A better approach for dealing with reproducibility and replicability in science

James D. Nichols[a,1], Madan K. Oli[a], William. L. Kendall[b], and G. Scott Boomer[c]

Science impacts our daily lives and guides national and international policies (1). Thus, results of scientific studies are of paramount importance; yet, there are concerns that many studies are not reproducible or replicable (2). To address these concerns, the National Research Council conducted a Consensus Study [NASEM 2019 (3)] that provides definitions of key concepts, discussions of problems, and recommendations for dealing with these problems. These recommendations are useful and well considered, but they do not go far enough in our opinion. The NASEM recommendations treat reproducibility and replicability as single-study issues, despite clear acknowledgement of the limitations of isolated studies and the need for research synthesis (3). We advocate a strategic approach to research, focusing on the accumulation of

To deal more effectively with reproducibility, replicability, and related problems, scientists should pursue a strategic approach to research, focusing on the accumulation of evidence via designed sequences of studies. Image credit: Dave Cutler (artist).

[a]Department of Wildlife Ecology and Conservation, University of Florida, Gainesville, FL 32611; [b]US Geological Survey, Colorado Cooperative Fish and Wildlife Research Unit, Fort Collins, CO 80523; and [c]US Fish and Wildlife Service, Division of Migratory Bird Management, Laurel, MD 20708.

The authors declare no competing interest.

[1]To whom correspondence may be addressed. Email: jamesdnichols2@gmail.com.

Published February 10, 2021.

evidence via designed sequences of studies, as a means of dealing more effectively with reproducibility, replicability, and related problems. These sequences are designed to provide iterative tests based on comparison of data from empirical studies with predictions from competing hypotheses. Evidence is then formally accumulated based on the relative predictive abilities of the different hypotheses as the sequential studies proceed.

In many disciplines, single studies are seldom adequate to substantially increase knowledge by themselves. Examples of Platt's (4) "crucial experiments," which are capable of definitively discriminating among competing hypotheses, can be found but are rare. Thus, we view individual study results as building blocks and the accumulation of evidence as requiring multiple studies of the same phenomena (5–7). This view can be incorporated strategically into research planning by developing sequences of studies to investigate focal hypotheses.

Here we emphasize the comparison of study results with model-based predictions as more useful to science than the comparison of results of different pairs of studies. The latter approach produces conclusions about whether two studies do or do not yield similar results, whereas the former leads to accumulated assessments of confidence in specific hypotheses and their predictions. When we entertain multiple plausible hypotheses (8), the task is to track the relative confidence in them as assessed by their relative predictive abilities as study results accumulate. We propose programs of inquiry designed to progressively and adaptively test model-based predictions for the purpose of accumulating evidence.

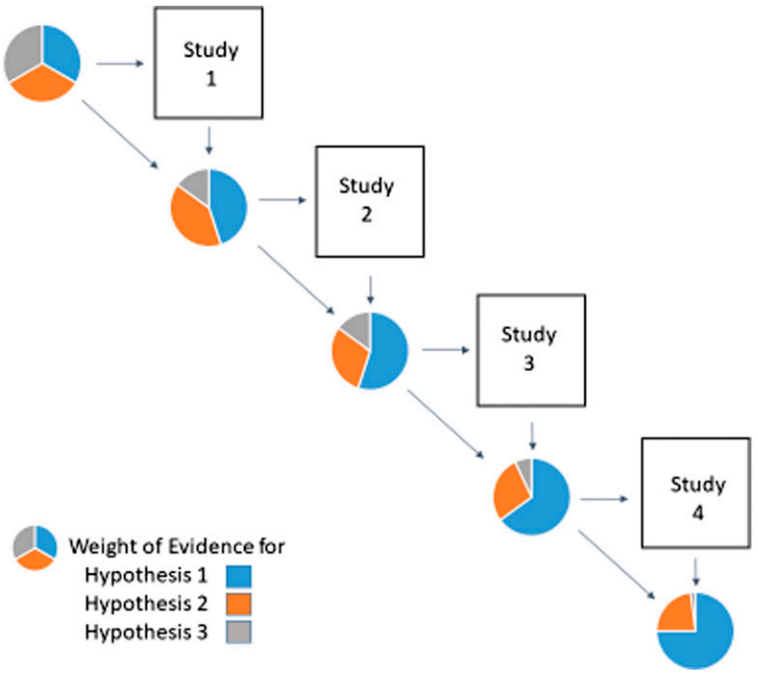

Fig. 1. In this schematic diagram of the "Evolving Information State" approach, the pie charts at each time step represent the evolution of the information state. The relative sizes of the different pie slices reflect the evolving model weights associated with the three hypotheses. Hypothesis 1 consistently provides predictions that are best supported by the data of the successive studies and thus attains more and more weight through time. Optimal design of each study depends on the current information state.

A Strategic Approach

Problems of reproducibility and replicability can be dealt with using strategic approaches that promote and accelerate the accumulation of evidence. One methodological approach, referred to as "evolving information state" [EIS (9)], is based on multiple hypotheses and their associated models. At any point in time, *t*, each model carries an associated "model weight," reflecting the relative degree of confidence in that model, based on its predictive ability assessed before *t* (Fig. 1). The information state is the vector of model weights, expressing the current confidence in the different models that has accumulated through that point in time. Weights, and thus relative confidence, are updated with the results of each new study or data set, using Bayes' theorem to combine the previous weight (confidence accrued before *t*) with a measure reflecting the degree to which each model predicted the new data (e.g., 9, 10). Furthermore, optimal design of each new study in the sequence is conditional on the current model weights. If a useful hypothesis is included in the set, then the weight associated with its model should approach 1 as evidence accumulates, whereas the weights associated with the other models should approach 0.

An example of the EIS approach is a 25-year management program for mid-continent North American mallard ducks (*Anas platyrhynchos*) led by the US Fish and Wildlife Service [USFWS (9, 11)]. The program was designed to simultaneously provide the information to make wise decisions about annual hunting regulations and discriminate among competing models (i.e., learn) about the specific population-dynamic effects of different regulations. For every year of this ongoing program, each of the four models is used to develop a prediction about breeding population size the next spring (year *t*+1), given the estimate of current (year *t*) breeding population size (based on a large-scale monitoring program) and the hunting regulations selected for the fall–winter hunting season of year *t*. In the following breeding season (year *t*+1), each of the model-based predictions from the previous year is compared with the new estimate of population size, and Bayes' theorem is used to update the previous model weights with the new information about predictive ability. Fig. 2 shows the annual population estimates and model-based predictions, as well as the evolution of the information state.

USFWS initiated the program with equal model weights (maximum uncertainty), and two of the models now have weights approaching 0. Of the remaining two models, one has more than twice the weight of the other. We note that the evolution of model weights shown in Fig. 2 occurred in a management program within which hunting regulations were established to achieve management objectives rather than to learn (9, 11). Application of this approach to programs focused on learning would entail making periodic (e.g., annual) decisions about system manipulations, focal parameters to estimate, covariates, etc., based on the objective of accumulating evidence to permit model discrimination.

Reproducibility and Replicability

NASEM 2019 defines reproducibility "to mean computational reproducibility—obtaining consistent computational results using the same input data, computational steps, methods, code, and conditions of analysis" (3). Recommendations include increased access to data and detailed descriptions of computational steps but do little to address such important sources of variation as model specification and selection. Reproducibility is defined for single studies, and if we follow a multiple-study approach to the accumulation of evidence, then a focus on reproducibility would minimally entail checking for consistency of some sample of component study results. This could be done, and discrepancies would likely be found. However, it is not clear how to use such information to inform future lines of inquiry or accumulate evidence.

We view lack of reproducibility as simply one of multiple sources of variation that influence replicability, the key concept underlying the accumulation of evidence. Other sources, such as specification of the focal hypotheses/models and appropriate deduction of predictions, are likely to be at least as important in reducing problems of replicability. Approaches to the accumulation of evidence should accommodate these different sources of variation and permit useful inference, even in their presence.

NASEM 2019 defines replicability as "obtaining consistent results across studies aimed at answering the same scientific question, each of which has obtained its own data," and discusses comparing results and "assessing replication between two results." In science, we believe that "results" should refer to the degree to which predictions of one or more models, and their associated hypotheses, are consistent with data collected either to assess the reasonableness of the model(s) or to discriminate between two or more competing models. Repeating a study provides additional evidence for or against the predictive abilities of the focal hypotheses. When considering results of two similar studies, we believe that the focus should not be on their comparison but rather on their respective consistency with hypothesis-based predictions, and on combining these assessments of consistency to obtain an overall weight of evidence. Indeed, replicated consistency (or not) with hypothesis-based predictions provides the ultimate evidence of replicability.

Preregistration of studies has been recommended as one approach to improving replicability (3). Under preregistration, a priori hypotheses and methods of assessing them are specified before initiation of the study to clarify whether a result is based on an "exploratory" versus a "confirmatory" analysis. We propose the additional proactive step of moving beyond individual studies to programs of inquiry in which researchers test specified hypotheses repeatedly in a progressive sequence of studies designed specifically to accumulate evidence.

The section of NASEM 2019 that deals with research synthesis represents an effort to focus on evidence from multiple studies. "In current use, the term *research synthesis* describes the ensemble of research activities involved in identifying, retrieving, evaluating, synthesizing, interpreting, and contextualizing the available evidence from studies on a particular topic and comprises both systematic reviews and meta-analyses" (3). This shift of focus to evidence from multiple studies is useful. However, the retrospective approach based on existing studies is limiting. Retrospective approaches to research synthesis are based primarily on existing publications and must deal with issues such as publication bias, in which results of published studies are not likely to be representative of all studies directed at selected hypotheses. Commonly cited approaches developed to detect and deal with publication bias have drawbacks. A shift to a prospective approach, in which researchers design and conduct sequences of studies to accumulate evidence on a specific topic, eliminates many of the problems with meta-analyses (9).

NASEM 2019 contains some statements acknowledging the importance of accumulating evidence from multiple studies, e.g.,

> "Some would argue that focusing on replication of a single study as a way to improve the efficiency of science is ill-placed. Rather, reviews of cumulative evidence on a subject, to gauge both the overall effect size and generalizability, may be more useful" (3).

Learning in geoscience and weather prediction is described as based on probabilistic forecasting and evidence accumulation (3). Thus, glimpses of our view of how science can be conducted are found scattered throughout the report, but they are not dominant, do not provide the basis for report recommendations, and provide no details about how programs to accumulate evidence might be implemented.

Approaches focused on accumulation of evidence are useful for addressing a number of issues associated with replicability. One source of nonreplicability is "prior probability (pre-experimental plausibility) of the scientific hypothesis" (3). EIS formalizes this concept and uses it explicitly in updating model weights and accumulating evidence. Rather than using "pre-experimental" beliefs, plausibility is based on previous predictive ability. The reporting of uncertainty associated with study results is often emphasized (3), but we should specify how such reports are to be used. The Bayes' theorem approach for updating model weights explicitly incorporates uncertainty associated with both the model-based predictions and the modeling process used to assess the degree of correspondence between data and predictions. Greater uncertainty may slow the rate of learning but does not otherwise alter the accumulation of evidence.

We propose the additional proactive step of moving beyond individual studies to programs of inquiry in which researchers test specified hypotheses repeatedly in a progressive sequence of studies designed specifically to accumulate evidence.

Preregistration of proposed investigations has been recommended to clarify interpretation of study results (3). EIS represents a proactive approach to sequences of studies, admitting the possibility of

design considerations to speed accumulation of evidence. When more than two models are being considered, the information state is an important determinant of optimal design (9), formalizing the use of "pre-experimental plausibility" (3) in study implementation.

EIS and related approaches to accumulate evidence largely deal with the issues of reproducibility and replicability. Important sources of uncertainty that decrease reproducibility (e.g., model selection, incorporated stochasticity) are dealt with directly during the updating process. Accumulation of evidence shifts the focus from replication of results of one study by those of another, to the ability of a hypothesis to consistently predict data arising from multiple studies. Study system characteristics that reduce replicability such as complexity, noise, and absence of stability (3) do not preclude accumulating evidence but simply slow the process. Avoidable sources of nonreplication (3) are either rendered irrelevant (e.g., publication bias) or else simply slow the rate of accumulation of evidence (e.g., poor design, errors).

Making Progress

In addition to the many useful recommendations that have already been made to researchers for improving reproducibility and replicability (3), we recommend that researchers deemphasize isolated studies and instead contribute to programs of sequential studies directed at accumulating evidence. We appreciate that this recommendation is easier said than done, but we believe that existing long-term research and monitoring programs are pre-adapted for such contributions, as illustrated in our mallard example. Researchers can form consortia focused on specific questions. Isolated researchers not integrated into such programs can participate by designing their own studies to contribute to model discrimination (12, 13). Differences in study designs selected by different investigators offer no conceptual problems, as long as the studies are directed at discrimination among the models (or a subset of them) in the specified set.

Existing recommendations for academic institutions and national laboratories emphasize provision of training (3). In addition, we believe it would be very useful if administrations of scientific institutions would shift their reward systems from emphasis on standalone studies to programs designed to accumulate knowledge.

NSF and other research funders have been asked to require more detailed descriptions of methods and data and to fund research exploring computational reproducibility, developing standard computational tools, and reviewing published work to assess reproducibility and replicability (3). We believe that NSF and other funders can also play an important role in shifting our scientific culture from that of one-and-done studies to carefully designed sequences of studies. Such a shift would provide motivation for researchers to become integrated into larger programs designed to accumulate evidence. Such top-down emphases on accumulating evidence could be very effective and would deal naturally with many of the problems of reproducibility and replicability.

The final recommendation made in NASEM 2019 is for policy makers and the general public: "Anyone making personal or policy decisions based on scientific evidence should be wary of making a serious decision based on the results, no matter how promising, of a single study." However, the appropriate advice to decision makers is not simply to be wary or to wait until uncertainty is resolved (unlikely in many cases) but rather to invoke decision algorithms and methods developed specifically to deal with uncertainty (12–14). Some approaches produce good decisions while simultaneously reducing uncertainty, and the concept of expected value of information identifies the value of such learning (10, 15).

In recent years there has been a growing recognition that scientific progress is slowed by problems in the conduct of science, as evidenced by the frequent inability to reproduce or replicate results of published studies. We believe that many problems with reproducibility and replicability are natural consequences of conducting isolated studies, as opposed to studies designed to contribute to an overall body of evidence. We recommend formal approaches for accumulating evidence by conducting planned sequences of studies. This overall recommendation leads to specific recommendations for individual researchers,

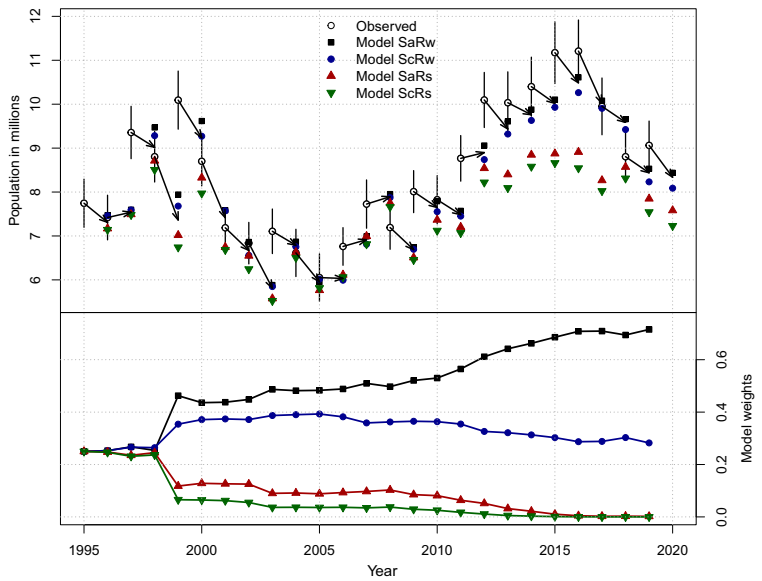

Fig. 2. The diagram shows time-specific comparisons of observations and model-based predictions, and the corresponding evolution of model weights, for mid-continent mallard ducks. Upper panel, population estimates of midcontinent mallards (in millions) compared with predictions of each member of the model set (SaRw = additive mortality and weakly density-dependent reproduction, ScRw = compensatory mortality and weakly density-dependent reproduction, SaRs = additive mortality and strongly density-dependent reproduction, ScRs = compensatory mortality and strongly density-dependent reproduction). Error bars represents 95% confidence intervals for observed population estimates. The arrow represents a weighted mean annual prediction based on the entire model set. Note that the mallard breeding population was not observed in the spring of 2020. Lower panel, annual changes in model weights for each member of the model set (the information state); weights were assumed to be equal in 1995.

associated institutions, funding agencies, and policy makers. We believe that a shift from a culture of isolated studies to a more integrated approach to science will lead to more rapid learning and, as a byproduct, improve and largely deal with the problems of reproducibility and replicability.

1 B. Ramalingam *et al.*, *Adaptive Leadership in the Coronavirus Response: Bridging Science, Policy, and Practice* (ODI Coronavirus Briefing Note, London, 2020).
2 M. Baker, 1,500 scientists lift the lid on reproducibility. *Nature* **533**, 452–454 (2016).
3 National Academies of Sciences, Engineering, and Medicine, *Reproducibility and Replicability in Science* (The National Academies Press, Washington, DC, 2019). 10.17226/25303.
4 J. R. Platt, Strong inference: Certain systematic methods of scientific thinking may produce much more rapid progress than others. *Science* **146**, 347–353 (1964).
5 H. Poincare, *Science and Hypothesis* (The Science Press, New York, 1905).
6 B. K. Forscher, Chaos in the brickyard. *Science* **142**, 339 (1963).
7 J. A. Nelder, Statistics, science and technology. *J. Roy. Stat. Soc. A* **149**, 109–121 (1986).
8 T. C. Chamberlin, The method of multiple working hypotheses. *J. Geol.* **5**, 837–848 (1897).
9 J. D. Nichols, W. L. Kendall, G. S. Boomer, Accumulating evidence in ecology: Once is not enough. *Ecol. Evol.* **9**, 13991–14004 (2019).
10 R. Hilborn, C. J. Walters, *Quantitative Fisheries Stock Assessment: Choice, Dynamics, and Uncertainty* (Chapman and Hall, New York, 1992).
11 U.S. Fish and Wildlife Service, *Adaptive Harvest Management: 2021 Hunting Season* (U.S. Department of the Interior, Washington, D.C., 2020).
12 P. Fackler, K. Pacifici, Addressing structural and observational uncertainty in resource management. *J. Environ. Manage.* **133**, 27–36 (2014).
13 B. K. Williams, Integrating external and internal learning in resource management. *J. Wildl. Manage.* **79**, 148–155 (2015).
14 J. D. Nichols, Confronting uncertainty: Contributions of the wildlife profession to the broader scientific community. *J. Wildl. Manage.* **83**, 519–533 (2019).
15 B. K. Williams, F. A. Johnson, Value of information in natural resource management: technical developments and application to pink-footed geese. *Ecol. Evol.* **5**, 466–474 (2015).

https://doi.org/10.1073/pnas.2100769118