The Place of Experimental Design and Statistics in the 3Rs

Richard M. A. Parker and William J. Browne

Abstract

The 3Rs—replacement, reduction, and refinement—can be applied to any animal experiment by researchers and other bodies seeking to conduct those studies in as humane a manner as possible. Key to the success of this endeavor is an appreciation of the principles of good experimental design and analysis; these need to be considered in concert before any data is collected. Indeed, many of the principles central to helping achieve the objectives of the 3Rs—such as conducting valid, reliable, and efficient experiments; clearly and transparently reporting findings; and ensuring that an appreciation and understanding of animal welfare plays a central role in laboratory practice—are to the betterment of research per se.

Key Words: 3Rs; animal welfare; experimental design; laboratory animals; reduction; refinement; replacement; statistics

Introduction

Many millions of nonhuman animals (hereafter "animals") are used every year as subjects in laboratory experiments. Whilst estimates of the number of animals used annually worldwide are very imprecise due to differences in how different nations collect and report statistics, estimates nevertheless extend into the tens and hundreds of millions. For example, in a review of the topic, Taylor and colleagues (2008) noted that prior published estimates ranged from 50 to 200 million, before deriving their own estimate of 115 million, a figure they thought likely to be a conservative one.

The uses to which these animals are put are manifold, from more fundamental biological research to applied studies concerning human health: be it the further understanding of disease, gauging the efficacy of potentially therapeutic interventions, or testing product toxicity. Public surveys largely indicate majority support for research using animals. In the United States, for example, 57% of respondents to Gallup's Values and Beliefs Survey in 2014 believed that medical testing on animals was morally acceptable (Gallup 2014). Similarly, the majority of the surveyed British public accept the use of animal experimentation, but conditional on factors such as there being no unnecessary suffering (66% agree), and there being no alternative to their use in medical research (63% agree) (Ipsos MORI 2012). As such, there is a commonly held tension between the perceived utility of animal experimentation and concerns regarding the cost to animals. This tension is reflected in legislation too: for instance, both the UK's Animals (Scientific Procedures) Act 1986 (ASPA) and the European Union's Directive 2010/63/EU on the Protection of Animals Used for Scientific Purposes explicitly apply the 3Rs, of **replacement**, **reduction**, and **refinement**, as indeed does the legislation of a variety of other jurisdictions, albeit implicitly.

The 3Rs

Proposed by Russell and Burch in *The Principles of Humane Experimental Technique* (1959), the 3Rs are akin to an ethical algorithm that can be applied to any animal experiment by researchers and other bodies seeking to conduct those studies in as humane a manner as possible.

Russell and Burch demarcated the first of the 3Rs, **replacement,** into relative and absolute. Absolute replacement, the optimal outcome, refers to changing an experimental protocol that uses animals in some capacity—be it as whole-animal study subjects, sources of tissue, etc.—into one that does not. For example, replacing the study of biological processes conducted in vivo, or from animal-derived tissue in vitro, to one conducted by the construction of virtual models on a computer (in silico). Otherwise, replacements may be relative, for example, using an established cell line that was originally derived from animals, or replacing a given animal with one thought less likely to suffer as a result of the proposed experimental conditions (in the terminology of the UK's legislation, this would be replacing a protected with an unprotected animal).

The second R, **reduction**, refers to reducing the numbers of animals used without meaningfully diminishing the amount

Richard M. A. Parker, PhD, is a research associate in the Animal Behaviour and Welfare Research Group, School of Veterinary Sciences, and the Centre for Multilevel Modelling, University of Bristol, UK. William J. Browne is professor of biostatistics in the Animal Behaviour and Welfare Research Group, School of Veterinary Sciences, and Director of the Centre for Multilevel Modelling, University of Bristol, UK.

Address correspondence and reprint requests to Dr. Richard M. A. Parker, Animal Behaviour and Welfare Research Group, University of Bristol, Langford House, Langford, Bristol, BS40 5DU, UK or email richard.parker@bristol.ac.uk.

ILAR Journal, Volume 55, Number 3, doi: 10.1093/ilar/ilu044

© The Author 2014. Published by Oxford University Press on behalf of the Institute for Laboratory Animal Research. All rights reserved.
For permissions, please email: journals.permissions@oup.com

Downloaded from https://academic.oup.com/ilarjournal/article/55/3/477/645295 by University of Strathclyde user on 04 March 2025

and quality of information gleaned from experiments. This may involve reducing the number of animals used in any given experiment but, perhaps counterintuitively, could also involve increasing the number if this results in a net reduction of numbers across a (potentially) longer series of experiments.

Finally, the third R of **refinement** refers to "any approach which avoids or minimises the actual or potential pain, distress and other adverse effects experienced at any time during the life of the animals involved, and which enhances their wellbeing" (Buchanan-Smith et al. 2005)—such concerns, either implicitly or explicitly, ultimately relate to the subjective experience of the animal (e.g., Dawkins 1990).

As an ethical algorithm, the 3Rs can be applied to a range of scenarios, such as field research, or the use of animals in veterinary or biological education, and of course most pertinently here, laboratory experiments. At present, however, the extent to which the 3Rs are successfully applied in laboratory research is a moot point. For example, a number of concerns have been raised recently about the manner in which animal experiments are designed, analyzed, or reported (i.e., whether the most efficient use is being made of the animals that are employed) (e.g., Button et al. 2013; Kilkenny et al. 2009; Sena et al. 2010), and the extent to which the experiments can be generalized to the target population about which inferences are to be made (e.g., Hackam and Redelmeier 2006; Perel et al. 2007; van der Worp et al. 2010). More generally, recent technological developments, such as methods of genetic modification, have resulted in increases in the numbers of animals used (e.g., after a downturn, the number of animals used in regulated procedures in Great Britain, at least, is increasing, chiefly due to the breeding of genetically modified and harmful mutants) (Home Office 2013). Similarly, member states contributing to a recent report on the number of animals used for experimental and other scientific purposes in the European Union reported an increase in research using transgenic mice (European Commission 2013).

Experimental Design and Statistics

Experiments are tools we use to empirically infer something about how the world works: to estimate quantities and test (potentially causal) effects we cannot directly observe or measure. So, for example, if a research team is interested in assessing the efficacy of a certain intervention in treating a particular medical condition, it's not possible, nor even necessary, to test the whole population of interest. Rather, a sample is studied, and inferences are made about the larger population from which that sample is drawn. Here, because that population has not been directly observed, inferences are made with uncertainty: they may not be what we would regard as a true statement if we had access to data from the full population. In many animal experiments, of course, an additional inferential leap is made, across species: a researcher may be unable to study a sample of the population of ultimate interest (e.g., humans with a particular medical condition) and so studies a sample from a population with key biological characteristics thought sufficiently analogous to those of humans to allow inferences to be fruitfully applied across species.

Experiments are often, of course, designed to test hypotheses: for example, if we hypothesize that a new intervention (A) has greater efficacy than an alternative (B) in treating a particular medical condition, we can gauge empirical support for this hypothesis by testing specific predictions generated from it. For instance, we may test a prediction that animals undergoing intervention A will have lower levels of a disease marker—a proxy measure of the underlying disease state of interest—than animals undergoing intervention B. Alternatively, rather than testing predictions, the focus of experiments may be to estimate relationships such as the shape of a dose-response curve. Naturally, different experimental designs and analyses are best suited to answer the different sorts of questions researchers can pose. One principle is common throughout, however, namely that experimental design and analysis go hand in hand, and this observation is central to fruitfully applying the 3Rs, be it in developing, and validating, alternatives to animal use; reducing the number of animals that are used with little or no cost to inferential gain; or refining procedures so that they avoid, or minimize, pain, distress, suffering, and lasting harm or otherwise improve animal welfare (NC3Rs 2014).

The relationship between experimental design and analysis in the 3Rs is a large topic, in part because any improvement in a study's design and analysis can be construed as an application of the 3Rs: ensuring animals are used as efficiently as possible in experiments that are valid and reliable may contribute to a net reduction in the number of subjects used. In what follows, therefore, we offer an overview of some of the main points to consider, and direct the reader to other resources for a deeper treatment of specific topics. We focus mostly on reduction, although naturally many of the principles that contribute to helping achieve this objective can be applied to choosing appropriate alternatives (replacement), or titrating the effect of refinement against any apparent costs of a modification in experimental protocol.

Reduction

Many research programmes need to carefully manage the number of observational units they study, be it for reasons of funding, logistics, time available, and so on. Laboratory-based research on animals shares these constraints, of course, but Russell and Burch's principle of reducing the number of animals used in such settings is naturally informed by something else, namely an ethical consideration of the costs, to those animals, of their use as subjects.

Reducing the number of animals used in an experiment can result from a change in experimental design and analysis such that the same amount of information is obtained from the study despite a reduction in the animal-level sample size (i.e., a researcher is able to provide as good, or better, an answer to his or her research question even though fewer animals are being used in an experiment). Depending on

Downloaded from https://academic.oup.com/ilarjournal/article/55/3/477/645295 by University of Strathclyde user on 04 March 2025

the nature of the experiment, reuse of animals may be a possibility, but since this has the potential to further compromise the welfare of any animal reused, it may run contrary to one of the other objectives of the 3Rs, namely that of refinement (e.g., Fenwick et al. 2009). Alternatively, a researcher might consider using more animals than initially anticipated in a given study, if that means that that experiment alone provides a considerably better test of their research question of interest. This might then lead to a net reduction in the total number of animals used when considering the experiment as part of a series (i.e., compared with conducting a larger number of smaller experiments, each a poorer test) (Button et al. 2013). Otherwise, careful consideration of the prospects of an experiment may result in the conclusion that it is so unlikely to answer the question that the use of animals is not justified, and the experiment is dropped.

Thus, in applying the principle of reduction to a given experiment, there is a tension between using as few animals as possible and ensuring the quality of evidence one can glean from that experiment remains sufficiently high, whilst respecting the other objectives of the 3Rs. This involves selecting an experimental design, and subsequent analysis, that can most help to reliably and validly answer the research question of interest, but efficiently so, using no more animals than necessary. Here, appropriate sample-size calculations play a key role (e.g., Festing et al. 2002).

Judging Appropriate Sample Sizes

Ideally, if there were no costs of doing so—ethical, financial, or otherwise—researchers wishing to uncover the nature of causal effects and relationships in the population of interest would test the whole population. Clearly, in the vast majority of cases, this is neither possible nor desirable, and so researchers must judge how small a sample size they can observe whilst still giving themselves sufficient probability of uncovering the truth about their research question of interest. As such, there are a variety of methods and formulae that researchers can use to help guide their choice of sample size (plus see suggested references at the end of this section). Whilst the relevant procedures can be technical, it's important to note that thinking about what aspects of the study influence the desired sample size, and how they interrelate, is rather intuitive.

For example, if we were to test the hypothesis that a new intervention (A) significantly decreases the level of a disease marker compared to an alternative intervention (B), then we might collect an outcome measure for a group of animals that are exposed to intervention A and calculate the mean value for the group. We would then compare this to the average value from animals in another group that are exposed to intervention B. When considering group sizes, clearly an observed difference based on sample averages derived from just two animals in each group might simply be due to the selected animals. We may have gotten a very different average had we repeated the procedure and sampled a different group of animals. However, the same observed difference based on a sample average of, say, thirty subjects would be clearer evidence of a real difference. More formally, each group mean has uncertainty that we measure through the standard error of the mean $(\mathrm{SEM}) = s/\sqrt{n}$, where s is our estimate of the standard deviation of the population (for which we use the standard deviation of the sample), and n is the sample size. Here as n increases we decrease the SEM and thus estimate the population mean with greater precision. Therefore we are better able to distinguish the outcome of one intervention group from the other.

Similarly, if we find that the sample mean for animals in group A is considerably lower than that for animals in group B, and the outcome measure varies little between the animals in each group (for example, only one animal in group A is above the sample mean of group B), then we will have more evidence of a significant difference than if the outcome measures from the animals exhibited large variability, with considerable overlap between the two groups. Relating this to the formula above, if s is increased for fixed n, then the SEM increases, and thus the population mean is estimated with less precision.

So the number of animals to be sampled (the **sample size)**, the anticipated difference between the group means (the unstandardized **effect size**), and the **variability** in the two samples are central to how much weight we give any evidence as being in support, or otherwise, of our hypothesis of interest.

Here we have been implicitly comparing our experimental, or alternative, hypothesis (H_1) with the null hypothesis (H_0) that it makes no difference at all to the outcome measure which intervention the subjects undergo. Rather than informally speculating to what extent our experimental evidence is in keeping with our alternative hypothesis, however, we naturally need to formalize this above a hunch (i.e., we need a rule/criterion for deciding between these two hypotheses). In this case, a natural rule would be to consider the value of the absolute difference in sample means and then reject the null hypothesis if this difference is greater than some chosen value that depends upon the variability in the samples and their respective sizes. The choice of value, or threshold, is a balance between making two types of error. The larger we make the threshold, the more often we will reject the null hypothesis both if it is false (true positive) but also if it is true (false positive), with the latter known as a type I error. Conversely the smaller we make the threshold, the more often we fail to reject the null hypothesis both if it is true (true negative) but also if it is false (false negative), with the latter this time known as a type II error. We can fix the probability of making a type I error (α) by choosing a specific threshold; this is known as the **significance level** of the test. As there is only one threshold but two types of error that are inversely related, having fixed α, the only way to reduce the probability of a type II error (β) is therefore to increase the precision of our estimates by increasing the sample size (aside from using a better design that might reduce variability).

We generally refer to the probability of rejecting the null hypothesis when it is false (i.e., 1- β), as the **power** of the

Downloaded from https://academic.oup.com/ilarjournal/article/55/3/477/645295 by University of Strathclyde user on 04 March 2025

test. Since increasing sample size will increase power, sample-size calculations are also referred to as power calculations (e.g., see Krzywinski and Altman 2013b for a succinct primer, including a supplementary interactive worksheet allowing the reader to graphically investigate the relationship between power, sample size, effect size, etc.).

So we have identified five pieces of information—the sample size, effect size, variability, significance level, and power—that determine whether we formally reject (or not) the null hypothesis. If we know four of these five pieces of information, we can, typically, derive the fifth via an appropriate calculation. For example, to estimate an advisable sample size we need to decide on a tolerable probability of yielding a false positive result (the significance level, by convention $p = 0.05$), an acceptable desired probability of rejecting the null hypothesis when it is false (e.g., a power of 0.8, 0.9, etc.), the effect size (how big a biological effect we wish to detect), and the variability within the data.

Although the significance level and power are generally set by convention, the effect size and, perhaps especially, the variability, are more challenging to estimate. Prior (e.g., published) studies can offer useful guidance, but often, of course, a given experimental protocol is novel (or, if it is not, it may be novel to a particular laboratory, and there may be unforeseen between-laboratory differences). In addition, the estimates of interest (in particular for the variability) might not actually be reported (e.g., Kilkenny et al. 2009), or, if they apparently are, there may nevertheless be some ambiguity as to whether they refer to the specific parameter of interest, for instance a component of the variance that is not of primary concern to you (e.g., Lenth 2001). In addition, if prior studies have low power (are underpowered)—a pervasive problem in many research disciplines—any true effect sizes that are found are more likely to be overinflated. This has been referred to as the "winner's curse": since underpowered experiments cannot detect small effect sizes, then in such cases only samples that, by the nature of random variation, find larger effect sizes (than the underlying true small effect sizes) will be statistically significant and reported, resulting in true effect size magnitudes being overestimated (Button et al. 2013).

Preliminary pilot studies may help researchers derive these estimates, under conditions comparable to those in which the proposed study will be undertaken, although of course there will be considerable uncertainty (given the small sample size) associated with any estimates of effect size and variance gleaned from them (e.g., Kraemer et al. 2006; Thabane et al. 2010). Note that one possible solution to deriving certain estimates is to consider the effect size relative to the variability, and set this to a particular value based on whether the effect is hypothesized to be "small," "medium," or "large" (following Cohen's [1988] original demarcation). If one uses a standardized effect size such as this, it is then not necessary to estimate the variability directly as it will cancel out in the power calculations, making it analytically attractive. However, note that some, including Cohen himself, recommend this approach be used with a degree of caution and as a last resort (e.g., Ellis 2010; see also Lenth 2001 for a robust critique), for example, in the case of a completely novel intervention where there is no pilot data available.

This may seem an unsatisfactory state of affairs, but a researcher can nevertheless profitably use his or her best judgment, having digested all relevant prior information, to estimate the appropriate sample size given a variety of possible scenarios. For example, for a given variability (as measured by a variance), what would be a reasonable sample size to test the hypothesis of interest if the effect size was the smallest it could be whilst remaining of scientific interest? Very small effect sizes might be real, but uncovering them might make only a trivial addition to our understanding of the biological world or the benefits of a new treatment being not practically important. If we think the effect size is actually likely to be larger than this, though, we naturally also need to take appropriate calculations based on this best estimate into account when making our decision, as otherwise we risk unnecessarily using too many animals. Similarly, sample size calculations could be made for a selection of possible variances. As such, the researcher can attempt to map estimated sample sizes for a range of possible scenarios (e.g., Lenth 2001), and use these to help make an informed judgment along with a consideration of the costs and benefits of the research (e.g., the costs to the subjects' welfare of the research, financial, logistical, and other personnel costs, and the likely benefits that will stem from having performed a good test of the hypothesis of interest).

The manner in which this is done (i.e., the actual sample size calculation) depends on the proposed experimental design, including the type of variables (continuous, categorical, etc.) to be collected, structure of dependency (for example, any nesting), probability of missing data, the inferences of key interest, and so on. For example, if the main concern is to test whether there is a significant difference between treatments on outcome variable y, then this requires a sample size calculation to be conducted specific to that inference. If, in the same experiment, the researcher is also interested in whether another explanatory variable significantly predicts y, or in estimating the between-animal variance, etc., then these will not be estimated with the same power as the treatment-related difference (i.e., each of these research questions, despite the fact they all relate to data from the same experiment, requires its own unique sample size calculation). The more complex the experiment and analysis required, the more unknown variabilities that need estimating and the more complex the sample size calculation.

There are a considerable range of useful pedagogic resources to aid the researcher in calculating samples sizes and effect sizes (e.g., including but not confined to Cohen 1992; Lenth 2001; Nakagawa and Cuthill 2007; Krzywinski and Altman 2013b; Howell 2013 [e.g., Chapter 8]). With regard to software, there are a variety of useful programs available, for instance, InVivoStat (Clark et al. 2012), G*Power (Faul et al. 2007, 2009), PiFace (Lenth 2006-9), PS (Dupont and Plummer 1990), and nQuery Advisor + nTerim (Statistical Solutions 2014), to name but a few. Amongst them is our own MLPowSim (Browne et al.

Downloaded from https://academic.oup.com/ilarjournal/article/55/3/477/645295 by University of Strathclyde user on 04 March 2025

2009), designed in the main for more complex experiments. MLPowSim calculates power by simulating data ("virtual studies") based on parameters set by the user. Whilst designed for research scenarios over which the researcher typically has considerably less control than in a laboratory environment, this does mean it is a useful resource for any researcher wishing to estimate sample sizes for certain more complex designs, including multilevel models with missing data and a lack of balance.

Increasing Power and the Use of Efficient Designs

To increase the power of a proposed experiment, then, the researcher can either (a) increase the sample size, (b) decrease the variation in the current experiment, and/or (c) investigate alternative measures that will answer the same research hypothesis but result in an increased (standardized) effect size.

As mentioned earlier, there may be occasions in which a sample-size calculation implies that a greater number of subjects should be used than a researcher may have assumed beforehand. On face value, this is contrary to the 3Rs' objective of reduction, but not necessarily so if it results in an appreciably better test of the hypothesis. For example, by increasing power, one increases the chance of rejecting the null hypothesis when it is false and increases the likelihood that any statistically significant result found actually relates to a true effect. Indeed, given the chronic problem of underpowered studies, the need to reconsider how resources are best deployed is pressing (see Button et al. 2013 for a discussion). Alternatively, if a sample size calculation implies that a greater sample size than originally envisaged might be necessary to achieve a certain power, this may imply—having weighed the costs and benefits—that the experiment is simply not worth conducting.

Whilst underpowered, rather than overpowered, experiments tend to characterize laboratory-based animal studies, it's nevertheless important to note that the benefits, in terms of the uncertainty with which any subsequent inferences are drawn, will diminish, per subject, as sample size is increased. Since the SEM is proportional to $1/\sqrt{n}$ for known standard deviation, precision will increase at a slower rate than data collection (Krzywinski and Altman 2013a), and so the matter is one of careful judgment (Bacchetti et al. 2005). The sample size can also be manipulated without increasing the number of animals, for example, by taking more samples from those animals. Any design in which observations are clustered in this manner needs to be modeled using appropriate statistical methods; all else being equal, measurements taken from the same animal are likely to be more correlated than measurements taken from different animals, and standard errors will be underestimated if this lack of independence is not taken into account. Hierarchical structures can be encountered in other scenarios too: for example, animals within cages, or even replicates across labs. In such instances, aggregating (i.e., taking a summary of an outcome measure) across higher-level units unnecessarily dispenses with information that can be more efficiently analyzed in the same analysis. Hierarchical designs such as these pose the question of what is the optimum level of replication at each level of the design to achieve a certain power (Bate and Clark 2014 [e.g., note the example on p.105]; Browne et al. 2009; RJ Tempelman, unpublished observations).

Increasing the (standardized) effect size might in certain experiments be achieved by choosing a more extreme treatment (e.g., Krzywinski and Altman 2013b). Such aspects of the design may be naturally determined by what is clinically relevant, or constrained by ethical considerations if a more extreme treatment is likely to induce an increasingly negative outcome for the animal. More generally, the choice of outcome measure will have a very important bearing on sample-size calculations. It naturally needs to be a sensitive proxy of the underlying construct of interest, and thus the choice will reflect the performance of outcome measures in prior work; scientific considerations of mechanism; as well as how invasive (with respect to the 3Rs' objective of refinement), and also logistically and financially feasible, taking the measure will be.

With regard to variation, known (or possible) sources of variance can be controlled for (and/or investigated) in the experimental design by appropriate use of blocking, inclusion of factors measuring them, choice of subjects, etc. (e.g., Bate and Clark, 2014; Festing et al. 2002). Preferentially targeting (i.e., controlling for) the sources of greatest residual variance is naturally a sensible strategy. Note that, in a nested design, the overall variance is made up of components at each level of the hierarchy, so if a researcher takes more than one measure from each animal, and the within-animal variance is greater than the between-animal variance, then this would suggest he or she is best advised to first address sources of within-animal variation (e.g.; Bate and Clark 2014; RJ Tempelman, unpublished observations). A change in design, such as employing combinations of factors rather than testing one factor at a time, can allow more information to be gleaned from the same number of animals, and can be designed with a relatively small number of animals per combination of factors by reaping the benefits of hidden replication (e.g., Bate and Clark 2014; Festing et al. 2002; Shaw et al. 2002).

Other Ways to Achieve Net Reduction in Animal Usage

As well as ensuring sufficient power, the researcher can use the tools of experimental design and statistics to address bias stemming from other sources. Naturally, if a treatment effect on an outcome variable is found in a given experiment, then this is only of interest if it doesn't represent the effect of unacknowledged factors (i.e., it is internally valid and not influenced by bias unwittingly introduced in the experimental design or analysis). This is obviously a considerable concern when considering how we might reduce the overall numbers of animals in experiments, as the better each test of a

Downloaded from https://academic.oup.com/ilarjournal/article/55/3/477/645295 by University of Strathclyde user on 04 March 2025

hypothesis is, the fewer the "false leads" (with associated animal usage) generated, the less corrective replication of previously compromised experimental work is needed, and so on.

There is a large amount of valuable literature detailing methods to ensure bias is minimized, including resources specific to laboratory animal studies (e.g., Bate and Clark 2014; Festing and Altman 2002; Festing et al. 2002). In essence, much of this concerns protecting us against ourselves (i.e., ensuring our own cognitive and behavioural biases do not jeopardize the integrity of our experimental results). For example, the importance of randomization, blinding (not only of researchers, but of technical and veterinary staff too, e.g., Bate and Clark 2014), and identifying and controlling for the influence of known confounding variables via the use of blocking is rightly stressed.

The researcher may have further sway over the inferences reported: so-called "researcher degrees of freedom" (Simmons et al. 2011), or "p-hacking" may be employed to render a result more significant, for example, than alternative analytical strategies might have found. This may be in the choice of variables reported, model selection, the choice of a statistical test, the treatment of outliers, choice of transformations, and so on. Some of the terminology used to describe such practices could be construed as implying a degree of underhandedness on the part of a researcher, fishing for significance, but whilst that may occasionally be the case, many such analytical choices are generally-accepted and appear reasonable (Gelman and Loken 2013). Given sufficient interrogation using commonplace methods of analysis, many datasets will yield p values under 0.05 (e.g., Bennett et al. 2009; Wagenmakers et al. 2011).

As such, phenomena such as the chronic problem of underpowered experiments (Button et al. 2013), a disproportionate prevalence of published p values that are marginally under 0.05 (Masicampo and Lalande 2012), results not in keeping with the experimental hypothesis residing in metaphorical file drawers rather than the pages of journals (e.g., Dwan et al. 2013; Rosenthal 1979; ter Riet et al. 2012) likely reflect a variety of investigator-level and more systemic factors. For instance journals' publication policies valuing positive results and novelty above null findings and study replication, the bearing of high-impact publications on the trajectory of scientists' careers, and also perhaps an investigator's own satisfaction that their efforts have indeed found evidence in keeping with their pet hypothesis, or the simple urge to construct a coherent narrative, rather than presenting a more confusing picture harder to explain (Anon 2013) all shape what is published. The trend for increasing transparency (see below) seeks to, at least partly, remedy this.

External validity, on the other hand, concerns the extent to which the results from a given experiment can be applied to the population about which the investigators ultimately wish inferences to be made. Sometimes this target population will be a more obviously close match to the sample studied, as in some ethological studies, the assessment of certain veterinary interventions, etc. In many other instances, generalizations will be sought across species (e.g., inferences drawn from a study of mice applied to humans). A study that has poor external validity is a study in which animals were used unnecessarily and, reflecting the ethical implications (with regard to both animals and humans), there is an important and lively debate concerning the success, or otherwise, of this venture (e.g., Hackam and Redelmeier 2006; Perel et al. 2007; van der Worp et al. 2010).

Finally, accurate and clear reporting of experimental protocols, analytical methods, and subsequent results ensures experiments are not reproduced unnecessarily (i.e., erroneously conducted under the assumption they have not been performed before), but also facilitates the faithful reproduction of experimental protocols when it is deemed desirable and a more informed critique of the study's findings. As such, there is a growing trend toward providing access, alongside a summarizing article, to the materials necessary to reproduce the analysis reported, such as an annotated dataset and command syntax used to analyze it (e.g., Diggle and Zeger 2010; Groves and Godlee 2012). Furthermore, in some disciplines there is a movement to publish protocols prior to data collection and analysis, helping guard against any undue sway attributable to researcher "degrees of freedom," and guaranteeing the publication of results regardless of whether they are null or not (e.g., Chambers 2013; Munafo and Strain 2014).

Transparency is also important in facilitating a critical appraisal of the experiment, as is the choice of which statistics and charts to present. The latter can, for example, help render any future meta-analysis considerably more tractable, thus ensuring the experiment can be put to further use as a data point in a larger retrospective systematic appraisal (e.g., Kilkenny et al. 2010; Nakagawa and Cuthill 2007). Furthermore, some intuitively reasonable urges, such as conducting and reporting post hoc power analyses, on closer inspection prove to offer nothing more than the p values already reported (e.g., Bate and Clark 2014; Hoenig and Heisey 2001; Nakagawa and Foster 2004).

To end on a pragmatic note, it is worth noting that whilst the very valuable body of work advocating, and advising on, better standards of experimental design, analysis, and reporting continues to grow, translating this into better practice is a considerable challenge. It is commonplace to hear researchers' disquiet at being advised, sometimes apocalyptically, that the methods they have dutifully and carefully learned might actually retard scientific progress, and that the larger community in which they work has gross systemic failures (e.g., Button et al. 2013; Ioannidis 2005; Nakagawa and Cuthill 2007). It is naturally important to alert the community to any fundamental problems that exist in the conduct of science, but also, of course, to remain mindful that improvements in experimental design, analysis, and reporting (and any positive impact on the 3Rs that ensues) can only be made if those conducting the work are carried along with that movement (e.g., Sharpe 2013). To this end, efforts to incentivize transparency and study replication, and to promote education in experimental design and statistical analysis, are increasingly being recognized as important drivers for change (Munafo et al. 2014).

Downloaded from https://academic.oup.com/ilarjournal/article/55/3/477/645295 by University of Strathclyde user on 04 March 2025

Replacement

To the extent that tests involving the absolute replacement of animal subjects are as reliable and valid as the animal-based experiments they replaced, this is the ultimate manifestation of a humane corrective to current procedure. If they are not as valid or reliable, however, then how humane the proposed replacement is becomes less certain, particularly if the experiment is designed as a model for medical or toxicity testing.

With regard to relative replacement with subjects thought less likely to suffer as a result of the proposed experimental conditions, establishing evidence for grades of subjective suffering across animals can be—philosophically (e.g., Nagel 1974) and scientifically (e.g., Sherwin 2001)—a challenging endeavor, but a necessary one, and so by-and-large consensus opinions are drawn based on a consideration of factors such as apparent behavioral and cognitive complexity and phylogeny. So, for example, certain legislation, such as the Animals Scientific Procedures Act (ASPA) in Great Britain, draws boundaries both between species, and between developmental stages within species, which divides those protected by the Act from the remainder.

The development of alternative methods in which protected animals are replaced as subjects is a rapidly developing area, somewhat dictated by the speed of technological advance (Balls 2007; Langley et al. 2007). Validation protocols assessing alternative methods, such as those of the Organisation for Economic Co-operation and Development (OECD 2005), assess the **relevance** and **reliability** of a method with regard to a defined **purpose**. Here relevance is akin to predictive validity (i.e., the extent to which the test predicts the effect of interest in the target population), whereas reliability refers to a test's reproducibility (across laboratories and across time). These definitions are substrate neutral. That is, they do not fall foul of the **high fidelity fallacy** (Balls 2007; Russell and Burch 1959): the assumption that (e.g.) placental mammalian animal models will necessarily be better models of within-human effects as a consequence of their greater apparent similarity (compared to alternatives) to us. All else being equal, the substrate of an experiment—be it in silico, in vitro, etc.—is of less importance than the ability of the experiment to reliably predict the effect of interest (i.e., "what is essential is not that a model system 'looks like' the system of interest, but that it behaves like it" [Richmond 2010]).

Refinement

With regard to refinement, Russell and Burch (1959) distinguished the **direct inhumanity** that may result from the experimental procedures themselves, from the **contingent inhumanity** that may result from "the infliction of distress as an incidental and inadvertent by-product of the use of the procedure, which is not necessary for its success."

With regard to the first of these, animal models designed to simulate human-based diseases are, by their very nature, likely to involve suffering. As such, and pending the possibility of any future replacement, there may be scope to at least alleviate the severity, or shorten the duration, of the induced disorder (e.g., Littin et al. 2008; Wolfensohn et al. 2013; Workman et al. 2010).

Otherwise, the reported success in training animals in a manner that facilitates collecting physiological data, or conducting general veterinary checks—likely reducing any stress associated with capture and restraint—has led to some advocating such techniques be more widely adopted (e.g., Perlman et al. 2012). Even relatively modest changes in the manner in which mice are handled, for example, can reduce behavior associated with anxiety and fear (Gouveia and Hurst 2013; Hurst and West 2010).

Stress may introduce noise into an experiment, and may even interact with treatments of interest; therefore, in addition to its obvious intrinsic value, refinement can result in an experiment being a better test of a hypothesis, insofar as variability and/or bias can be reduced. Trying to understand the laboratory world from the animals' point of view can not only further the aim of refinement from a welfare perspective but can also help the researcher detect potential confounders and sources of noise, be it auditory frequencies that we are not privy to but rodents are (e.g., Burman et al. 2007); fluorescent light flicker that, unlike captive birds, we cannot perceive (e.g., Greenwood et al. 2004); and so on. Indeed, the subtle sensitivity of laboratory animals' responses to the environment around them can often surprise and fascinate (e.g., Sorge et al. 2014).

Conclusion

A carefully considered appreciation of experimental design and statistics prior to data collection is key when seeking to successfully apply the 3Rs, and indeed many of the principles central to helping realize their objectives—for instance, conducting valid, reliable, and efficient experiments; clearly and transparently reporting findings; and ensuring that an appreciation and understanding of animal welfare plays a central role in laboratory practice—are to the betterment of research per se.

Acknowledgments

We thank two anonymous referees for helpful and constructive comments.

References

Anon. 2013. Should scientists tell stories? Nat Methods 10:1037–1037.

Bacchetti P, Wolf LE, Segal MR, McCulloch CE. 2005. Ethics and sample size. Am J Epidemiol 161:105–110.

Balls M. 2007. Alternatives to animal experiments: Time to focus on replacement. AATEX 12:145–154.

Bate ST, Clark RA. 2014. The Design and Statistical Analysis of Animal Experiments. Cambridge, UK: Cambridge University Press.

Downloaded from https://academic.oup.com/ilarjournal/article/55/3/477/645295 by University of Strathclyde user on 04 March 2025

Bennett CM, Baird AA, Miller MB, Wolford GL. 2009. Neural correlates of interspecies perspective taking in the post-mortem Atlantic Salmon: An argument for multiple comparisons correction. Poster Presented at the 15th Annual Meeting of the Organization for Human Brain Mapping, San Francisco, USA.

Browne WJ, Golalizadeh Lahi M, Parker RMA. 2009. A Guide to Sample Size Calculations for Random Effect Models via Simulation and the MLPowSim Software Package. Bristol, UK: Centre for Multilevel Modelling, University of Bristol.

Buchanan-Smith HM, Rennie AE, Vitale A, Pollo S, Prescott MJ, Morton DB. 2005. Harmonising the definition of refinement. Anim Welfare 14:379–384.

Burman OHP, Ilyat A, Jones G, Mendl M. 2007. Ultrasonic vocalizations as indicators of welfare for laboratory rats (Rattus norvegicus). Appl Anim Behav Sci 104:116–129.

Button KS, Ioannidis JPA, Mokrysz C, Nosek BA, Flint J, Robinson ESJ, Munafo MR. 2013. Power failure: Why small sample size undermines the reliability of neuroscience. Nature Rev Neurosci 14:365–376.

Chambers CD. 2013. Registered Reports: A new publishing initiative at Cortex. Cortex 49:609–610.

Clark RA, Shoaib M, Hewitt KN, Stanford SC, Bate ST. 2012. A comparison of InVivoStat with other statistical software packages for analysis of data generated from animal experiments. J Psychopharmacol 26:1136–1142.

Cohen J. 1988. Statistical Power Analysis for the Behavioral Sciences. Hillsdale, NJ: Erlbaum.

Cohen J. 1992. A power primer. Psychol Bull 112:155–159.

Dawkins MS. 1990. From an animal's point of view - motivation, fitness, and animal welfare. Behav Brain Sci 13:1–9.

Diggle PJ, Zeger SL. 2010. Editorial. Biostatistics 11:375–375.

Dupont WD, Plummer WD. 1990. Power and sample size calculations: A review and computer program. Control Clin Trials 11:116–128.

Dwan K, Gamble C, Williamson PR, Kirkham JJ; Reporting Bias Group. 2013. Systematic review of the empirical evidence of study publication bias and outcome reporting bias - an updated review. Plos One 8:e66844.

Ellis PD. 2010. The Essential Guide to Effect Sizes: Statistical Power, Meta-Analyses and the Interpretation of Research Results. Cambridge, UK: Cambridge University Press.

European Commission. 2013. Report from the Commissin to the Council and the European Parliament: Seventh Report on the Statistics on the Number of Animals used for Experimental and other Scientific Purposes in the Member States of the European Union. Brussels: European Commission.

Faul F, Erdfelder E, Buchner A, Lang A-G. 2009. Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. Behav Res Meth 41:1149–1160.

Faul F, Erdfelder E, Lang A-G, Buchner A. 2007. G*Power 3: A flexible statistical power analysis program for the social, behavioural, and biomedical sciences. Behav Res Meth 39:175–191.

Fenwick N, Griffin G, Gauthier C. 2009. The welfare of animals used in science: How the "Three Rs" ethic guides improvements. Can Vet J Rev 50:523–529.

Festing MFW, Altman DG. 2002. Guidelines for the design and statistical analysis of experiments using laboratory animals. ILAR J 43:244–258.

Festing MFW, Overend P, Gaines Das R, Cortina Borja M, Berdoy M. 2002. The Design of Animal Experiments: Reducing the Use of Animals in Research Through Better Experimental Design. London: Royal Society of Medicine Press.

Gallup. 2014. Gallup News Service - Gallup Poll Social Series: Values and Beliefs - Final Topline. Washington, DC: Gallup.

Gelman A, Loken E. 2013. The garden of forking paths: why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hyptothesis was posited ahead of time. [Internet]. Available from: http://www.stat.columbia.edu/~gelman/research/unpublished/p_hacking.pdf

Gouveia K, Hurst JL. 2013. Reducing mouse anxiety during handling: Effect of experience with handling tunnels. Plos One 8:e66401.

Greenwood VJ, Smith EL, Goldsmith AR, Cuthill IC, Crisp LH, Walter-Swan MB, Bennett ATD. 2004. Does the flicker frequency of fluorescent lighting affect the welfare of captive European starlings? Appl Anim Behav Sci 86:145–159.

Groves T, Godlee F. 2012. Open science and reproducible research. Br Med J 344:e4383.

Hackam DG, Redelmeier DA. 2006. Translation of research evidence from animals to humans. JAMA 296:1731–1732.

Hoenig JM, Heisey DM. 2001. The abuse of power: The pervasive fallacy of power calculations for data analysis. Am Stat 55:19–24.

Home Office. 2013. Annual Statistics of Scientific Procedures on Living Animals, Great Britain 2012. London: The Stationery Office.

Howell DC. 2013. Statistical Methods for Psychology. Canada: Wadsworth Cengage Learning.

Hurst JL, West RS. 2010. Taming anxiety in laboratory mice. Nat Methods 7:825–U1516.

Ioannidis JPA. 2005. Why most published research findings are false. Plos Med 2:696–701.

Ipsos MORI. 2012. Views On the Use of Animals in Scientific Research. London: Ipsos MORI.

Kilkenny C, Browne WJ, Cuthill IC, Emerson M, Altman DG. 2010. Improving bioscience research reporting: The ARRIVE guidelines for reporting animal research. Plos Biol 8:e1000412.

Kilkenny C, Parsons N, Kadyszewski E, Festing MFW, Cuthill IC, Fry D, Hutton J, Altman DG. 2009. Survey of the quality of experimental design, statistical analysis and reporting of research using animals. Plos One 4:e7824.

Kraemer HC, Mintz J, Noda A, Tinklenberg J, Yesavage JA. 2006. Caution regarding the use of pilot studies to guide power calculations for study proposals. Arch Gen Psychiatry 63:484–489.

Krzywinski M, Altman N. 2013a. Importance of being uncertain. Nat Methods 10:809–810.

Krzywinski M, Altman N. 2013b. Power and sample size. Nat Methods 10:1139–1140.

Langley G, Evans T, Holgate ST, Jones A. 2007. Replacing animal experiments: choices, chances and challenges. BioEssays 29:918–926.

Lenth RV. 2001. Some practical guidelines for effective sample size determination. Am Stat 55:187–193.

Lenth RV. 2006–9. Java Applets for Power and Sample Size [Internet]. Available from: http://www.stat.uiowa.edu/~rlenth/Power

Littin K, Acevedo A, Browne W, Edgar J, Mendl M, Owen D, Sherwin C, Wurbel H, Nicol C. 2008. Towards humane end points: behavioural changes precede clinical signs of disease in a Huntington's disease model. Proc R Soc Lond [Biol] 275:1865–1874.

Masicampo EJ, Lalande DR. 2012. A peculiar prevalence of p values just below .05. Q J Exp Psychol 65:2271–2279.

Munafo M, Noble S, Browne WJ, Brunner D, Button K, Ferreira J, Holmans P, Langbehn D, Lewis G, Lindquist M, Tilling K, Wagenmakers E-J, Blumenstein R. 2014. Scientific rigor and the art of motorcycle maintenance. Nat Biotech 32:871–873.

Munafo MR, Strain E. 2014. Registered Reports: A new submission format at drug and alcohol dependence. Drug and Alcohol Depend 137:1–2.

Nagel T. 1974. What is it like to be a bat. Philosoph Rev 83:435–450.

Nakagawa S, Cuthill IC. 2007. Effect size, confidence interval and statistical significance: a practical guide for biologists. Biol Rev 82:591–605.

Nakagawa S, Foster TM. 2004. The case against retrospective statistical power analyses with an introduction to power analysis. Acta Ethologica 7:103–108.

NC3Rs. 2014. The 3Rs [Internet]. Available from: http://www.nc3rs.org.uk/the-3rs

OECD. 2005. Guidance Document no. 34 on the Validation and International Acceptance of New or Updated Test Methods for Hazard Assessment. Paris: OECD.

Perel P, Roberts I, Sena E, Wheble P, Briscoe C, Sandercock P, Macleod M, Mignini LE, Jayaram P, Khan KS. 2007. Comparison of treatment effects between animal experiments and clinical trials: systematic review. Br Med J 334:197–200.

Perlman JE, Bloomsmith MA, Whittaker MA, McMillan JL, Minier DE, McCowan B. 2012. Implementing positive reinforcement animal

Downloaded from https://academic.oup.com/ilarjournal/article/55/3/477/645295 by University of Strathclyde user on 04 March 2025

training programs at primate laboratories. Appl Anim Behav Sci 137:114–126.

Richmond J. 2010. The Three Rs. In: Hubrecht R, Kirkwood J, editors. The UFAW Handbook on the Care and Management of Laboratory and Other Research Animals. 8th ed. p. 5–22.

Rosenthal R. 1979. The file drawer problem and tolerance for null results. Psychol Bull 86:638–641.

Russell WMS, Burch RL. 1959. The Principles of Humane Experimental Technique. London: Methuen & Co. Ltd.

Sena ES, van der Worp HB, Bath PMW, Howells DW, Macleod MR. 2010. Publication bias in reports of animal stroke studies leads to major overstatement of efficacy. Plos Biol 8:e1000344.

Sharpe D. 2013. Why the resistance to statistical innovations? bridging the communication gap. Psychol Meth 18:572–582.

Shaw R, Festing MFW, Peers I, Furlong L. 2002. Use of factorial designs to optimize animals experiments and reduce animal use. ILAR J 43:223–232.

Sherwin CM. 2001. Can invertebrates suffer? Or, how robust is argument-by-analogy? Anim Welfare 10:S103–S118.

Simmons JP, Nelson LD, Simonsohn U. 2011. False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. Psychol Sci 22:1359–1366.

Sorge RE, Martin LJ, Isbester KA, Sotocinal SG, Rosen S, Tuttle AH, Wieskopf JS, Acland EL, Dokova A, Kadoura B, Leger P, Mapplebeck JC, McPhail M, Delaney A, Wigerblad G, Schumann AP, Quinn T, Frasnelli J, Svensson CI, Sternberg WF, Mogil JS. 2014. Olfactory exposure to males, including men, causes stress and related analgesia in rodents. Nat Methods 11:629–632.

Statistical Solutions. 2014. nQuery Advisor + nTerim. Cork, Ireland: Statistical Solutions.

Taylor K, Gordon N, Langley G, Higgins W. 2008. Estimates for worldwide laboratory animal use in 2005. Altern Lab Anim: ATLA 36:327–342.

ter Riet G, Korevaar DA, Leenaars M, Sterk PJ, Van Noorden CJF, Bouter LM, Lutter R, Elferink RPO, Hooft L. 2012. Publication bias in laboratory animal research: A survey on magnitude, drivers, consequences and potential solutions. Plos One 7:e43404.

Thabane L, Ma J, Chu R, Cheng J, Ismaila A, Rios LP, Robson R, Thabane M, Giangregorio L, Goldsmith CH. 2010. A tutorial on pilot studies: the what, why and how. BMC Med Res Methodol 10:1.

van der Worp HB, Howells DW, Sena ES, Porritt MJ, Rewell S, O'Collins V, Macleod MR. 2010. Can animal models of disease reliably inform human studies? Plos Med 7:e1000245.

Wagenmakers EJ, Wetzels R, Borsboom D, van der Maas HLJ. 2011. Why psychologists must change the way they analyze their data: the case of psi: comment on Bem (2011). J Pers Soc Psychol 100:426–432.

Wolfensohn S, Hawkins P, Lilley E, Anthony D, Chambers C, Lane S, Lawton M, Robinson S, Voipio H-M, Woodhall G. 2013. Reducing suffering in animal models and procedures involving seizures, convulsions and epilepsy. J Pharmacol Toxicol Meth 67:9–15.

Workman P, Aboagye EO, Balkwill F, Balmain A, Bruder G, Chaplin DJ, Double JA, Everitt J, Farningham DAH, Glennie MJ, Kelland LR, Robinson V, Stratford IJ, Tozer GM, Watson S, Wedge SR, Eccles SA, Navaratnam V, Ryder S. 2010. Guidelines for the welfare and use of animals in cancer research. Br J Cancer 102:1555–1577.

Downloaded from https://academic.oup.com/ilarjournal/article/55/3/477/645295 by University of Strathclyde user on 04 March 2025