this figure illustrates that meta-analyses can inform about when further study of a particular treatment or

phenomenon would be unproductive. Heeding such information would reduce the number of animals used in experimental research. This figure was reproduced from Sena et al. (2010) and edited with permission.

Figure 1



JUSTIFY STUDY

Is the study beneficial to humans or animals? Benefits can be medical, veterinary, economic, biological or educational.

Will it generate novel knowledge or have applied relevance?

Has the design been logged prior to commencing the experiment?

Is the experimental design appropriate to address the research question e.g. blinding, randomisation?

REPLACEMENT

Necessity for using whole animals? Can an immature form or invertebrate* model be used? Human volunteers, human cells and tissues or animal cell and tissue preparations? Is the model Relevant?

Mathematical or computational modelling of existing data sets rather than a new study using animals?

REDUCTION

Is the sample size just large enough to give sufficiently informative results but avoiding the use of too many animals?

Will the outcomes be published and/or included in a future meta-analysis?

REFINEMENT

Husbandry and housing – are animals kept in good health?

Will health be appropriately monitored and action taken quickly to improve welfare before and during experiments?

Are the least invasive techniques being employed to promote good welfare during experiments?

Where procedures compromise welfare are protocols in place to improve this e.g. pain-relief?

Figure 2

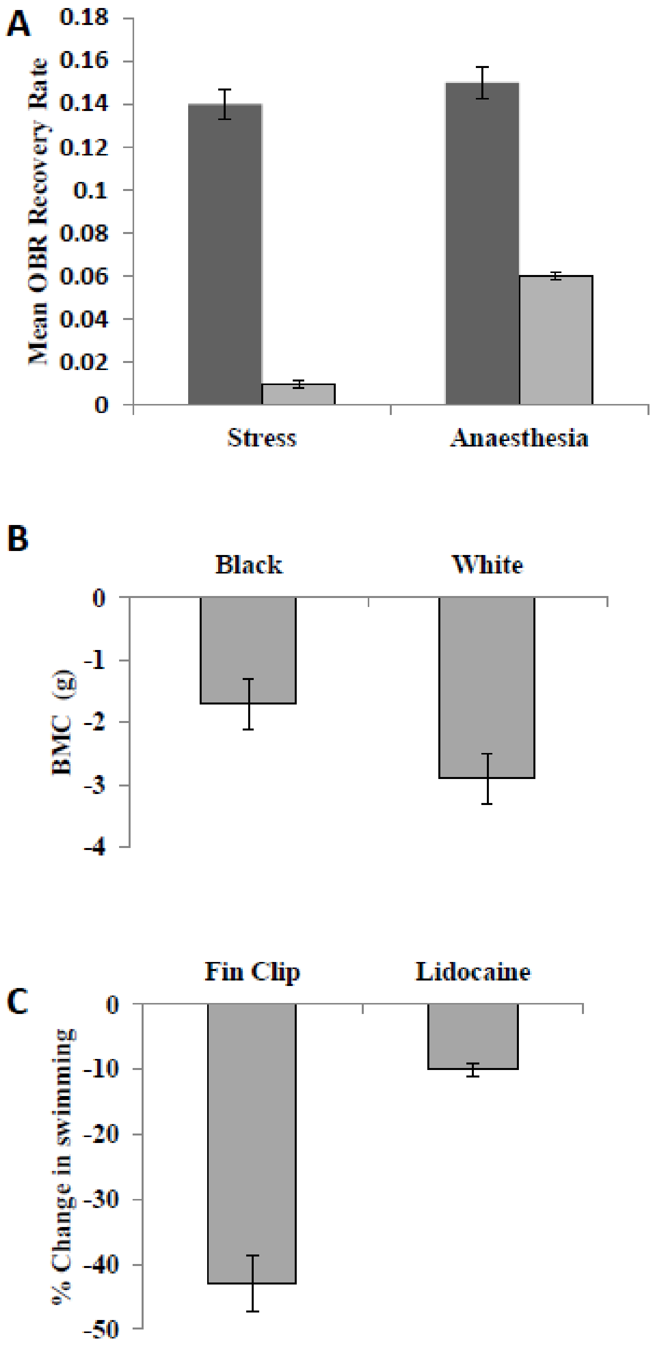
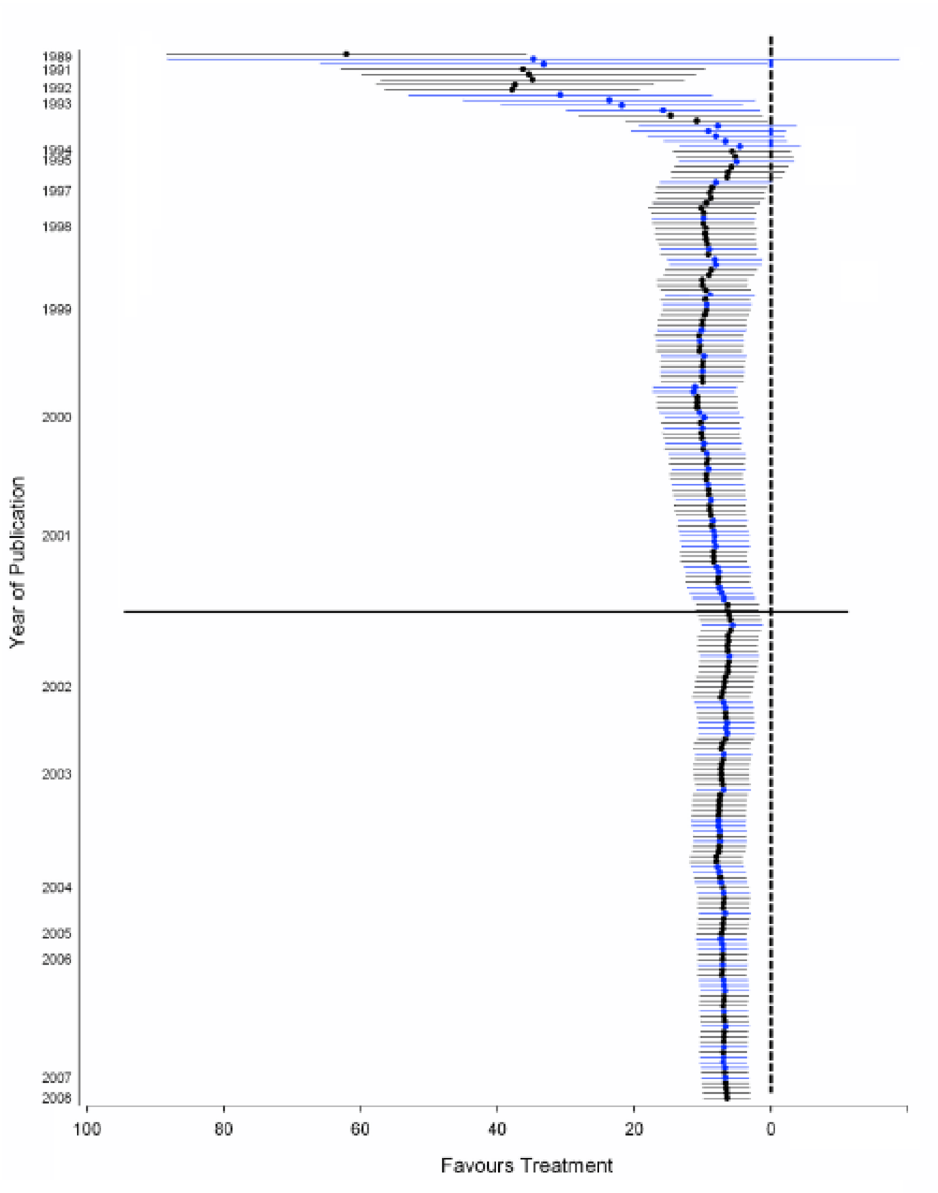

A
Mean OBR Recovery Rate
0.18
0.16
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
Stress
Anaesthesia
B
Black
White
BMC (g)
0
-1
-2
-3
-4
C
Fin Clip
Lidocaine
% Change in swimming
0
-10
-20
-30
-40
-50

Figure 3


1989
1991
1992
1993
1994
1995
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
Year of Publication
100
80
60
40
20
0
Favours Treatment