Framework for Testing the Effectiveness of Bat and Eagle Impact-Reduction Strategies at Wind Energy Projects

Technical Monitors:

Karin Sinclair and Elise DeGeorge
National Renewable Energy Laboratory

NREL is a national laboratory of the U.S. Department of Energy
Office of Energy Efficiency & Renewable Energy
Operated by the Alliance for Sustainable Energy, LLC

This report is available at no cost from the National Renewable Energy Laboratory (NREL) at www.nrel.gov/publications.

Technical Report
NREL/TP-5000-65624
April 2016

Contract No. DE-AC36-08GO28308

NREL
NATIONAL RENEWABLE ENERGY LABORATORY

Framework for Testing the Effectiveness of Bat and Eagle Impact-Reduction Strategies at Wind Energy Projects

Technical Monitors:

Karin Sinclair and Elise DeGeorge
National Renewable Energy Laboratory

Prepared under Task No. WE11.1005

NREL is a national laboratory of the U.S. Department of Energy
Office of Energy Efficiency & Renewable Energy
Operated by the Alliance for Sustainable Energy, LLC

This report is available at no cost from the National Renewable Energy Laboratory (NREL) at www.nrel.gov/publications.

National Renewable Energy Laboratory
15013 Denver West Parkway
Golden, CO 80401
303-275-3000 • www.nrel.gov

Technical Report
NREL/TP-5000-65624
April 2016

Contract No. DE-AC36-08GO28308

This report is available at no cost from the National Renewable Energy Laboratory (NREL) at www.nrel.gov/publications.

Available electronically at SciTech Connect http:/www.osti.gov/scitech

Available for a processing fee to U.S. Department of Energy and its contractors, in paper, from:

U.S. Department of Energy
Office of Scientific and Technical Information
P.O. Box 62
Oak Ridge, TN 37831-0062
OSTI http://www.osti.gov
Phone: 865.576.8401
Fax: 865.576.5728
Email: reports@osti.gov

Available for sale to the public, in paper, from:

U.S. Department of Commerce
National Technical Information Service
5301 Shawnee Road
Alexandra, VA 22312
NTIS http://www.ntis.gov
Phone: 800.553.6847 or 703.605.6000
Fax: 703.605.6900
Email: orders@ntis.gov

Cover Photos by Dennis Schroeder: (left to right) NREL 26173, NREL 18302, NREL 19758, NREL 29642, NREL 19795.

NREL prints on paper that contains recycled content.

Abstract

The objectives of this framework are to facilitate the study design and execution to test the effectiveness of bat and eagle impact-reduction strategies at wind energy sites. Through scientific field research, the wind industry and its partners can help determine if certain strategies are ready for operational deployment or require further development. This framework should be considered a living document to be improved upon as fatality-reduction technologies advance from the initial concepts to proven readiness (through project- and technology-specific testing) and as scientific field methods improve.

Acknowledgments

This work was supported by the U.S. Department of Energy (DOE) under Contract No. DE-AC36-08GO28308 with the National Renewable Energy Laboratory (NREL). Funding for the work was provided by DOE's Office of Energy Efficiency and Renewable Energy, Wind and Water Power Technologies Office.

NREL thanks the numerous individuals who contributed to the development of this document, including (in alphabetical order):

- Taber Allison, American Wind Wildlife Institute
- Robert C. Beason, Accipiter Radar Corporation
- Oliver Behr, University of Erlangen, Germany
- Jocelyn Brown-Saracino, DOE[1]
- Paul Cryan, U.S. Geological Survey[1]
- Wallace Erickson, Western EcoSystems Technology, Inc.
- Cris Hein, Bat Conservation International[1]
- Christy Johnson-Hughes, U.S. Fish and Wildlife Service[1]
- Brian Millsap, U.S. Fish and Wildlife Service
- Michael Morrison, Texas A&M University[2]
- Laura Nagy, Iberdrola Renewables
- Christian Newman, Normandeau Associates, Inc.
- Kaj Skov Nielsen, Siemens
- Steve Pelletier, Stantec, Inc.
- Michael Schirmacher, Bat Conservation International[2]
- Lynn Sharp, Consultant
- Shawn Smallwood, Consultant[2]
- Heidi Souder, University of Colorado
- Dale Strickland, Western EcoSystems Technology, Inc.
- Crissy Sutter, Normandeau Associates, Inc.
- Julie Yee, U.S. Geological Survey[2]

[1] Served as a peer reviewer of the draft document
[2] Major contributor in the editing of this document

Table of Contents

List of Figures

List of Tables

1 Introduction

As the wind energy industry in the United States continues to expand, there is an increased need for proven and cost-effective tools that reduce impacts on wildlife. To date, the suite of impact-reduction strategies (e.g., wildlife operational curtailment) and techniques, many of which are in various stages of testing for effectiveness, has been limited for most species, primarily focusing on collision-related avian and bat fatalities. But other wildlife may be at risk as a result of habitat impacts. This section includes the background, purpose, and outline of the framework.

Examples of potential impact-reduction strategies for wind energy projects include (in alphabetical order):

- **Blade-painting schemes**. Paints or colors may improve the visual contrast of wind turbine blades against a terrain or the sky, or color schemes may reduce motion smear.
- **Detect-and-curtail approaches**. Automated or human detections of target species can result in shutting down wind turbines. This is also known as *informed curtailment*.
- **Detect-and-deter approaches.** Automated or human detections of target species can result in activating deterrents, such as noise emitters or lights.
- **Deterrents**. Acoustic or visual devices can be used to discourage wildlife from approaching (to within a certain distance of) a wind turbine or its rotor.
- **Wildlife operational curtailment**. Wildlife operational curtailment is the process of stopping or greatly reducing the rotor rotation rate to eliminate wildlife fatalities, and it can be implemented in a couple of different ways. The turbine control system can be adjusted to change the blade pitch angle (e.g., pitching the blades out of the wind) so that the rotor blades are stationary or only rotate slowly due to wind variability. Alternately, rotor rotation can be stopped by adjusting the control system to apply the parking brake to stop the rotor. Multiple scenarios are under consideration or being field-tested, including a fixed stepped change to the blade pitch angle during seasons of the year when bats are highly active and an adjustable step change to the blade pitch angle wherein the size of the step increase will be determined by algorithms based on variables used to predict bat activity at the site.
- **Wildlife seasonal curtailment**. This is the process of employing wildlife operational curtailment of wind turbine operation during one or more seasons of the year to eliminate fatalities.
- **Wind turbine design modifications**. Evaluation of novel turbine features designed to reduce fatalities by limiting or preventing wildlife from entering the rotor plane from various flight directions depending on the probability of collision should be pursued. These include minimizing potential attraction, minimizing perching, or deterring or alerting wildlife to the hazard (via noise and/or lights).

1.1 Background

Although experiments that adjust feathering and higher cut-in speeds have been shown to reduce bat mortality, curtailing turbine operation above manufacturer-specified cut-in speeds may reduce energy production and may pose technical challenges for the turbines. Promising

deterrent systems are under development, but more research is needed to clearly demonstrate their efficacy. For eagles and other raptors, wind energy project owners and operators have exercised manual (triggered by human observers) and automated turbine shutdowns in the presence of target species or under conditions thought to increase collision risk. Manual shutdowns require human observers and are therefore costly, and automated detect-and-deter systems, which may also be costly, are in need of peer-reviewed studies to clearly demonstrate their effectiveness.

Wind energy project developers and regulators need to be assured of the biological efficacy of impact-reduction measures. To accomplish this, robust, transparent field tests must be performed, and results, if warranted, must be published in peer-reviewed journals (with an understanding about whether there could be environmental implications from conducting large-scale studies). Methodologies for such trials must address multiple associated challenges, prominently including the often rare recording of the events these measures seek to reduce (e.g., golden eagle fatalities found on the order of once per year among 10 to 20 wind turbines or the finding of only 5 to 10 bat carcasses for every 100 bat fatalities estimated to have occurred). These studies must be designed to statistically demonstrate the efficacy of the impact-reduction measure and thus need a sufficient number of observed carcasses to do so. Expanding the sample size by including more wind facilities or field seasons can greatly increase the cost and duration of studies; however, determining whether a particular study should be conducted will be a decision for each individual project team. Although researchers can and should design methods to increase the detection rate of carcasses, doing so for some species of concern may be insufficient to overcome large variances in fatality among sampling units. In such cases, unless a convincing argument can be made for why the results of a test performed using a broader taxonomic group or using a behavioral indicator of risk will be relevant to the target species, proceeding with the test would be imprudent.

Field tests should be based on experimental design principles that include a number of statistical considerations: clearly articulated research questions, clearly defined experimental units with care to avoid or manage pseudoreplication, consideration of anticipated levels of variation and whether these can be controlled, effect size, potential for confounding variables to influence response, randomization and interspersion of treatments, representation, and appropriate spatial and temporal scales. Even research protocols that take such issues into consideration may have design flaws; thus, experimental designs should be subject to peer reviews prior to the outset of any fieldwork.

1.2 Purpose

Given that strategies to reduce negative impacts to birds and bats at operational wind energy projects are evolving, NREL developed this framework for testing the efficacy of these strategies. To accomplish this, the laboratory enlisted input from a panel of experts, including wildlife biologists, biostatisticians, and others (see contributors included in the acknowledgments section). This framework is intended to facilitate consistent implementations of experimental design principles and methods for field testing impact-reduction strategies so that test results are convincing and, to the degree feasible, comparable to those of other impact-reduction studies.

Potential users of this framework include:

- State and federal agencies
- Impact-reduction technology developers and manufacturers
- Scientists and statisticians
- Nongovernmental organizations
- Wind industry developers and consultants
- Impact-reduction technology investors.

1.3 Outline

The framework contains the following sections:

- **Problem formulation.** Section 2 is intended to help the user identify the questions they are trying to answer and the results they are seeking to obtain.
- **Analysis and characterization.** Section 3 provides concepts for the user to consider when compiling their testing plan, analyzing results (e.g., characterizing exposure and/or ecological effects), and determining next steps.
- **Reporting.** Section 4 offers suggestions for conducting proposal and peer reviews and disseminating research results.
- **Discussion.** Section 5 provides guidance on communicating data, results, and next steps.
- **References.** The references listed in this section correspond to the in-text citations.
- **Bibliography.** The resources listed in this section may provide additional useful information for the reader.
- **Glossary.** The glossary defines some of the terms used throughout this report.
- **Appendix.** The appendix includes a hypothetical example of a study designed to test the efficacy of a deterrent technology.

2 Problem Formulation

2.1 Study Feasibility and Testing Combinations

In this section, we discuss basic considerations in designing a study that investigates strategies to reduce fatalities at wind energy projects. It is critical to recognize that the design of a study is the foundation upon which any rigorous and thus reliable inference will be based. Our objective is to outline an approach to the study design that will provide reliable, comparable results. A clear articulation of the research question is generally the first step.

Testing the effectiveness of impact-reduction strategies requires statistical power to determine a treatment effect. Getting this statistical power can be challenging for many species of interest at wind energy projects because fatalities of most species are a statistically rare occurrence. Another challenge is that the conditions under which an experiment can be attempted are dictated by the layout and operation of a wind energy project. It can be difficult for the researcher to design an energy project to meet the objectives of an experiment. More often, the researcher will design an experiment around an existing wind project or one that has already been planned, if not yet built. The researcher will then attempt to control the variation in data collected from a manipulative experiment that was formulated at a wind project out of convenience or in less-than-ideal circumstances.

In the context of this document, we define "rare events" as events that are expected to be observed too infrequently to readily determine a treatment effect. For example, we may propose studying an impact-reduction strategy that is expected to reduce a fatality rate by 50%. If the study is conducted at a site that has an inherent rate of 10 fatalities per wind turbine but a detection rate of only 10%, then the average number of fatalities actually observed per turbine (10 fatalities × 0.1 = 1) might be too small to determine a treatment effect without including a prohibitively high number of turbines in the study. On the other hand, the same study conducted at a site that has an expected rate of 30–50 fatalities per turbine (i.e., an expected average detection rate of 3–5 per turbine) might produce enough carcasses with fewer replicate turbines. For example, many research questions have yet to be answered regarding the impacts of wind turbines on Indiana bats (*Myotis sodalis*). Because fatalities of Indiana bats are considered rare events, it would be unreasonable to use fatalities of this species of bat as a direct measure of the effectiveness of a proposed impact-reduction strategy.

The decision trees that follow are intended to help identify a path to:

- Assess the feasibility of testing an impact-reduction strategy by determining if there is a reasonable chance of detecting a treatment effect
- Determine the steps necessary to identify a treatment effect and, if applicable, the behavioral response of the species to the impact-reduction strategy being tested.

The decision trees are designed to determine if an impact-reduction strategy is feasible. Measuring fatalities is the most appropriate method to assess the effectiveness of a proposed impact-reduction strategy. If fatality events are too rare (e.g., few fatalities of golden eagles [*Aquila chrysaetos*], Indiana bats, or Hawaiian hoary bats [*Lasiurus cinereus semotus*]) to detect an immediate treatment effect, then alternative metrics or fatality events of a broader taxonomic group may be needed. However, using alternative metrics or broader taxonomic groups may

impinge on the comparability of the results to other studies unless the other studies use the same metrics and taxonomic groups.

Alternative metrics may include measuring passage rates or avoidance behaviors that may correlate with fatalities. Broadening the taxa under study is another potential option to increase a study's feasibility. An appropriate broadening of a taxonomic group might be from golden eagles to large raptors (e.g., eagles and Buteos) that have similar behavioral patterns and similar use of the wind facility habitat, or from Indiana bats to other bats that overlap in ecomorphology; however, this approach will likely vary depending on the impact-reduction strategy being tested, and it may require additional experiments to determine the appropriateness of the larger taxonomic group. When possible, it may be appropriate to confer with other investigators to agree on alternative metrics, and then include those metrics in the study design. The justification for using an alternative metric should be clearly articulated.

In situations when power analysis indicates that detecting treatment effects will be unlikely when based on a fatality metric, and when there are no appropriate expansions of the taxa and alternative metrics are unavailable or inappropriate, it may be infeasible to test an impact-reduction strategy at a wind energy project. These situations would likely require much more robust experimental design incorporating multiple projects throughout an extended period of time; however, an experiment that spans multiple projects would require extensive coordination.

Figure 1 shows the steps needed to determine if field-testing a proposed impact-reduction strategy is feasible. If conducting a field test is feasible, Figure 2 provides study design considerations.

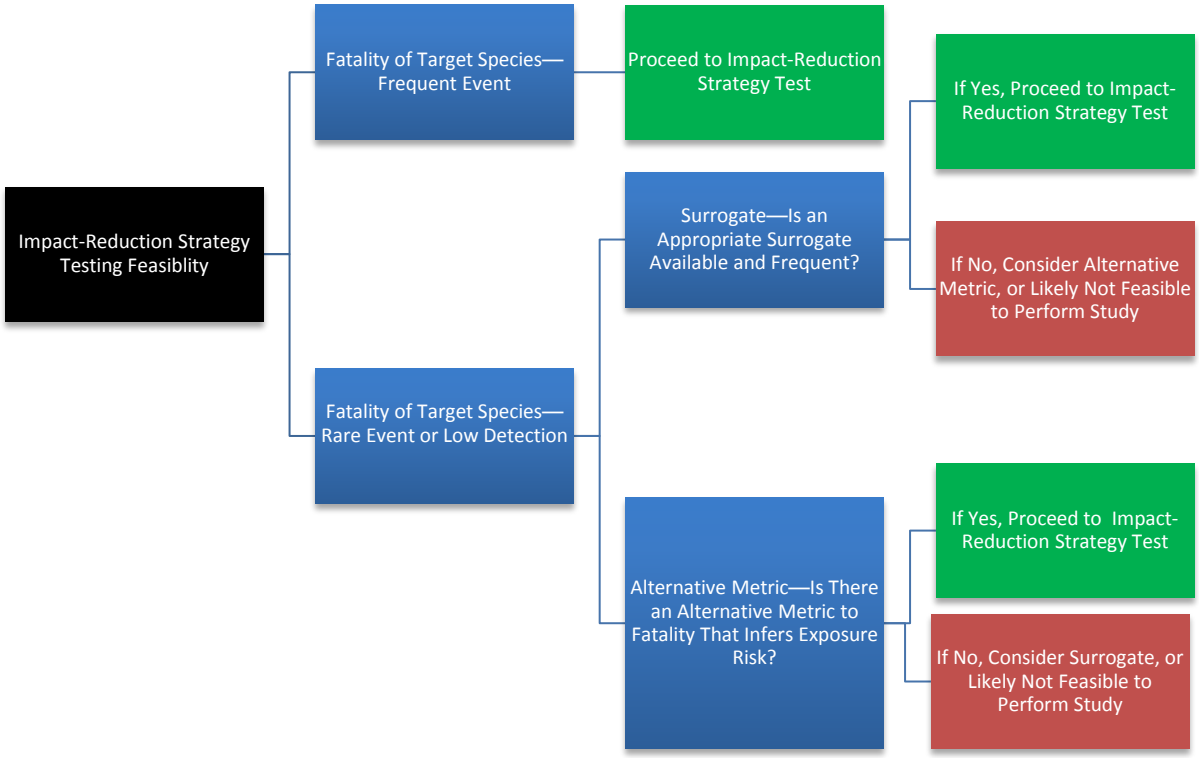

Figure 1. Steps to determine the feasibility of conducting an impact-reduction strategy test as determined by frequency of fatality event or type of fatality metrics

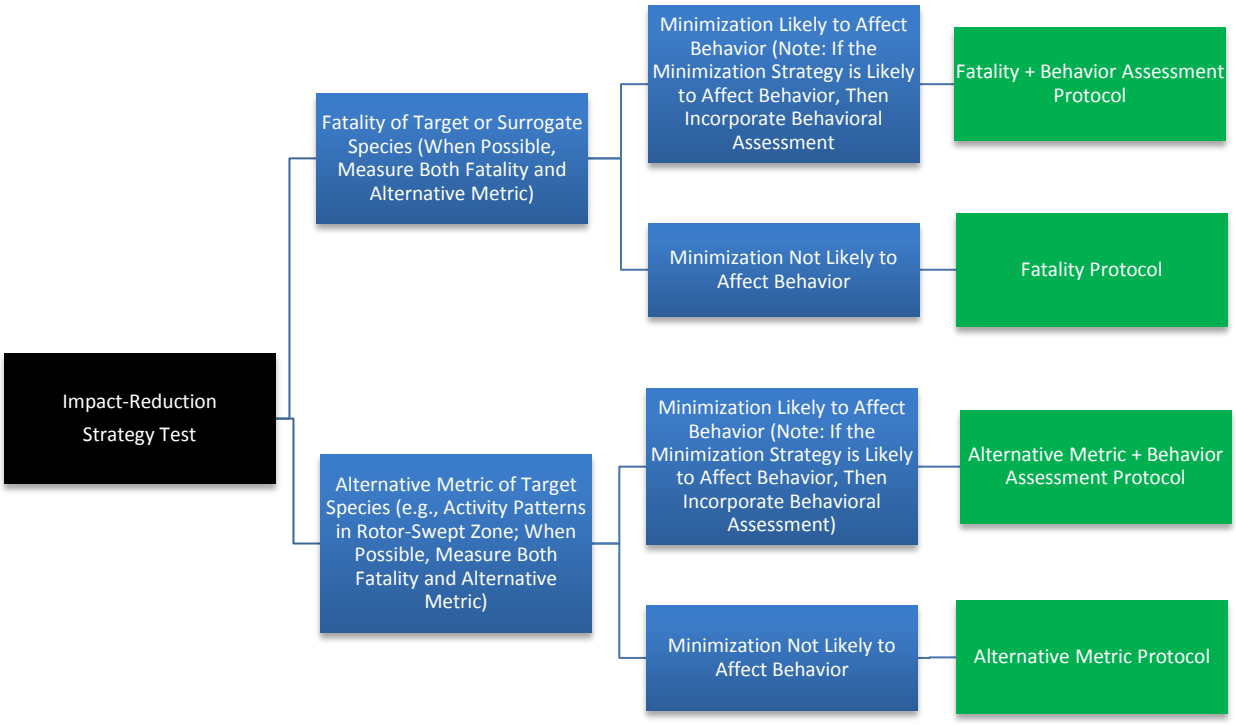

Figure 2. Study design considerations for preparing an impact-reduction strategy test as determined by anticipated frequency of events

2.2 Experimental Design

The ideal experiment has adequate power to (almost always) reject the null hypothesis when it is false and to (almost always) fail to reject it when it is true. It is unlikely that the results of a single experiment will be definitive; thus, experiments typically need to be repeated and the results compared for consistency or to reveal insightful differences. Results of every experiment are most useful when they are simple, decisive, and convincing.

To ensure that an impact-reduction strategy is convincingly tested, to ensure comparability to other test results involving the same or similar impact-reduction strategy and based on the same metrics, and to provide enough information for future comparisons, principles of experimental design should be heeded along with elements of experimental design and field-study methods that help isolate the treatment effect. For a more detailed discussion of study design considerations, see Strickland et al. (2011). For specific examples of study designs used to test impact-reduction strategies on bats, see Arnett et al. (2011, 2013a, 2013b). Additionally, see the appendix of this document for an example of a potential study design to investigate whether a particular deterrent treatment will reduce bat fatalities at a wind energy project. In some situations, the preferred experimental design for testing impact-reduction strategies will be a before-after, control-impact design.

2.3 Experimental Design Principles

Multiple steps are involved in developing an experimental design. The first is to clearly articulate the research question, and the second is to consider a suite of experimental design elements. These experimental design elements can be organized around a few principles—namely, conducting the study at appropriate spatial and temporal scales, using experimental controls, and replicating and interspersing treatments.

2.3.1 Clearly Articulate the Research Question

The research question is central to the experimental design, and it is needed to develop hypotheses, define variables, and ensure that the field-testing methods are the most appropriate for the study. When considering an impact-reduction strategy, an example research question might be:

> At study site X, is the fatality rate of species Y lower at turbines treated with strategy Z than at wind turbines without Z?

The Zth strategy could consist of curtailment, a deterrent, a detect-and-deter protocol, laser lights, a blade-painting scheme, or any number of strategies that modify the structure or appearance of the turbine or that work in tandem or independently of the structures. The leading phrase of the question—"At study site X"—helps define the scope of inference. The question also identifies the treatments, which will consist of applying strategy Z to a subset of turbines and withholding strategy Z from other turbines as the control treatment. The question also identifies the experimental units—i.e., the units of inference—and the response variable, which is the fatality rate of the Yth species. The question is only a starting point, however, because statistical methods for testing hypotheses need to be formulated along with desired effect sizes and statistical power. As these latter details are developed, so too will the sample sizes, number of replications, and levels of interspersing the treatments that will be needed to estimate the variance among treatments and prevent confounding results caused by gradient effects across the study site (which could be a project, a portion of a project, or even multiple projects within a wind resource area).

After formulating some or all of the experimental design to test whether the Zth strategy reduces fatality rates of species Y, it might be determined that the fatalities of species Y will not be detected often enough within the budget and time frame for achieving the desired statistical power. If this is the case, then it might warrant replacing species Y with a broader taxonomic group, such as all raptors or all bats. Another option might be to ask a different research question that focuses on an alternative response variable for which sufficient numbers of events could be obtained. An example of an alternative research question might be:

> At study site X, do bats fly less often through the rotor zone of wind turbines treated with strategy Z than the rotor zone of wind turbines without Z?

This alternative response variable might suffice as long as a relationship can be established between passage rates and fatality rates. Even if no such relationship can be established, it might sometimes be reasonable to assume the relationship. For example, if bats no longer passed through the rotor, then it might be reasonable to assume that bats would no longer be killed by the wind turbines treated with strategy Z. It is more likely, however, that a much-reduced passage rate will be the best that can be achieved, and this decrease will not necessarily result in

a reduced fatality rate. A reduced passage rate could translate into no change in the fatality rate if the passages that continue also involve behaviors or environmental conditions that result in fatalities. Whether the experimental outcome of an indirect response variable is convincing will be conditional on an accurate assumption of a relationship or reliance on an established relationship between the alternative response variable (i.e., metric) and fatality rates.

2.3.2 Determine the Appropriate Scale

The spatial and temporal scales of the study must suit the response variable and the intended scope of inference. For example, if the response variable is fatalities per megawatt (MW) per year, then the temporal scale of the study should span at least one full year or the relevant seasons. The temporal scale could include multiple years or seasons to cover interannual variations in the response variable. The spatial scale should also start with the space that is appropriate for the response variable of a species of interest. For some species of bats that might experience 30+ fatalities per wind turbine per year at a study site, an appropriate scale might involve a portion of a wind project; whereas for golden eagles, which might experience 1 fatality per 20 turbines per year at a particular study site, an appropriate scale might involve an entire wind project, all of the wind projects within a wind resource area, or an even larger scale than that. After first considering the appropriate spatial scale for the species and response variable, the spatial scale must also be sufficient to accommodate adequately replicating and interspersing the treatments, including allowing sufficient buffer spacing between experimental units to minimize contamination that may be introduced by fatalities resulting at one turbine in a treatment group being found and associated with another turbine in a different treatment group. The temporal scale must also be sufficient to ensure that any temporal phasing of treatments, such as in a crossover design, are sufficiently separated in time to prevent a fatality caused during one treatment from being attributed to another treatment applied in a subsequent phase. The temporal scale should also be sufficient for each phase of a crossover design to last long enough to adequately represent the response variable.

Clearly defining the spatial and temporal scales of the study will determine the extent, duration, and intensity (e.g., search interval, inter-transect spacing, and number of personnel) needed to answer the research question. It is helpful to know as much as possible about the study objectives to ensure that the appropriate scale is incorporated. The study should provide spatially replicated data within and among different landscapes. For example, for bats and eagles the characteristics of the sites selected for study are an important part of the experimental design. In addition, wildlife can habituate to deterrents, especially when the deterring stimulus is neutral and does not cause or warn of discomfort or disorientation. Thus, the study protocol should provide for temporal replication to sufficiently test the effectiveness of the deterrent against the possibility of habituation. When determining the scale of the study, consider the following:

- Can the research question be addressed at the turbine, project, or regional scale?
- Does the proposed project site have adequate data to inform the sampling period?
- Does the study require year-round or multiple periods of monitoring?
- Is daily fatality monitoring, nightly video, or radar monitoring required?
- Is monitoring required during adverse weather (e.g., snowstorms)?

The duration of monitoring that represents each treatment period should at least minimize the effects of confounding that can be caused by seasonal variation in activity levels and behaviors or by trends in fatality rates that follow multiannual population cycles.

Given the fatality detection rate anticipated with the fatality search methodology, the number of turbines searched and the monitoring duration needs to be established to obtain a sufficient number of event detections (e.g., fatalities).

2.3.3 Include Control Treatments

At least one of the experimental treatments needs to be a control treatment against which the effects of the manipulated treatment are compared. In a wind energy project, wind turbines serving as controls will be those in which the impact-reduction strategy is not applied but where the response variable is measured using the same methods employed at the wind turbines where the impact-reduction strategy is applied. For example, an acoustic deterrent installed on a subset of wind turbines might be monitored for fatalities in the same manner as a subset of wind turbines used in the control treatment where the acoustic deterrent was not installed.

In another example, a detect-and-deter strategy for reducing golden eagle fatalities would likely involve a deterrent device that would reach multiple wind turbines. Wind turbines serving as controls might need to be located so far away from the deterrent device that they are on the opposite end of a wind project, or perhaps in an entirely different yet representative project. The spatial scale needed to support an experimental test of a detect-and-deter system may need to be larger than many single wind projects could support because it is likely that the detect-and-deter system would be implemented project-wide. The same problem would likely apply to detect-and-curtail strategies. For these types of strategies, wind projects might need to serve as either manipulated or control sites and be incorporated into a larger experiment involving many wind projects.

2.3.4 Replicate and Intersperse Treatments

Experimental treatments, including the control treatment, need to be replicated so that the variances can be estimated for the response variable measured under each treatment. Measures of variance are needed to determine whether the mean values of the response variable differ significantly among treatments. Also, to ensure that the variances measured for each treatment apply to the same scope of inference, the replicates of each treatment must be interspersed across the study area and measured simultaneously (i.e., during the same study period). The treatments require simultaneous measurements to ensure that the temporal scope of inference does not differ from one treatment to another. For example, measuring the control treatment in one year and measuring the manipulative treatment the next year would not be appropriate. When measured in separate years, the scope of inference between the control and manipulative treatments would be different.

Treatments need to be interspersed across the study area and examined during the same time period to prevent confounding caused by gradient effects across space or time. Randomization is the most common method for interspersing treatments, but it can result in an arrangement of experimental treatments that is far from interspersed when the sample size of the experimental units is small (Hurlbert 1984), and this can be a problem in wind projects. For example, if an experiment will involve only 10 wind turbines, then randomization could result in 5 manipulated

wind turbines clumped together on one side of the study area and the 5 control turbines clumped together on the other side of the study area. Even if the arrangement of treatments was randomized, the experiment would still be vulnerable to pseudoreplication because the 5 control turbines might be on the less windy side of the project, the 5 manipulated turbines might be on the windier side, and wind might matter to the outcome of the experiment.

Another approach for achieving interspersion is systematically assigning experimental units to treatment groups, starting from a random selection. For example, if an experiment will involve two treatments—the impact-reduction strategy and controls—assigned to 64 wind turbines, and if it was decided that the appropriate buffering to minimize contamination effects will require applying each treatment to groups of 4 adjacent turbines, then the 64 turbines could be divided into 16 groups of 4, and the groups could be numbered sequentially. A random selection from among these 16 groups might be Group 9, so Group 9 would receive the impact-reduction strategy, Group 10 would serve as the control, Group 11 would also receive the impact-reduction strategy, and so on. Note, however, that if a treatment is assigned to a group of 4 turbines, then there would be only 16 experimental units in the study—i.e., 15 total degrees of freedom not 63. Grouping decisions should be made with full awareness of the implications to subsequent analysis and inference.

Another similar approach for interspersing two treatments would be to divide the turbines or turbine plots into nearest-neighbor pairs or blocks and randomly assign the control treatment to one member of each pair or to one plot, leaving the other member of each pair or plot to serve in the impact-reduction strategy. However, if neighboring turbines behave much more similarly than any two turbines chosen at random, this pairing may serve as an effective variance-reduction method and would be worth the cost in degrees of freedom. Different designs account for known variations (e.g., blocking designs, stratified random samples), but no matter which approach is used to suitably intersperse treatments, it is important to leave a sufficient buffer between treatments, which may include controls, to minimize or avoid contamination.

2.4 Experimental Design Elements

2.4.1 Scope of Inference

"The population to be sampled (the *sampled* population) should coincide with the population about which information is wanted (the *target* population)" (Cochran 1977), which in the case of wind turbine impacts might be a particular species encountering a certain type of wind turbine within a specific region. The ability to infer that effects observed in a designed experiment apply to other sites or turbines at other times (i.e., the scope of inference) depends on how representative the sampled site is relative to the population of interest. Judgments about the differences among sampled sites and other wind facilities are needed to determine the scope of inference.

It is common to extend the conclusions beyond a specific study to an unstudied area or period of time; however, it is important to articulate any assumptions and state clearly how the extrapolation was based on experimental design principles versus expert opinion and rhetorical argument. A single study is rarely, if ever, enough to provide unequivocal evidence of an effect (Abelson 1995). And although some studies are much more influential than others because of their design, scope, and results—and they may be considered landmark studies—achieving the

goal of an ideal experiment may not always be possible. Inferences that extend beyond a specific study might become valid if enough independent, carefully designed studies at different wind facilities identify similar effects. The ability of the fatality rate or other response variable to represent the target population may be even more important than the standardization of the field method. To represent a target population, it is important to avoid altering the collision risk for any reason other than the fatality-reduction strategy. For example, picking up and removing found carcasses at monitored wind turbines might prevent certain bird species from approaching to scavenge on the carcasses. Additionally, mowing to improve ground visibility and carcass detection might alter the way certain bird and bat species use the study area. In addition, mowing could increase scavenging if access to carcasses is easier for scavengers as well.

2.4.2 Unit of Inference

The unit of inference is the unit about which the inference regarding the efficacy of a treatment is made. The unit of inference determines the level at which the analysis is conducted (Klar and Donner 2007). The unit of inference and the experimental unit are often, but not always, the same. For example, if a bat-deterrent treatment is expected to affect fatality rates at an individual turbine and the treatment can be independently applied to an individual turbine, then the unit of inference is the turbine. Alternatively, if a deterrent can act only on a site as a whole and can be applied only to an entire site, then the experimental unit and the unit of inference is the site.

2.4.3 Response Variable

The response variable is what is being measured and how it is measured. It is the number of events that are biologically meaningful, such as fatalities, passages through or by a rotor, or changes in behavior that bear on collision risk. Events are often expressed in the context of the unit of inference (e.g., wind turbines) and temporal scale (e.g., years) to arrive at rates of events, or metrics. Examples of possible response metrics by different units include:

- Number of fatalities per turbine per year
- Number of fatalities per megawatt-hour
- Number of fatalities per megawatt per year
- Number of fatalities per rotor-swept area
- Number of flight paths per hour (passage rates)
- Number of bat calls per hour
- Proportion of eagles or bats that changed direction in response to a deterrent.

Note that megawatts and rotor-swept area both typically express the size of a wind turbine. If the turbines used in a study differ in size, then a size bias can be introduced if the metric is the number of fatalities per turbine. Similarly, if the size of the wind turbines used in an experiment varies, then passage rates might be biased because larger turbines will present greater rotor areas for passage rates to be counted. Care is needed when choosing the response variable and understanding its potential biases and sources of uncertainty. If a unique or rarely used metric is employed in a particular study, reporting sufficient information would help improve the comparability of the study's results to those of other studies and enable other investigators to calculate the more widely used metrics. For example, a study relying on fatalities per turbine per

year ought to report the rotor-swept area so that other investigators can convert the results to fatalities per rotor-swept area.

Important steps in deciding on one or more metrics include defining potential sources of measurement error and bias. In the case of fatality rates, sources of error can include the fatalities not found because of searchers missing available carcasses, carcasses falling outside the searchable area, or carcasses being removed from the search area before the next search. Sources of bias can also be introduced, such as placing large bird carcasses in searcher detection trials to measure the detection rate attributed to small birds, or deriving mean days to carcass removal from a trial that lasts twice as long as the average interval between searches. How these biases and sources of error are addressed is critical.

Another consideration when selecting a metric is the degree to which the response variable will vary as a result of the biological events compared to the other components of the metric. For example, the variation in a metric, such as the number of fatalities per megawatt-hour, might be dominated by variations in megawatt-hours, especially when the biological events are relatively rare and occur during a wide range of megawatt-hours. If the risk of collision is influenced more by the existence of the wind turbine structure and less by the wind turbine's moving parts, then the biological events (i.e., collisions, in this case) would be nearly independent of megawatt-hours, and the variations in the metric (i.e., fatalities per megawatt-hour) would be caused by variations in megawatt-hours rather than the event of interest.

A similar problem arises in the temporal basis of a rate metric, such as fatalities per megawatt per year. If the experimental units were measured during different periods of time and the events of interest varied because of season or interannual trends, then the temporal basis of the metric could introduce bias. Converting the response variable to a rate metric does not necessarily standardize the response variable. The most reliable response variables are the simplest, unless one is certain that the denominators in a rate metric have no influence on the probability of events occurring or being detected.

Response variables that are influenced by field methods can also introduce biases and large uncertainties (Smallwood 2007; Smallwood et al. 2013). Experiments requiring extra studies or trials to adjust response variables for missed detections will be more prone to bias and large uncertainties because each adjustment factor introduces variance and one or more of them can introduce bias (Korner-Nievergelt et al. 2011). Each additional trial that accompanies an experiment for the purpose of adjusting the response variable also introduces a source of variability in methods among studies, which thereby impinges on the comparability of the results. For example, carcass persistence trials have varied greatly in species used, carcass condition, number and schedule of carcass placements, duration, on-site versus off-site placements, and whether the best persistence distribution was fit to the data (Bispo et al. 2013). Experimental tests of impact-reduction strategies will be more comparable among studies when response variables are standardized and their adjustments are both minimized and standardized. Korner-Nievergelt et al. (2011) demonstrated that different methods have different bias depending on the conditions at a site, which makes it difficult to standardize these methods. Currently, a "universal estimator" is not available, and until it is, projects should consider following the recommendations of Korner-Nievergelt et al. (2011), which suggest selecting the most appropriate methods given site-specific conditions. At minimum, researchers should collect

and make available enough information so comparable results can be generated, particularly if researchers are implementing novel methods. This would also allow for reanalysis if a universal estimator is developed, which is something that Korner-Nievergelt et al. (2011) suggested might be possible.

2.4.4 Experimental Unit

In general, the experimental unit is the unit to which a treatment can be independently (randomly) assigned. An experimental unit can be, but is not limited to, a turbine, set of turbines, a facility (or facilities), or a set of days or nights. Important considerations associated with experimental units include controlling or minimizing variability, randomly allocating treatments to experimental units, and allocating treatments to multiple experimental units.

2.4.5 Controlling Variation

A number of methods can control or minimize variation. The appropriate use of control, randomization, replication, and interspersion of treatments all address sources of variation. Additional advanced methods should be considered, including blocking, stratification, analysis of covariance, hierarchical linear models, and others.

2.4.6 Treatments

Treatment is a general term that refers to a condition or an action applied independently to an experimental unit, and it includes the control (nothing done) as well as the action applied. For example, a treatment could be changing turbine operations (e.g., feathering), and the control would be making no changes to turbine operations.

2.4.7 Sample Size

An adequate sample size is the number of experimental units needed to detect a meaningful effect size given the assumed (or estimated) magnitudes of sources of variation. When possible, consider adding experimental units (e.g., turbines and projects) to what was otherwise regarded as an adequate sample in the event that individual units within the original sample are no longer functional. Also, it may be beneficial to have replacement test devices (e.g., deterrents) available in the event that one or more of them malfunction.

2.4.8 Anticipated Effect Size

In the context of an experiment, the effect size is the magnitude of effect caused by an experimental treatment. The investigator needs to decide whether the anticipated effect size, which may not be known until the experiment is conducted and the results are analyzed, will be sufficient to warrant deploying the impact-reduction strategy. Questions to consider might be: Would the strategy be practical if the effect size were only 5%? Would 50% qualify as a practical effect size? Or, does the effect size need to be > 90% to justify the cost of implementation?

Anticipating the detectable effect size (i.e., the statistical significance of the effect size) can help guide the experimental design, including considerations such as where to locate the study, sample size, difference among treatment means, and variance for each treatment. Some of this information can come from a pilot study or the literature until the experiment is actually performed. For the purpose of testing an impact-reduction strategy, it might be worthwhile to

increase the likelihood of detecting an effect size by designing the experiment around the portions of a wind energy project that are known to have high activity, fatality, or risk as long as the inference is not projected onto the portions of the wind project where those factors are known to be lower.

2.4.9 Statistical Power of the Study

In simple terms, statistical power is the statistical likelihood that a study will detect a treatment effect given that a treatment effect is present. Power calculations based on the desired effect size, anticipated variance among experimental units, sample size, and preferred alpha and beta levels should be carried out to determine if the study will have adequate statistical power. Alternatively, given the desired statistical power and effect size, a sample size analysis can be conducted to determine the level of replication necessary to detect the preferred effect size.

3 Analysis and Characterization

This section describes prestudy action items as well as field-practice considerations for fatality events, alternative metrics, and behavioral studies.

3.1 Field Practices/Study Implementation

3.1.1 Prestudy Considerations

Meeting with wind energy operators is necessary when developing field practices to determine which strategies are feasible, ensure adherence to safety protocols, and meet permitting conditions of the site. It also provides an opportunity to articulate expectations and determine workload capacities of all partners. Some considerations for this part of the process are as follows.

Researchers should determine whether the study design would be affected by project restrictions, such as the number of turbines available for the study, plot size, accessibility, and so on. Restrictions might come in the form of existing wind project permits, such as noise and visual effects thresholds, or in expectations associated with power purchase agreements or grid compliance. Other restrictions might be related to permits required to implement the study, such as those required by the U.S. Fish and Wildlife Service and/or state fish and wildlife agencies. There might be mitigation agreements that limit site access or access by time of day. For example, there may be a seasonal restriction on driving on-site to minimize the risk of driving over endangered herpetofauna. There might be construction or turbine maintenance plans scheduled for the site that might interfere with the study, or rancher or farmer activities might interfere with or confound the study results. Examples can include scavenger or prey control, herbicide applications, disking, and hunting leases. Researchers should also determine whether experimental treatments require turbine-specific or project-wide changes in turbine operations. If operational changes are needed, then it should be established who is responsible for implementing the changes and how the implementation will be verified during the study and corrected if needed.

Researchers should establish whether the impact-reduction strategy being tested requires access to the turbines and, if applicable, who is responsible for installing, monitoring, and removing equipment. They should also determine if the study requires weather data or operational data, how these data will be gathered and standardized, and whether there are any restrictions or limitations regarding availability, use, or publication of these data.

3.1.2 Considerations When Measuring the Fatality Rate of the Target Species

3.1.2.1 Sampling Scheme

Prior to initiating the study, a rigorous experimental design needs to be developed (see Section 2.2). In addition to general design questions, researchers should determine whether the study period will be appropriate to meeting the study objectives and whether the plot size effectively balances carcass detection rates with the cost of searching. They should determine the intertransect separation distance and the implication of this distance on carcass detection. The use of dogs and professional dog handlers to achieve carcass detection goals and the appropriate search interval for the species of interest should also be considered. In addition, researchers

should consider whether and how to measure potential covariates, such as carcass size, ground visibility, vegetation, topography, precipitation regime, and season.

Researchers need to determine how detection trials should be implemented (more details below) and which fatality rate estimator would most effectively adjust fatality rates for the proportion of fatalities not found. Researchers should consider the consequences of incorrectly estimating time since death of the carcass and of misidentifying the species of decomposed or partial carcasses. They should take into account the consequences of including or excluding carcasses found incidentally to routine searches as well as carcasses discovered during routine searches but outside the maximum survey radius. Researchers should consider the effectiveness of performing "clearing searches" at the start of monitoring periods, and researchers should evaluate whether all available carcasses can truly be found during a single search. They should take into account the consequences of missed surveys as a result of weather, safety, or unexpected issues and whether periodic searches need to be equally spaced throughout the monitoring period. In addition, researchers should decide on the attributes to record for each carcass, such as age, sex, carcass condition, and hair and tissue samples, and they should evaluate the cost and informational value of each of those attributes.

Researchers should also consider the possible consequences of removing found carcasses, which has been a routine practice among monitoring studies. Some experts believe that removing carcasses might change the local scavenger ecology and therefore could impinge on the scope of inference of the experiment. Alternative approaches might have carcasses redistributed as they are found to still assess detection bias but avoid dramatically changing carcasses availability and therefore scavenger ecology. In addition, researchers need to be mindful of the agency requirements for submitting carcasses.

3.1.2.2 Detection Bias

Fatality surveys are complicated by the inability to detect all fatalities (i.e., probability of detection < 1) because of cost and logistics; therefore, to the greatest extent practicable, field practices should be implemented to balance carcass discovery with cost and logistical feasibility. Depending on the experimental design, quantitative adjustments are needed to estimate the proportion of carcasses that are not discovered by fatality searches and to account for differences in detection probabilities among the experimental units. The reasons for carcasses not being discovered can include remains having been deposited farther from the turbine than the searches are performed (beyond the maximum search radius; see Hull and Muir [2010], Smallwood [2013], Huso and Dalthorp [2014]); remains that are available to be found by searchers but are missed (i.e., searcher efficiency); remains removed by scavengers prior to the next carcass search (i.e., carcass persistence; see Bispo et a. 2013); and remains occurring within the search area but invisible to the searchers as a result of impenetrable vegetation, water, or hazardous terrain, although in this situation these areas should be considered unsearchable and accounted for based on the distribution of carcasses. In the context of an experiment, it is important to remember that adjustments to metrics such as fatality rates are needed to account for different detection probabilities among experimental units, but these adjustments can assist only with the comparisons of the metric and might not be possible in situations of large numbers of missed detections. For example, if fatality searches in an experiment were performed at 100% of the search areas at some wind turbines but only 5% at others, then adjustments needed to the others

would add much more uncertainty to the fatality rates, and thus comparisons might not be possible, which would affect the ability to determine a treatment effect.

Strategies for increasing the detection rates of available carcasses can include searching at a slower pace, decreasing the average time between searches, reducing the distance between transects, and increasing the maximum search radius, but each of these improvements will come with increases in labor costs, particularly if searcher fatigue is a concern. An additional strategy can include using trained dogs and dog handlers because this method has resulted in higher detection rates (Arnett 2006; Mathews et al. 2013). Searches using dogs will differ logistically and potentially in coverage from searches using humans. For example, the attention span of dogs is shorter, and their work time is optimal during the coolest portions of the day in arid environments. In addition, dog searches might impose other sources of bias, such as impacting scavenger ecology or higher carcasses discovery for older carcasses that have stronger odor. This would likely need to be balanced with treatment rotation, particularly if dogs are not trained to find "fresh" carcasses.

Another strategy can include restricting searches to along roads and pads where ground visibility is superior, or within areas of relatively lower vegetation, as long as the fraction of carcasses expected to land within the searched area is accounted for (Hull and Muir 2010; Smallwood 2013; Huso and Dalthorp 2014); however, carcasses on exposed ground are likely to be removed more quickly by scavengers. Therefore, searches on these exposed areas might be started earlier in the day to find carcasses before diurnal scavengers do, but this strategy would not mitigate the removal rates by nocturnal scavengers. In addition, differences in carcass distribution among control and treatment turbines should also be considered, particularly if the impact-reduction strategy might affect behavior or conditions when fatalities occur (i.e., higher or lower wind speeds or blade tip speeds). By not accounting for this potential bias and given the fact that turbines usually have gravel located around the base, researchers might conclude that an impact-reduction strategy is more effective than it actually is, or, in the worst case, fail to reject a false null hypothesis (i.e., Type II statistical error).

Another strategy would be to manage vegetation by mowing or other means, but this would need to be implemented uniformly at all wind turbines and all treatments and would increase costs. Another problem with this strategy is that managing vegetation to increase detection rates might also alter the local ecology of the scavenger community and thus impinge on the scope of inference. An experiment may not be appropriate where vegetation is tall within the prospective fatality search areas.

To achieve sample size objectives, it might also be necessary to search more wind turbines, if more wind turbines are available for inclusion in the experiment. But doing so would increase costs. Tools to explore trade-offs among search area (coverage), searcher efficiency, and search interval while targeting a particular probability of detection are available at from Dalthorp et al. (2014).

3.1.2.3 Detection Bias Trials

Estimating the proportion of fatalities not found during fatality monitoring requires implementing trials intended to simulate detection probabilities experienced by the searchers. Ideally, the same types of animals that compose the response variable should be volitionally and

periodically placed within the search areas during the fatality-monitoring period. The periodicity, number of carcasses, and locations of carcasses should simulate the patterns of carcass deposition as realistically as possible. Unfortunately, researchers usually lack prior knowledge of the spatial and temporal patterns of carcass deposition from wind turbines. Therefore, placements should be randomized to the degree practicable, including by location, by day between routine searches, and by time of day. Carcasses used should include no chemicals or artifacts that are harmful to scavengers, and they should have been frozen immediately after death or placed just after death to minimize the impact of decay on the attractiveness of the carcasses to scavengers. Human scent should also be minimized on the remains. Carcasses should be marked discreetly and placement locations should be carefully recorded by the trial administrator, and searchers should be blind to the placements.

3.1.3 Alternative Metric: Passage Rates

3.1.3.1 Prestudy Decisions

When fatality detections are likely to be too few or detectability too low to determine a treatment effect, alternative metrics of fatality risk might suffice in a test of an impact-reduction strategy. Examples of alternative metrics may include passage rates and dwell time of the target species within the rotor zone, where dwell time is the sum of the time the animal is vulnerable to collision and can be influenced by ground speed (flight speed plus or minus wind speed) and behavior. Passage rates and dwell time can be quantified using field observers, acoustic and video (e.g., thermal or near-infrared) equipment, or radar. Understanding and stating the assumptions and limitations is important when using an alternative metric as an indicator of fatality risk. Note that methods to assess behavior (e.g., video imagery) may be similar to methods used for alternative metrics of risk, but these observations would likely need additional analysis to determine behavioral changes rather than presence or absence (see Section 3.1.4). Prior to initiating the study, a rigorous experimental design needs to be developed (see Section 2.2). In addition to general design questions, researchers should reference previously published guidance documents (e.g., Kunz et al. [2007]; Strickland et al. [2011]) or specific studies that have used the same alternative metric.

The researchers should state their assumptions and anticipated limitations of using measurements of passage rates or dwell time as indicators of fatality risk. An example limitation of acoustic detector surveys may be the bias associated with species that do not echolocate or vocalize as often as other species. An example limitation of thermal image surveys might be dampened thermal signatures of owls caused by their feathers. An example limitation of use surveys might be low correlation between activity levels and fatality rates. There also might be limitations associated with identifying the target species and judging the location of individuals when relying on thermal imaging and identifying individuals when using radar. For example, thermal imaging probably cannot differentiate bat species except to the general size class (e.g., small, large). The researcher should consider the consequences of any of these limitations as well as of missing or excluding events or missing planned survey sessions as a result of weather, safety issues, or other unforeseen circumstances. The use of multiple tools may reduce bias associated with using a single methodology. For example, species identification utilizing thermal cameras may be difficult; however, when paired with acoustic detectors, it may become possible to identify to the species level. The consequences of repeat passes by the same individual should be considered. To help determine these consequences, researchers might ask: does it make any

difference to estimates of collision risk whether one individual passes through a rotor 10 times or 10 other individuals pass through the rotor once each?

The researchers should also consider the spatial scale needed to determine a treatment response. For example, if the field of view is too small to record a response to the treatment, then it will likely be difficult to determine a treatment effect. Moreover, given any detection biases caused by the sampling methods, the researchers need to determine whether a treatment effect can still be detected. Going into the study, the researchers should have some idea of the activity available to be measured, such as numbers present and during what time periods the activity levels will be measureable.

3.1.3.2 Detection Bias

Similar to fatality surveys, measurements of passage rates are likely complicated by the inability to detect all events (i.e., probability of detection < 1) and may be further complicated by cost and/or logistics. Detection bias should also incorporate the ability to identify the target species. Therefore, field practices should be implemented to balance target detection with cost and logistical feasibility.

Researchers should identify any detection biases associated with the sampling method used to measure dwell time or passage rates and any shifts in bias as a result of weather, moon phase, turbine location, and so on. They should identify the detection biases that most influence determinations of treatment effect, such as target size, acoustic emission rate, distance between the observer and the targets, angle of view affecting visibility, flight speed, or behavior. It should be known whether the technology used to measure activity might directly or indirectly affect dwell time or passage rates. For example, the visual spectrum of illumination might attract a prey species, such as insects. The presence of an observer might reduce eagle use of an area.

Going into the study, the researchers should consider the consequences of achieving low detection rates or target species identification rates. Field and analytical methods that might increase detection rates or improve target species identification could include positioning cameras or acoustic detectors to reduce background clutter or noise; changing the camera lens or type of microphone to increase the field of view or detection cone; increasing the number of megapixels per area sampled for improved resolution while maintaining field of view; increasing the frames per second for a greater sampling rate; using different sensor types such as thermal versus near-infrared cameras; adding supplemental technology, such as acoustic detectors to confirm target species or locations of individuals; and relying on software versus manual detection to increase efficiency and precision of postprocessing. The potential advantages of all of these methods need to be weighed against increases in data processing time and cost.

3.1.4 Alternative Metric: Behavioral Avoidance Rates

Effective impact-reduction strategies will change the behaviors of target species such that passage rates, dwell time, and fatalities decrease. An effective deterrent will cause bats or eagles to turn away from the rotor-swept area of a turbine, and this behavioral shift, also often referred to as avoidance (Band, Madders, and Whitfield 2005; Chamberlain et al. 2006; Smales et al. 2013), can be quantified as another alternative metric to fatalities or passage rates, or preferably as a supplemental metric to fatality rates. Avoidance rates can be related to fatality rates to help understand whether and how a fatality-reduction strategy works.

The origins of fatality-reduction strategies have been rooted in species' sensory perception and behavioral patterns (May et al. 2015). For example, Cryan et al. (2014) observed a high frequency of leeward turbine approaches by bats, suggesting that this may be an area to target the application of acoustic deterrents. Observational studies have quantified behavioral patterns to wind, terrain, and wind turbines of birds (Hoover and Morrison 2005; Barrios and Rodriguez 2004; Smallwood, Rugge, and Morrison 2009; Smallwood et al. 2009) and bats (Horn, Arnett, and Kunz 2008; Ahlen, Gaagoe, and Bach 2009; Cryan et al. 2014), and these observations have led to multiple candidate impact-reduction strategies, including carefully siting new wind turbines. Also, an effective detect-and-curtail system will accurately predict that a flying eagle (or other type of bird or bat) is going to approach one or more particular wind turbines or dwell within a wind project for a certain period of time and then curtail turbine operations appropriately. A curtailment algorithm will accurately predict peak bat passage rates based on wind speeds and other factors, and it will shut down wind turbines throughout the duration of those conditions. Whether changing a behavior or predicting and responding to a behavior, an impact-reduction strategy will be effective only if the relevant behaviors have been sufficiently quantified and understood through prior study or demonstrated to correlate strongly with fatality rates. Relevant behaviors can also include habituation, which might eventually defeat an impact-reduction strategy that initially looked promising.

This document focuses on the experimental testing of impact-reduction strategies; therefore, the remainder of this discussion focuses on behavioral monitoring as an alternative to fatality monitoring in experimental tests of impact-reduction strategies. As mentioned earlier, behavioral researchers should consider experimental design principles and associated design elements.

3.1.4.1 Prestudy Decisions

Researchers should state their assumptions and identify the limitations of behavior responses as indicators of fatality risk as well as any limitations of the methods used to measure and quantify behavioral responses. Consideration should be given to the appropriate times of day, seasonal coverage, weather, and overall behavior-monitoring effort needed to adequately measure rates of behavioral responses. Additionally, researchers should determine how to measure habituation and what time period is needed to detect it.

3.1.4.2 Bias Considerations

Many studies have been published on bias associated with behavioral studies (e.g., Altmann [1974]; Marsh and Hanlon [2007]; Burghardt [2012]). As mentioned earlier, behavioral studies should consider the metrics discussed in Section 3.1.3 to help determine and, if necessary, address detection bias associated with the sampling method. Behavioral studies are much more comprehensive and likely more variable than what is described in this document; therefore, below are general concepts on which to reflect with the understanding that much more detailed information is available and should be referenced when designing these types of experiments. When software is used to quantitatively process video data, it would be useful to store all input parameters used to classify a target to allow for future reviews by observers to further assess specific behaviors.

Researchers should be wary of observer bias, wherein systematic differences exist among observers in estimating distances or heights above ground, identifying species, characterizing behavior, or recording data. Observer bias can be reduced through frequent comparisons of

observations, such as species identifications, estimated height above ground, flight direction, the entirety of the flight path, flight behavior, and reactions to turbines or impact-reduction strategies. These communications can improve observer skills and standardize observations.

Additional ways to standardize observations are to clearly describe thresholds that define behavior responses and restrict recorded observations to what was seen rather than attempting to interpret the motivation behind the behavior. To prevent inattentiveness, survey sessions should be limited in duration and number per day. In a before-after, control-impact study, observers should be kept blind to the treatments during the before phase, but they probably cannot be kept blind during the after phase. The effects of observer bias might be reduced by interchanging observers at each station.

3.2 Standardizing and Preparing for Comparative Studies

A secondary goal of this framework is to facilitate the comparison of study results of impact-reduction strategies. Whether future comparisons will consist of meta-analysis or simple reviews, researchers should think about how their experimental results can or will be used when comparing results. Lessons can be drawn from recent efforts to compare fatality rates among wind projects (Loss, Will, and Marra 2013; Smallwood 2013; Zimmerling et al. 2013; Erickson et al. 2014; Johnson et al. 2016), understanding, of course, that these studies were not the same as randomized controlled tests of impact-reduction strategies. As researchers began comparing fatality rate estimates among wind projects during the studies mentioned above, metadata gaps in monitoring study reports led to some fatality estimates being omitted and others being summarized more crudely than originally reported. Researchers inconsistently measured and reported covariates of fatality rates, including average fatality search intervals, sizes, and condition of placed carcasses in persistence rates; ground visibility in both detection trials and routine monitoring; maximum search radius; wind turbine tower height; rotor diameter; turbine operation details (e.g., cut-in speed); and rated capacity. Standards of data collection and measurement inevitably change as more data are gathered in emerging scientific topic areas and as politics and legal decisions dictate, but the impacts of this change on future comparisons of study results can be mitigated by measuring and reporting on likely covariates and by providing appendices of the core data (e.g., fatality finds, placed trial carcasses, detailed detection trial outcomes, and fatality search dates).

4 Reporting

This section includes suggestions for conducting proposal and peer reviews and disseminating research data and results.

When developing the experimental design for a project, a Technical Advisory Committee (TACs) is often used to review the proposed level of effort, safety considerations, logistical constraints, statistical analysis, and dissemination process. This initial step has the advantage of coordinating the stakeholders (e.g., research organizations, industry partners, government agencies) to confirm that all involved parties understand the goals of the project and ensure that the work products are of professional quality and delivered in a timely manner; however, maintaining a TAC depends on the availability of experts, technicians, and other participants. There are also legal considerations when including federal agencies in a TAC. The use of a TAC should be thoroughly researched and consider the local situation.

An essential component to any scientifically credible research is the peer-review process. Peer reviews are intended to assess the scientific validity of a project and provide constructive criticism to improve the end product. Reviewers can indicate whether the research question, methods, and experimental design are articulated clearly and if the analysis is of appropriate rigor. If the TAC is qualified and formed early enough, it might provide this type of review at the outset; if not, then a special peer review could be solicited from qualified scientists to ensure that the experimental design and other study elements are sound. Ultimately, the study should be published in a peer-reviewed scientific journal.

A critical, but often overlooked, aspect to wind energy and wildlife research is disseminating the data and results to other researchers and the public. This dissemination can be done by making reports publicly available, presenting at meetings, and/or publishing data and results in a scientific journal. This key step is necessary to advancing the overall understanding of the issue and reducing unwarranted replication of research projects. The intent to disseminate research data and findings should be discussed during initial meetings and contract negotiations so that all parties are in agreement about which information will be made public and which information will remain confidential (e.g., raw operational or weather data provided by an industry partner).

5 Discussion

The primary goal of this work is to produce decisive, convincing results. A secondary goal is to conduct each study using the framework and thereby facilitate repeatability as well as comparability among sites that are testing the same or similar impact-reduction strategies. The data collected during the studies should be analyzed and the results should be presented in a way that maximizes their comparability. For example, an experimental test of an eagle deterrent should generate metrics that can be measured from data collected at any existing wind project where the deterrent might be installed experimentally. Example metrics might include the number of eagles passing through wind turbine rotor planes per hour or the proportion of approaching eagles that turned away within a specified distance range from the turbine. Likewise, for bats, it is also expected that the data will help calculate one or more metrics, such as data describing bat activity and/or fatalities before and during testing.

During the reporting stage, the following issues should be discussed:

- Performance of the impact-reduction strategy, including:
    - Reliability of the equipment when tested under operational conditions
    - Ease of use of the equipment, including the logistics of deployment and operations
    - Costs associated with manufacturing, installing, and maintaining the equipment
    - Cost-effectiveness of the equipment relative to other impact-reduction strategies or operational minimizations.
- Remaining uncertainties and potential biases
- Knowledge gained that will help improve the technology.

A successful study should demonstrate whether the tested fatality-reduction measure was effective. In the event it was not, the study should provide likely reasons for the lack of demonstrated effectiveness. These reasons should rarely include inappropriate metrics, insufficient detection of events, or inadequate sample size because these should have been considered in depth prior to implementing the study. Lack of demonstrated effectiveness should be reasonably interpretable as evidence of inadequacies of the technology.

Thought should be given to the methods researchers use to identify and assess the presence of potential secondary or downstream effects. Some measures, such as active deterrent devices, may have secondary effects (e.g., exclusion from a previously utilized habitat). Although the benefit of these impact-reduction measures may outweigh such secondary effects, considering and evaluating those effects is important when making cost-benefit evaluations.

The study should provide spatially replicated data within and among different landscapes. Consequently, for bats and eagles the characteristics of the sites selected for study are an important part of the experimental design. Because eagle and bat fatalities at some wind facilities may be relatively rare events, the site selected to test the impact-reduction strategy should be expected to have a fatality rate related to the species of interest that is sufficient for detecting an effect.

Wildlife can habituate to deterrents, especially when the deterring stimulus is neutral and does not cause or warn of discomfort or disorientation. Thus, the study protocol should provide for temporal replication to sufficiently test the effectiveness of the deterrent against the possibility of habituation.

Glossary

Analysis of covariance	A measure of the linear association between two variables (i.e., how much a change in one variable is linearly associated with a change in another variable); analysis of covariance is a general linear model that blends the analysis of variance and regression
Analysis of variance	A collection of statistical models used to analyze the differences among group means and their associated procedures (such as variation among and between groups)
Before-after, control-impact	A design for impact assessment wherein the potential ecological impacts are addressed by collecting data in a control and impact zone both before and after a potential impact begins
Blocking	A source of variability that is not of primary interest to the researcher and is not expected to influence the effect of treatments; an example of a blocking factor might be turbine height—e.g., by blocking the turbine height, the source of variability is controlled for, which thus leads to greater precision
Completely randomized design	A design for studying the effects of one primary factor without the need to take other nuisance variables into account; the experiment compares the values of a response variable based on the different levels of that primary factor; for completely randomized designs, the levels of the primary factor are randomly assigned to the experimental units
Confounding variables	A variable other than the independent or explanatory variable of interest that may affect the response or dependent variable
Crossover design	A repeated measurement design such that each experimental unit receives different treatments during varying time periods
Cut-in speed	The wind speed needed to begin generating electricity to the grid; note that some curtailment methods may cause turbines to generate electricity at a higher cut-in speed, resulting in a loss of energy production, which is sometimes referred to as “raising the cut-in speed”
Detection probability	The probability of observing a carcass killed at a wind facility; the detection probability is < 1 because some carcasses land in unsearched areas, some are removed by scavengers, and some are missed in the search process

Deterrent	Something (e.g., sound, noise, light) that discourages, restrains, or causes a shift in an act, proceeding, or behavior; these strategies often produce a stimulus that can be neutral, discomforting, or disorienting (note that wildlife can habituate to deterrents, especially when the deterring stimulus is neutral and does not cause or warn of discomfort or disorientation)
Elements	An item for which some measurement is made, such as an animal, roost site, snag, or other item of interest; note that the definition of *element* is different when it is used within the term *experimental design elements*
Fatality	An individual event, occurrence, or instance resulting in death; a tendency to result in death
Fatality rate	The number of fatalities in relation to a specific temporal scale (e.g., number of fatalities per year or season)
Feathering or feathered	Adjusting the angle of the rotor blade parallel to the wind, or turning the whole unit out of the wind, to slow or stop blade rotation; normally operating turbine blades are angled perpendicular to the wind at all times
Free-wheeling	Blades that are allowed to slowly rotate even when fully feathered and parallel to the wind; in contrast, blades can be locked and cannot rotate, which is a mandatory situation when turbines are being accessed by operations personnel
Habituation	Diminishing response to a stimulus, such as that produced by a determent after repeated exposure
Independent variable	A variable (e.g., treatment) that may cause a change in the response variable (see *response variable*); also referred to as a predictor, explanatory, or exposure variable
Mortality	The death rate; the ratio of the total number of deaths to the total population or the ratio of deaths in an area to the population in that area
Observer bias	Systematic differences among observers in estimating and recording data
Response variable	A variable (e.g., fatality, passage rate) used to measure the potential influence or effect caused by the independent variable (see *independent variable*); also referred to as an outcome or dependent variable

Raising the cut-in speed	The turbine's computer system (referred to as the supervisory control and data acquisition, or SCADA, system) is programmed to a cut-in speed higher than the manufacturer's set speed, and turbines are programmed to stay feathered at 90° until the increased cut-in speed is reached during some average number of minutes (usually 5–10 min), thus triggering the turbine blades to pitch back "into the wind" and begin to spin normally
Randomized block design	A design wherein the experimenter divides subjects into blocks such that the variability within the blocks is less than the variability among the blocks; subjects within each block are then randomly assigned to treatment conditions
Statistical power	The statistical likelihood that a randomly chosen sample, satisfying the model assumptions, will detect a difference of the specified type when the procedure is applied if the specified difference does indeed occur in the population being studied; in simple terms, statistical power is the statistical likelihood that a study will detect a treatment effect given that a treatment effect is present (the power is 1-beta [i.e., probability of a Type II error])
Type I error	A statistical error that results in the rejection of a true null hypothesis. The probability of a Type I error is the alpha or significance level (sometimes referred to as a false positive)
Type II error	A statistical error that results in the failure to reject a false null hypothesis; the probability of a Type II error is the beta (sometimes referred to as a false negative)
Wildlife operational curtailment	Stopping or greatly reducing a turbine's rotor rotation rate to eliminate wildlife fatalities; it can be implemented in a couple of different ways: 1) the turbine control system can be adjusted to change the blade pitch angle (e.g., pitching the blades out of the wind) so that the rotor blades are stationary or only rotate slowly due to wind variability, and 2) rotor rotation can be stopped by adjusting the control system to apply the parking brake to stop the rotor (there is a potential loss of power when this strategy is implemented at or above a turbine's cut-in speed)
Wildlife seasonal curtailment	Employing wildlife operational curtailment of wind turbine operation during one or more seasons of the year to eliminate fatalities

References

Abelson, R. 1995. *Statistics as Principled Argument*. Hillsdale, NJ: Laurence Erlebaum Associates.

Ahlen, I., H.J. Gaagoe, and L. Bach. 2009. "Behavior of Scandinavian Bats During Migration and Foraging at Sea." *Journal of Mammalogy* 90:1,318–1,323. Accessed October 15, 2015.

Altmann, J. 1974. "Observational Study of Behavior: Sampling Methods." *Behavior* 49(3):227–266. Accessed October 15, 2015.

Arnett, E.B. 2006. "A Preliminary Evaluation on the Use of Dogs to Recover Bat Fatalities at Wind Energy Facilities." *Wildlife Society Bulletin* 34:1,440–1,445. Accessed October 15, 2015.

Arnett, E.B., M. Schirmacher, M.M.P. Huso, and J.P. Hayes. 2011. "Altering Turbine Speed Reduces Bat Mortality at Wind-Energy Facilities." *Frontiers in Ecology and the Environment* 9:209–214. Accessed October 20, 2015. doi:10.1890/100103.

Arnett, E.B., G.D. Johnson, W.P. Erickson, and C.D. Hein. 2013a. A Synthesis of Operational Mitigation Studies to Reduce Bat Fatalities at Wind Energy Facilities in North America (Subcontract Report). Austin, TX: Bat Conservation International. http://www.batsandwind.org/pdf/Operational%20Mitigation%20Synthesis%20FINAL%20REPORT%20UPDATED.pdf

Arnett, E.B., C.D. Hein, M.R. Schirmacher, M.M.P. Huso, and J.M. Szewczak. 2013b. "Evaluating the Effectiveness of an Ultrasonic Acoustic Deterrent for Reducing Bat Fatalities at Wind Turbines." *PLOS One* 8(9). Accessed November 1, 2015. doi: 10.1371/journal.pone.0065794.

Band, W., M. Madders, and D.P. Whitfield. 2005. "Developing Field and Analytical Methods to Assess Avian Collision Risk at Wind Farms." In *Birds and Wind Power*, edited by M. De Lucas, G. Janss, and M. Ferrer. Barcelona, Spain: Lynx Edicions.

Barrios and Rodriguez. 2004. "Behavioural and Environmental Correlates of Soaring-Bird Mortality at On-Shore Wind Turbines." *Journal of Applied Ecology 41:* 72-81. doi: 10.1111/j.1365-2664.2004.00876.x.

Bispo, R., J. Bernardino, T.A. Marques, and D. Pestana. 2013. "Modeling Carcass Removal Time for Avian Mortality Assessment in Wind Farms Using Survival Analysis." *Environmental and Ecological Statistics* 20:147–165. doi: 10.1007/s10651-012-0212-5.

Burghardt, G.M. 2012. "A Behavioral Biology for the Future." *Ethology* 118 (3):222–225. Accessed October 15, 2015. doi: 10.1111/j.1439-0310.2012.02024.

Chamberlain, D.E., M.R. Rehfisch, A.D. Fox, M. Desholm, and S.J. Anthony. 2006. "The Effect of Avoidance Rates on Bird Mortality Predictions Made by Wind Turbine Collision Risk Models." *IBIS International Journal of Avian Science* 148:198–202. Accessed November 15, 2015.

Cochran, W.G. 1977. *Sampling Techniques*. 3rd edition. New York: Wiley & Sons.

Cryan, P.M., P.M. Gorresen, C.D. Hein, M.R. Schirmacher, R.H. Diehl, M.M. Huso, D.T.S. Hayman, P.D. Fricker, F.J. Bonaccorso, D.H. Johnson, K. Heist, and D.C. Dalton. 2014. "Behavior of Bats at Wind Turbines." *Proceedings of the National Academy of Sciences in the United States of America* 111(42): September.

Dalthorp, D.H., M.M.P. Huso, D. Dail, and J. Kenyon. 2014. *Evidence of Absence Software*. Corvallis, OR: United States Geological Survey. http://pubs.usgs.gov/ds/0881/.

Erickson, W.P., M.M. Wolfe, K.J. Bay, D.H. Johnson, and J.L. Gehring. 2014. "A Comprehensive Analysis of Small-Passerine Fatalities from Collision with Turbines at Wind Energy Facilities." *PLOS One* 9. Accessed November 15, 2015. doi:10.1371/journal.pone.0107491.

Hoover, S.L., and M.L. Morrison. 2005. "Behavior of Red-tailed Hawks in a Wind Turbine Development." *Journal of Wildlife Management* 69:150–159. Accessed October 15, 2015.

Horn, J.W., W.B. Arnett, and T.H. Kunz. 2008. "Behavioral Responses of Bats to Operating Wind Turbines." *Journal of Wildlife Management* 72:123–132; 2008. Accessed October 15, 2015.

Hull, C.L., and S. Muir. 2013. "Behavior and Turbine Avoidance Rates of Eagles at Two Wind Farms in Tasmania, Australia." *Wildlife Society Bulletin* 37:49–58.

Hurlbert, S.H. 1984. "Pseudoreplication and the Design of Ecological Field Experiments." *Ecological Monographs* 54:187–211. Accessed November 5, 2015. http://www.esajournals.org/doi/abs/10.2307/194266.

Huso, M.M.P., and D. Dalthorpe. 2013. "Accounting for Unsearched Areas in Estimating Wind Turbine-Caused Fatality." *Journal of Wildlife Management* 78:347–358.

Johnson, D.H., A.R. Loss, K.S. Smallwood, and W.P. Erickson. Forthcoming. "Avian Fatalities at Wind Energy Facilities in North America: A Comparison of Recent Approaches." *Human–Wildlife Interactions* 10(1).

Klar, N., and A. Donner. 2007. "Cluster Randomization." In *Wiley Encyclopedia of Clinical Trials*. New York: John Wiley & Sons.

Korner-Nievergelt, F., P. Korner-Nievergelt, O. Behr, I. Niermann, R. Brinkmann, and B. Hellriegel. 2011. "A New Method to Determine Bird and Bat Fatality at Wind Energy Turbines from Carcass Searches." *Wildlife Biology* 17:350–363. Accessed November 15, 2015.

Kunz, T.H., E.B. Arnett, B.M. Cooper, W.P. Erickson, R.P. Larkin, T. Mabee, M.L. Morrison, M.D. Strickland, and J.M. Szewczak. 2007. "Assessing Impacts of Wind-Energy Development on Nocturnally Active Birds and Bats: A Guidance Document." *Journal of Wildlife Management* 71:2,449–2,486.

Loss, S.R., T. Will, and P. Marra. 2013. "Estimates of Bird Collision Mortality at Wind Facilities in the Contiguous United States." *Biological Conservation* 168:201–209. Accessed November 15, 2015.

Mathews, F., M. Swindells, R. Goodhead, T.A. August, P. Hardman, D.M. Linton, and D.J. Hosken. 2013. "Effectiveness of Search Dogs Compared with Human Observers in Locating Bat Carcasses at Wind-Turbine Sites: A Blinded Randomized Trial." *Wildlife Society Bulletin* 37:34–40. Accessed November 15, 2015.

Marsh, D.M., and T.J. Hanlon. 2007. "Seeing What We Want to See: Confirmation Bias in Animal Behavior Research." *Ehology* 113(11):1,089–1,098. Accessed November 15, 2015. doi 10.1111/j.1439-0310.2007.01406.x.

May, R., O. Reitan, K. Bevanger, S.-H. Lorensten, and T. Nygård. 2015. "Mitigating Wind-Turbine Induced Avian Mortality: Sensory, Aerodynamic and Cognitive Constraints and Options." *Renewable and Sustainable Energy Reviews* 42:170–181. Accessed November 15, 2015.

Smales, I., S. Muir, C. Meredith, and R. Baird. 2013. "A Description of the Biosis Model to Assess Risk of Bird Collisions with Wind Turbines." *Wildlife Society Bulletin* 37:59–65. Accessed November 15, 2015.

Smallwood, K.S. 2007. "Estimating Wind Turbine-Caused Bird Mortality." *Journal of Wildlife Management* 71:2,781–2,791. Accessed November 15, 2015.

Smallwood, K.S. 2013. "Comparing Bird and Bat Fatality-Rate Estimates Among North American Wind-Energy Projects." *Wildlife Society Bulletin* 37:19–33. Online Supplemental Material. Accessed November 15, 2015.

Smallwood, K.S., L. Rugge, and M.L. Morrison. 2009. "Influence of Behavior on Bird Mortality in Wind Energy Developments." *Journal of Wildlife Management* 73:1,082–1,098. Accessed November 15, 2015.

Smallwood, K.S., L.A. Neher, D.A. Bell, J.E. DiDonato, B.R. Karas, S. Snyder, and S. Lopez. 2009. *Range Management Practices to Reduce Wind Turbine Impacts on Burrowing Owls and Other Raptors in the Altamont Pass Wind Resource Area, California* (Subcontract Report). Oakland, CA: East Bay Regional Park District. http://docs.wind-watch.org/CEC-500-2008-080.PDF.

Smallwood, K.S., D.A. Bell, B. Karas, and S.A. Snyder. 2013. "Response to Huso and Erickson's comments on novel scavenger removal trials." *Journal of Wildlife Management* 77:216–225. Accessed November 15, 2015.

Strickland, D., A. Edward, W. Erickson, D. Johnson, G. Johnson, M. Morrison, J. Shaffer, W. Warren-Hicks. 2011. *Comprehensive Guide to Studying Wind Energy/Wildlife Interactions* (Subcontract Report). Washington, D.C.: National Wind Coordinating Collaborative. https://nationalwind.org/wp-content/uploads/assets/publications/Comprehensive_Guide_to_Studying_Wind_Energy_Wildlife_Interactions_2011_Updated .pdf.

Zimmerling, J.R., A.C. Pomeroy, M.V. d'Entremont, and C.M. Francis. 2013. "Canadian Estimate of Bird Mortality Due to Collisions and Direct Habitat Loss Associated with Wind Turbine Developments." *Avian Conservation & Ecology* 8:10. Accessed November 15, 2015. doi: 5751/ACE-00609-080210.

Bibliography

Allison, T.D. 2012. *Eagles and Wind Energy: Identifying Research Priorities* (White Paper). Washington, D.C.: The American Wind Wildlife Institute.

American Wind Wildlife Institute. 2014. *Developing a Research Framework for Increasing Understanding of Interactions between Eagles and Wind Energy*. Washington, DC. http://awwi.org/wp-content/uploads/2014/01/AWWI-Eagle-Research-Framework_Final-01-23-14-2.pdf

Bernardino, J., R. Bispo, H. Costa, and M. Mascarenhas. 2013. "Estimating Bird and Bat Fatality at Wind Farms: A Practical Overview of Estimators, Their Assumptions and Limitations." *New Zealand Journal of Zoology* 40: 63–74. doi: 10.1080/03014223.2012.758155.

Gehring, J., P. Kerlinger, and A.M. Manville. 2009. "Communication Towers, Lights, and Birds: Successful Methods of Reducing the Frequency of Avian Collisions." *Ecological Applications* 19:505–514. http://mnfi.anr.msu.edu/reports/2009-27%20Ecol%20Applications%20article-J%20Gehring-communication%20towers.pdf.

Gehring, J., P. Kerlinger, and A.M. Manville II. 2011. "The Role of Tower Height and Guy Wires on Avian Collisions with Communication Towers." *Journal of Wildlife Management* 75:848–855. Accessed November 15, 2015. http://onlinelibrary.wiley.com/doi/10.1002/jwmg.99/full.

Hodos, W. 2003. *Minimization of Motion Smear: Reducing Avian Collisions with Wind Turbines, Period of Performance: July 12, 1999–August 31, 2002*. College Park, MD: University of Maryland.

Johnston, N.N., J.E. Bradley, A.C. Pomeroy, and K.A. Otter. 2013. "Flight Paths of Migrating Golden Eagles and the Risk Associated with Wind Energy Development in the Rocky Mountains." *Avian Conservation and Ecology* 8:12–28. Accessed November 15, 2015. doi: 5751/ACE-00608-080212.

U.S. Fish and Wildlife Service. 2013. *Eagle Conservation Plan Guidance: Module 1—Land-Based Wind Energy, Version 2*. http://www.fws.gov/ecological-services/es-library/pdfs/Eagle_Conservation_Guidance-Module%201.pdf.

U.S. Fish and Wildlife Service. 2012. *Land-Based Wind Energy Guidelines*. http://www.fws.gov/ecological-services/es-library/pdfs/WEG_final.pdf.

Appendix: Example Experimental Design for Field-Testing Deterrents Intended to Reduce Impacts on Bats at Wind Energy Facilities

5.1 Introduction

This appendix describes one example of a study design to investigate whether deterrent treatments will reduce bat fatalities at a wind energy facility. The design incorporates controls, randomization, replication, and the use of blocking to control for variability. In this scenario, there are 25 wind turbines at the site, but only 12 are available for the experimental study. The study will be conducted during a 75-night period. The experiment is described in general below, which was modeled after the experimental design covered in Section 2.2, Section 2.3, and Section 2.4.

5.2 Approach

Randomly select 12 turbines for the 75-night experiment. Use a randomized block design (RBD) with the turbine as the blocking factor and night-within-turbine as the sampling unit for treatment. Each night, randomly assign 4 turbines to each of the 3 treatment groups (i.e., Control, Treatment A, or Treatment B), but ensure that full balance (i.e., each turbine receives each treatment group an equal number of times) is achieved every 15 nights during the entire study period. During the course of 75 nights, each treatment will occur within each block (i.e., turbine) on 25 nights.

Search all 12 turbines on a daily basis to recover the maximum number of carcasses estimated to have died the night before. Daily searches improve accuracy in estimating time since death; hence, place the fatalities in the appropriate treatment period. Classifying the time of death becomes increasingly difficult as the carcass ages, and it is easier to attribute time of death to the previous night than to determine whether a carcass was killed 2 rather than 3 days prior. In addition, daily searches likely increase the number of carcasses available to be found, assuming carcass removal has a temporal component.

The total number of carcasses attributed to each treatment within each turbine will be the response variable, so a total of 36 observations will be used. The analysis will be carried out as a generalized mixed model, with carcass count modeled as a Poisson-distributed random variable (with potential for overdispersion), turbine as the random effect, and treatment as the fixed effect.

Deviations from this design are certainly acceptable and might be necessary because of logistical or cost constraints. For example, it might not be possible to rotate treatments among turbines, and the study design may consist of fixed treatments as in the completely randomized design (CRD) described in Section 5.5 (see Arnett et al. 2013b). In this case, additional turbines are likely needed to increase the sample size, and bias trials must be conducted. Changes to this study design must be articulated, and a rationale must be provided for how the alternative design will control for variability and possess enough power to address the research question. In addition, the researchers will need to show how the statistical model applied to the data will reflect the proposed design changes.

5.3 Experimental Design Principles

It is important to clearly articulate the research question. For example:

- When acoustic deterrents are activated, are bat fatalities < 60% of fatalities when no deterrents are activated?
- When acoustic deterrents are activated only from sunset to midnight, are bat fatalities < 60% of fatalities when no deterrents are activated?

5.3.1 Appropriate Scale

This is an experimental, not observational study. The scale of the study is at the turbine level during bat migration (i.e., inferences will be made to the effectiveness of deterrents implemented on a particular turbine during the fall migration period).

5.3.2 Use of Controls

A control treatment will comprise one of the three treatments. In an RBD example, control turbines will have deterrents but will not be operating.

5.3.3 Replication and Interspersion of Treatments

If implemented as planned, each treatment will be replicated three times during the 75-night study, and fully interspersed among the wind turbines randomly selected for the study 5 times during the 75-night study.

5.4 Experimental Design Elements

5.4.1 Scope of Inference

The scope of inference is related to areas wherein the complement of bat species at risk of being killed by turbines is similar to that of the study area and during the same period of time. Other factors such as the turbine model (e.g., turbine size, cut-in speed) might limit the scope of inference.

5.4.2 Unit of Inference

The unit of inference is a turbine. This is the level at which the treatment is applied and a response can be expected. Note that the placement and orientation of devices might vary depending on the turbine model, which should be considered when developing the experimental design.

5.4.3 Response Variable

The response variable is the number of bats found in searches immediately following the night on which the treatment was applied, or in rare cases the ability to relate carcass age to a particular turbine night.

5.4.4 Experimental Unit

In the CRD described below, the experimental unit is the turbine. It is the unit to which a treatment is applied. The sample size (number of replicates) is the number of turbines receiving each treatment.

In the RBD, the experimental unit is the set of nights to which a treatment was applied (e.g., 25 nights in the example above). The sample size (number of replicates) is the total number of turbines in the study because each will receive each treatment.

5.5 Controlling the Variation

Two appropriate designs for this research question are the CRD and RBD. The simplest experimental design for this research question with three proposed treatments would be a CRD. A total of 3**r* turbines would be randomly selected from the turbines at a site, and each of these turbines would be randomly assigned to one of the three treatments, resulting in a sample size of *r* replicates per treatment. This design accounts for many sources of variation (e.g., variation caused by weather, insect eruptions, and migration patterns) because treatments are randomly assigned to turbines and no set of turbines receiving a treatment is expected to be substantially different than any other with respect to these sources of variation. There is a small risk that the random assignment of turbines with the same treatment will, by chance, assign turbines that inherently kill fewer (or more) bats.

The advantages of the CRD are as follows:

- Deterrents need to be mounted on turbines that are assigned the deterrent treatments only.
- Because treatment assignment is static throughout the study period, searches can be carried out at an interval that is optimal relative to the scavenging rate; therefore, daily searches may not be necessary.
- All carcasses found can be included in the analysis, not only those that are "fresh" (likely killed the previous night).

The disadvantages of the CRD are as follows:

- It requires a larger number of turbines in the experiment than the RBD to achieve the same statistical power.
- The greater the inherent turbine-to-turbine variation in fatality, the greater the sample size necessary to detect the hypothesized effect.
- It requires adjusting observed carcass counts to account for imperfect detection.
- It requires some resources to estimate factors related to detection probability, such as searcher efficiency, carcass persistence, and proportion of carcasses expected to land in the sampled area.
- The greater the imprecision in the estimate of detection probability, the greater the sample size necessary to detect the hypothesized effect.

A slightly more complicated design is an RBD, wherein the block is the turbine and each treatment is assigned to an equal number of nights within each turbine. The experimental unit now becomes the set of nights taken together, not the individual night. With this design, inherent differences in the fatality response among turbines are accounted for in the blocking factor, and treatment differences are calculated after removing this potentially large source of variation. In addition, because the detectability of carcasses at a turbine will not change, the response can be

the simple count of carcasses the following day, with no adjustments needed for imperfect detection.

The advantages of the RBD are as follows:

- It requires a smaller number of turbines in the experiment than the CRD because much of the among-turbine variation is controlled.
- It requires no resources to estimate factors related to detection probability.

The disadvantages of the RBD are as follows:

- Deterrents must be mounted on all turbines because all treatments will be implemented at all turbines.
- Because treatment assignment is nightly, searches must be carried out daily.
- Searchers will need to determine in the field whether a found carcass was "fresh," i.e., likely killed the previous night
- Not all carcasses found can be included in the analysis, only fresh ones.
- It increases the likelihood of contamination of missed carcasses being found later and assigned to the wrong treatment because estimating time since death is imperfect.

Important sources of variation in the response can be identified in various ways. For example, if numbers of bats killed is related to activity, then the variation in the number of bats is easily controlled by a researcher, so the experimental design must attempt to equalize the effect of each of these among experimental units and treatments. This balance can be achieved by appropriate randomization. Another important source of variation is not in the number of bats killed but in the number actually observed. If the searchable area beneath one turbine is substantially different in both extent and configuration than another, then even though the average number killed may be the same for both turbines, the number of bats observed may differ substantially. Researchers can account for these differences by using models that estimate the probability of detection at each turbine and then adjusting the observed count by this probability; however, these models are themselves imprecise, and any measure of fatality that also includes a model-based component to adjust for probability of detection will have an additional source of variation and ultimately require a larger sample size to detect the same effect. In addition, if the treatment could affect behavior or conditions when fatalities occur (e.g., different wind speeds), and hence carcass distribution, then the searchable area needs to be large enough to avoid statistical error.

5.6 Treatments

Three experimental treatments are included in this study. All turbines are free to operate normally, and the treatments are defined only by the activation of the deterrents. The experimental treatments are as follows:

- Control: acoustic deterrents inactive
- Treatment A: acoustic deterrents activated throughout the night

- Treatment B: acoustic deterrents activated only from a set period of time between sunset and midnight.

5.7 Sample Size

In the CRD mentioned earlier, the experimental unit is the turbine—the unit to which a treatment was applied. The sample size (number of replicates) is the number of turbines receiving each treatment.

In the RBD, the experimental unit is the set of nights for a particular turbine to which a treatment was applied. The sample size (number of replicates) is the total number of turbines in the study because each will receive each treatment.

5.8 Anticipated Effect Size

The research question determines the minimum anticipated effect size at 60% fewer bats killed in either deterrent approach relative to the undeterred approach. For comparison purposes, previous studies have shown a reduction of 50% or more from operational minimization. An effect similar to this was set as the target of this study. Note that neither experimental design is intended to provide an estimate of fatality per turbine. The design is intended to detect a consistent change in fatality rate that can be attributed to the operation of the deterrents.

5.9 Statistical Power of the Study

In the CRD, a turbine receives the same (randomly assigned) treatment for the full 75-night duration of the study, resulting in 4*75 = 300 turbine nights for each treatment (Table A-1). The nights are only the sampling unit, though, and the observed carcasses from the 75 nights at each turbine are adjusted for detection probability to form a single response value for that turbine. Total degrees of freedom (DOF) for this design are 12-1=11.

In the RBD, each turbine receives each treatment for 25 of the full 75 nights in the study, resulting in 12*25 = 300 turbine nights for each treatment. The nights are only the sampling unit, though, and the total observed carcasses from the 25 nights receiving a treatment at each turbine are summed to form a single response value for that treatment for that turbine. Because there are three response values (one per treatment) for each turbine, total DOF for this design are 12*3-1 = 35.

Table A-1. Analysis of Variance Table for the CRD and RBD Using a Total of 12 Turbines in the Experiment

	CRD		RBD	
Source	df	df	df	df
Turbine	N/A	N/A	(b-1)	11
Treatment	(*t*-1)	2	(t-1)	2
Residual Error	*t**(*r*-1)	9	(b-1)*(t-1)	22
Total	(*r***t*-1)	11	(b*t-1)	35

b = number of turbines
t = number of treatments
r = number of turbines in each treatment

When assigning a treatment to a turbine night in this design, it is critical that 4 turbines receive each treatment on any given night, and during the course of 3*t nights, each treatment is applied three times at each turbine (i.e., each treatment is assigned 25 times to each turbine). This rebalancing is important to ensure that the treatments are spread relatively evenly throughout the season.

5.10 Study Execution

In both of these designs, it is very important that the searchers are not aware of the treatment assignment, whether a treatment is assigned to a turbine for the duration of the project or reassigned each night. As the study progresses, if there is a strong treatment effect, it might be difficult for a searcher in a CRD to remain naïve to the treatment assignment.

In the RBD, random assignment and rebalancing are critical. Programming constraints in the Supervisory Control and Data Acquisition system might make it tempting to develop a hybrid design wherein the 12 study turbines are divided into 3 sets and each set of 4 is randomly assigned a treatment on each night. The same 4 turbines always receive the same treatment. This might appear to be an equivalent design to the one mentioned earlier, but it is not. In effect, it becomes a CRD with no replication because the treatment assignment is always to the set of turbines, not the individual turbine. Thus, the set is the experimental unit. Any inherent differences among sets will be confounded with inherent differences in environmental conditions.

An alternative might be to reassign treatments every 3 nights. This approach reduces the chance of spreading uncontrolled variation among sampling units, but it does not completely compromise the design as the above example would. Searches will nevertheless need to be conducted daily, not at the end of the 3 nights of treatment, to accurately assign a carcass to a treatment. As mentioned earlier, a searcher is very unlikely to be able to accurately assign the time of death between 3 nights prior versus 4 nights prior. On the other hand, if searcher efficiency is high and the proportion of carcasses persisting is high, then few carcasses will be

missed, and the inability to accurately assign time of death may be an issue for only a few carcasses. If reassigning treatments every 3 nights and searching every 3 days will allow for a threefold increase in the number of turbines included in the study, then this approach should be considered because the increase in power might offset the measurement error induced.