Invited review: Reproducible research from noisy data: Revisiting key statistical principles for the animal sciences

Nora M. Bello*†‡¹ and David G. Renter‡§

*Department of Animal Science, University of Wisconsin, Madison, WI 53706

†Department of Statistics,

‡Center for Outcomes Research and Epidemiology, and

§Department of Diagnostic Medicine and Pathobiology, Kansas State University, Manhattan 66506

ABSTRACT

Reproducible results define the very core of scientific integrity in modern research. Yet, legitimate concerns have been raised about the reproducibility of research findings, with important implications for the advancement of science and for public support. With statistical practice increasingly becoming an essential component of research efforts across the sciences, this review article highlights the compelling role of statistics in ensuring that research findings in the animal sciences are reproducible—in other words, able to withstand close interrogation and independent validation. Statistics set a formal framework and a practical toolbox that, when properly implemented, can recover signal from noisy data. Yet, misconceptions and misuse of statistics are recognized as top contributing factors to the reproducibility crisis. In this article, we revisit foundational statistical concepts relevant to reproducible research in the context of the animal sciences, raise awareness on common statistical misuse undermining it, and outline recommendations for statistical practice. Specifically, we emphasize a keen understanding of the data generation process throughout the research endeavor, from thoughtful experimental design and randomization, through rigorous data analysis and inference, to careful wording in communicating research results to peer scientists and society in general. We provide a detailed discussion of core concepts in experimental design, including data architecture, experimental replication, and subsampling, and elaborate on practical implications for proper elicitation of the scope of reach of research findings. For data analysis, we emphasize proper implementation of mixed models, in terms of both distributional assumptions and specification of fixed and random effects to explicitly recognize multilevel data architecture. This is critical to ensure that

experimental error for treatments of interest is properly recognized and inference is correctly calibrated. Inferential misinterpretations associated with use of *P*-values, both significant and not, are clarified, and problems associated with error inflation due to multiple comparisons and selective reporting are illustrated. Overall, we advocate for a responsible practice of statistics in the animal sciences, with an emphasis on continuing quantitative education and interdisciplinary collaboration between animal scientists and statisticians to maximize reproducibility of research findings.

Key words: experimental replication, model specification, statistical errors, multiple testing

INTRODUCTION

Reproducible results define the very core of scientific integrity in modern research. Yet, legitimate questions have been recently raised about the robustness and reliability of new scientific knowledge (Ioannidis, 2005) based on a reported inability to reproduce research results (Ioannidis et al., 2009; Begley and Ellis, 2012; Begley and Ioannidis, 2015). Collectively, these concerns have come to be known as research reproducibility (**RR**) issues and seem to be quite pervasive across scientific disciplines (Begley and Ioannidis, 2015; Nuzzo, 2015; Baker, 2016).

To illustrate the magnitude of the RR crisis, consider the following evidence. In 2009, Ioannidis et al. (2009) reevaluated 18 microarray-based gene expression studies, of which only 2 were fully reproduced. Soon afterward, in 2012, scientists reported replicating only 6 out of 53 landmark studies on preclinical cancer efforts for drug development (Begley and Ellis, 2012). Meanwhile, a large collaborative project attempted replication of 100 experiments published in high-ranking psychology journals and succeeded only one-third to one-half of the time, depending on the criteria (Open Science Collaboration, 2015). A recent compilation of biomedical studies showed that approximately 75 to 90% of results in preclinical research, and as much as 85% of research

Received October 11, 2017.

Accepted March 8, 2018.

¹ Corresponding author: nbello@ksu.edu

results across the biomedical spectrum, were deemed irreproducible (Begley and Ioannidis, 2015). A survey of *Nature* readers revealed that as many as 70% of researchers have tried and failed to reproduce another scientist's research results and over 50% have failed to reproduce their own results (Baker, 2016). Interest in RR is further evidenced in special dedicated issues in high-profile journals, including *Science* (<http://www.sciencemag.org/site/special/data-rep/>) and *Nature* (<http://www.nature.com/news/reproducibility-1.17552>). Even lay publications (Young and Karr, 2011) and mainstream newspapers (Shaywitz, 2009; Naik, 2011) have reported on the subject of RR. Taken together, the extent and scope of the reproducibility crisis is broad, encompassing many basic scientific fields as well as more applied disciplines (reviewed by Begley and Ioannidis, 2015; Baker, 2016). In the animal sciences, the full extent of the RR crisis has not yet been characterized in detail, though the issue has been discussed previously (Bello et al., 2016; Tempelman, 2016).

Admittedly, the multifaceted and increasingly complex nature of many of the research problems currently at the forefront of science adds to the complications of the RR issue. Specific to agriculture, one may consider the daunting challenge of meeting demands for a safe and secure food supply for an exponentially growing world population under competing demands for environmental sustainability and limited natural resources in a changing climate. Meanwhile, the research process is becoming more and more quantitative across scientific domains. This seems largely attributed to increasingly large data sets and new data types, which in turn call for more and more sophisticated quantitative methods that often require specialized expertise. Overall, this state of affairs argues for dynamic multidisciplinary integration of scientific disciplines, for which a common, presumably quantitative, language is imperative. In this context, statistical practice is increasingly becoming a critical component of research efforts across scientific disciplines, including the animal sciences.

By its very interdisciplinary nature, statistics is uniquely poised to help address the RR crisis. Indeed, statistics, as a discipline, provides both a common quantitative language to help bridge scientific domains and a sophisticated formal framework that, when properly implemented, allows for the recovery of signal from the inherently noisy nature of data. In this review, we focus on the role of statistical practice in ensuring RR in animal health and production. Often defined as the science of learning from data, the statistical sciences offer both conceptual infrastructure and practical tools to deal with the many sources of variability naturally embedded in complex systems. However, for

all its sophistication, statistics is not a silver bullet for dealing with noisy data; the most one can expect is proof beyond a reasonable doubt, though hardly 100% certainty. As famously discussed by statistician George Box (1919–2013), “all models are wrong...,” as, even in best-case scenarios, statistical models are often simplistically naive and only provide rough approximations of a much more complex reality. Nevertheless, so continues Box's quote, when properly implemented, “... some [models] can be useful,” mainly by introducing a level of robustness and rigor within which the process of learning from data can be tackled more objectively. Particularly with large data sets, foundational statistical concepts take renewed relevance, as flaws in experimental design and bias of inference and prediction cannot be overcome by large sample sizes, as conspicuously illustrated by the so-called parable of Google Flu (Lazer et al., 2014). While there undoubtedly is valuable signal to be extracted from large data sets, “big” is not always better. That is, more data does not necessarily produce more information; rather, it may muddy the waters with irrelevant information. We argue that it is precisely in the context of big data that the ideas of sound experimental design and well-implemented data analysis may turn out to be as, or more, important than ever in the research process. To this end, the 2016 joint meeting of the American Dairy Science Association and the American Society of Animal Sciences had a special session on “Big Data in Animal Science: Uses for Models, Statistics and Meta-Approaches” (<http://www.jtmtg.org/JAM/2016/abstracts/612.pdf>).

Misunderstanding and misuse of statistical concepts are contributing factors to the RR crisis, both due to their nature (Ioannidis, 2005; Nuzzo, 2015) and to their widespread ubiquity across the sciences (Reinhart, 2015). Though with fewer documented examples, production agriculture (Sargeant et al., 2009, 2011; Kramer et al., 2016), and more specifically the animal sciences (Tempelman, 2009; Bello et al., 2016), are no exception. Therefore, our specific objectives in this review were (1) to reexamine foundational statistical concepts relevant to RR in the context of the animal sciences, (2) to educate the animal science community about common statistical misuses that undermine RR, and (3) to outline guidelines and introduce tools for statistical practice that maximize legitimacy and credibility of new scientific knowledge.

Before proceeding, a couple of disclaimers are in order. First, we emphasize foundational statistical principles, both for experimental design and for data analysis and interpretation. Admittedly, most of these principles are hardly novel in and of themselves; yet, they can be highly nuanced and have implications not immediately obvious to the animal scientist in the evolving land-

scape of modern scientific research, as will be described in later sections. Second, and just for clarification, we note that this article circumvents issues of scientific fraud and does not deal with other means of ill-willed overhauls of the scientific method. In fact, our working assumption in this article is that the vast majority of the scientific community cares deeply about sound research practices that lead to robust and reliable findings (Baker, 2016), and in so doing is committed to setting a solid foundation for further advances that get us closer to the truth.

DEFINING RESEARCH REPRODUCIBILITY

Ironically, a formal definition for what specifically constitutes RR has been elusive in the scientific literature (Baker, 2016), with terms such as replicability, repeatability, reliability, and reproducibility being used interchangeably and defined locally at best, yet inconsistently in any broader framework. Efforts from professional societies and funding agencies are on going to provide reference benchmarks and streamline formal definitions, including the American Statistical Association (<https://www.amstat.org/asa/files/pdfs/POL-ReproducibleResearchRecommendations.pdf>) and the National Academy of Sciences at the request of the National Science Foundation (<http://www8.nationalacademies.org/cp/projectview.aspx?key=49906>).

While we recognize the fluid status of these concepts, in this article we purposely follow the approach of Begley and Ioannidis (2015) and propose as a basic sensible tenet of RR that one should reasonably expect the “main ideas and major conclusions” of a research study to withstand both “close interrogation” and “independent validation.” The first concept refers to the ability to sustain close scrutiny and reproduce a study’s findings from its original data, whereas the second refers to the repetition of a study independent of the original investigator.

Admittedly, as a scientist, one poses research questions of which the answers are presumably unknown; as a result, even the best conceived, designed, and executed studies may yield unexpectedly negative results. In other words, by its very nature research is not a finished product, as it continually raises as many or more questions as it addresses. Furthermore, controversy and debate are not only routine practice, but in fact required for the progress of scientific knowledge. Thus, it is not reasonable to expect that each and every result of a study will replicate perfectly (Begley and Ioannidis, 2015). Yet, the unwritten covenant of RR lies at the core of science, with the ultimate expectation that scientific findings and advances will stand the

“test of time” (Begley and Ioannidis, 2015). At stake is not only our professional credibility as scientists, but, more importantly, the overall credibility of science and, thus, its role to society as an evidence-based conduit to inform decision- and policy-making. Indeed, the societal perception of science is of direct relevance to public support for the sciences and, as a consequence, to federal, state, and private funding allocation for research pursuits.

EXPERIMENTAL REPLICATION AND THE DATA GENERATION PROCESS

Relevant Concepts and Practical Tools

Central to RR is the idea that independent repetitions of an experimental condition, say a dietary treatment or a drug, under similar conditions should yield comparable results. Thus, we argue that, by its very nature, the action of recognizing the *actual unit of independent experimental repetition* is an indispensable tenet for RR. In statistics, such units are formally defined as units of experimental replication, or experimental units (EU; Kuehl, 2000; Casella, 2008; Milliken and Johnson, 2009; Mead et al., 2012). In fact, experimental replication is a cornerstone statistical principle of design of experiments in its most traditional form, because it sets a probabilistic foundation for inference on treatment effects distinguishable from those of any other sources of variation (i.e., noise). In fact, a study characterized by only 1 EU per treatment is said to lack experimental replication, as it is not possible to discriminate the source of effects in any legitimate way. From an RR perspective, experimental replication is critical to support the validity of any recommendations for, or against, treatment application.

Distinct from EU, consider the concept of observational units (OU; Kuehl, 2000; Casella, 2008). These are formally defined as physical entities on which the response variable, sometimes referred to as the dependent variable, is observed or measured. As such, OU define the level of observation in the data, whereas EU define the level of experimental replication. This distinction is critical, because the number of EU is not often the same as the number of observations (Mead et al., 2012; Stroup, 2013). Moreover, the distinction between EU and OU is not always obvious and can be elusive, particularly when the wording used to describe a research study is not precise enough, as often results from increasingly limited publication space. Indeed, omissions and deficiencies in the reporting of animal-related research are quite common, often apparent in 40% or more of research articles surveyed for quality of reporting (Kilkenny et al., 2009; Sargeant et al., 2009;

Sargeant et al., 2011). In a recent case report, Bello et al. (2016) illustrated how failure to provide a clear and complete description of how a study was conducted can lead to incorrect analytic decisions that substantially impact inference and conclusions. Overall, proper recognition of the EU requires a detailed understanding of the randomization process or, more generally, of the process of data generation (Milliken and Johnson, 2009; Mead et al., 2012; Stroup, 2013).

To illustrate, consider a study designed to compare 3 treatments in cattle. A total of 36 animals arranged in 12 pens of 3 animals each are available. Assuming that the response of interest (e.g., BW gain or milk yield) is observed on individual animals, one may ask: Is there any experimental replication available to legitimately enable inference on treatment effect from this study? If so, how much? To answer these questions, additional information is required, as more than one experimental arrangement is possible with the details provided this far. Namely, treatments may be randomly assigned and applied to animals individually within a pen (scenario A), or, alternatively, treatments may be randomly assigned and applied to a group of animals in the same pen (scenario B). Without more precise description, it is not possible to ascertain the actual experimental arrangement and thus the proper EU. Of note, regardless of the experimental scenario, the study requires similar resources and will yield a total of 36 observations, 1 per animal (i.e., the OU). Yet, scenarios A and B differ substantially in their data generation process due to distinct logistical restrictions embedded in the randomization process (i.e., can animals be treated and managed independently?).

To formalize the characterization of the data generation process, and thus facilitate proper identification of experimental replication, we work through an exercise called “What Would Fisher Do?” (WWFD) introduced by Stroup (2013). The WWFD exercise is a remodeled moniker of the thought process originally proposed by statistician R. A. Fisher. It is a general strategy intended to translate the description of an experimental design encompassing the process of data collection to an ANOVA shell and to an actual statistical model. In the context of data assumed to be normal, the WWFD exercise is equivalent to the traditional mean squares ANOVA exercise (Milliken and Johnson, 2009), essential to identifying the proper experimental error, and thus distinguishing between EU and OU. Further, the WWFD exercise can easily handle considerably more complex designs and, different from ANOVA, is amenable to non-normal responses (Stroup, 2013, 2015). Anecdotally, we have found the WWFD exercise to be surprisingly user-friendly, and thus well received in both teaching and collaborative settings.

Figures 1 and 2 depict the implementation of the WWFD exercise to our study arranged under scenarios A and B, respectively. To properly characterize the process giving rise to data, it is important that the WWFD exercise outlines the relative position of rows corresponding to treatment structure (i.e., central column of the WWFD table) and rows corresponding to elements of the experimental design (i.e., left-most column of the WWFD table), as well as their combination (i.e., right-most column of the WWFD table). For technical details on the WWFD exercise, the interested reader is referred to Stroup (2013, 2015).

Experimental design or design structure		Treatment structure		Combined	
Source	df	Source	df	Source	df
Pen	$12 - 1 = 11$			Pen = Block	11
		Treatment	$3 - 1 = 2$	Treatment	2
Animal(Pen)	$(3 - 1) \times 12 = 24$			Pen $\times$ Treatment = EU = OU	$24 - 2 = 22$
Total	$36 - 1 = 35$				35

Figure 1. “What Would Fisher Do?” exercise for scenario A: treatments are independently assigned to individual animals within a pen, whereby each treatment is represented in a pen only once. This experiment has the general design structure of a randomized complete block design (RCBD). The arrow indicates the position of the row corresponding to the treatment structure relative to a row of the experimental design that identifies the unit of randomization for treatment and, thus, its experimental unit (EU), as shown in the combined table section. The level of observation of the response variable is indicated by the observational unit (OU). Specific to this experimental design (i.e., RCBD), EU and OU are matched to the same physical entity, namely, an individual animal within a pen, as recognized by the term “pen $\times$ treatment.”

Example: Randomized Complete Block Design

For scenario A, assignment and application of treatment is separate for each animal within a pen. Hence, pen serves as clustering or blocking structure whereas the individual animal nested within a pen is the unit of randomization and, thus, the EU for treatment; this is reflected in the position of the term “animal(pen)” (to be read as animal nested within pen) immediately below the term “treatment” in the left-most and central columns of the WWFD table, respectively (Figure 1). In the combined right-most section of the WWFD table, the term “pen $\times$ treatment” uniquely identifies each individual animal as the combination of a treatment implemented within a pen, with degrees of freedom specified by subtracting the “treatment” degrees of freedom from those of the term “animal(pen).” Simultaneously, the term “pen $\times$ treatment” recognizes the individual animal as the EU for treatment, with corresponding degrees of freedom. In this scenario, the individual animal (as recognized by the term “pen $\times$ treatment”) is also the OU on which the response is measured. The level of observation, and thus the OU, is always identified by the bottom row of the WWFD table; in our example, that is “pen $\times$ treatment.” Notable for scenario A is that both the EU and the OU are matched to the same physical entity; that is, the individual animal identified by the term “pen $\times$ treatment” defines both the OU and the EU, such that the level of observation and the level of experimental replication are one and the same. As a result, in scenario A, the amount of experimental replication for treatment is directly aligned with the number of observations. From an experimental design standpoint, scenario A is referred to as a randomized

complete block design (Kuehl, 2000; Milliken and Johnson, 2009; Mead et al., 2012).

Example: Completely Randomized Design with Subsampling

Consider now scenario B, whereby treatment was randomly assigned to pens, with all animals in a pen receiving the same treatment. Here, the WWFD table (Figure 2) shows the same actual source terms as those in scenario A, though their relative positions in the table are modified to reflect differences in randomization and, thus, in the process of data collection. More specifically, scenario B differs from scenario A in the relative position of the “treatment” row, now immediately above the row for “pen” in the design structure. This reflects the fact that random assignment of treatments was to pen, so that it is the pen that defines the unit of randomization and thus the EU for treatment in scenario B. In the combined section of the WWFD table, “pen(treatment)” (read as pen nested within treatment) has degrees of freedom specified by subtracting the “treatment” degrees of freedom from those of the “pen” row. However, it is still the individual animal that defines the OU in scenario B, as apparent from the bottom row of the corresponding WWFD table (Figure 2). This mismatch between EU and OU in scenario B leads each unit of experimental replication (i.e., EU = pen) to contain multiple units of observation (i.e., OU = animal). As a result, the number of EU (i.e., pens) is clearly smaller than the number of observations (i.e., animals within pens), with the latter defining nonindependent units of technical replication, also known as

Experimental design or design structure		Treatment structure		Combined	
Source	df	Source	df	Source	df
		Treatment	$3 - 1 = 2$	Treatment	2
Pen	$12 - 1 = 11$			Pen(Treatment) = EU	$11 - 2 = 9$
Animal(Pen)	$(3 - 1) \times 12 = 24$			Animal(Pen(Treatment)) = OU	24
Total	$36 - 1 = 35$				35

Figure 2. “What Would Fisher Do?” exercise for scenario B: treatments are independently assigned to pens, whereby all animals in a pen receive the same treatment. This experiment has the general design structure of a completely randomized design (CRD) with subsampling. The arrow indicates the position of the row corresponding to the treatment structure relative to a row of the experimental design that identifies the unit of randomization for treatment and, thus, its experimental unit (EU). The level of observation of the response variable is indicated by the observational unit (OU). Specific to this experimental design (i.e., CRD with subsampling), EU and OU are not matched to the same physical entity.

pseudoreplicates or subsamples (Kuehl, 2000; Casella, 2008). Indeed, if animals within a pen share the same treatment, their observations will, by definition, be correlated thereby failing the criterion of independence required to define the EU. From an experimental design standpoint, scenario B is referred to as a completely randomized design with subsampling. Subsampling, also referred to as pseudoreplication or technical replication, seems to be a common occurrence in animal science research (St-Pierre, 2007; Tempelman, 2009), as number of EU is not generally the same as number of observations. The next paragraphs discuss subsampling in more detail.

Example: Unreplicated Design with Subsampling

A study could, and sometimes does, contain multiple observations for each treatment (e.g., hundreds, thousands, even more), yet fails to provide true experimental replication due to lack of independent repetitions of the experimental conditions. Particularly susceptible are large data sets (i.e., big data), whereby sheer size, either in the way of number of observations (i.e., tall data) or number of variables (i.e., wide data) may be confused with experimental replication. This problem can be further compounded with multiple testing issues due to a large number of different response variables, as will be discussed in a later section. Examples of big data in the animal sciences include high-throughput phenotyping, such as mid-infrared spectroscopy for compositional evaluation of physiological tissues or fluids, video monitoring for tracking animal activity and assessing behaviors, and infrared thermography for evaluations of lameness, inflammation, and heat stress. Further, interest is growing in precision management of livestock by continuous automated real-time monitoring through sensors and image analysis (Berckmans, 2017). Any of these technologies can easily yield massive data sets within short periods of time, even if only a single or a few animals are being measured. To emphasize, the sheer size of a data set is not necessarily informative of experimental replication, and thus cannot replace substantive knowledge or make up for flaws in an experimental design. In other words, simply by observing EU (e.g., animals or pens, depending on the case) more closely or more frequently, one does not increase experimental replication nor observe more EU. More data does not necessarily mean more information (Lazer et al., 2014); therefore, detailed attention to the data generation process and identification of EU and OU becomes all the more important in the context of large data sets. In fact, foundational concepts of experimental design may be more important here than ever, particularly in the interest of ensuring that research

results can reproduce reliably beyond the confines of a study, no matter how large a specific data set.

To illustrate failure in experimental replication despite multiple observations, consider hypothetical scenario C, in which the 36 animals of our study are now arranged in 3 pens, such that all animals within each pen receive the same treatment. Figure 3 implements the WWFD exercise to scenario C following a similar logic to that used for scenario B, recognizing that the unit of randomization, and thus the EU, is the pen. With only 1 pen for each treatment in scenario C, no independent repetition of the experimental condition exists, and thus no experimental replication. This is true whatever the number of animals inhabiting a pen and is formalized in the WWFD table by means of 0 df available for the term “pen(treatment)” on the combined section, which specifies the EU for treatment (Figure 3). In fact, the experimental arrangement in scenario C is irrevocably flawed because, by design, there is no legitimate way to separate treatment effects from any other sources of variability. That is, treatment is confounded with any other effects that might be common to all animals in the pen. Granted, 36 observations can hardly be considered big data; however, the point here is not about the number of observations, but rather the number of EU per treatment. For as long as scenario C is arranged in only 3 pens, a fatal design flaw due to lack of experimental replication will hold regardless of whether each pen consists of 12 animals (as in scenario C), 100, or 1 million. More observations in a data set does not necessarily translate into more information, at least not the appropriate information to address the research question of interest. Most importantly, the problem of lack of experimental replication may not always be obvious, particularly in the context of more complex experimental designs, as discussed by Bello et al. (2016). Admittedly, subsampling can prove beneficial in the way of increased inferential precision and thus improved power (Kuehl, 2000; Tempelman, 2009), particularly if measurement error is substantial or if variability between EU is high (Tempelman, 2005). Yet, these trade-offs show diminishing returns to increasing inputs, so that benefits decrease rapidly with amount of subsampling. Thus, subsamples may be best used as logistical complements to true experimental replication as long as they do not compete for resources with, or deter from, true EU. Overall, regardless of their number, subsamples cannot replace true experimental replication (Stroup, 2015; Bello et al., 2016).

Comparison and Contrast of Scenarios

Clearly, scenarios A and B differ dramatically from C on the presence or absence of experimental replication,

Experimental design or design structure		Treatment structure		Combined	
Source	df	Source	df	Source	df
		Treatment	$3 - 1 = 2$	Treatment	2
Pen	$3 - 1 = 2$			Pen(Treatment) = EU	$2 - 2 = 0$
Animal(Pen)	$(12 - 1) \times 3 = 33$			Animal(Pen(Treatment)) = OU	33
Total	$36 - 1 = 35$				35

Figure 3. “What Would Fisher Do?” exercise for scenario C: each treatment was independently assigned to a single pen, whereby all animals in a pen received the same treatment. This experiment has the general design structure of an unreplicated completely randomized design (CRD) with subsampling. The arrow indicates the position of the row corresponding to the treatment structure relative to a row of the experimental design that identifies the unit of randomization for treatment and, thus, its experimental unit (EU). The level of observation of the response variable is indicated by the observational unit (OU). Specific to this experimental design (i.e., CRD with subsampling), EU and OU are not matched to the same physical entity.

and thus inferential viability. In turn, scenarios A and B differ in the relative amount of experimental replication for treatment. This is despite the fact that each of scenarios A, B, and C have the same number of observations. For A and B, the terms immediately underneath the respective “treatment” row in the combined section of their WWFD tables (Figures 1 and 2) show 22 versus 9 error degrees of freedom for treatment, respectively. This illustrates how mismatches between EU and OU can create design inefficiencies. In scenario A, EU and OU are matched to the same physical entity, so that level of replication and level of observation are one and the same. This, in turn, maximizes the error degrees of freedom for treatment. Instead, in scenario B, the EU is of greater physical size than the OU (i.e., pen > animal), inevitably curtailing the amount of replication for treatment (i.e., smaller degrees of freedom) as sample size is fixed. Although suboptimal, these inefficiencies are sometimes inescapable due to logistical constraints in the implementation of the experiment or to design considerations (i.e., split-plots; Kuehl, 2000; Littell et al., 2006; Milliken and Johnson, 2009; Mead et al., 2012; Stroup, 2013).

Any discrepancies between EU and OU in the data can, and should, be planned for and explicitly recognized as part of the data generation process. Failure to properly recognize pen as EU in scenario B (and even more concerning, in scenario C) would incorrectly default experimental replication to the unit of subsampling (i.e., OU). This would, in turn, artificially enlarge error degrees of freedom for treatment, thereby increasing the chances of type I error (Milliken and Johnson, 2009; Stroup, 2013) and making the study unduly prone to false positives. As will be discussed later, false

positives constitute errors and thus are, by definition, not reproducible. In addition, incorrect specification of EU further misrepresents the reach of generalizability of study results, as will be discussed in the next section.

Also apparent from scenarios A, B, and C, and reflected in the left-most column of their corresponding WWFD tables (Figures 1, 2, and 3, respectively), is the inherently hierarchical arrangement of the data produced by the study design. This is often referred to as data structure or data architecture (Stroup, 2013; Bello et al., 2016), and it reflects a multilevel configuration of the data whereby animals are managed jointly in pens, whereas groups of pens may be under a business framework defined at a higher, say farm, level. Implied by this hierarchical data structure is that observations from individual animals are not mutually independent but rather correlated in some multilevel clustered way, maybe within pens and possibly also within farms. These correlation patterns may be due, for example, to a shared management or environmental background (e.g., feed batch, assigned personnel, weather conditions), to within-pen dynamics (e.g., dominance, competition), or to unforeseen synergisms or antagonisms specific to treatment and pen combinations (Tempelman, 2009). It is worth realizing, though, that data architecture and its implications for experimental replication may not be overtly apparent from a data spreadsheet (Stroup, 2013), particularly in large data sets, and can be easily overlooked (Sargeant et al., 2009, 2011).

Recap

From the standpoint of RR, we emphasize that observations do not necessarily represent replication. Yet,

the specification of EU is not a one-size-fits-all, but rather is uniquely dictated by how the experiment was designed and how the randomization was carried out. That is, the specification of EU is not a matter of opinion. Undoubtedly, subtle changes in how an experiment was run can have profound effects on the amount of experimental replication available for treatment (Milliken and Johnson, 2009; Mead et al., 2012; Stroup, 2013). This is explicitly illustrated in scenarios A, B, and C, for all of which 36 observations were available but offered very different amounts of experimental replication, as illustrated in the corresponding WWFD tables.

SCOPE OF INFERENCE

Key Concepts

The concepts of experimental replication and hierarchical data architecture discussed thus far underlie an overarching statistical principle that constitutes a central tenet of RR. We are referring to scope of inference, a concept directly concerned with *defining the population to which results from a research study are applicable*. In so doing, scope of inference addresses the question “How generalizable are my results?” or, alternatively, “Where can I reasonably expect results to reproduce?” (McLean et al., 1991; Robinson, 1991).

The process of defining scope of inference for a given study reveals an inherent tension between the tightly controlled conditions in which research is often conducted and the real world of commercial livestock production to which research results are ultimately intended to be applicable (Mead et al., 2012). Ideally, a population of interest would be sampled at random to ensure proper representation and, thus, straightforward back transfer of results. However, in practice, logistical constraints and a need for efficient management of the research often lead scientists to work with purposely selected samples (Mead et al., 2012). For instance, rather than randomly selecting dairy cows from a herd, recruitment of animals may be explicitly aligned with farm protocols that make certain groups of animals more easily accessible to the researcher (e.g., incoming cohorts of cattle into a feedlot, or weekly clusters of animals scheduled to complete a voluntary waiting period). Similarly, commercial farms recruited for research are often targeted based on high standards of management and recordkeeping, as well as a reasonable proximity from the sponsoring institutions, particularly if the data collection process calls for multiple farm visits. Sometimes, sample selection may be directed even further to tightly control variation by using populations narrowly defined for homogeneity (e.g., only primiparous or only multiparous cows). Ironically, research

results are likely to be useful only if applicable to a wide population, often beyond the specifications of a research setting. This discrepancy is relevant because research results can only be expected to reproduce within the confines of the population sampled, however it may have been defined. That the sampled subset used in a study be, at least, reasonably representative of the population of interest is a legitimate consideration and can justify having a proof-of-concept study originally conducted in a highly controlled research site be repeated in commercial operations under real-world business-like management. Indeed, this has been an explicit motivation for repeating proof-of-concept studies in commercial settings (Cull et al., 2012; Gonçalves et al., 2016) and across multiple herds in a region (Stevenson et al., 2006, 2008). Further, operational data regularly recorded in commercial herds also can be used to gain realistic broader-reaching insight beyond the confines of research settings (Rosa and Valente, 2013).

Hierarchical Systems and Scope of Inference

When addressing the question of “How generalizable are my results,” scope of inference also delineates which levels of organization in the production system are relevant to elicit research conclusions. This is a worthwhile consideration, particularly under the framework of systems biology (Sinoquet, 2014), whereby multiple levels of organization may be simultaneously at play. Levels might range from fine-grained (e.g., molecules or cells) to physiological mechanisms in whole living organisms (Sinoquet, 2014), or from individual animals to management clusters (e.g., pens or herds) to a whole livestock production industry across national or even international landscapes (Tempelman, 2010). In other words, scope of inference frames the inherently multilevel hierarchy of biological systems and thus outlines the expected reach of results. Ideally, a study would balance representativeness over multiple such levels of organization, thus enabling meaningful conclusions across all relevant scales of inference. In practice, though, data collection at such a broad scope is seldom, if ever, feasible, and studies often require focused selection of a working subset of such levels of organization (e.g., animal- and herd-data).

In principle, it is the highest level of organization represented in a data set that ultimately outlines the reference population represented and, thus, the largest possible scope of inference for any conclusion drawn from the study (McLean et al., 1991; Robinson, 1991). For instance, the inferential space for a study may be narrowly defined and extremely local, limited to a single animal (e.g., if data were obtained from multiple tissues sampled from that individual), to a manage-

ment-defined group of animals (e.g., pen), or even to a single herd (e.g., if multiple animals from that single herd were sampled; Tempelman, 2009, 2010). The latter is a common occurrence in the animal sciences, as many studies are conducted within the confines of an individual herd, be it a research site or a commercial operation. Although likely selected for legitimate reasons, the selected herd represents a single realization from a population of herds. Even if considered typical of the population, the single selected herd grants no true replication from such population and, as a result, the inherent variability among herds cannot be estimated. Data collected from a single herd delineates a strictly narrow inferential space (McLean et al., 1991; Tempelman, 2010) limited to just that herd. Admittedly, single-herd studies may serve as pilot and provide preliminary evidence for further multiherd experimental studies or even to guide exploration of observational field data from commercial livestock operations (Rosa and Valente, 2013). Yet, it is important to realize that results from narrowly scoped studies represent just “local truths” with little external validity (Richter et al., 2009) beyond the specifics of a given study. Results may not generalize to other herds, let alone the overall industry. Presuming otherwise is a disservice to RR. Indeed, empirical evidence indicates that treatment effects as observed in a single herd may not necessarily reproduce in others, either in relative magnitude or even in relative ranking of performance (Stevenson et al., 2008).

The Role of Multiple-Site Studies in Scope of Inference

From an experimental design standpoint, one potential solution around too narrow an inference space might be to enlarge the research platform and introduce additional controlled variation that allows representation of the desired scope of inference. For example, a broader applicability may entail replicating the same experiment in multiple herds and multiple locations, thereby enabling generalization of treatment effects beyond an individual herd. This is a compelling argument for concerted efforts of multiple research institutions and commercial entities with matching funding from federal and industry sources (Stevenson et al., 2006, 2008; Tempelman et al., 2015). Replicating a study in multiple sites across a region, for example, could avoid the danger that any observed differences are determined by the unique, though not necessarily generalizable, experimental conditions inherent to a narrowly defined group of animals in a very precise set of circumstances at a single location. An even broader scope of inference may involve national and international efforts

to include data from across geographical boundaries (Tempelman et al., 2015). In this regard, one should recognize the possibility of treatment-by-study interactions (St-Pierre, 2001), as well as the potential utility of study-specific explanatory variables (e.g., ambient temperature) that might be useful in inferring the nature of that interaction. Observational data routinely collected for operational purposes in commercial herds, albeit with other inferential limitations (Rosa and Valente, 2013), can naturally support a broader scope of inference provided it includes multiple operations.

Even of greater practical relevance, a multiple-site study could, and should, be used to assess the extent to which differences between treatments are consistent over different herd environments. This is a legitimate question with a solid basis both from theory (Gates, 1995) and from empirical evidence (Stevenson et al., 2008). In this context, scope of inference also can be framed as an issue of herd $\times$ treatment interactions driven by herd-specific treatment effects. This is analogous to the concept of genotype-by-environment interaction in quantitative genetics (Tempelman, 2010) or study-by-treatment interaction within the context of the increasingly popular meta-analysis (St-Pierre, 2001).

Example: Randomized Complete Block Design with Subsampling

To illustrate, recall our proposed study concerned with comparing 3 treatments in cattle. Consider now a scenario D, in which 10 herds from across the US Midwest are recruited, with a total of 150 animals, 50 per treatment, in each herd. Figure 4 depicts the implementation of the WWFD exercise for scenario D. Note the term “subset(herd)” (read subset nested within herd) on the left-most column of the table, whereby the subset identifies each group of 50 animals within a herd environment and subjected to a common treatment. As discussed in the previous section, *observations from animals within a common environment, be it a pen or a herd, that also share the same treatment will, by definition, fail the EU criteria for independent repetitions of the experimental condition* (i.e., treatment). Indeed, observations from shared environment-by-treatment combinations can be expected to be correlated. It then follows that, for multiherd studies, observations at the animal level serve, at best, as subsamples. Instead, it is the subset of 50 animals assigned to a treatment within a herd that, as a whole, defines the level of experimental replication and, thus, the EU for treatment in a multiherd study. This is consistent with the definition of experimental error for block designs based on traditional derivation of expected mean squares (Gates, 1995), and has been

recognized in the animal sciences as a randomized complete block design with subsampling (Tempelman, 2009). Indeed, the cross-product term “herd × treatment” on the right-most column of the WWFD exercise recognizes the subset of 50 animals assigned to a treatment within a herd as the EU for treatment. In turn, it is still the individual animal that defines the OU in scenario D, as apparent from the bottom row of the WWFD table in Figure 4. If applicable, one may consider additional partitions of this OU term to recognize any intermediate within-herd organization, say pens within a herd and animals clustered within such pens. Although conceptually aligned with data architecture, further partitioning of WWFD terms below the EU is of little practical relevance, as doing so only dissects subsampling without any contribution to experimental replication (Gates, 1995). The EU for treatment in scenario D still would be defined at the level of “herd × treatment,” regardless of how individual animals were organized within a herd-by-treatment combination.

In addition, the term “herd × treatment” formalizes a scope of inference of broader reach, enabled by the inclusion of multiple herds into the study (Tempelman, 2009), similar to that of study-by-treatment in the context of meta-analysis (St-Pierre, 2001). It is then feasible to assess the consistency of treatment effects across a wider range of management or climatic condi-

tions or both, characteristic of the population of herds that is the target of inference. Note that the term “herd × treatment” is just another level in the hierarchical architecture of the data. As such, scope of inference elegantly crystalizes as an additional component of data architecture that clearly delineates the expected reach of results within livestock production systems.

DATA ANALYSIS OF STRUCTURED DATA

From a practical standpoint, appropriate learning from structured data requires that foundational principles of experimental design be closely linked with data analysis. Arguably, it is not enough for a study to produce data that provide experimental replication for treatments of interest and that are adequately representative of the desired levels of organization. To elicit sound conclusions, it is also necessary that data be analyzed with rigorous statistical methods that are carefully selected and soundly implemented to recognize the data generation process and the resulting structure of the data. In so doing, one incorporates into the analysis any dependencies between observations and carefully dissects EU from OU (Milliken and Johnson, 2009; Stroup, 2013). Yet, a staggering number of standard statistical methods commonly used for data analysis, particularly those taught in introduc-

Experimental design or design structure		Treatment structure		Combined	
Source	df	Source	df	Source	df
Herd	10 – 1 = 9			Herd = Block	9
		Treatment	3 – 1 = 2	Treatment	2
Subset(Herd)	(3 – 1) × 10 = 20			Herd × Treatment = EU	20 – 2 = 18
Animal(Subset (Herd))	(50 – 1) × 3 × 10 = 1,470			Animal(Herd × Treatment) = OU	1,470
Total	1,500 – 1 = 1,499				1,499

Figure 4. “What Would Fisher Do?” exercise for scenario D: treatments are independently assigned to individual animals within a herd and each treatment is represented in a herd more than once. This multierd experiment has the general design structure of a randomized complete block design (RCBD) with subsampling. The arrow indicates the position of the row corresponding to the treatment structure relative to a row of the experimental design that identifies the unit of randomization for treatment and, thus, its experimental unit (EU). The level of observation of the response variable is indicated by the observational unit (OU). Specific to this experimental design (i.e., RCBD with subsampling), EU and OU are not matched to the same physical entity.

tory statistics courses (e.g., correlations, *z*-tests, *t*-tests, traditional ANOVA, regression), implicitly assume that observations are mutually independent and that data structure is nonexistent. Most concerning, these critical assumptions often go understated, thus effectively disregarding any correlation patterns in the data induced by the hierarchical structure of data architecture. As a direct consequence, experimental replication is naively matched one-to-one with level of observation in the data (Stroup, 2013) and the inferential space is misrepresented (Tempelman, 2009). This oversight can have serious downstream implications for inference (Aitkin and Longford, 1986) and provide a false sense of security on results, meanwhile undermining RR.

Available Statistical Methods and Implementation Tools

The flexible statistical framework of mixed models offers one plausible strategy uniquely suited to these challenges. Mixed models are inherently hierarchical models, and thus can seamlessly encode a multilevel data generation process into the analysis phase (Littell et al., 2006; Milliken and Johnson, 2009; Stroup, 2013) and naturally delineate the inferential scope (McClean et al., 1991). Through the specification of fixed and random effects, mixed models can simultaneously assess multiple sources of variability in a data set while properly recognizing the EU, especially when mismatched with the OU (Mead et al., 2012). Fixed effects reflect the treatment structure that characterize the effects of selected factors (i.e., treatments) or explanatory covariates on the response of interest (Milliken and Johnson, 2009). Meanwhile, random effects reflect the design structure or data collection process (Milliken and Johnson, 2009). Simultaneously, random effects partition variability across hierarchical levels of data architecture, such as animal, pen, and farm levels, while recognizing one or the other as the EU for a treatment of interest, thereby calibrating hypothesis testing appropriately (Tempelman, 2009). In so doing, the random effects in a mixed model naturally characterize levels of organization in a biological system to the extent represented in a given study.

Despite their flexibility, mixed models are also sophisticated tools and their proper implementation can seem cumbersome at times, especially in cases of limited statistical expertise and complex experimental designs. Granted, the computational aspects of mixed model implementation are now greatly facilitated through widely available software, such as SAS procedures (e.g., GLIMMIX, MIXED, NLMIXED), R packages (e.g., lme4, nlme), and ASREML routines, among others. However, ease of computation may at times be

a double-edged sword, as the actual specification of a mixed model is not a task directly amenable to software automation. In fact, advanced statistical expertise may be required for a model to properly reflect the many subtleties of the data generation process and the resulting data architecture.

Especially controversial in the animal sciences seems to be the decision of what effects should be included in the linear predictor, particularly those specified as random (Bello et al., 2016). The importance of this issue is worth emphasizing, as it has explicitly been recognized that

“...improper attention to the presence of random effects is one of the most common and serious mistakes in the statistical analysis of data,” (Littell et al., 2002)

and that

“...translating the study design into a statistical model is perhaps the most important facet in the practice of statistical modeling ... Also the most undervalued and overlooked” (Stroup, 2013).

Therefore, the standing recommendation is that data analysis be based on a model specification that directly reflects data architecture and, thus, is justified both by the data generation process and by how randomization was implemented (Milliken and Johnson, 2009; Mead et al., 2012; Stroup, 2013). The WWFD exercise can again serve as a practical tool to facilitate the process of model specification given that it fully characterizes the process of data collection. Design flaws notwithstanding (e.g., recall scenario C), one may then collect the row terms of the combined section of the WWFD table and transfer them into a linear predictor that specifies the mixed model. Those row terms originated from the central column of the WWFD table characterize selected groups of interest for direct comparison (e.g., treatments, demographics) and should thus be fitted as fixed effects (McClean et al., 1991). In turn, terms originated from the left-most column of the WWFD table represent the “plot plan” or design structure of the experiment (Stroup, 2013) and should be specified as random effects (McClean et al., 1991). The only exception might be the bottom row of the combined WWFD table, which, as the reader might recall from a previous section, reflects the OU. In the specific case of normally distributed data, any leftover variation at the level of observation is implicitly captured by residuals; thus, there is no need to specify the bottom term of the WWFD table into the linear predictor. Supplemental File S1 (<https://doi.org/10.3168/jds.2017-13978>)

outlines the linear predictors for specification of mixed models corresponding to scenarios A, B, C, and D.

Notably, the WWFD exercise is not one of automation to specify a mixed model. As explained in Supplemental File S1 (<https://doi.org/10.3168/jds.2017-13978>), no appropriate model specification could salvage the lack of experimental replication in scenario C. That is, for all their sophistication, the use of mixed models for data analysis cannot undo flaws incurred at the level of experimental design. As stated by statistician Ronald A. Fisher (1890–1962), “To consult the statistician after an experiment is finished is often merely to ask him to conduct a post-mortem examination. He can perhaps say what the experiment died of.” Further, the connection between appropriate model specification and large data sets should also be mentioned, as the consequences of a misspecified model are likely to be especially harmful in situations with many variables in which errors could easily be propagated. Further, bias resulting from misspecification of statistical models cannot be overcome with a large sample size (Van der Laan and Rose, 2011).

Common Hurdles

In the context of RR, specification of mixed models is often faced with conundrums, some of which are common enough to merit a more detailed discussion. The first one refers to whether the effect of any general blocking structure should be specified as a fixed effect or as a random effect in a statistical model. Particularly contentious examples in the animal sciences include the specification of blocking effects of herds in multisite studies (Tempelman, 2009) or of pens in single-herd studies (St-Pierre, 2007). In general terms, specifying blocks as random effects allows for the recovery of interblock information, thus enhancing inferential efficiency (Yates, 1940). Random blocks also allow for a broader scope of inference of treatments of interest beyond the blocks specifically involved in a study, encompassing the population of blocks from which those enrolled in a study are presumably representative (Tempelman, 2010). Instead, a fixed effects specification of blocks (i.e., herds, pens) formally impedes such broad interpretation, despite any generalizability intended by the researcher (McLean et al., 1991). That is, with herd blocks or pen blocks treated as fixed effects, the scope of inference is strictly narrow and is limited to only those herds or pens in the study, regardless of the number of animals recruited into the study (Tempelman, 2009), with direct implications for impaired RR. Further, whereas modeling of herds as fixed effect does account for mean differences between herds,

interblock information is lost and correlation patterns of observations collected within a herd are effectively disregarded. As a result, model specification is oversimplified, leading to biased estimates and inefficient inference (Aitkin and Longford, 1986). One should keep in mind, though, that specification of blocks as random variables with a normal distribution, as opposed to fixed effect constants, requires estimation of a block-level variance component. This is arguably an onerous task that naturally calls for careful reflection on whether the number of blocks available in a study is sufficient to estimate any between-block variability. If not, one should probably reconsider the practicality of fitting blocks as random and the ensuing reach (or lack thereof) of the inferential space, along with the expected RR of any subsequent findings beyond the blocks recruited for a study.

A second hurdle in the specification of mixed models is often encountered in experimental designs characterized by blocks, within which treatments are represented multiple times. For example, studies in which different treatments are assigned to multiple animals within a pen or within a herd. In this case, the conundrum lies on whether the term “block $\times$ treatment” (i.e., “herd $\times$ treatment” or “pen $\times$ treatment,” as appropriate in each case) should be specified at all in the statistical model. In other words, whether the design arrangement should be treated as a randomized complete block design with subsampling (which calls for explicit specification of the term “block $\times$ treatment” in the linear predictor) or as a so-called generalized block (which bypasses the term “block $\times$ treatment” and implicitly pools it with the level of observation; Gates, 1995). Beyond statistical interactions, the question underlying “block $\times$ treatment” refers to the extent to which differences between treatments are consistent over different environments, as defined by the blocking structure (i.e., herds, pens). This is undoubtedly a worthwhile question, particularly under the umbrella of RR. First, if a treatment is represented more than once within each block, then, different from scenario A, the term “block $\times$ treatment” is separately identifiable from the OU (i.e., the individual animal) and can be explicitly recognized in the architecture of the data, in a way analogous to the term “herd $\times$ treatment” in scenario D (Figure 4) with block defined by herd. Second, if the blocking structure is modeled as a random effect (as advocated in the previous paragraph), the term “block $\times$ treatment” will by definition also be random. This, in turn, ensures that the experimental error term for testing treatments comprises the combination of block and treatment as a source of variability, as supported by the formal derivations of mean square expectations (Gates,

1995). This, in turn, removes any need to assume that treatment effects are consistent across blocks; in fact, it allows for its testing.

INTERPRETATION OF *P*-VALUES

The Underlying Mechanisms of Hypothesis Testing

A properly specified mixed model automatically determines the appropriate experimental error and calibrates hypothesis testing and elicitation of *P*-values, thus enabling us to carry this discussion forward to the arena of statistical inference. Albeit rather technical in nature, a proper discussion of *P*-values is critical to RR. Indeed, the seriousness, persistence, and widespread occurrence of misinterpreted *P*-values prompted the American Statistical Association, for the first time in its 177-year history, to release an official “Statement on statistical significance and *P*-values” (Wasserstein and Lazar, 2016). The statement warns about the misuse of *P*-values and outlines principles for good practice with the ultimate goal to “inform the growing emphasis on reproducibility of science research.”

We start by considering the primary goal of a research study, where the researcher wants to know from the data about the probability that, for a response of interest (e.g., milk yield), a difference between treatments A and B truly exists. In statistical notation, we write this question as

$$P(\mu_A - \mu_B = 0 \mid \mathbf{y}), \quad [1]$$

where $P(\cdot)$ indicates a probability value between 0 and 1; μ_A and μ_B indicate the true mean response under treatments A and B, respectively; the vertical bar is read “conditional on”; and $\mathbf{y}$ is the observed response. We briefly note that this interpretation in terms of “probability of truth given observed data” constitutes a Bayesian formulation of the research question (Gelman et al., 2013), though further elaboration is beyond the scope of this review.

Perhaps surprisingly to some readers, the expression in Eq. [1] is not a *P*-value. A *P*-value is defined as the probability that the response in the data show treatment differences at least as large as those actually observed in the current study. More formally, a *P*-value is

$$P \left(\begin{array}{c} (\bar{y}_A - \bar{y}_B) \text{ at least as extreme} \\ \text{as the value observed} \end{array} \middle| \begin{array}{c} H_0 : \mu_A = \mu_B, \\ \text{ORS} \end{array} \right), \quad [2]$$

where $P(\cdot)$ is defined as above; $\bar{y}_A$ and $\bar{y}_B$ indicate data-based sample averages in milk yield for treatments A

and B, respectively; $H_0 : \mu_A = \mu_B$ states the null hypothesis of nonexistent true treatment differences; and ORS stands for “over repeated sampling.” This is precisely how the concept of a *P*-value was originally proposed by statistician R. A. Fisher early in the 20th century (Fisher, 1925; p 46–47).

Despite their striking differences, it is an all-too-common fallacy to confuse the probability in Eq. [1] with the concept of a *P*-value. To explicitly highlight their discrepancies, we note the following. First, a *P*-value assumes that the experiment will be repeated many times, technically an infinite number of times (i.e., ORS) under exactly the same conditions. Thus, a *P*-value projects what would have happened under other possible data sets (not the observed data) in a make-believe world of infinite repetitions. Second, from inception, *P*-values assume that treatments are truly ineffective (i.e., treatment differences are nonexistent or $\mu_A = \mu_B$) and any evidence to the contrary is due to pure chance. This is hardly a reasonable assumption as, realistically, researchers often design experiments using strong substantive rationale for treatment effects so that frequently the actual goal of the experiment is to quantify those effects. Taken together, these discrepancies highlight the notoriously counterintuitive information conveyed by *P*-values and how far removed they are from the notions of truth and knowledge of interest to the researcher (Berger and Sellke, 1987). Indeed, so defined, a *P*-value is just a measure of incompatibility between observed data and hypothetical postulates; that is, an indication of implausibility of what was actually observed given a set of working assumptions.

Technicalities aside, a more colloquial, though legitimate, interpretation of a *P*-value might be that of an indicator of weirdness or a measure of surprise in the data given expectations from a set of assumptions (Reinhart, 2015). The smaller the *P*-value, the more surprising the results obtained from the data can be considered to be. One might then interpret the surprise associated with a small *P*-value as a rare random occurrence due to chance [i.e., a(n) (un)lucky hit?] or, as the statistical argument goes, as empirical evidence against the working assumption. Specifically, a small *P*-value may then be interpreted as an open-minded suspicion that the underlying assumption of no treatment effects was wrong to start with and that the treatment actually works. Although counter-intuitive, this is exactly how *P*-values work: by raising suspicion on status-quo assumptions.

Probably just as important as defining what a *P*-value is might be to clarify increasingly common misconceptions of what a *P*-value is not (Goodman, 2008). Specifically, a *P*-value is not a gauge of how right a researcher is on his or her conclusions, neither is it an

indicator of how likely he or she is of having made a mistake (this is actually related to statistical errors; see next section). Indeed, the actual value of P does not indicate the chance that results are incorrect. Further, a P -value is not a mark of how important a treatment difference is, nor a criterion of strength or magnitude of a treatment effect of interest. Simply put, a P -value serves just as a measure of surprise on results obtained from observed data relative to the working assumption of no treatment differences.

Thresholds for P -Values

As a follow-up question, one might further ask how surprising is surprising enough? Without much formality, R. A. Fisher opted for 1 in 20 as “frequent enough” (Fisher, 1925), as only 5% of possible values of an event of interest are more extreme than this benchmark. Thus, with a P -value at or below 0.05, all one knows is that if treatment effects are truly nonexistent, treatment differences at least as large as those observed in the data would be obtained in approximately 1 (or fewer) in 20 experimental repetitions. The cutoff value 0.05 has since received disproportionate attention from the scientific community as a practical rule of thumb, being embraced at a level akin to doctrine in the “culture of $P < 0.05$ ” that is prevalent even to this day. However, there is nothing mathematically rigorous or otherwise particularly elegant about a 0.05 threshold (Cowles and Davis, 1982). In fact, the threshold for significance, be it 0.05, 0.01, or, as recently proposed, 0.005 (Benjamin et al., 2017), is little more than an arbitrary choice. Even with tiny treatment differences, one could obtain small P -values, below any arbitrary threshold, by increasing recording precision (i.e., sophisticated technological equipment) or by collecting a very large number of observations on a given physical unit (i.e., subsampling). Even chance differences may be incorrectly labeled as significant due to unduly large sample sizes (Granger, 1998). In other words, without further context, a statistically significant conclusion on a treatment effect may not have any practical significance.

As a side note, we briefly mention the misleading misnomer of a “trend” or “tendency towards significance” that is sometimes used incorrectly to interpret a P -value that hovers close to a selected threshold but does not quite make the cut (e.g., somewhere between 0.05 and 0.10). Whereas the significance threshold itself is admittedly arbitrary, wording of “trend” or “tendency” inappropriately designates directionality to the P -value. This interpretation lacks any foundation, either from underlying theory or from observed data, and is thus unwarranted. At best, such hovering

P -values may suggest a greater willingness to accept a false-positive result (though there are also caveats to this interpretation); this is sometimes referred to as marginal significance.

Interpreting Nonsignificant P -Values

Although the practical relevance of some statistically significant results may be questionable, nonsignificant results are inconclusive at best. That is, claiming practical relevance of numerical differences in the absence of statistical significance is uninformative in any way. *Lack of statistical significance prevents discrimination between a potentially real treatment effect for which data were insufficiently informative and a spurious numerical difference due to chance.* As such, any claims based on numerical, though nonsignificant, differences are, by definition, not reproducible.

Similarly, nonsignificant results do not indicate nonexistent differences (Reinhart, 2015). In other words, *absence of evidence* (i.e., failure to reject H_0 based on nonsignificant P -values) is not the same as *evidence for absence* (i.e., acceptance of H_0). The rationale behind this premise lies on the inherently asymmetric mechanism of hypothesis testing. Recall that in the context of classical hypothesis testing, a hypothesis set is elicited as a null hypothesis of no treatment differences (i.e., $H_0 : \mu_A = \mu_B$) posed against an alternative hypothesis that treatments differ in some way (i.e., in general terms, $H_1 : \mu_A \neq \mu_B$, though in some cases $H_1 : \mu_A > \mu_B$ may be appropriate). Despite the duality of the question posed by H_0 and H_1 , the statistical test itself is not a balanced proposition between the 2 and neither is the resulting P -value. That is, the choice between H_0 and H_1 is not a symmetrical one. By design, the null hypothesis H_0 already is the benchmark for assessing surprise from data, so that, from the outset, the odds of statistical inference are stacked with H_0 and against H_1 . Therefore, H_0 serves as a fallback default that characterizes the assumed status quo; H_0 is already accepted and it can only be disproved based on data. Indeed, data that tilts the scale from H_0 to H_1 provides evidence to suspect the former and instead reconsider the latter (i.e., reject H_0 in favor of H_1). However, data that fails to tilt the scale does not necessarily constitute proof for H_0 , because, as indicated above, H_0 is the start-line working assumption. Failure to reject H_0 may be due to nonexistent treatment effects (implying equivalent treatment behavior), but this possibility is not distinguishable from a scenario in which treatment differences actually exist but the study lacks power to detect them. Further, it is not possible to discern one possibility from the other using the same data used for hypothesis testing, thereby rendering moot the so-

called notion of “post-hoc power” (Hoenig and Heisey, 2001; Lenth, 2007), also known as retrospective power or observed power. Sometimes incorrectly advocated to aid in the interpretation of nonsignificant results, post-hoc power is a fundamentally flawed misconception due to its one-to-one relationship between P -values (Hoenig and Heisey, 2001); that is, nonsignificant P -values are by definition associated with low power (Lenth, 2007). Trying to justify one with the other is little more than a circular argument. Put simply, in the context of classical hypothesis testing, lack of evidence to tilt the scale (i.e., nonsignificant results) only indicates *absence of evidence*, not to be confused with *evidence for absence* of an effect of interest.

As such, nonsignificant results may be reported in terms of “no evidence for treatment differences” or even “nonsignificant treatment differences.” We highlight the difference with wording of “no treatment differences,” “no impact or no effect of treatment on outcome,” or, even worse, “no change due to treatment,” all of which incorrectly imply evidence of absence and are thus akin to inappropriately accepting the null hypothesis. Similarly, conclusions worded in terms of “similar treatment outcomes” or “equivalent treatment effect” are also incorrect; inconclusive results is the most that can be said from data. This seemingly semantical disparity on the wording of nonsignificant results has practical implications that are extremely powerful, particularly for ensuring transparent communication of science to society. Painstaking word choice is of utter importance, particularly for results associated with nonsignificant P -values. Whenever an assessment of treatment equivalence is of legitimate interest, classical hypothesis testing is not the appropriate statistical methodology (Tempelman, 2004), as one can never accept the null hypothesis of no treatment differences; instead, equivalence testing procedures should be considered (Hoenig and Heisey, 2001; Tempelman, 2004).

Recap on P -Values

In summary, P -values should be carefully considered and not be taken at face value when making scientific claims, regardless of the actual thresholds selected to claim significance. This is the case when test statistics yield nonsignificant results and extra care is needed to ensure that results are properly worded. Just as important is to avoid blind and misguided use of significant P -values, paradoxically even if associated with a claim that is consistent with existing theory. This has been referred to as “hypothesis myopia” (Nuzzo, 2015), whereby expected results are more likely to get a pass and only nonintuitive ones are subjected to scrutiny.

Rather, P -values should be used in combination with rigorous experimental design, properly specified statistical models, and sound estimation of parameters and their uncertainty (e.g., SE, CI; Ioannidis, 2005), along with a keen understanding of mechanisms and assumptions underlying inference. Only when so implemented and interpreted can P -values play a meaningful role for, rather than against, RR.

STATISTICAL ERRORS

In the statistical process of learning from noisy data, legitimate errors are not only possible, but rather bound to happen. That is because the entire population of interest is seldom accessible and, instead, researchers work with data from a subset of such population, with the subset hopefully being a random sample or at least a purposely representative one. In either case, samples are not perfect replicates of the population and, as a result, the information conveyed by sample data is imperfect at best. Therefore, even in best-case scenarios of sound experimental design, properly specified statistical models, and well-implemented hypothesis testing, it is still not possible to unambiguously know the truth. For better or for worse, statistics cannot provide 100% assurance, so it is not possible to know whether the decision made from data on a given research question is correct or not. Statistical errors are inevitable and cannot be eliminated, but they can be controlled within acceptable boundaries. Doing so is critical to ensuring that research results are reproducible, given that statistical errors are, by definition, circumstantial mistakes and thus not necessarily reproducible. Supplemental File S2 (<https://doi.org/10.3168/jds.2017-13978>) provides a refresher of the traditional types of statistical errors; namely, type I and type II.

Multiple Testing and Inflation of Type I Error

Setting a type I error rate α of say, 5%, pertains to an *individual* hypothesis test representing a single specific research question posed from a data set. Realistically, most experiments are purposely designed to efficiently address multiple questions simultaneously. For example, a study consisting of a 2-way factorial treatment structure poses an implicit interest not only on the main effects of levels of a factor A or a factor B, but on potential synergisms or antagonisms arising from combinations of these factors (i.e., interaction). As a result, the researcher is faced with many possible comparisons either between levels of, say, factor A averaged across levels of B (i.e. marginal effects) or between combinations of A-by-B levels (i.e. simple effects), not

to mention potential tailored contrasts that further dissect specific aspects of the question. In addition, there are often several outcomes of interest for which data were collected and questions are being asked. In this context, numerous questions are being asked and the opportunity for statistical error increases. Thus, a legitimate concern is whether any observed significant differences might be due to the compounded effect of a larger number of questions on chance. This is often referred to as multiplicity or a multiple testing situation and is explained by an inevitable inflation of type I error rates resulting from many tests. Informally, just by asking many questions one increases the probability of “finding something” in the data, even if there is nothing to be found. Error inflation due to multiplicity seems to be particularly risky when effects are subtle, variability is large, and sample sizes are small (Gelman and Loken, 2014), all of which describe research conditions in modern animal agriculture, characterized by limited funding despite research questions of increasing complexity.

We illustrate inflation of the probability of false positives due to multiplicity (Figure 5). Assume a comparison-wise type I error rate $\alpha = 0.05$ for any individual statistical test posed. As multiple tests are conducted, each with $\alpha = 0.05$, one may tentatively quantify the so-called experiment-wise type I error rate α^* as

$$\alpha^* = 1 - (1 - \alpha)^c, \quad [3]$$

with c indicating the number of tests posed from a given experimental data set (adapted from Milliken and Johnson, 2009). Here, α^* describes the probability of “finding something” in a study if there is nothing to be found. Note from Figure 5 the rapid increase of α^* as a function of increasing number of tests, whereby as few as 14 comparisons suffice to yield an approximately 50% coin toss-like chance of finding something significant in the data; with 45 tests, one is almost sure ($\sim 90\%$ probability) to do so. In other words, specifying a significance threshold of, say, $\alpha = 0.05$ in each of any number of statistical tests amounts to allowing for an unreasonably high probability of false positives in the overall study. In the context of animal health and production, an experiment with this many or even more comparisons legitimately implicit in the treatment structure is hardly uncommon. Consider, for example, an experimental setup consisting of a 2×3 factorial treatment structure (e.g., all combinations of 2 diets and 3 drugs, or a total of 3 selected treatments evaluated in steers and heifers or in primi- and multiparous animals) for which 5 response variables may be measured weekly over the course of 4 wk. So designed, the treatment

Figure 5. Inflation of experiment-wise type I error rate α^* due to multiple comparisons, assuming a prespecified comparison-wise type I error rate $\alpha = 0.05$, whereby $\alpha^* = 1 - (1 - \alpha)^c$, where c indicates the number of independent comparisons. Dotted line demonstrates a 50% chance of a type I error ($\alpha^* = 0.50$) with approximately $c = 14$ multiple comparisons. Dashed line indicates $\alpha^* = 0.90$ with approximately $c = 45$ comparisons.

structure implicitly calls for at least 9 directly interpretable simple-effect pairwise comparisons of treatment combinations (i.e., 3 treatment comparisons within each of the 2 groups, plus a group comparison within each of the 3 treatments). These are to be repeated for each of the 5 response variables at each of the 4 time points, thereby yielding $(9 \times 5 \times 4)$ 180 pairwise comparisons. Based on a comparison-wise type I error rate $\alpha = 0.05$ on each individual test, one may conservatively expect at least one false positive with almost-certain probability [i.e., $\alpha^* = 1 - (1 - 0.05)^{180} = 0.9999$]. Even at more stringent comparison-wise type I error thresholds, say $\alpha = 0.01$ or, as recently proposed (Benjamin et al., 2017), $\alpha = 0.005$, the experiment-wise type I error rate α^* will still be inflated at approximately 84 and 60%, respectively.

Additionally sobering is the realization that these estimates of error inflation are only an approximation, as the multiple tests are incorrectly assumed to be mutually independent. The shared source of the data used for multiple testing renders this assumption untenable, thus potentially worsening the problem. Granted, significant results can, and in some cases will, imply correct decisions (i.e., if H_1 is true); however, by definition, the truth ultimately remains unknown so it is impossible to assess which, if any, of the potentially many significant results can be trusted.

Strategies to Control Inflation of Type I Error due to Multiplicity

With multiplicity being inevitable, practices to control experiment-wise type I error α^* can and should be routinely implemented to minimize the occurrence of false positives, and thus of irreproducible findings in the experiment as a whole. The statistical literature on procedures for simultaneous inference and multiple testing corrections to control experiment-wise type I error α^* within acceptable boundaries is large and well developed for many types of experimental settings (summarized by Milliken and Johnson, 2009). Procedures with reportedly reasonable performance include Tukey(-Kramer)'s adjustment for marginal factorial comparisons and Bonferroni's procedure for simple-effect comparisons, selected contrasts, and quantitative regression-type predictors; also performing reasonably are Dunnett's adjustment for comparisons against a reference category or even the most conservative Scheffé's strategy for exploring questions raised after observing the data (i.e., data snooping), to name a few. Bonferroni-type corrections also may be used to correct for the total number of different response variables, particularly in the context of wide data. By contrast, in scientific fields that involve thousands of comparisons for screening and future exploration, such as genetics and genomics, it may be more relevant to focus on false discovery rates (Benjamini and Hochberg, 1995). False discovery rate controls the expected proportion of falsely rejected hypotheses among those that were rejected, as opposed to among those that are truly negative. It should be noted that in so doing, false discovery rate does not control for inflation of type I error per se, but rather for false positives among those declared significant.

In the context of multiple comparisons, it is important to recognize that a single overarching research question driving a study can legitimately decompose into many possible subquestions and comparisons on which error inflation can be prevented. This is to note that multiple testing is not necessarily "*P*-hacking" (Simmons et al., 2011) or the unethical practice of outright fishing for significance, both ethically questionable practices outside of the umbrella of this article.

Selective Reporting

Unfortunately, statistical adjustments for multiple testing to prevent inflation of type I error often go misused or, even worse, entirely overlooked in scientific practice (Kramer et al., 2016). These concerns are only exacerbated by the apparently widespread practice of selective reporting of results (Baker, 2016), whereby an

unknown number of tests likely preceded the subset of significant results that actually make it to publication.

It is conceivable for series of alternative analyses, all of them hindsight justifiable in some way, to be conducted in search for significance (Simmons et al., 2011). Examples of such alternative decisions, so-called "researcher degrees of freedom" (Simmons et al., 2011), might include analyses conducted with (or without) data transformations, or alternatively with (or without) outlier exclusion or even with (or without) selected covariate adjustments. Consider also the all-too-common practice of splitting up big data sets into multiple subsets for analysis or arbitrarily combining experimental conditions (e.g., age groups defined by alternative categorizations of the corresponding continuous covariate). Moreover, these issues can easily be compounded together in data mining exercises, particularly with observational data in which the specific questions may be only vaguely defined and analysis plans may not be predefined (MacKinnon and Glick, 1999). Multiple analytic attempts yield an unprecedented flexibility in the search for significance, and this, in turn, grows exponentially the opportunity for false positives. A "how bad can it be?"-type simulation showed that the practice of combining just a few of these researcher degrees of freedom could lead to a stunning ~60% false-positive rate despite a 5% specification on comparison-wise type I errors (Simmons et al., 2011). In other words, just by trying multiple analytic strategies, a researcher can exert great influence on statistical significance, so much so as to make significance almost inevitable.

Even more concerning, these alternative analytic attempts often go undisclosed and unreported. So when (as opposed to if) significance is attained, it is impossible for the reader or the scientific community at large to make informed decisions about error inflation and credibility of findings.

Recent years have seen a widespread availability of statistical software tools, particularly those of the point-and-click nature, thus facilitating access of animal scientists and other domain researchers to sophisticated statistical methods. Some of these tools showcase a plug-and-play implementation that, although user-friendly, can ultimately serve as a double-edged sword. Unless responsibly used, it is possible for these tools to encourage unwarranted analytic attempts without a full understanding of the underlying methodology, thus escalating the search for significance (Nuzzo, 2015).

One may argue that the problem of researcher degrees of freedom in the search for significance goes well beyond statistics and permeates the academic culture, with its growing pressures to publish or perish (Baker, 2016). In this increasingly competitive environment, the probability of publication seems to depend directly

on novelty and *P*-values (Rothstein et al., 2005; Ridley et al., 2007), with celebration of significance and poor reporting of nonsignificance. As a consequence, consciously or unconsciously, the academic system of evaluation and reward, including investigators' promotion and grant potential, seems to be inherently dependent on whether or not the null hypothesis is rejected. Even without an explicit financial conflict of interest or any ill-willed efforts to mislead, researchers have a stake in the outcome as peer-reviewed publications ultimately define the currency of research (Hirsch, 2005). In addition, little incentive exists to publish studies that reproduce results and, also, an understandable reluctance to publish nonsignificant results (Baker, 2016) due to a perceived high prevalence of underpowered studies (Ioannidis, 2005).

RANDOMIZATION AND THE NATURE OF INFERENCE

Experimental vs. Observational Studies

Randomization constitutes a critical and integral part of the data generation process in research studies both in terms of randomizing the selection of subjects from the population and in terms of randomizing their allocation to (treatment) groups for comparisons (Kuehl, 2000; Mead et al., 2012). Notably, randomization involves a formal mechanism (not a haphazard one) where entities have a known probability of recruitment into the study and of treatment assignment before the randomization process, and the actual treatment allocation or selection is determined by a chance process that cannot be predicted (Dohoo and Waltner-Toews, 1985). Of special interest to us is the process of random treatment allocation, which sets a qualitative distinction between studies described as experimental and those described as observational. Specifically, problems with randomization seem to be common in the experimental design of animal research (McCance, 1995; Kilkenny et al., 2009; Sargeant et al., 2009), reportedly affecting as many as 80% of the studies. In experimental studies, the random allocation of physical entities to different treatments or groups balances out (on average) any systematic differences between the groups, which might, knowingly or unknowingly, favor the outcome of interest in one treatment over another (Kuehl, 2000; Milliken and Johnson, 2009; Mead et al., 2012). Randomization can thus help circumvent so-called confounding bias (Dohoo et al., 2009) by averaging out the confounding factor so that it is equally distributed among the groups being compared. It is the process of random treatment allocation that enables causal inference in experimental studies and defines the gold standard for eliciting cause-and-effect relationships (Dohoo and Waltner-Toews, 1985).

However, randomization may not always be logistically feasible, ethical, or economically viable. In such cases, it may only be possible to collect so-called observational data from selected entities, such as animals or herds already subjected to a condition of interest, and observe their performance outcomes (Dohoo et al., 2009). Due to the lack of randomization inherent to observational data, it is often not possible to unequivocally exclude the potential causal effects of other, confounding, factors, thereby limiting standard inference on relationships to nondirectional associations. Granted, recent developments in the field of causal inference have provided further insight into advanced methodologies and specific conditions under which causal claims from observational data may be warranted (Pearl, 2009; Morgan and Winship, 2017); those such topics are beyond the scope of this review.

Importantly, one should recognize that most biases in observational data cannot be overcome simply by increasing the volume of data (Dohoo and Waltner-Toews, 1985; Theurer et al., 2015). This is particularly relevant for large operational data sets from livestock production systems (Rosa and Valente, 2013), as are commonly used in epidemiological applications (Dohoo et al., 2009).

Wording of Cause and Effect vs. Associations

From a practical RR standpoint, understanding the difference between data originated from experimental or observational studies is critical to prevent unwarranted causal claims from observational data that are unlikely to reproduce. Notably, when 52 causal claims originated from 12 reputable observational studies in the medical literature were subjected to randomized trials, a total of zero of such claims were replicated in the direction claimed by the observational studies (Young and Karr, 2011); in fact, 5 claims were reversed as randomized trials yielded opposite results (Young and Karr, 2011). Clearly, one should maintain an appropriate awareness of the practical implications that the mechanism of randomization has on the wording of inference. Randomized experimental studies inherently warrant cause-and-effect conclusions worded as "effect of" or "impact of" a given treatment on an outcome of interest. In turn, claims from observational studies call for a more careful choice of words, whereby non-directional "associations" or "relationships" between a condition of interest and a response variable is more likely to be appropriate. Stronger, even causal, claims from observational data may be feasible in the more advanced context of modern causal inference (Pearl, 2009; Morgan and Winship, 2017), though this topic is beyond the scope of this review.

ADDITIONAL CONSIDERATIONS

Many more aspects of the principles and practice of statistics can be cited in the context of reproducibility of research findings. We briefly mention a few additional considerations of relevance to the animal sciences and direct the reader to the specific literature for details.

First, consider the phenomenon of interactions between treatment factors, typically in the context of multiple-factorial experimental arrangements. For example, suppose a study designed to evaluate both the effect of an injectable drug and that of a dietary supplement on some measure of productive performance (i.e., ADG). Each of these treatment factors would necessitate an untreated control for reference, thereby defining a 2×2 factorial design with 4 treatment combinations defined by whether the drug or the supplement are administered. After data collection, a rather simplistic exploration of the data might assess the average effect of drug or that of dietary supplement, without regard of the other; these are commonly referred to as main effects or marginal effects. However, one would be remiss to overlook a potential combined effect of drug and dietary supplement, which may or may not have been anticipated. This combined effect is referred to as an interaction and defines a condition beyond linear additivity. More specifically, interactions capture synergistic (or antagonistic) effects between treatments, whereby the whole is more (or less, respectively) than the sum of the parts. In our example, this amounts to assessing whether any effect of the injectable drug on the outcome depends on, or is modified by, simultaneous administration of the dietary supplement. In the most extreme case, interactions will indicate whether any drug effect is disproportionally increased or, alternatively, cancelled out or even reversed by the dietary supplement. Beyond enriching the process of learning from data, interactions can render main-effects-based (i.e., average) inference to be not only insufficient, but actually misleading (Kutner et al., 2004; Stroup, 2013). This is of concern, as a recent study (Kilkenny et al., 2009) showed that approximately 40% of animal studies posing questions amenable to a factorial design failed to use it. When so disregarded, interactions can yield quantitatively or, even worse, qualitatively misleading results that mischaracterize the effect of one treatment by failing to provide context about the other treatment. This can, in turn, impair RR by oversimplifying conclusions in meaningless ways that fail to reflect the mechanisms at play. More intuitively, the old adage should resonate: averages do not always tell the story.

Second, during data analysis, it is worth recognizing the proper nature of the response variable of interest, particularly if it is beyond the realm of continuity and

symmetry of the normal distribution (Stroup, 2013, 2015). Specifically, discrete outcomes are commonly used in the animal sciences in the form of categorical classifications, ordered scores, or counts. Examples include assessments of fitness and fertility (e.g., pregnancy and embryonic losses, health events, BCS) as well as behavior and well-being (e.g., temperament scores, frequency of behaviors). Conventional analytic recommendations for such discrete data often call for “normalizing practices,” such as variance-stabilizing transformations or normal approximations based on the Central Limit Theorem and on the robustness of standard tests. In decades past, these practices were arguably inevitable, even desirable, practical solutions. Alternatively, modern statistical methods can explicitly recognize the non-normal discrete nature of a response through user-friendly software tools. Examples include generalized linear mixed models, which can explicitly accommodate categorical or count responses for data analysis (Gbur et al., 2012; Stroup, 2013). Doing so circumvents tenuous approximations, power losses, and interpretation blunders associated with the cost of normalization (Larrabee et al., 2014), thereby enhancing inferential efficiency in a way consistent with research reproducibility.

PROPOSED SOLUTIONS

Statistics is a vibrant scientific discipline in continuous development to keep pace with the new quantitative challenges posed by technological developments and the ensuing creation of novel data types and bigger data sets. In this ever-evolving landscape, it may be hardly relevant to elicit a list of statistical rules that attempt to cover the breadth of all possible situations in scientific research. In fact, one should beware of recommendations for tight standardizations that straight-jacket data analyses into stereotypic methods, particularly if outdated. Indeed, blanket suggestions to decrease significance thresholds or increase sample size by prespecified amounts (Simmons et al., 2011; Benjamin et al., 2017) can easily miss the complexity of the issue and are thus unlikely to be helpful. Statistical practice needs to be tailored to the specifics of a research question of interest. From experimental design and the data generation process through analysis to interpretation of results, there is no rule-of-thumb or software automation that can replace careful thought. So, what to do?

Rather than a static laundry list of statistical recommendations, it is our opinion that a principled approach centered on statistical education that is modern and domain-tailored is vital to both established researchers and upcoming scientists alike. This opinion is consis-

tent with responses from over 1,500 scientists surveyed about approaches to improve RR in science, of which nearly 90% ranked “more robust experimental design” and “better statistics” as the top 2 approaches (Baker, 2016).

Established researchers effectively serve as the gatekeepers of science, mostly through their roles as journal editors, peer reviewers, and grant panelists, thereby ultimately deciding the fate of published and funded research. However, in this high-paced information age, it can be difficult for established researchers to keep up with statistical developments (either new methods or new tools) in addition to advances in their own scientific disciplines. In fact, in our experience, it is often the case that established researchers often continue conducting statistics within the tradition that they were trained on, presumably using the same software, despite the multi-year, even multidecade span of many research careers, with inevitable outdating. For this reason, continuing quantitative education can be a valuable resource for established researchers, particularly if facilitated by professional societies. For example, the American Dairy Science Association has offered professional workshops on mixed models on an approximately biennial basis for the past 19 yr; most often, this workshop runs at capacity. Another examples is the American Society of Agronomy and its sister societies, who recently sponsored the targeted development of a modern statistical textbook for their membership (Gbur et al., 2012). Modern statistical training is arguably also imperative for upcoming scientists, such as graduate students. Many land-grant academic institutions offer graduate-level courses in experimental design and mixed models. In as much as these courses keep relevant and updated to modern developments, we believe they should be given serious consideration as standard requirements for graduate programs in the animal sciences. In addition, recent funding recommendation have focused on the development of targeted coursework (Broman et al., 2017) that promotes a working knowledge of statistical principles and methods aligned with research needs. Peer-reviewed journals also have a role to play and can steer change by asking authors to adhere to the newest appropriate best methods and by enforcing such standards in the review process (Erb, 2010). Specifically, explicit standards are required to ensure that reporting of data collection and analysis is comprehensive, accurate, and transparent, thereby avoiding reporting omissions and enabling reviewers and the readership in general to better evaluate the credibility of research findings. Initiatives of this nature are already in place in related disciplines and can serve as templates for developing of reporting standards in the animal sciences. Examples include the REFLECT statement (Reporting Guide-

lines for Randomized Controlled Trials for Livestock and Food Safety; <http://www.reflect-statement.org/>), the STROBE statement (Strengthening the Reporting of Observational Studies in Epidemiology; <http://www.strobe-statement.org>), the ARRIVE guidelines (Animal Research: Reporting of In Vivo Experiments; <https://www.nc3rs.org.uk/arrive-guidelines>), and the MIAME guidelines (Minimum Information About a Microarray Experiment; <https://www.ncbi.nlm.nih.gov/geo/info/MIAME.html>).

This multipronged approach to enhanced quantitative literacy of domain scientists is likely to uplift the quality of scientific discourse and create a sense of awareness of when animal scientists should seek collaboration with a statistician to address more complex questions. To promote such interdisciplinary collaborations, it is critical that scholarly credit be properly recognized to all involved. Specifically, the role of applied statisticians as collaborators and partners in the research effort needs to be clearly realized and rewarded, as recently articulated by the American Statistical Association Board (<http://www.amstat.org/asa/files/pdfs/POL-Statistics-as-a-Scientific-Discipline.pdf>).

Similarly, an honest conversation is needed on how to provide scholarly credit to attempts to replicate standing research results. The current academic culture and funding model provide little incentive to even attempt any such replications, despite the fact that, ultimately, the only way to infer that a study is reproducible is to attempt to reproduce it. The scientific community should consider withholding judgment on any study until its findings have been verified independently. The findings of a single study should be considered preliminary regardless of how well the study may have been conducted or analyzed, even if decision making is on the line. Arguably, science advances on the weight of accumulated evidence.

Admittedly, the problem of irreproducible research findings goes well beyond statistics. Attempts to mitigate the RR crisis are extensive and multipronged, as summarized elsewhere (Gelman and Loken, 2014; Nuzzo, 2015; Broman et al., 2017). Efforts include pre-registration of studies, open science (i.e., open access, open data, open code, open source), and blind data analysis, along with an educated awareness of common caveats and problems, among others.

Along the road toward reproducible research, we find specially compelling the quest of Nobel laureate physicist Richard Feynman (1918–1988) for “utter scientific integrity” (Feynman, 1974); that is, extra conscientious care to “not fool ourselves—and [oneself is] the easiest person to fool.” Beyond honesty, Feynman argued for a deep sense of reflected skepticism and self-scrutiny that may well be of service in the on-going efforts to ensure

the reproducibility of research findings and, ultimately, the advance of science and its public credibility.

CONCLUSIONS

This review article focuses on key foundational principles and practices of statistics that are of critical relevance to ensure that research findings in the animal sciences are reproducible. The current RR crisis poses a compelling problem that strikes at the very core of science and its role to society. In this context, we argue that statistical principles and practices are now as, if not more, relevant as ever to scientific progress, especially in the context of big data and modern interdisciplinary research. Statistics as a discipline is uniquely poised to play a role in the inherently messy process of learning from noisy data in a way that is consistent with RR. To this end, the benefits of a study that is well designed, executed, analyzed, and interpreted are self-evident. By contrast, the harm of proceeding otherwise becomes more insidious, ultimately resulting in wasted resources, complicating future research, and jeopardizing the credibility of science.

In the interest of RR, we emphatically highlight the need for a keen and in-depth understanding of the data generation process as foundational to sound statistical practices downstream. Insight into how data were collected and how randomization was implemented is critical to all stages of the research endeavor. It is imperative that thoughtful experimental design ensures experimental replication independent of level of observation, thereby enabling efficiently powered studies. Emphasis should be placed on carefully deployed data analysis that is clear about its purpose and rigorous about how models are specified and how error is calibrated for inference. Embedded in the data generation process is the concept of data architecture that further delineates scope of inference to elicit the proper reach of research results while preventing overgeneralizations. Additional considerations include appropriate inferential interpretation of *P*-values, both significant and not, as well as recommendations for routine implementation of corrective remedial measures to prevent error inflation from multiple testing. Also in need of careful consideration are comprehensive reporting practices, including detailed description of data (e.g., collection, editing, management, storage) and step-by-step reporting of the methods implemented for analysis. Finally, we highlight the need for conscientious interpretation and responsible wording of conclusions, particularly when communicating with the public. Critical distinctions in wording include causal claims as opposed to association-based conclusions, as well as statistical significance and lack thereof.

Whereas statistical practice should be tailored to the specific complexities of a research question, rigorous implementation of statistical principles and methods should be standard practice and not considered a matter of opinion. The infamous quote, popularized by Mark Twain, “there are three kinds of lies: lies, damn lies, and statistics,” is sometimes colloquially misused to undermine the power and rigor of statistical practice. Although “it is easy to lie with statistics,” it is certainly “easier to lie without them,” replied Andrew Gelman (Gelman and Loken, 2014), probably one of the most influential contemporary statisticians in paraphrasing a quote by his late colleague Frederick Mosteller (1916–2006). To this end, we strongly believe the discipline of statistics has much to offer to the scientific community in confronting the research reproducibility crisis. We argue for a responsible practice of statistics in the animal sciences, with a vital emphasis on continuing statistical education of established researchers and modern training of graduate students, as well as close collaboration of animal scientists and statisticians. Taken together, we believe these efforts can go a long way in maximizing reproducibility of research findings and, ultimately, the very credibility of the scientific endeavor associated with animal health and production.

ACKNOWLEDGMENTS

The authors appreciate insightful discussions with Guilherme Rosa, professor of animal sciences at University of Wisconsin-Madison. We also recognize the thoughtful and focused input of two anonymous reviewers.

REFERENCES

- Aitkin, M., and N. Longford. 1986. Statistical modeling issues in school effectiveness studies. *J. R. Stat. Soc. A* 149:1–43.
- Baker, M. 2016. Is there a reproducibility crisis? *Nature* 533:452–454.
- Begley, C. G., and L. M. Ellis. 2012. Raise standards for preclinical cancer research. *Nature* 483:531–533.
- Begley, C. G., and J. P. A. Ioannidis. 2015. Reproducibility in science: improving the standard for basic and preclinical research. *Circ. Res.* 116:116–126.
- Bello, N. M., M. Kramer, R. J. Tempelman, W. W. Stroup, N. R. St-Pierre, B. A. Craig, L. J. Young, and E. E. Gbur. 2016. Short communication: On recognizing the proper experimental unit in animal studies in the dairy sciences. *J. Dairy Sci.* 99:8871–8879.
- Benjamin, D., J. O. Berger, M. Johannesson, B. A. Nosek, E. J. Wagenmakers, R. Berk, K. A. Bollen, B. Brembs, L. Brown, C. Camerer, D. Cesarini, C. D. Chambers, M. Clyde, T. D. Cook, P. De Boeck, Z. Dienes, A. Dreber, K. Easwaran, C. Efferson, E. Fehr, F. Fidler, A. P. Field, M. Forster, E. I. George, R. Gonzalez, S. Goodman, E. Green, D. P. Green, A. Greenwald, J. D. Hadfield, and L. V. Hedges., L. Held, T. H., H. Ho, J. Hoijsink, D. J. Holland Jones, K. Hruschka, G. Imai, J. P. A. Imbens, M. Ioannidis, M. Jeon, D. Kirchner, J. Laibson, R. List, A. Little, and Lupia., EM-achery, S. E., M. Maxwell, D. McCarthy, S. L. Moore, M. Morgan, S. Munafó, B. Nakagawa, T. H. Nyhan, L. Parker, M. Pericchi, J. Perugini, J. Rouder, V. Rousseau, F. D. Savalei, T. Schönbrodt, B.

- Sellke, D. Sinclair, and Tingley., TVan Zandt, S., D. J. Vazire, C. Watts, R. L. Winship, Y. Wolpert, C. Xie, J. Young, Zinman, and V. E. Johnson. 2017. Redefine statistical significance. *Nat. Hum. Behav.* 115:1–2. <https://doi.org/10.3410/f.731156677.793540852>.
- Benjamini, Y., and Y. Hochberg. 1995. Controlling the false discovery rate—A practical and powerful approach to multiple testing. *J. R. Stat. Soc. B* 57:289–300.
- Berckmans, D. 2017. General introduction to precision livestock farming. *Anim. Front.* 7:6–11.
- Berger, J. O., and T. Sellke. 1987. Testing a point null hypothesis—The irreconcilability of p-values and evidence. *J. Am. Stat. Assoc.* 82:112–122.
- Broman, K., M. Centikaya-Rundel, A. Nussbaum, C. Paciorek, R. Peng, D. Turek, and H. Wickham. 2017. Recommendations to funding agencies for supporting reproducible research. American Statistical Association. Accessed Jan. 23, 2018. <http://www.amstat.org/asa/files/pdfs/pol-reproducible/researchrecommendations.pdf>.
- Casella, G. 2008. *Statistical Design*. 1st ed. Springer, Gainesville, FL.
- Cowles, M., and C. Davis. 1982. On the origins of the .05 level of statistical significance. *Am. Psychol.* 37:553–558.
- Cull, C. A., Z. D. Paddock, T. G. Nagaraja, N. M. Bello, A. H. Babcock, and D. G. Renter. 2012. Efficacy of a vaccine and a direct-fed microbial against fecal shedding of *Escherichia coli* O157:H7 in a randomized pen-level field trial of commercial feedlot cattle. *Vaccine* 30:6210–6215.
- Dohoo, I., W. Martin, and H. Stryhn. 2009. *Veterinary Epidemiologic Research*. 2nd ed. VER Inc., Charlottetown, Prince Edward Island, Canada.
- Dohoo, I. R., and D. Waltner-Toews. 1985. Interpreting clinical research. 1. General considerations. *Comp. Cont. Educ. Pract. Vet.* 7:S473–S478.
- Erb, H. N. 2010. Changing expectations: Do journals drive methodological changes? Should they? *Prev. Vet. Med.* 97:165–174.
- Feynman, R. P. 1974. Cargo cult science. *Eng. Sci.* 37:10–13.
- Fisher, R. A. 1925. *Statistical Methods for Research Workers*. 1st ed. Oliver and Boyd, London, UK.
- Gates, C. E. 1995. What really is experimental error in block-designs. *Am. Stat.* 49:362–363.
- Gbur, E. E., W. W. Stroup, K. S. McCarter, S. Durham, L. J. Young, M. Christman, M. West, and M. Kramer. 2012. Analysis of Generalized Linear Mixed Models in the Agricultural and Natural Resources Sciences. American Society of Agronomy, Madison, WI.
- Gelman, A., J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. Rubin. 2013. *Bayesian Data Analysis*. 3rd ed. Chapman & Hall/CRC Press, Boca Raton, FL.
- Gelman, A., and E. Loken. 2014. The statistical crisis in science. *Am. Sci.* 102:460–465.
- Gonçalves, M. A. D., K. M. Gourley, S. S. Dritz, M. D. Tokach, N. M. Bello, J. M. DeRouche, J. C. Woodworth, and R. D. Goodband. 2016. Effects of amino acids and energy intake during late gestation of high-performing gilts and sows on litter and reproductive performance under commercial conditions. *J. Anim. Sci.* 94:1993–2003.
- Goodman, S. 2008. A dirty dozen: Twelve p-value misconceptions. *Semin. Hematol.* 45:135–140.
- Granger, C. W. J. 1998. Extracting information from mega-panels and high-frequency data. *Stat. Neerl.* 52:258–272.
- Hirsch, J. E. 2005. An index to quantify an individual's scientific research output. *Proc. Natl. Acad. Sci. USA* 102:16569–16572.
- Hoening, J. M., and D. M. Heisey. 2001. The abuse of power: The pervasive fallacy of power calculations for data analysis. *Am. Stat.* 55:19–24.
- Ioannidis, J. P. A. 2005. Why most published research findings are false. *PLoS Med.* 2:e124.
- Ioannidis, J. P. A., D. B. Allison, C. A. Ball, I. Coulibaly, X. Q. Cui, A. C. Culhane, M. Falchi, C. Furlanello, L. Game, G. Jurman, J. Mangion, T. Mehta, M. Nitzberg, G. P. Page, E. Petretto, and V. van Noort. 2009. Repeatability of published microarray gene expression analyses. *Nat. Genet.* 41:149–155.
- Kilkenny, C., N. Parsons, E. Kadyszewski, M. F. W. Festing, I. C. Cuthill, D. Fry, J. Hutton, and D. G. Altman. 2009. Survey of the quality of experimental design, statistical analysis and reporting of research using animals. *PLoS One* 4:e7824.
- Kramer, M. H., E. T. Paparozzi, and W. W. Stroup. 2016. Statistics in a horticultural journal: Problems and solutions. *J. Am. Soc. Hortic. Sci.* 141:400–406.
- Kuehl, R. O. 2000. *Design of Experiments: Statistical Principles of Research Design and Analysis*. 2nd ed. Brooks/Cole Cengage Learning, Belmont, CA.
- Kutner, M., C. Nachtsheim, J. Neter, and W. Li. 2004. *Applied Linear Statistical Models*. 5th ed. McGraw-Hill/Irwin, Columbus, OH.
- Larrabee, B., H. M. Scott, and N. M. Bello. 2014. Ordinary least squares regression of ordered categorical data: Inferential implications for practice. *J. Agric. Biol. Environ. Stat.* 19:373–386.
- Lazer, D., R. Kennedy, G. King, and A. Vespignani. 2014. Big data. The parable of Google Flu: Traps in big data analysis. *Science* 343:1203–1205.
- Lenth, R. V. 2007. Statistical power calculations. *J. Anim. Sci.* 85(Suppl):E24–E29.
- Littell, R. C., G. A. Milliken, W. W. Stroup, R. D. Wolfinger, and O. Schabenberger. 2006. *SAS for Mixed Models*. SAS Institute Inc., Cary, NC.
- Littell, R. C., W. W. Stroup, and R. J. Freund. 2002. *SAS for Linear Models*. 4th ed. SAS Institute Inc., Cary, NC.
- MacKinnon, M. J., and N. Glick. 1999. Data mining and knowledge discovery in databases - An overview. *Aust. N. Z. J. Stat.* 41:255–275.
- McCance, I. 1995. Assessment of statistical procedures used in papers in the Australian Veterinary Journal. *Aust. Vet. J.* 72:322–328.
- McLean, R. A., W. L. Sanders, and W. W. Stroup. 1991. A unified approach to mixed linear-models. *Am. Stat.* 45:54–64.
- Mead, R., S. G. Gilmour, and A. Mead. 2012. *Statistical Principles for the Design of Experiments: Applications to Real Experiments*. 1st ed. Cambridge University Press, Cambridge, UK.
- Milliken, G. A., and D. J. Johnson. 2009. *Analysis of Messy Data. Volume I: Designed Experiments*. 2nd ed. Chapman & Hall/CRC Press, Boca Raton, FL.
- Morgan, S. L., and C. Winship. 2017. *Counterfactuals and Causal Inference: Methods and Principles for Social Research*. 2nd ed. Cambridge University Press, New York, NY.
- Naik, G. 2011. Scientists' elusive goal: Reproducing study results. *The Wall Street Journal*. Accessed Jan. 23, 2018. <https://www.wsj.com/articles/SB10001424052970203764804577059841672541590>.
- Nuzzo, R. 2015. Fooling ourselves. *Nature* 526:182–185.
- Open Science Collaboration. 2015. Estimating the reproducibility of psychological science. *Science* 349:aac4716.
- Pearl, J. 2009. *Causality: Models, Reasoning and Inference*. 2nd ed. Cambridge University Press, New York, NY.
- Reinhart, A. 2015. *Statistics Done Wrong: The Woefully Complete Guide*. 1st ed. No Starch Press, San Francisco, CA.
- Richter, S. H., J. P. Garner, and H. Wurbel. 2009. Environmental standardization: Cure or cause of poor reproducibility in animal experiments? *Nat. Methods* 6:257–261.
- Ridley, J., N. Kolm, R. P. Freckelton, and M. J. G. Gage. 2007. An unexpected influence of widely used significance thresholds on the distribution of reported P-values. *J. Evol. Biol.* 20:1082–1089.
- Robinson, G. K. 1991. That BLUP is a good thing: The estimation of random effects. *Stat. Sci.* 6:15–51.
- Rosa, G. J. M., and B. D. Valente. 2013. Breeding and Genetics Symposium: Inferring causal effects from observational data in livestock. *J. Anim. Sci.* 91:553–564.
- Rothstein, H. R., A. J. Sutton, and M. Borenstein. 2005. *Publication Bias in Meta-Analysis—Prevention, Assessment and Adjustments*. John Wiley & Sons, Chichester, UK.
- Sargeant, J. M., R. Elgie, J. Valcour, J. Saint-Onge, A. Thompson, P. Marcynuk, and K. Snedeker. 2009. Methodological quality and completeness of reporting in clinical trials conducted in livestock species. *Prev. Vet. Med.* 91:107–115.
- Sargeant, J. M., A. M. O'Connor, D. G. Renter, D. F. Kelton, K. Snedeker, L. V. Wisener, E. K. Leonard, A. D. Guthrie, and M. Faires. 2011. Reporting of methodological features in observational studies of pre-harvest food safety. *Prev. Vet. Med.* 98:88–98.

- Shaywitz, D. A. 2009. When science is a siren song. *The Washington Post*. Accessed Jan. 23, 2018. <http://www.washingtonpost.com/wp-dyn/content/article/2009/03/13/AR2009031302910.html>.
- Simmons, J. P., L. D. Nelson, and U. Simonsohn. 2011. False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychol. Sci.* 22:1359–1366.
- Sinoquet, C. 2014. Probabilistic graphical models for next-generation genomics and genetics. In *Probabilistic Graphical Models for Genetics, Genomics and Post-genomics*. 1st ed. C. Sinoquet and R. Mourad, ed. Oxford University Press, Oxford, UK.
- St-Pierre, N. R. 2001. Invited review: Integrating quantitative findings from multiple studies using mixed model methodology. *J. Dairy Sci.* 84:741–755.
- St-Pierre, N. R. 2007. Design and analysis of pen studies in the animal sciences. *J. Dairy Sci.* 90:E87–E99.
- Stevenson, J. S., J. R. Pursley, H. A. Garverick, P. M. Fricke, D. J. Kesler, J. S. Ottobre, and M. C. Wiltbank. 2006. Treatment of cycling and noncycling lactating dairy cows with progesterone during Ovsynch. *J. Dairy Sci.* 89:2567–2578.
- Stevenson, J. S., D. E. Tenhouse, R. L. Krisher, G. C. Lamb, J. E. Larson, C. R. Dahlen, J. R. Pursley, N. M. Bello, P. M. Fricke, M. C. Wiltbank, D. J. Brusveen, M. Burkhart, R. S. Youngquist, and H. A. Garverick. 2008. Detection of anovulation by Heat-mount detectors and transrectal ultrasonography before treatment with progesterone in a timed insemination protocol. *J. Dairy Sci.* 91:2901–2915.
- Stroup, W. W. 2013. *Generalized Linear Mixed Models: Modern Concepts, Methods and Applications*. 1st ed. Chapman & Hall/CRC Press, Boca Raton, FL.
- Stroup, W. W. 2015. Rethinking the analysis of non-normal data in plant and soil science. *Agron. J.* 107:811–827.
- Tempelman, R. J. 2004. Experimental design and statistical methods for classical and bioequivalence hypothesis testing with an application to dairy nutrition studies. *J. Anim. Sci.* 82(E-Suppl.):E162–E172.
- Tempelman, R. J. 2005. Assessing statistical precision, power, and robustness of alternative experimental designs for two color microarray platforms based on mixed effects models. *Vet. Immunol. Immunopathol.* 105:175–186.
- Tempelman, R. J. 2009. Invited review: Assessing experimental designs for research conducted on commercial dairies. *J. Dairy Sci.* 92:1–15.
- Tempelman, R. J. 2010. Addressing scope of inference for global genetic evaluation of livestock. *Rev. Bras. Zootec.* 39:261–267.
- Tempelman, R. J. 2016. The frontier spirit and reproducible research in animal breeding. *J. Anim. Breed. Genet.* 133:441–442.
- Tempelman, R. J., D. M. Spurlock, M. Coffey, R. F. Veerkamp, L. E. Armentano, K. A. Weigel, Y. de Haas, C. R. Staples, E. E. Connor, Y. Lu, and M. J. VandeHaar. 2015. Heterogeneity in genetic and nongenetic variation and energy sink relationships for residual feed intake across research stations and countries. *J. Dairy Sci.* 98:2013–2026.
- Theurer, M. E., D. G. Renter, and B. J. White. 2015. Using feedlot operational data to make valid conclusions for improving health management. *Vet. Clin. North Am. Food Anim. Pract.* 31:495–508.
- Van der Laan, M. J., and S. Rose. 2011. *Targeted Learning: Causal Inference for Observational and Experimental Data*. 1st. ed. Springer, New York, NY.
- Wasserstein, R. L., and N. A. Lazar. 2016. The ASA's statement on p-values: Context, process, and purpose. *Am. Stat.* 70:129–133.
- Yates, F. 1940. The recovery of inter-block information in balanced incomplete block designs. *Ann. Hum. Genet.* 10:317–325.
- Young, S. S., and A. Karr. 2011. Deming, data and observational studies: A process out of control and needing fixing. Pages 116–120 in *Significance*. Vol. 8. American Statistical Association, Alexandria, VA, and Royal Statistical Society, London, UK.