The Journal of Wildlife Management; DOI: 10.1002/jwmg.21413

Commentary

Increased Scientific Rigor Will Improve Reliability of Research and Effectiveness of Management

SARAH N. SELLS,[1] *Montana Cooperative Wildlife Research Unit, 205 Natural Sciences Building, Wildlife Biology Program, University of Montana, Missoula, MT 59812, USA*

SARAH B. BASSING, *Montana Cooperative Wildlife Research Unit, 205 Natural Sciences Building, Wildlife Biology Program, University of Montana, Missoula, MT 59812, USA*

KRISTIN J. BARKER, *Montana Cooperative Wildlife Research Unit, 205 Natural Sciences Building, Wildlife Biology Program, University of Montana, Missoula, MT 59812, USA*

SHANNON C. FORSHEE, *Montana Cooperative Wildlife Research Unit, 205 Natural Sciences Building, Wildlife Biology Program, University of Montana, Missoula, MT 59812, USA*

ALLISON C. KEEVER, *Montana Cooperative Wildlife Research Unit, 205 Natural Sciences Building, Wildlife Biology Program, University of Montana, Missoula, MT 59812, USA*

JAMES W. GOERZ, *Montana Cooperative Wildlife Research Unit, 205 Natural Sciences Building, Wildlife Biology Program, University of Montana, Missoula, MT 59812, USA*

MICHAEL S. MITCHELL, *U.S. Geological Survey, Montana Cooperative Wildlife Research Unit, 205 Natural Sciences Building, Wildlife Biology Program, University of Montana, Missoula, MT 59812, USA*

ABSTRACT Rigorous science that produces reliable knowledge is critical to wildlife management because it increases accurate understanding of the natural world and informs management decisions effectively. Application of a rigorous scientific method based on hypothesis testing minimizes unreliable knowledge produced by research. To evaluate the prevalence of scientific rigor in wildlife research, we examined 24 issues of the *Journal of Wildlife Management* from August 2013 through July 2016. We found 43.9% of studies did not state or imply *a priori* hypotheses, which are necessary to produce reliable knowledge. We posit that this is due, at least in part, to a lack of common understanding of what rigorous science entails, how it produces more reliable knowledge than other forms of interpreting observations, and how research should be designed to maximize inferential strength and usefulness of application. Current primary literature does not provide succinct explanations of the logic behind a rigorous scientific method or readily applicable guidance for employing it, particularly in wildlife biology; we therefore synthesized an overview of the history, philosophy, and logic that define scientific rigor for biological studies. A rigorous scientific method includes 1) generating a research question from theory and prior observations, 2) developing hypotheses (i.e., plausible biological answers to the question), 3) formulating predictions (i.e., facts that must be true if the hypothesis is true), 4) designing and implementing research to collect data potentially consistent with predictions, 5) evaluating whether predictions are consistent with collected data, and 6) drawing inferences based on the evaluation. Explicitly testing *a priori* hypotheses reduces overall uncertainty by reducing the number of plausible biological explanations to only those that are logically well supported. Such research also draws inferences that are robust to idiosyncratic observations and unavoidable human biases. Offering only *post hoc* interpretations of statistical patterns (i.e., *a posteriori* hypotheses) adds to uncertainty because it increases the number of plausible biological explanations without determining which have the greatest support. Further, *post hoc* interpretations are strongly subject to human biases. Testing hypotheses maximizes the credibility of research findings, makes the strongest contributions to theory and management, and improves reproducibility of research. Management decisions based on rigorous research are most likely to result in effective conservation of wildlife resources.

KEY WORDS hypotheses, management, philosophy of science, reliable knowledge, research questions, rigorous science, scientific method, wildlife biology.

Researchers, managers, commissioners, legislators, and the public rely on the credibility of scientific research. Appropriate use of a rigorous scientific method strengthens inference and reduces potential for drawing misleading or spurious conclusions (Platt 1964, Romesburg 1981, Williams 1997). As a result, rigorous science helps researchers contribute to the body of scientific knowledge and build credibility for their work, for their research groups

Received: 16 September 2017; Accepted: 6 December 2017

[1]*E-mail: sarahnsells@gmail.com*

and organizations, and for science as a whole (Gill 1985). Furthermore, rigorous science produces what Romesburg (1981) termed reliable knowledge (i.e., the set of ideas that provide accurate understanding of nature), whereas non-rigorous science more readily contributes to unreliable knowledge (i.e., the set of inaccurate ideas falsely accepted as knowledge). Reliable knowledge informs management decision-making and guides appropriate application of results to other places and times. Management decisions may be ineffective or detrimental if based on spurious conclusions generated by non-rigorous research. Reliable knowledge therefore contributes to effective conservation of wildlife resources (Leopold 1933, Gill 1985).

The inherent efficiency of rigorous science (Platt 1964, Romesburg 1981, Williams 1997) improves biological understanding while reducing unnecessary use of limited research dollars, time, and personnel. Rigorous science iteratively builds support for or against hypotheses, reducing uncertainty over time. By building on previous studies and producing results generalizable to other places and times, rigorous science helps reduce the need for repetitive research. Appropriate use of a rigorous scientific method guides efficient study design and provides a solid foundation for addressing the inevitable, unforeseen challenges common to all research projects (e.g., technical problems, severe weather, decreased funding).

All best practices are most effectively implemented when the motivation and reasoning behind those practices are clearly understood. For rigorous science, this understanding includes an appreciation for the historical growth of scientific thinking and the resulting philosophy and logic inherent to drawing reliable inferences. The basic concepts of rigorous science are far from new, but we are not aware of a paper that concisely reviews and summarizes the major components of rigorous science in wildlife biology. Although authors have previously emphasized the importance of producing reliable knowledge (Platt 1964; Romesburg 1981, 2009; Nichols 1991; Williams 1997), primary literature does not provide clear, succinct guidance on designing and carrying out rigorous research in wildlife biology.

We had 2 objectives. First, we sought to determine the extent to which our field has answered the call Romesburg (1981) made >3 decades ago to increase the rigor of scientific studies. Second, we sought to develop concrete guidance for maximizing the rigor of wildlife science.

IS WILDLIFE RESEARCH PRODUCING RELIABLE KNOWLEDGE?

Romesburg (1981) asserted that wildlife scientists tended to retroductively generate research hypotheses (i.e., plausible biological explanations) from patterns and correlations but rarely used rigorous science to explicitly test these hypotheses and derive reliable inference (i.e., the hypothetico-deductive or H-D method). Unreliable knowledge is produced and perpetuated when untested hypotheses are misinterpreted as rigorously derived conclusions rather than speculative explanations, a practice Williams (1997) ascribed to most wildlife research. Adding untested hypotheses to a body of knowledge does not reduce uncertainty, whereas testing hypotheses can reduce uncertainty by eliminating possible explanations for a given phenomenon.

We investigated the reliability of knowledge produced by recent wildlife science by evaluating peer-reviewed research articles intended to produce biological inferences in the *Journal of Wildlife Management* (*JWM*) from August 2013 to July 2016. We chose *JWM* because research published therein is intended to "assist management and conservation" (www.wildlife.org, accessed 4 Sep 2017). Effective management depends on reliable knowledge, arguably setting a higher standard for scientific rigor in wildlife research than in disciplines where unreliable knowledge has less-tangible consequences. If Romesburg's (1981) and Williams's (1997) assertions that wildlife science often generates but fails to test biological hypotheses remain true, we predicted we would find that many studies continue to present what appear to be retroductively derived, untested hypotheses.

We evaluated 287 research articles after excluding commentary articles, most human dimensions articles, and articles in which the research was designed to improve or develop estimation techniques and analyses, because such studies generally do not test biological hypotheses ($n = 92$). Six observers evaluated 4 journal issues each, resulting in 40–59 articles/observer. Based on Romesburg's (1981) arguments, we assumed that presence of explicitly stated *a priori* hypotheses was a sufficient indicator that Romesburg's H-D methodology was followed. We classified articles into 3 categories:

1) Reliable knowledge: ≥1 biological hypotheses were explicitly stated for each research question being addressed, most commonly within the introduction or methods, using language (e.g., we hypothesized, predicted, expected, thought) representing a biologically plausible answer to the research question being asked (see examples in Supporting Information, available online).
2) Possibly reliable knowledge (i.e., benefit of the doubt): hypotheses and their biological reasoning were implicit (i.e., the authors omitted the language described above), but enough detail was provided that *a priori* hypotheses could be plausibly inferred.
3) Unreliable knowledge: no *a priori* hypotheses were stated or implied for analyses that were presented, and inferences appeared to be derived retroductively from statistical analyses.

We did not evaluate papers fully to confirm complete application of the H-D method but assumed that the presence of explicitly stated hypotheses was an accurate indicator of its use. Violation of our assumption would have no effect on the proportion of studies we identified as unreliable, but it would inflate the proportion of studies identified as reliable or possibly reliable if any of these studies did not fully apply the H-D method. The protocol we used to classify studies was simple and objective; we therefore assumed potential effects of observer bias were minimal.

Only 41.8% of the 287 studies we reviewed stated biological hypotheses explicitly. We gave the benefit of the doubt to 14.3% of the studies for which we could infer the probable biological hypotheses based on what was presented, although no explicit hypotheses were stated. The remaining 43.9% of articles we reviewed stated neither hypotheses nor biological justification for study designs and analyses.

Our results suggest that, at best, slightly more than half of the studies we evaluated followed Romesburg's (1981) H-D methodology, providing knowledge that is reliable for research and management. This represents a liberal estimate because we gave the benefit of the doubt to studies where hypotheses and supporting biological reasoning appeared discernable but were not explicitly stated. Implicit or vaguely stated hypotheses, however, create opportunity for confusion and misinterpretation of results because information critical to full understanding of inferences is absent. We suggest most readers will not invest the time needed to discern such information, perhaps concluding erroneously the inferences lack reliability.

Consistent with our prediction, we found that many studies (43.9–58.2%, depending on how often we gave benefit of the doubt incorrectly) did not follow Romesburg's (1981) H-D methodology; their inferences thus appeared to be based on retroductively derived, untested hypotheses, which are inherently unreliable for research and management (Romesburg 1981). A rigorous scientific method may have been implicit (but undiscernible) in an unknown proportion of these studies, but credibility of their findings was voluntarily compromised because the scientific method used was not made explicit. Conceivably, some of these studies may have been sufficiently novel that *a priori* hypotheses could not be formulated or tested, but such novelty should be rare in wildlife research where theoretical and empirical precedent is abundant.

Arguably, the prevalence of studies that only generate untested hypotheses may be considered one of wildlife science's major problems (J. D. Nichols, U.S. Geological Survey [retired], personal communication); the high proportion of studies we evaluated that generated but did not test hypotheses should therefore be of concern to researchers and managers. Remedying this problem requires a common understanding among wildlife researchers of what rigorous science entails and why it is important. The remainder of this paper aims to establish such an understanding. We draw from the breadth of available historical, philosophical, and scientific literature to synthesize the key concepts underpinning production of reliable knowledge, and discuss implications for research and management.

HOW IS RELIABLE KNOWLEDGE PRODUCED?

Understanding biological causes of observed effects is of prime interest to researchers who seek to understand wildlife ecology and to managers who seek to manipulate causes to achieve desired effects. Although numerous means of establishing cause and effect exist, producing reliable knowledge about biological causation requires an understanding and application of a scientific method that develops and tests hypotheses (Romesburg 1981, 2009). Whereas detailed information is available in books suitable for in-depth study (Gauch 2003, Copi and Cohen 2005, Romesburg 2009, Curd et al. 2012), more readily-accessible papers in the primary literature that argue the importance of scientific rigor (Platt 1964, Romesburg 1981, Williams 1997) offer limited explanations for the defining steps of rigorous science. The lack of clear, succinct justification for employing the full series of steps is an understandable obstacle to acceptance by skeptics and critics, and a significant hurdle to graduate students and wildlife professionals developing research. Such justification has deep roots in history and logic.

Historical and Logical Roots of Scientific Methodology

Scientific methodology uses logic and observation to answer questions about the natural world. Although no historically or philosophically unified idea of scientific methodology applies to all applications of science, it is generally agreed that a rigorous scientific method for understanding biological causation consists of the following steps (Fig. 1; Platt 1964, Romesburg 1981, Hilborn and Mangel 1997, Williams 1997): 1) generate a research question from theory and prior observations, 2) develop hypotheses (i.e., plausible biological answers to the question), 3) formulate predictions (i.e., facts that must be true if the hypothesis is true), 4) design and implement research to collect data potentially consistent with predictions, 5) evaluate whether predictions are consistent with collected data, and 6) draw inferences based on the evaluation.

The roots of scientific methodology are ancient. Aristotle (384−322 BCE) arguably had the greatest impact on the history of biology (Mayr 1982), in part by developing a logical framework for drawing inferences about the physical world that "got 70% of scientific method right" (Gauch 2003:48). Aristotle's method of reasoning remains the

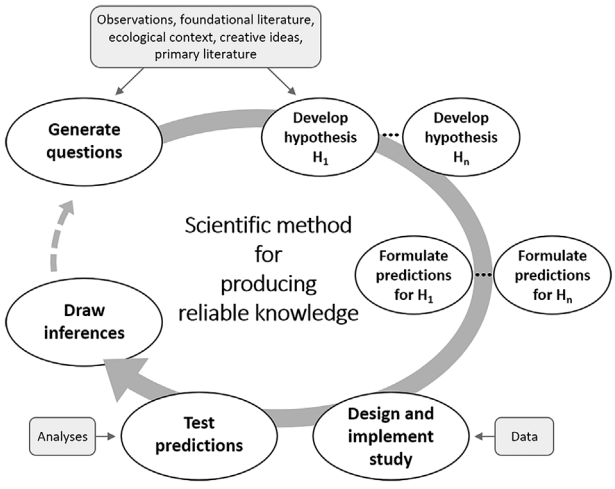

Figure 1. In wildlife science, a rigorous scientific method for producing reliable knowledge follows a series of logical steps to answer questions about the natural world. Each step is fundamental to the next. Inferences help inform new questions in future studies.

fundamental backbone of rigorous science to this day (Losee 1993:6–9, 29–44; Gauch 2003). Although modern empiricism is commonly attributed to F. Bacon (1561–1626), the work of R. Grosseteste (~1168–1253) and others in his era had solidified a "basically correct and complete" empirical scientific method by the thirteenth century (Gauch 2003:163). Grosseteste refined Aristotle's method and emphasized experimentation and falsification "in search of true causes" (Crombie 1962:84). His work influenced other scholars, who continued to spread these new, experimental approaches to science across medieval universities and through each subsequent century (Crombie 1962).

Aristotle's assertion that every belief arises through either inductive or deductive reasoning remains fundamental to the logic behind rigorous science (Gauch 2003:161). Each type of reasoning draws different types of conclusions with differing degrees of certainty. Conclusions reached through inductive logic represent generalizations inferred from specific observations, whereas those reached through deductive logic represent specific predictions derived from general concepts (Gauch 2003). The primary strength of inductive logic lies in its ability to use observations to generate broadly applicable hypotheses, an important component of scientific research (Williams 1997). In extrapolating something that is unknown from something that is known, however, induction draws strictly on association (i.e., correlation), not mechanism (i.e., causation; Romesburg 1981, Gauch 2003). Deduction is inherently mechanistic and does not rely on extrapolation; thus, conclusions drawn from deduction are more logically sound than those drawn from induction (Gauch 2003).

How the Steps of a Rigorous Scientific Method Produce Reliable Knowledge

A rigorous scientific method alternates between induction and deduction, using the strengths of each to compensate for the shortcomings of the other (Losee 1993, Gauch 2003). Making observations, detecting patterns and relationships, and developing potential answers to questions typically relies on inductive logic to develop general explanations (i.e., biological hypotheses) from specific observations (Williams 1997, Gauch 2003). Alternatively, hypotheses can be deductively generated completely *de novo* (e.g., Einstein's theory of relativity had almost no basis in the empirical physics of his time; Isaacson 2007), potentially leading to scientific revolutions (i.e., paradigm shifts; Kuhn 1962). In practice, however, science is generally normative such that new hypotheses proceed inductively from existing theory and empirical precedent (Kuhn 1962). Developing predictions associated with biological hypotheses uses deductive logic to formulate a prediction that must be true if the hypothesized explanation is true. Comparing collected data to predictions requires a return to inductive logic to draw inferences from analytical results to the population or system being studied. A synthesis of the conceptual underpinnings and applications of each of these steps follows.

Generate research questions from theory and observations.—Science is inherently question-driven, and each subsequent step of a rigorous scientific method proceeds directly from the initial research question. Scientists develop research questions by drawing from the broader context of previous scientific observations (e.g., by considering detected patterns and relationships) and theory (i.e., the body of knowledge operationally accepted as true; Romesburg 1981, Williams 1997). In wildlife biology, scientific studies often originate from a management need; suitable research questions are therefore developed by considering this management need within the context of related ecological questions of interest.

Wildlife biology primarily seeks to understand relationships between biological mechanisms and the effects they produce. Accordingly, wildlife research typically asks research questions about whether, how, or why certain effects occur (alternatively, research may focus on developing or refining techniques to measure those effects). Asking appropriate research questions can lead to increased understanding of the biological system and how it can be manipulated to achieve management goals. The complexity of questions, and their utility for predicting system responses to management actions, generally increases as the ease of answering them decreases.

The simplest research question asks, "is something happening?" For example, do beavers (*Castor canadensis*) gnaw cottonwood (*Populus* spp.) trees? Answering this question documents presence or absence of a pattern but does not reveal how or why the pattern occurs. Thus, although such answers can provide precursors to new biological hypotheses for future studies, they usually provide limited capacity for predicting system responses to management actions or predicting presence or absence of the pattern beyond the spatiotemporal scope of the study.

A common type of research question in wildlife management seeks to describe an observed pattern by asking, "what is happening?" For example, what species of trees do beavers gnaw? Answers to this type of question can help address management needs for the study system from which the data were collected, but because they do not identify how or why the pattern occurs, they can neither be used to confidently predict system responses to management actions nor to accurately determine whether the pattern will occur elsewhere. Beavers in one place, for example, may gnaw cottonwood trees more commonly than other tree species, yet without knowing the reason for this pattern (e.g., if it is a product of availability or preference), it is impossible to know how cottonwoods can be manipulated to affect beavers or whether beavers gnaw cottonwood trees at the same frequency in other places.

A research question that helps identify a plausible mechanism causing a pattern asks, "how is something happening?" For example, how do beavers select which trees to gnaw? Answers to this type of question describe (without explaining) causal mechanisms. They can therefore be used to predict how beavers would respond to management actions and to predict similar patterns in other times and places with greater confidence. Managers can use this knowledge to manipulate the possible mechanism influencing the pattern to help achieve management objectives

within and beyond the study system. One could find, for example, that beavers select trees based on nutritional quality. This observation would be expected for beavers in other places as well, allowing managers to manipulate the causes rather than correlates of beaver behavior (e.g., to manipulate gnawing, managers could fence trees of relatively high nutrition, instead of assuming a particular species such as cottonwoods will always be selected).

The research question "why is something happening?" seeks to explain evolutionary or ecological causal mechanisms that created the pattern. For example, why do beavers select certain species to gnaw? The why question is the brass ring of ecological research. It is the most difficult question to address, but explaining the means by which biological processes produce effects maximizes understanding of the system and provides the most predictive power to managers for reliable application to other times and places. For example, beavers may choose trees that maximize the energy gained from food resources over the energy lost to obtaining them; perhaps in some areas beavers select tree species that are more abundant and more easily obtained than cottonwoods even though they are of inferior nutritional quality.

These 4 types of research questions comprise an inclusive hierarchy. For example, a question asking, "why do beavers selectively gnaw certain species?" may reveal that beavers choose trees that maximize energetic benefits over costs. This answer would simultaneously reveal how beavers choose trees (foraging selectively on trees of high energetic value that are easily accessed and handled), what trees beavers will choose at a particular location (young hardwood trees close to water), and whether or not a particular species will be chosen (cottonwoods are gnawed).

Rigorous research asks and answers questions appropriate to the intended use of study results. Generality and reliability of these results to external application increases across the question spectrum, from describing a pattern unique in space and time (i.e., is or what questions) to identifying likely ecological mechanisms that could be consistent across space and time (i.e., how and especially why questions). Reliable extrapolation beyond the spatial or temporal scope of the study thus requires answering a how or why question, but project objectives and limited resources can preclude the ability to answer these types of questions. Where this occurs, a study can be redesigned to answer simpler questions, but in doing so researchers must recognize consequential limits on inferential scope and application of resulting inferences. If answers to an is or what question are insufficient for management needs, or if a goal is to provide knowledge for reliable extrapolation to other times and places, a study can be redesigned to answer more complex questions through creative thinking or increases in scope or funding.

Develop hypotheses.—Hypotheses are plausible biological answers to a research question (Romesburg 2009). A hypothesis typically posits a plausible biological cause for an observed effect. Biological hypotheses and statistical hypotheses are commonly conflated, but they are logically very different for all but the most basic questions. A statistical hypothesis represents a pattern predicted to be present in collected data if a biological hypothesis is true (Romesburg 1981, Johnson 1999). Using the term hypothesis to refer only to a biological hypothesis, not a statistical hypothesis, can reduce confusion and lack of clarity in scientific writing. Hypotheses developed prior to being tested are *a priori* hypotheses, whereas those developed based on results of data analyses but not yet tested are speculative, *a posteriori* hypotheses.

Although a single hypothesis may sufficiently address a research question, rarely does only one plausible explanation for an observed pattern exist (Pirsig 1974). Developing good hypotheses requires reducing potential explanations to a limited number of the most realistic, compelling, and useful biological answers to the research question (Williams 1997). Hypotheses of the greatest management utility address potential causal factors that management actions can influence (Nichols and Williams 2006). Strong *a priori* hypotheses typically build on past insights rather than reproduce them; evaluating a well-supported or well-refuted hypothesis is generally unproductive unless doing so would likely expose a flaw in current theory or interpretations of empirical precedent. Developing an understanding of key theories and concepts underlying published research can reduce the considerable difficulty of developing strong hypotheses. Particularly for students, reading an authoritative book or synthesis article or taking a course that summarizes relevant fundamental concepts can provide a foundation from which to synthesize primary literature and develop good hypotheses.

Having developed a candidate set of hypotheses, a study may test $\geq$1 hypotheses from that set. Whereas some philosophers have viewed hypothesis testing as sequential tests of single explanatory hypotheses (Popper 1959), testing multiple hypotheses simultaneously is more efficient (Chamberlin 1890, Platt 1964) and allows explicit consideration of the fact that multiple factors may simultaneously contribute to an observed pattern (Hilborn and Mangel 1997, Williams 1997, Belovsky et al. 2004).

Testing multiple hypotheses also greatly improves reliability of findings by reducing the influence of cognitive bias. Cognitive bias is an inescapable part of human thinking; all people inherently tend to reach conclusions consistent with their existing beliefs and biases (i.e., confirmation bias; Kahneman 2011). Developing only a single hypothesis can thus blind researchers to additional explanations for an observed phenomenon, increasing the likelihood of finding support for their pet hypothesis. Chamberlin (1890:755) put it colorfully:

> "The moment one has offered an original explanation for a phenomenon which seems satisfactory, that moment affection for his intellectual child springs into existence. ... There is an unconscious selection and magnifying of the phenomena that fall into harmony with the theory and support it and an unconscious neglect of those that fail of coincidence."

Testing multiple alternative hypotheses requires thinking through multiple plausible answers to a research question, including those inconsistent with a personal favorite

(Chamberlin 1890, Platt 1964), thereby reducing the potential for confirmation bias. Of the 41.8% of *JWM* papers that produced reliable information, only half tested multiple hypotheses, producing stronger inferences and advancing understanding more efficiently than the other studies, with less potential for influence from cognitive biases.

Formulate predictions.—Because biological hypotheses represent processes that can rarely be tested directly (Romesburg 1981, 2009), researchers develop predictions they can directly refute or support, commonly through statistical analysis (Nichols et al. 2011). Predictions are logical deductions of what one expects to observe in collected data if a biological hypothesis is true (Romesburg 1981, Williams 1997). For example, to evaluate a hypothesis that beavers preferentially gnaw softer wood because it is energetically more efficient to consume, a researcher may test the prediction that gnawed trees will be softer than a representative sample of trees available to beavers. Developing predictions from hypotheses necessarily requires researchers to make assumptions (e.g., assuming trees >20 m from water are not available to beavers and therefore not part of a representative sample).

When testing multiple hypotheses, predictions for each hypothesis should be unique. Multiple hypotheses that produce similar predictions cannot be differentiated and are therefore conflated. For example, the hypotheses that beavers prefer cottonwoods, and beavers prefer relatively soft woods, would be conflated if beavers selectively gnawed cottonwoods whose wood happened to be softest among the tree species available. An attempted test of the 2 hypotheses would thus be meaningless and could potentially lead to misleading inferences.

Design and implement research to collect data.—Study design, which ranges from experimental to observational, allows a researcher to gather appropriate data to find support for or against hypotheses (Romesburg 1981, 2009; Eberhardt and Thomas 1991; Sinclair 1991; Gotelli and Ellison 2012). Experiments allow researchers to directly test for the presence of an effect in the presence (treatment) and absence (control) of a hypothesized cause while using replication and randomization to control for other potential causes, site-specific idiosyncrasies, and potentially biased sampling (Sinclair 1991, Gotelli and Ellison 2012). In wildlife research, experiments can be manipulative (e.g., fencing off highly nutritious trees, paired with an unfenced control, could be used to test the hypothesis that beavers select trees based on nutrition) or natural (e.g., testing the same hypothesis where wildfire changed forest structure at a beaver pond, paired with an unburned control). Experiments are challenging to conduct in the natural world because of the logistical difficulties of manipulation and the paucity of natural experiments, and because it is rarely possible to test for or control all potential causes for the complex patterns that can be observed. Observational studies are therefore more common in wildlife biology (Sinclair 1991, Gotelli and Ellison 2012). Observational studies do not test for hypothesized effects in the presence of a control and instead use associations between hypothesized causes and effects to draw inferences. Reducing the potential effects of unmeasured causes, site-specific characteristics, and sampling bias on results is arguably most important when hypotheses are tested in the absence of a control. Thus, the reliability of inference drawn from observational studies benefits from study designs that include adequate replication and randomization.

Study design has important implications for the ability to build support for or against hypotheses (Romesburg 1981, 2009; Eberhardt and Thomas 1991; Nichols 1991; Gotelli and Ellison 2012). Under an experimental study design, presence of the effect only in presence of the hypothesized cause, and lack of the effect in absence of the cause, builds support for causation. In an observational study, positive association between hypothesized cause and effect demonstrates correlation, not causation, because other potential causes were not controlled for and thus could also result in the same effect (Hilborn and Stearns 1982, Nichols 1991, Gotelli and Ellison 2012). Therefore, a lack of correlation typically implies the hypothesized cause did not lead to the effect (or that it was masked by other effects), whereas correlation suggests only potential causation. The potential to reduce uncertainty through finding poor support for a hypothesis is the chief strength of a hypothesis-driven observational study. Evaluating multiple competing hypotheses efficiently amplifies this inferential power because tests will typically find poor support for some hypothesized cause-effect relationships while supporting others (Williams 1997).

Test predictions.—Testing biological hypotheses involves mathematically formalizing predictions (e.g., $\bar{x}_\alpha > \bar{x}_0$ or $\beta_1 < 0$) and comparing them to empirical observations (i.e., collected data). Simple observations (e.g., tooth marks from beavers on cottonwoods) or summaries of data (e.g., $\bar{x}$ and SD for number of different tree species gnawed by beavers) may be sufficient to determine whether a prediction has been met. In wildlife science, however, testing predictions commonly entails statistics; therefore, rigorous science depends on logical rigor and statistical rigor. Statistical rigor comprises separate, fundamental considerations beyond the scope of this paper (Hilborn and Mangel 1997, Taper and Lele 2004, Bolker 2008, Gotelli and Ellison 2012). We note, however, that despite some disagreement among statistical theorists (Bayarri and Berger 2004), any statistical approach (including frequentist and Bayesian frameworks) can be used to find support for or against biological hypotheses if analytical assumptions are met (Hilborn and Mangel 1997, Bolker 2008, Gotelli and Ellison 2012). Importantly, statistical tests evaluate predictions from biological hypotheses, not the hypotheses themselves (Romesburg 1981, Hilborn and Mangel 1997, Williams 1997, Nichols et al. 2011). Therefore, inferring whether results of objective statistical tests indicate support for or against biological hypotheses necessarily requires subjective interpretation.

Draw inferences.—The logic of drawing inferences from hypothesis testing is based on the deductive reasoning of material implication, a form of Aristotle's syllogism

(Williams 1997). Syllogistic logic specifies that a conclusion is logically supported only if multiple premises are true (Aristotle ~350 BCE). For hypothesis testing, this logic can be represented as:

Premise 1: $\{T\} + H \rightarrow P$

Premise 2: $O \rightarrow \sim P$

Conclusion: $\sim H$

where {T} = theory (conceptual framework accepted as true), H = hypothesis (an amendment to theory) being tested, P = prediction (based on the hypothesis and theory), O = observations (i.e., data), $\rightarrow$ = implication, and $\sim$ = negation. This argument makes logically explicit the steps of a rigorous scientific method. It states that if an accepted theory is amended by a true hypothesis, then specific predictions produced by the amendment should be accurate (premise 1), but if observations show that predictions are not accurate (premise 2; generally established statistically) then the hypothesis is not supported (conclusion). Under this form of syllogistic logic, known as *modus tollens*, inconsistency between observations and predictions is logically sufficient to reject the hypothesis, assuming theory and observations are valid (Copi and Cohen 2005).

This same syllogistic logic shows why hypothesis confirmation is logically invalid. In the above argument, observations could be consistent with predictions (premise 2: $O \rightarrow P$) so that the hypothesis is not falsified (Williams 1997). In this case, the hypothesis is considered supported. Repeated failures to falsify a hypothesis reflect accumulation of supporting evidence that the hypothesis may be useful and predictive. The hypothesis cannot be confirmed, however, because other plausible but untested hypotheses could have equal or better explanatory power. Furthermore, inferring confirmation of the hypothesis represents circular reasoning because the truth of the hypothesis is posited in premise 1 and affirmed in the conclusion. This logical fallacy is known as affirming the consequent (Williams 1997, Copi and Cohen 2005) and is consistent with the cognitive error of confirmation bias. This inferential error is most problematic when hypotheses are not stated explicitly in a study, because audiences cannot detect the error. Williams (1997) asserted inferences based on affirming the consequent dominate wildlife research; if true, such studies make up an unknown proportion of the *JWM* articles we evaluated.

Based on the syllogistic logic of *modus tollens*, studies that subject *a priori* hypotheses to falsification produce the most reliable knowledge. Hypotheses that remain unfalsified through repeated testing eventually gain support and can be operationally treated as true (e.g., theories of gravity and natural selection). Inductively derived *a posteriori* explanations for observed patterns do not benefit from the logical rigor of *modus tollens* and therefore are untested hypotheses.

A posteriori hypotheses do, however, have an important place in the final step of rigorous science. Inferences drawn from hypothesis tests inevitably lead to *post hoc*, speculative explanations for unexplained variation in observed effects. For example, if a study finds support for the hypothesis that beavers choose trees based on nutritional quality, a new *a posteriori* hypothesis might be that the same will be true for selection of trees by porcupines (*Erethizon dorsatum*). Although uncertainty due to speculation increases with extrapolation beyond the current study and *a priori* hypotheses tested, *a posteriori* hypotheses suggest modifications to existing theory, question reigning paradigms, and constitute the backbone of management implications. Importantly, untested *a posteriori* hypotheses must be clearly represented as speculative to make their value to subsequent studies and limited inferential strength clear.

Humans are incapable of perceiving and understanding full reality so can only use science to evaluate a simplified version of it; therefore, research by definition is based on assumptions that have implications for the reliability of inferences drawn. Premise 1 of *modus tollens* assumes the theory and hypothesis are true and that predictions are accurately derived from the hypothesis; premise 2 assumes sampling design and statistical analyses are appropriate for determining the relationship between predictions and observations. Support for or against the hypothesis is valid only if all other components of the premises are safely assumed to be true. The only assumption typically tested in science, however, is the validity of the hypothesis. The assumption that the theory is true could be violated; the hypothesis could therefore be incorrectly tested because the theory was misused, misrepresented, or false. Similarly, the hypothesis could be incorrectly tested because the assumption that predictions were accurate could be violated (e.g., predictions were inappropriate to temporal or spatial scales of the study). Finally, the hypothesis could be incorrectly tested if there was a violation of the assumptions that data collected for the study were sufficient to test the hypothesis (e.g., sample size was too small) or that analyses used to compare observations to predictions were correctly applied. Power analyses and consideration of statistical assumptions (e.g., normally distributed data) can be used to evaluate potential violations of logical assumptions.

Violation of assumptions can bias results, leading to assessing the degree of support for a hypothesis incorrectly (i.e., finding support for an incorrect hypothesis or falsifying a correct one, contributing to type I and II errors, respectively; Gotelli and Ellison 2012). Inferences are inherently more reliable when violation of an assumption is unlikely, or when effects of a violation on results would be minimal because the assumption is trivial. The reliability of inferences therefore requires an explicit consideration and communication of each important assumption made, potential effects of its violation on results, and the assessed likelihood of violation.

EXPLORATORY ANALYSES

Exploratory research may seek to provide new hypotheses without actually testing them. Such research generally uses statistical analyses to detect patterns in datasets, producing

results that are interpreted *post hoc*. Platt (1964:348) argued that scientists have "become 'method-oriented' rather than 'problem-oriented'" by allowing such methods to become "ends in themselves." The ongoing development of statistical approaches, increasing ease of data analysis, and the growing prevalence of large and complex data sets likely compound this problem. Using statistics to search for patterns without first developing questions, hypotheses, and predictions (i.e., data dredging; Burnham and Anderson 2002) represents non-rigorous science. These inductively identified patterns are untested hypotheses because they are not subjected to the rigorous logic of *modus tollens*. Importantly, interpreting results of such studies is inherently subjective because biological concepts and empirical precedents that lead to inferences are derived from *post hoc*, idiosyncratic interpretation; the lack of objective logical support for such inferences means the reader's unique interpretation of the results is as valid as the authors'. Finally, searching for patterns can yield misleading *post hoc* inferences due to spurious correlations, particularly when covariates lack clear biological justifications (Vigen 2017).

The capacity of exploratory studies to contribute to the accumulation of evidence for or against existing theory or management practices is strongly limited. When no theory or empirical precedent is available to generate *a priori* hypotheses, offering *a posteriori* explanations for observations is necessary to provide the foundation for *a priori* hypotheses that subsequent studies can test (i.e., normal science; Kuhn 1962). Because a complete lack of existing theory or empirical precedent is rare, however, generating solely *a posteriori* explanations from observations should be similarly rare. More commonly, sufficient information is available to researchers prior to a study that they will have informal ideas of a research question, hypotheses, and predictions in mind (this is clear when the results of an exploratory study are shown to be consistent with those of prior studies, negating the need for exploration). In such cases, failing to formalize questions, hypotheses, and predictions ignores exploratory work already done and sacrifices opportunity to build on it to produce new insights. For example, formally evaluating the causation implied by each biological covariate in an exploratory model establishes the model's generality more efficiently and effectively than waiting for the results of multiple exploratory models to converge on the same insights.

Failing to present *a posteriori* hypotheses produced by exploratory research as speculative exaggerates the usefulness of findings and could mislead future research and management. The accumulation of speculative, untested findings over time presents a significant impediment to scientific understanding and effective management because repetition contributes to a mistaken impression of reliability (Romesburg 1981, Sinclair 1991).

RESEARCH IMPLICATIONS

Rigorous science increases the reliability of scientific knowledge by using sound logic to gradually reduce the set of plausible biological explanations to a smaller subset of the most likely and well-supported hypotheses (Platt 1964, Williams 1997). Thus, rigorous science reduces overall uncertainty. In contrast, non-rigorous research generates hypotheses but does not test them. Ultimately, both rigorous and non-rigorous science may have the capacity to reveal explanations for biological phenomena through accumulation of evidence (i.e., convergence of inferences across multiple studies), but efficiency and reliability differ strongly between the 2 approaches. Rigorous science accumulates evidence for competing hypotheses and discards incorrect explanations by falsifying hypotheses. Non-rigorous science does not discard incorrect explanations. The number of plausible explanations to a research question therefore increases as more *a posteriori* hypotheses are added to a body of knowledge but never tested. Although some have argued that deriving inference from induction alone accumulates evidence when conclusions from multiple studies converge on the same untested hypothesis (e.g., F. Bacon; Dick 1955), it is rare that sufficiently similar, independent studies are numerous enough for convergence to occur. Non-rigorous science therefore provides relatively little contribution to reducing uncertainty or advancing scientific understanding.

Researchers maximize the reliability of knowledge they provide by not only adhering to, but also explicitly presenting, their use of a rigorous scientific method. Clearly introducing the research question, hypotheses, and predictions up front (i.e., in the introduction or methods) allows audiences to accurately assess the broad applicability of the study and to easily identify and understand the biological reasoning of the *a priori* hypotheses that the study tested. Failure to explicitly state hypotheses voluntarily sacrifices rigor of research findings and limits robust contributions to theory or management. When reporting results and drawing inference, clearly differentiating tested *a priori* hypotheses from untested *a posteriori* hypotheses distinguishes reliable knowledge from speculative *post hoc* inference. Explicitly discussing assumptions further allows the audience to assess credibility of conclusions made.

Lack of clarity in presentation of scientific research hinders applicability of the research and its reproducibility. Because science is an iterative pursuit of knowledge, future studies rely on reproducibility of past research. Reproducing research can entail repeating exact methods, but more often reproduction is based on the concepts and hypotheses driving the original research. Research is most reproducible when it follows a rigorous, explicitly stated scientific method.

MANAGEMENT IMPLICATIONS

Management decisions based on rigorous science are most likely to result in effective conservation of wildlife resources (Leopold 1933, Gill 1985). Risk is associated with any application of research results to management because of the uncertainty inherent in any scientific inference. The decision to use research results to inform management is therefore partly a function of a manager's risk attitude. Managers with

low risk tolerance are best served by research results that explicitly reduce uncertainty about reliability of results through application of a rigorous scientific method. Although research that minimizes uncertainty would clearly benefit any management decision, managers with high risk tolerance may find the greater uncertainty associated with speculative findings of non-rigorous research acceptable if the consequences of a misinformed decision are not substantial. Our analysis of research presented in *JWM* suggests that managers with relatively low risk tolerance can rely on slightly more than half of the studies we evaluated to make their decisions. Whether this proportion represents the distribution of risk attitudes among managers, or whether the non-rigorous research presented in *JWM* was designed to inform low-risk management decisions, is worthy of further investigation.

Many researchers and managers incorrectly assume an inherent disconnect exists between advancing basic scientific theory and meeting applied management needs (Gavin 1989, Nudds and Morrison 1991, Belovsky et al. 2004). Rigorous science that answers how or why questions allows studies to consider management-related questions explicitly within the context of biological theory. Results of such research therefore reliably inform basic ecology and its application to wildlife management. Furthermore, applied research based on how and why questions provides better understanding of biological mechanisms; management decisions based on the results of such research are therefore more likely to prove effective beyond the scope of the original study system.

Basing management decisions on inferences gained from rigorously tested *a priori* hypotheses increases the likelihood of those decisions resulting in the desired effect (Sutherland et al. 2013). Additionally, discerning the type of question a study addressed allows managers to determine potential management actions and how reliably knowledge can be extrapolated beyond the original study system to other systems with similar management needs. Management implications offered by any study are *post hoc* inferences, just like *a posteriori* biological hypotheses, unless effects of management practices were explicitly tested. As such, they should be treated in management practice as untested hypotheses. Where regular monitoring is a part of management, it can be tailored to test these hypotheses over time to reduce uncertainty within an adaptive management framework (Nichols and Williams 2006).

ACKNOWLEDGMENTS

We thank J. D. Nichols, R. W. Mannan, J. J. Nowak, P. M. Lukacs, J. D. DeVoe, K. L. Eneas, T. A. Hayes, C. J. Peterson, K. E. Loonam, C. L. Waters, C. M. Vester, and an anonymous reviewer for their thoughtful feedback on previous versions of this manuscript. We also thank J. B. Stetz for assisting the authors in analyzing the *JWM* articles, and S. Le Bihan for contributing to our discussions about the philosophy of science.

LITERATURE CITED

Aristotle. ~350 BCE. Prior analytics. Translated by R. Smith, 1989. Hackett Publishing Company, Inc., Indianapolis, Indiana, USA.

Bayarri, M. J., and J. O. Berger. 2004. The interplay of Bayesian and frequentist analysis. Statistical Science 19:58–80.

Belovsky, G. E., D. B. Botkin, T. A. Crowl, K. W. Cummins, J. F. Franklin, M. L. Hunter, Jr., A. Joern, D. B. Lindenmayer, J. A. MacMahon, C. R. Margules, and J. M. Scott. 2004. Ten suggestions to strengthen the science of ecology. BioScience 54:345–351.

Bolker, B. M. 2008. Ecological models and data in R. Princeton University Press, Princeton, New Jersey, USA.

Burnham, K. P., and D. R. Anderson. 2002. Model selection and multimodel inference: a practical information-theoretic approach. Second edition. Springer-Verlag, New York, New York, USA.

Chamberlin, T. C. 1890. The method of multiple working hypotheses. Science 15:92–96 (reprinted in Science 148:754–759, 1965).

Copi, I. M., and C. Cohen. 2005. Introduction to logic. Twelfth edition. Pearson Education, Inc., Upper Saddle River, New Jersey, USA.

Crombie, A. C. 1962. Robert Grosseteste and the origins of experimental science 1100–1700. Oxford University Press, London, United Kingdom.

Curd, M., J. A. Cover, and C. Pincock. 2012. Philosophy of science: the central issues. Second edition. W. W. Norton & Company, New York, New York, USA.

Dick, H. G. 1955. Selected writings of Francis Bacon. Modern Library, New York, New York, USA.

Eberhardt, L. L., and J. M. Thomas. 1991. Designing environmental field studies. Ecological Monographs 61:53–73.

Gauch, H. G. 2003. Scientific method in practice. Cambridge University Press, Cambridge, United Kingdom.

Gavin, T. A. 1989. What's wrong with the questions we ask in wildlife research? Wildlife Society Bulletin 17:345–350.

Gill, R. B. 1985. Wildlife research: an endangered species. Wildlife Society Bulletin 13:580–587.

Gotelli, N. J., and A. M. Ellison. 2012. A primer of ecological statistics. Second edition. Sinauer Associates, Inc., Sunderland, Massachusetts, USA.

Hilborn, R., and M. Mangel. 1997. The ecological detective: confronting models with data. Princeton University Press, Princeton, New Jersey, USA.

Hilborn, R., and S. C. Stearns. 1982. On inference in ecology and evolutionary biology: the problem of multiple causes. Acta Biotheoretica 31:145–164.

Isaacson, W. 2007. Einstein: his life and universe. Simon and Schuster, New York, New York, USA.

Johnson, D. H. 1999. The insignificance of statistical significance testing. Journal of Wildlife Management 63:763–772.

Kahneman, D. 2011. Thinking, fast and slow. Farrar, Straus and Giroux, New York, New York, USA.

Kuhn, T. S. 1962. The structure of scientific revolutions. University of Chicago Press, Chicago, Illinois, USA.

Leopold, A. 1933. Game management. C. Scribner's Sons, New York, New York, USA.

Losee, J. 1993. A historical introduction to the philosophy of science. Third edition. Oxford University Press, London, United Kingdom.

Mayr, E. 1982. The growth of biological thought: diversity, evolution, and inheritance. Belknap Press, Cambridge, Massachusetts, USA.

Nichols, J. D. 1991. Science, population ecology, and the management of the American black duck. Journal of Wildlife Management 55:790–799.

Nichols, J. D., K. U. Karanth, and A. F. O'Connell. 2011. Science, conservation, and camera traps. Pages 45–56. *in* A. F. O'Connell, J. D. Nichols, and K. U. Karanth, editors. Camera traps in animal ecology: methods and analyses. Springer, New York, New York, USA.

Nichols, J. D., and B. K. Williams. 2006. Monitoring for conservation. Trends in Ecology and Evolution 21:668–673.

Nudds, T. D., and M. L. Morrison. 1991. Ten years after "reliable knowledge": are we gaining? Journal of Wildlife Management 55:757–760.

Pirsig, R. M. 1974. Zen and the art of motorcycle maintenance: an inquiry into values. William Morrow and Company, New York, New York, USA.

Platt, J. R. 1964. Strong inference. Science 146:347–353.

Popper, K. 1959. The logic of scientific discovery. Routledge, New York, New York, USA.
Romesburg, H. C. 1981. Wildlife science: gaining reliable knowledge. Journal of Wildlife Management 45:293–313.
Romesburg, H. C. 2009. Best research practices: how to gain reliable knowledge. Lulu Enterprises, Morrisville, North Carolina, USA.
Sinclair, A. R. E. 1991. Science and the practice of wildlife management. Journal of Wildlife Management 55:767–773.
Sutherland, W. J., D. Spiegelhalter, and M. A. Burgman. 2013. Twenty tips for interpreting scientific claims. Nature 503:335–337.
Taper, M. L., and S. R. Lele. 2004. The nature of scientific evidence: statistical, philosophical, and empirical considerations. University of Chicago Press, Chicago, Illinois, USA.
Vigen, T. 2017. Spurious correlations. http://tylervigen.com/spurious-correlations. Accessed 4 Sep 2017.
Williams, B. K. 1997. Logic and science in wildlife biology. Journal of Wildlife Management 61:1007–1015.

Associate Editor: Quresh Latif.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of this article at the publisher's website.