AVIAN/WIND STATISTICAL PEER REVIEW PROJECT

CONSULTANT REPORT

Prepared for:
California Energy Commission

Prepared by:
California Institute for Energy and Environment

DECEMBER 2006
CEC-500-2006-114

Prepared By:
California Institute for Energy and Environment
Terry Surles and Edward Vine
Oakland, California
Contract No. 500-02-004

Prepared for:

California Energy Commission

Kelly Birkinshaw
Project Manager

Laurie ten Hope
Manager
Environmental Research Office

Martha Krebs
Deputy Director
Energy Research and Development Division

B.B. Blevins
Executive Director

DISCLAIMER

This report was prepared as the result of work sponsored by the California Energy Commission. It does not necessarily represent the views of the Energy Commission, its employees or the State of California. The Energy Commission, the State of California, its employees, contractors and subcontractors make no warrant, express or implied, and assume no legal liability for the information in this report; nor does any party represent that the uses of this information will not infringe upon privately owned rights. This report has not been approved or disapproved by the California Energy Commission nor has the California Energy Commission passed upon the accuracy or adequacy of the information in this report.

TABLE OF CONTENTS

EXECUTIVE SUMMARY ..... **II**

CHAPTER 1 – INTRODUCTION **1**

DEVELOPMENT OF THE RFQ 1
SELECTION OF PEER REVIEW TEAMS 2
TECHNICAL EXPERTISE OF PEER REVIEW TEAMS 3
PEER REVIEW GUIDANCE 4
KICKOFF MEETINGS 5
PREVIOUS PEER REVIEWS 6

CHAPTER 2 – PEER REVIEW FINDINGS **7**

INTRODUCTORY COMMENTS 7
WAS THE STATISTICAL METHODOLOGY USED ON THE ANALYSIS CONSISTENT WITH ACCEPTED METHODS USED IN OTHER BIO-STATISTICAL ANALYSES? 7
WERE THE TECHNICAL APPROACHES USED IN THE RESEARCH APPROPRIATE FOR ACHIEVING THE STATED GOALS? 8
WERE THE DATA COLLECTION METHODS AND ASSUMPTIONS CLEARLY STATED, RELIABLE AND VALID? WERE THERE FLAWS, ERRORS, OR RELEVANT FACTORS MISSING? 9
WAS THE STUDY DESIGN SUFFICIENTLY SOUND? WAS THERE SUFFICIENT TIME TO CONDUCT THE STUDY? 9
WERE THE UNCERTAINTIES DESCRIBED, EITHER QUANTITATIVELY OR QUALITATIVELY? 10
WERE THE FINDINGS STATISTICALLY SIGNIFICANT? 11
WERE THE CONCLUSIONS SUPPORTED? 12

CHAPTER 3 – FUTURE RESEARCH DIRECTIONS **13**

THE POLICY CONTEXT 13
RESEARCH SUGGESTIONS 13

Attachments

A. Team 1 Peer Review — Attached file
B. Team 2 Peer Review — Attached file
C. Team 3 Peer Review — Attached file
D. Response by Smallwood and Thelande — Attached file

Executive Summary

On behalf of the California Energy Commission (Energy Commission), the California Institute for Energy and Environment (CIEE) conducted a peer review – with a primary focus on statistics - of the following report submitted to the Energy Commission: "Developing Methods to Reduce Bird Mortality in the Altamont Pass Wind Resource Area," prepared by K. Shawn Smallwood and Carl Thelander. CIEE first prepared a Request for Qualifications (RFQ) and used its main mailing list and other list servers to distribute the RFQ. A Peer Review Selection Committee reviewed four proposals and selected three teams of reviewers. The peer review teams submitted preliminary peer review reports that were reviewed by the authors of the report (Smallwood and Thelander). The authors provided a detailed response to the peer review reports. After reviewing the authors' response, the peer reviewers finalized their reports. The key findings from the final peer reviews are presented in this report. The peer reviews are attached (Attachments A-C), along with the authors' response (Attachment D).

In general, all of the reviewers were explicit in pointing out that the authors had taken on an important issue and had done a credible job with the resources that were available to them. The original report was clearly an exploratory study meant to set the stage for future work in this area. As such, the report serves to provide a basis for continuing research on the topic of avian/wind turbine interactions.

The report should not, however, be considered as the basis for developing siting requirements for future wind energy projects. It is clear from the peer reviewers' comments that there are significant problems with this paper and that additional studies are needed. The positive aspects of the report, coupled with the constructive criticism of the reviewers, could form the basis of future work to better define siting requirements and guidelines that should be put in place by permitting agencies. As the reviewers noted, it will be important to evaluate the model developed in the original study with new wind projects. Future research will need to minimize some of the confounding problems in the reviewed study, since remaining disagreements require additional research for their resolution. Also, utilizing new data sets will better serve to determine whether the model is effective. If not, the results should suggest improvements to the model that would help in its function as a predictive tool. More research is needed to identify the causes of collisions and what measures need to be taken to reduce mortality caused by wind turbines.

Chapter 1 – Introduction

On behalf of the California Energy Commission (Energy Commission), the California Institute for Energy and Environment (CIEE) conducted a peer review – with a primary focus on statistics - of the following report submitted to the Energy Commission: "Developing Methods to Reduce Bird Mortality in the Altamont Pass Wind Resource Area," prepared by K. Shawn Smallwood and Carl Thelander.[1] The primary objectives of this report were to: (1) quantify bird use, including characterizing and quantifying perching and flying behaviors exhibited by individual birds around wind turbines; (2) evaluate the flying behaviors and the environmental and topographical conditions associated with flight behaviors; (3) identify possible relationships between bird mortality and bird behaviors, wind tower design and operations, landscape attributes and prey availability; and (4) develop predictive empirical models that identify areas or conditions that are associated with high vulnerability.

CIEE first prepared a Request for Qualifications (RFQ) and used its main mailing list and other list servers to distribute the RFQ. A Peer Review Selection Committee reviewed four proposals and selected three teams of reviewers. The peer review teams submitted preliminary peer review reports that were reviewed by the authors of the report (Smallwood and Thelander). The authors provided a detailed response to the peer review reports. After reviewing the authors' response, the peer reviewers finalized their reports. The key findings from the final peer reviews are presented below. The peer reviews are attached (Attachments A-C), along with the authors' response (Attachment D).

Development of the RFQ

CIEE held a meeting with key stakeholders and consultants on April 4, 2006 to discuss the preparation of a Request for Qualifications (RFQ) for recruiting peer reviewers. The meeting participants represented the Energy Commission, CIEE, the National Audubon Society (Audubon), and the California Wind Energy Association (CalWEA). The principal objectives of the meeting were to agree on a scope for the peer review and the criteria for choosing peer reviewers.

<u>Scope for peer review and RFQ:</u>

1. The scope of the peer review was discussed and, after reviewing several options, the group agreed that the peer review should focus on the Energy Commission's avian/wind report.

2. The preliminary questions for conducting the statistical review were discussed, and there was agreement that these were the appropriate general questions to ask. After the meeting, the revised RFQ was circulated to all individuals attending

[1] The report (Publication #500-04-052) can be downloaded from the following web site: http://www.energy.ca.gov/pier/final_project_reports/500-04-052.html

this meeting, and a final RFQ was subsequently approved by the Energy Commission (Attachment A).

3. The group agreed that the selected peer reviewers would remain anonymous to the authors of the report (but that their review comments would be publicly available) and that the only people knowing the selected reviewers would be CIEE technical staff and members of the selection committee.

4. There was agreement that peer reviewers would be compensated for their work, and that, ideally, three reviewers would be selected, depending on the quality of the proposals and available budget.

Criteria for choosing peer reviewers and RFQ:

The preliminary criteria for selecting reviewers were discussed, and there was agreement that these criteria were the appropriate criteria. Several modifications were proposed and agreed upon and were incorporated into the final RFQ (Attachment A).

Selection of Peer Review Teams

The Peer Review Selection Committee convened via teleconference on Thursday, May 18, 2006. The Committee members were Terry Surles (Chair; CIEE), Al Manville (US Fish and Wildlife Service), and Richard Myhre (Bevilacqua Knight, Inc.). The purpose of the meeting was to select two or more peer reviewers from the four proposals that responded to the RFQ.

The Committee first discussed two “gates” as part of the selection process. The first gate was cost. All of the four proposals that were submitted were acceptable in terms of cost. In fact, the four proposals were within $500 of each other. At this point, it was agreed that we had sufficient funds to select three of the proposals to conduct the review. The Committee concluded that three reviewers would provide greater consistency than two. The second gate was that no reviewer had any conflict of interest with either of the report authors. The Committee agreed that all four respondents explicitly addressed this requirement in their cover letters, as well as being reflected in their resumes. The Committee was satisfied that all four proposals met this requirement. After selecting two review teams, one review team was excluded because the specific statistical and overall breadth of expertise provided by the remaining review team was slightly, but noticeably, better. The Committee then reviewed the selections we had made and the reasons for making these selections to ensure that we agreed that these were the correct selections. The agreement was unanimous among the three Committee members that the three proposals chosen for the peer review of the Smallwood and Thelander report were appropriate.

The Committee also discussed whether these reviewers would be able to address the Commission Executive Director’s concern about the appropriateness of using the Chi-

Square test and analysis in the Smallwood and Thelander report. The Committee unanimously agreed that all of the selected reviewers had the technical expertise to address this question. In fact, the Committee expected that the reviewers would be directly responding to this issue, since this was one of the requests made in the Request for Qualifications: “Was the statistical methodology used on the analysis consistent with accepted methods used in other bio-statistical analyses?” The Committee felt that this question (and the remaining RFQ questions) should be part of the guidance letter sent to the selected reviewers. Notification and guidance letters were then sent to the three review teams.

Technical Expertise of Peer Review Teams

The three peer review teams significantly exceeded minimum technical requirements needed to conduct the peer review. We describe the three teams below.

Review Team 1. There were three individuals in this team.

Team member #1 has a Ph.D. in biostatistics and is a biostatistician at a university, with specific interests and expertise in multivariate statistical methods, analysis of variance, experimental design, and survival analysis. He has applied these interests to a wide array of applications, from content analysis to large avian mortality studies, and he is widely published. He teaches a diverse group of courses in statistics and has consulted extensively to the public and private sectors on statistical issues, including providing reviews of study design, analysis, and conclusions.

Team member #2 is a consultant and has designed and conducted many avian studies as a federal research biologist and field station leader. These studies were primarily been done to determine population trajectory and to identify causes of mortality of waterfowl and game birds. He has extensive field experience and is widely published. He is a Certified Wildlife Biologist and has been recognized by the U.S. Fish and Wildlife Service for exemplary performance and by The Wildlife Society for his leadership.

Team member #3 is a population biologist at a university. He has a thorough knowledge of the scientific literature on avian collisions, especially nocturnal collisions.

All of the team members have conducted independent and unbiased technical reviews.

Review Team 2. There were three individuals in this team.

Team member #1 has a Ph.D. in biostatistics and teaches at a university. His recent research interests focus on monitoring, sampling, generalized linear models, and statistics education.

Team member #2 has a Ph.D. in ecology and is a wildlife population ecologist specializing in ornithology and teaching at a university. He has recently taught courses in ornithology, land-bird conservation and management, habitat ecology, wildlife techniques and scientific method, wildlife ecology and conservation, and principles of wildlife management.

Team member #3 has a Ph.D. in environmental engineering and is a co-director of a university energy research center.

All of the team members have conducted independent and unbiased technical reviews.

Review Team 3. There was one individual in this team.

This person has a Ph.D. in biomathematics and is a statistics professor at a university. His area of expertise addresses sampling and estimation problems in ecology and wildlife biology, and he has published extensively on these topics. He has participated in several ornithological field research studies. He has also served as an editor for the Wildlife Society Bulletin and Journal of Wildlife Management. He is also president of a statistical consulting company, and he has conducted many independent and unbiased technical reviews.

Peer Review Guidance

The peer reviewers were asked to address the following questions:

- Was the statistical methodology used on the analysis consistent with accepted methods used in other bio-statistical analyses?
- Were the technical approaches used in the research appropriate for achieving stated goals?
- Were the data collection and analysis methods and assumptions clearly stated, valid, and reliable? Were there any errors or, flaws? Were any relevant factors missing?
- Was the study design scientifically sound? Was there sufficient time to conduct the study (e.g., time for conducting searches, time for assessing seasonal effects)?
- Were uncertainties described, either qualitatively or quantitatively?
- Were findings statistically significant?
- Were the conclusions supported?
- Other observations and comments?

Kickoff Meetings

CIEE held three kickoff meetings – one with each review team. The principal objectives of the kickoff meetings were to review the scope of work, key issues, and to answer any questions from the reviewers. The following topics were covered:

1. Conflict of Interest. One of the key criteria in selecting the peer reviewers was that no members of the peer review team had any conflict of interest (as stated in the RFQ). All review teams verbally confirmed that there had been no change to their statements in their proposals, and that there was no conflict of interest.

2. Anonymous Review. The review process was generally modeled after the National Academy of Sciences' review process where the names of the reviewers are kept anonymous. The only people knowing the names of the reviewers were the members of the Peer Review Selection Committee and CIEE staff working on the contracts with the reviewers. The Energy Commission and the report authors do not know the names of the reviewers. The reports from the peer reviewers would be part of the public record, but the names of the reviewers would not be on the reports. On one specific issue, there was a divergence from typical review processes: there was explicitly no opportunity available for the authors to update their report. Therefore, where authors and peer reviewers were in agreement on necessary changes to the report, there was no mechanism to effect those changes.

3. Focus of Peer Review. It was emphasized that, in order to maintain an objective and impartial review, the peer reviewers would review only the Smallwood and Thelander report, and not the reviews by other parties. The focus would be on the statistical and technical validity of the report. The reviewers would also focus on the research design, data collection, and data analysis components of the report, and not on the policy recommendations. The statistical concerns of the Executive Director (as reflected in the note in the Technical Scope of Work provided to the review teams) were reviewed. All of the review teams indicated that they understood these issues and that they would be addressed in their review (but not be the primary focus of their review).

4. Report Organization and Format. It was emphasized that the peer review reports should be understandable to the layperson and non-statisticians, especially in the Executive Summary. While the main report would highlight the key statistical findings from the review, the authors of the reports would consider placing some text in an Appendix if it is perceived as too detailed. The reviewers would highlight the overall impressions of the report and then indicate if there were any "fatal flaws" in the report. This would be followed by a critical discussion of the report (indicating page and line numbers) and opportunities for improvement.

5. Interaction Between Reviewers and Authors. Two types of interaction between reviewers and authors were anticipated. The first type of interaction would occur during the review process where the reviewers would want to obtain references or seek clarifying questions from the authors. All questions were to be sent by the lead contact from the peer review team to CIEE; CIEE would then copy and paste these questions in a new email that would be sent to the authors of the report. The authors would then respond to CIEE; CIEE would then forward their response to the reviewers. We only received one request from one review team for a document. One reviewer noted that he would have liked to talk to the authors but that the discussion would have been very lengthy and would have used up many of the hours allocated to the peer review.

   The second type of anticipated interaction would occur after the draft report had been submitted to CIEE. CIEE would forward the draft reports to Smallwood and Thelander who would be given the opportunity to respond to the draft reports. Their comments would then be forwarded to the peer reviewers. The peer reviewers would respond to the authors' comments and make changes to their report if warranted. A record of these discussions would be attached as an appendix to the final report. This process did occur. Smallwood and Thelander were given three weeks to respond to the comments from the peer review teams. The response from Smallwood and Thelander is attached (Attachment E). The peer review teams were given two weeks to make any changes to their preliminary report. Changes to the preliminary report were marked in italics and the individual peer review reports are attached (Attachments B, C, and D). While it may be difficult for the reader to follow all of the edits in the peer review reports, we believe it is important to keep these edits in the reports, so that the reader can follow the entire review process and understand how the peer reviewers responded to Smallwood and Thelander's response to the original peer review reports.

Previous Peer Reviews

While not specifically pertinent to the mechanics of this review, we believe that it is important to point out that this is the third peer review of the Smallwood and Thelander report. Prior to its release by the Energy Commission, this report was peer reviewed by two scientists with expertise in the issue of bird collisions with wind turbines. The report was then revised and reviewed by the wind turbine owners and their consultants, and by biologists from the California Department of Fish and Game and the U.S. Fish and Wildlife Service. The report was revised again before the Energy Commission released it. The Energy Commission administered a second peer review by three more scientists with expertise in the issue of bird collisions with wind turbines. However, Smallwood and Thelander were not given the opportunity to revise the report. The Energy Commission requested a third review – to be conducted by scientists with expertise in biostatistics and with no conflict of interest with the authors of the report. The findings from this third review are presented in the next section.

Chapter 2 – Peer Review Findings

Introductory Comments

Prior to providing a summary of the three sets of peer review, a few general comments are necessary. In general, all of the reviewers were explicit in pointing out that the authors had taken on an important issue and had done a credible job with the resources that were available to them. The reviewers also recognized study difficulties related to the limited ability to manipulate the site to meet the data collection requirements for statistical analyses. The original report was clearly an exploratory study meant to set the stage for future work in this area. As such, the report serves to provide a basis for continuing research on the topic of avian/wind turbine interactions. Absent future research, it is clear from the peer reviewers' comments that there were sufficient problems to preclude the use of this paper for regulatory decision-making. Rather, the positive aspects of the report, coupled with the constructive criticism of the reviewers, should form the basis of future work to better define what siting requirements and guidelines could be put in place by permitting agencies.

The following discussion has been designed to explain the reviewers' comments in a manner that will be understandable to the lay reader. The following discussion is structured based on the questions that the reviewers were requested to answer per the RFQ (see Section 1.4). For an explanation of terminology, please refer to Review Team #2 Comments in Appendix C, pages 2 – 4. The remainder of Section 2 is essentially an extraction and summary of the reviewers' comments, unless noted otherwise. An additional section has been added at the end, which focuses on recommendations for future work. These recommendations are based on reviewer comments as well as those of the original authors who were asked to respond to the draft reviews.

Was the Statistical Methodology Used on the Analysis Consistent with Accepted Methods Used in Other Bio-statistical Analyses?

The three review teams unanimously felt that there were four important statistical flaws: (1) use of chi-square tests on confounding variables; (2) pseudo-replication in the way bird behavior was reported; (3) presentation of hypothesis results from a study design that did not carefully control for external factors; and (4) the high probably of Type I errors from conducting hundreds of univariate tests.

Chi-squared analyses with measured variables of time were not appropriate. The analysis of a measured variable, such as minutes, using a Chi-square analysis is invalid. The Chi-squared tests are sensitive to changes in scale, i.e. the results would change if data were expressed in seconds or hours. Chi-squared analysis must be used with counts, not time. The authors of the study used Chi-squared analysis to assess the significance of timed bird behavior. Also, Chi-square analysis assumes that counts are exact and not estimated. Adjusted counts (adjusted for scavenging and detection rates)

were frequently used in the report. It was not obvious to all reviewers whether adjusted or raw counts were utilized to calculate the statistics. If adjusted counts were used, it was not clear how the uncertainty in these counts would influence the conclusions reached on numerous chi-square hypothesis tests. Finally, using a very large number of univariate chi-square tests (as used by the authors) is not common in bio-statistical analyses.

The reviewers were concerned about possible pseudo-replication. Pseudo-replication is possible when samples are not independent, but are treated independently. The reviewers observed that the turbine strings were surveyed using transects and the turbine strings were the sampling unit. The authors used both turbine strings and individual turbines as independent samples. Thus, it is possible that pseudo-replication is a source of statistical uncertainty.

The author's study design was observational, but the authors presented results from hypothesis testing as if the study was carefully controlled for external factors. Therefore, the authors proceeded to present results from hypothesis testing as if the study was carefully controlled for external factors and results were confirmatory. Presentation of results should be clearly reported as exploratory.

Multiple comparisons with inter-correlated variables were made without appropriate corrections. One reviewer pointed out that the use of alternative analytical techniques would still not address all of the problems associated with the data sets, such as non-random sampling. While all reviewers noted that it is likely some number of the reported tests were statistically significant, they would have more confidence in the analysis if other tests were performed (e.g., multiple regression). In addition, performing many statistical tests increases the probability that "significant" results would be found when there is none – a Type I error. The probability of Type I errors increased by conducting multiple tests with confounding independent variables. The reviewers noted that the threshold for significance (P values of 0.05 or smaller) could average about 1 Type I error in every 20 tests. Therefore, the hundreds of tests produced by the authors could result in numerous Type I errors. The peer reviewers felt that the study failed to use appropriate corrections to account for the probability of Type I errors.

Were the Technical Approaches Used in the Research Appropriate for Achieving the Stated Goals?

The stated goals for this study were to: (1) quantify bird use; (2) evaluate flying behaviors and conditions associated with flying behaviors; (3) identify the relationships between bird mortality and various explanatory variables; and (4) develop predictive, empirical models that can be utilized to identify areas or conditions associated with high vulnerability. The reviewers believed that limitations imposed on the study design precluded the ability to effectively address all four goals. In particular, a design for addressing #1 and #2 goals would not be the appropriate design for developing predictive models (#4). All reviewers felt that the major problems with the approach

were that the individual turbines were not statistically independent and turbines were not given equal levels of survey effort. The reviewers noted that while observational studies are commonly performed, there are limitations in the degree to which defensible inferences can be made in studying biological systems. The issues associated with the technical approaches are best addressed in the responses to questions serving as Section headers from 2.4 to 2.8.

Were the Data Collection Methods and Assumptions Clearly Stated, Reliable and Valid? Were There Flaws, Errors, or Relevant Factors Missing?

The reviewers noted that numerous assumptions were made over the course of the study. While it may have been necessary under budget and time constraints to make these assumptions, these assumptions were not readily defensible and, in some cases, were not clear to the reviewers. Some examples follow.

The authors used recently developed protocols for carcass searches, standard bird observations, rodent surveys, etc., to obtain data and, thus, were appropriate. However, issues arose in terms of methods used to measure bird mortality. This is because differences in observer ability were not incorporated into the evaluation, nor were they demonstrated to be insignificant. For example, differences in scavenger removal were not incorporated. Since scavenger removal is distributed unevenly across a landscape, its influence could confound patterns caused by other variables. Instead, the authors chose to adopt adjustments to published scavenging and detection rates based on assumptions that were inadequately supported with information contained in the report. Finally, the degree of interference from the operators of the wind turbines (site operators) was probably not taken into account. For example, site operators would find carcasses on site and bury the carcasses – clearly an unanticipated form of scavenger removal. Further, scavenger removal is spatially variable and difficult to correct for. Thus, there are many sources of error that make accurate measurement difficult as part of this study.

Was the Study Design Sufficiently Sound? Was There Sufficient Time to Conduct the Study?

The study design had several flaws that could compromise the reliability of the results. All of the reviewers agreed that one of the biggest issues was the non-random sampling of turbine strings. Constraints imposed by the site operators precluded the authors from implementing a truly random sampling design (e.g., not all sites were accessible to the authors at the same time). Although approximately 75% of the turbine strings were eventually sampled, their results cannot be extrapolated to those strings not sampled. The order in which turbine strings were, somewhat piecemeal, added to the study could have affected the results. While the authors did not bias their sampling locations, unknown biases may have been present due to the non-random selection of turbines.

Individual turbines were not statistically independent because they were surveyed as strings (i.e. strings of turbines). Turbines were not sufficiently separate to unquestionably assign a carcass to the turbine that caused a death. The appropriate unit of analysis was the turbine string. Thus, the statistical analyses of the impacts due to individual turbines may be inappropriate due to their lack of separation from one another. This can make the effects of the turbines appear to be more significant than they are. Thus, all analysis of variance (ANOVA) tests using the distance of birds to turbines as the experimental unit suffer from this analytical flaw.

The study design was hindered, as described in a preceding section, by the difficulties encountered in obtaining access to the site. This lack of access – or access granted by site operators on an incremental basis – led to the difficulties associated with non-random sampling and the statistical assumptions and analyses that followed from this basic problem.

This led to an observation by the reviewers on what the study was (and wasn't). There is the acknowledgement that the authors put considerable effort into collecting a very large amount of data. One reviewer observed that, as a result, this focus led to less attention on what to do with information once it was obtained. Thus, these analyses could have been more thoughtful and sophisticated (statistically). As it is, the statistical analyses are applied in a seemingly automated manner that fail to utilize the data at hand and ignores the probability of confounding variables. As mentioned previously, large number of tests likely result in Type I errors.

An important conclusion from one set of reviewers is that the authors performed an observational study. While these are commonly performed in biological sciences, they do have limitations. Non-random sampling is not the fault of the authors (due to their lack of site access), but the results may lead to an "unrepresentative" set of data. Therefore, this study should be seen as a significant addition to the literature, due to the volume of information obtained. However, any conclusions on the statistical significance of the findings should simply be treated as indicators of what should be explored in future studies in which funders, scientists, statisticians, and site owners are clear on the study design and their respective roles from the onset of the study.

Were the Uncertainties Described, Either Quantitatively or Qualitatively?

The reviewers noted that uncertainties were described in many cases. However, uncertainties were either insufficiently explained or not explained in other cases. The authors made no attempt to adjust, quantify, or describe the probability of a number of positive results across the project due to the vary large number of univariate tests that were performed.

The authors made significant use of extrapolation of mortality estimates, thus leading to additional uncertainty. While this may be reasonable for observational studies, it is not

appropriate in cases where these extrapolations may form the basis of statistical analyses.
The reviewers believed that many estimates of rates were provided with no attempt to describe associated uncertainties. For example, estimates of the percentage of mortality increases associated with a given variable are provided with no assessment of the uncertainties associated with these estimates. The lack of consistency in seasonal sampling introduced additional uncertainty.

Additionally, the authors attempted to connect rodent eradication practices to reduced mortality. While this effort may be admirable, it also introduces one additional variable that can further confound the overall study's conclusions. The issue of confounding variables, as in this example, is not properly addressed in the study's analyses. Thus, some of the conclusions that are drawn may be inappropriate due to lack of attention to this issue. Future studies will need to address which (and how) certain variables may be confounded. It is also possible that confounding cannot always be reduced or eliminated. When this fact is determined, the results should be stated clearly that this is the case.

As described in previous sections related to study design, there was no attempt to address differences in observer abilities, although the authors have the opinion that mortality will be underestimated. However, since the results of the study rely heavily on comparisons of bird kills at different turbine types, observer ability could bias fatality rates and study conclusions.

A similar problem occurs when attempting to provide reasonable estimates for scavenger removal. As described previously, the site operator's procedures for removing carcasses as they were discovered through normal operation and maintenance activities adds a variable that may not normally be encountered. In not addressing these and other variables, the concomitant uncertainties arising from the activities and procedures at different locations across the study area were not addressed.

Were the Findings Statistically Significant?

The discussions in the preceding sections make clear that the findings are not necessarily statistically significant. Small sample size (in some cases) and over-use of a variety of statistical tests (such as Chi-square test) can lead to a potentially large number of Type I errors – to name just two examples of why some findings would not be statistically significant. All reviewers agreed with this observation. As one reviewer pointed out, with so many Chi-square tests being performed, the probability of a false positive being seen as "real" became 1. In addition, the combination of flawed sample design and analytical methods used make it difficult to identify which results are *biologically significant* and which are artifacts of process weaknesses. All reviewers pointed out that many of the findings may in fact be significant but that a re-analysis using multiple regression would give them more confidence in the results.

Were the Conclusions Supported?

This study and the related analyses cannot be accepted as one that has rigorously tested hypotheses regarding bird mortality. Thus, it is unreasonable to assume that this document could, by itself, form the basis of informed decision-making. Instead, this project should be considered as an exploratory analysis that has identified a number of variables that are positively associated with increased mortality rates. Thus, the product of this research endeavor is an educated list of working hypotheses associated with bird mortality. Additionally, having developed a model that attempts to predict mortality with the study data, this study can be followed up with studies to examine the validity and robustness of the model through a rigorous sampling design that will work to control confounding variables.

Chapter 3 – Future Research Directions

The reviewers and the authors themselves provided some excellent suggestions for conducting future research in the area of avian/wind turbine interactions. Smallwood and Thelander, as requested by the Energy Commission, devoted a portion of their response to the draft review comments to pointing out specific needs for future studies (see Appendix E). Prior to considering these specific suggestions, some other points need to be made.

The Policy Context

The ratepayers of the investor-owned utilities (IOUs) in California fund the Energy Commission's Public Interest Energy Research (PIER) program. As such, any work funded by PIER must meet the needs of ratepayers and provide the opportunity of future benefits to these ratepayers. Proper resolution of the issues raised in PIER studies is appropriately in the interest of the ratepayers, for at least two reasons. First, the State of California has implemented legislation that promotes the use of renewable energy technologies for the production of electricity. In order to meet the goals of these legislative initiatives, a substantial portion of renewable-powered electricity generation must come from wind turbines. Until these issues have been resolved, the full development of the wind turbine industry will be delayed. Second, the State of California has been remarkably consistent in addressing serious environmental concerns in a manner that meets with public approval. Thus, the understanding of and the development of mechanisms for the protection of threatened and endangered avian species is an important facet of these types of programs and is expected to be a high priority for the PIER program.

Research Suggestions

As the reviewers noted, the lack of cooperation with the operators of any set of facilities will be a showstopper for conducting future research studies that are scientifically sound. There will need to be a clear understanding for future studies between all parties as to the study design and site access to ensure the conduct of a robust program. As Smallwood and Thelander point out, the installation of expensive retrofits by existing wind turbine operators may be too expensive. Thus, the only measures left to researchers at existing sites may be painting schemes, lighting, acoustical devices, and land use practices. They also suggest that the more important factors for future development may be siting, wind farm configuration, the type of turbine, and rotor blade height. Analyses of these factors should be considered as part of future research.

As the reviewers noted, it will be important to evaluate the model developed in the original study with new wind development projects. The projects will need to be designed to minimize some of the confounding problems that were raised in the previous study. Also, utilizing new data sets will better serve to determine whether the

model is effective. If not, the results should suggest improvements to the model that would help in its function as a predictive tool. More research is needed to identify the causes of collisions and what measures need to be taken to reduce mortality caused by wind turbines.

The installation of a wind turbine farm clearly modifies the environment. Based on these modifications, more needs to be known about perching and flight patterns that have been modified due to wind farm installation. This effort would necessarily need to include several bird species. Where possible, it would be very useful to design a study examining these behavioral patterns in a region prior to construction of the facility and following the start up of facility operations. The reviewers and authors also agree that adequately characterizing avian behaviors in different seasons is critical.

The reviewers were concerned about the validity of human observations. Smallwood and Thelander have suggested using advanced integrated radar and camera systems (AIRCAMS) for properly characterizing a site. Human observation would remain an important component of any project, but data obtained by both methods should be more robust (i.e., the human observations would be validated by the AIRCAMS technology).

The need to obtain peer review and approval of future study designs will be important. For example, there continues to be debate on the validity of using turbine strings versus individual turbines as the sampling point. Some evaluation will need to be done in the future concerning the amount of separation between turbines to allow truly independent analysis for each turbine.

The reviewers raised several issues regarding data analysis. In their response, the authors agreed with many of these comments. The authors suggest that a means to address this issue will be to collaboratively share data with the developer of the Avian Risk of Collision (ARC) model. Each research budget should include some funding for this coordination.

Other information that would need to be collected as part of a more robust set of studies in the future would include that for diurnal raptors, raptor nest survey data, nocturnal raptors, grassland songbirds, and bats. Mechanisms and parameters for fatality searches must be clear to an advisory board or external group. And, as described previously, the potential for access to all turbines on site must be unlimited.

Since there was debate on this issue, there must be agreement on the mortality metric used for these studies that will be consistent with metrics for similar studies in other locations. The authors have suggested fatalities/kWh. This may be a reasonable suggestion, since site operators will be thinking in terms of the output of their facility. Also, other environmental emissions, such as sulfur dioxide, are measured in terms of electricity output.

A better analysis of ecological relationships will be important for understanding impacts to avian species. Thus, a better understanding of other species behavior, such as cattle,

may be an important set of data to obtain in the future. Also, it will be important to better understand the relationships of animals with their surrounding ecosystem. Therefore, better information on ground cover, grass height, and any invasive species introduction will be valuable.

The preceding discussion forms the basis of future studies. It is critical to re-iterate that these studies must be done in the public interest: these studies must provide a benefit to the ratepayers that are funding these programs. All study designs must keep in mind that, ultimately, these programs should enhance the understanding of public decision-makers and provide a more robust basis for developing siting, construction, and operating permitting guidelines.

ATTACHMENT A

Review Team #1

Review of "Developing Methods to Reduce Bird Mortality in the Altamont Pass Wind Resource Area" by BioResource Consultants

Executive Summary

The report "Developing Methods to Reduce Bird Mortality in the Altamont Pass Wind Resource Area" represents a large effort collecting data about the association of avian mortality with wind turbine type, topography, rodent management, and other variables. Despite the extensive survey effort, flaws in study design and statistical analysis hamper the interpretation of results.[1]

The study has three major statistical flaws. These are:

Pseudoreplication. Statistical inference depends on samples being randomly selected and measured. Pseudoreplication is when samples are not independent, but are treated independently. In the study, turbine strings were sampled as a unit, but individual turbines were then treated as independent samples. This causes results to appear to be more significant than they actually are.

Inappropriate use of Chi-squared analyses with measured variables. Chi-squared analysis must be used with counts, not measurements of time. The authors of the study used Chi-squared analysis to assess significance of timed bird behaviors.

Multiple comparisons with inter-correlated variables without appropriate corrections. Conducting many statistical tests increases the chance that "significant" results will be found that are actually not significant (called a Type I error). The probability of Type I errors is also increased by conducting multiple tests with correlated independent variables. The authors of the study failed to use the appropriate corrections to account for this.

The study design also has several flaws that could compromise the reliability of results.

Nonrandom sampling of turbine strings. Constraints imposed by wind farm operators precluded the authors of the study from implementing a random sampling design. Although they ultimately sampled 75% of the turbine strings, their results cannot be extrapolated to turbine strings that were not sampled and the order turbine strings were added to the surveys could have affected results.

Differences in observer ability were not incorporated. The authors failed to incorporate differences in observer ability, or to demonstrate that they are insignificant.

Differences in scavenger removal were not incorporated. Assessments of avian mortality by locating carcasses should account for carcass removal by scavengers. Because scavenging is distributed unevenly across landscapes in response to vegetation conditions, its influence could confound patterns caused by other variables (e.g., turbine type, topography, etc.). The authors

---

[1] ~~Until these flaws are addressed, the conclusions of the study are premature.~~

assumed that scavenging rates were equal across the whole study area, regardless of local vegetation conditions. This may have affected the results of the study.

Technical Overview

This document reviews the methodological and statistical adequacy of "Developing methods to reduce bird mortality in the Altamont Pass Wind Resource Area" prepared by BioResource Consultants and published in August 2004.

The study has three major statistical flaws: 1) Pseudo-replication; 2) Inappropriate application of Chi-squared analyses to measured variables; 3) High probability of Type I errors resulting from using uncorrected post-hoc comparisons and inter-correlated independent variables.

[2]Individual turbines are not statistically independent *because they were surveyed in strings and turbines were not sufficiently separate to unquestionably assign a carcass to the turbine that caused the death. The appropriate unit of analysis is the turbine string. Consequently*, all analyses at the turbine level must account for this hierarchical structure in the data. Failure to do so, as occurred in this study, artificially inflates error degrees of freedom and F-ratios and makes effects appear more significant than they actually are. All of the ANOVA tests *using turbines as the experimental unit* performed in Chapter 3 suffer from pseudo-replication. Though less obvious, the Chi-squared analyses in both Chapter 7 and Chapter 8 should have also accounted for the structure of the data.

Chi-squared tests are often viewed as designed to compare observed data values to expected values derived from some model. This notion is only partly correct. The data must be counts of sample units possessing a particular attribute. The analyses in Chapter 8[3] apply Chi-squared frequency analyses to a measurement variable (i.e., *minutes of activity, top of p 254*). This is an incorrect application of the Chi-squared test.[4] *Consequently, the results presented in Tables 8-6 (minutes of perching), 8-7 (minutes of flight), 8-8 (mean flight height), 8-9 (mean distance from nearest wind turbine), 8-10 (flight time within 50 meters), 8-11 (minutes of perching, minutes of flying) are not valid. The appropriate analysis would have been to use an Analysis of Variance.*

The study furthermore has three methodological flaws that may alter the conclusions drawn from the study: 1) turbine strings were sampled haphazardly, 2) results were not adjusted for observer ability, and 3) adjustments for scavenger removal relied on other studies and did not account for differences in vegetation type or height.

Investigators added turbines to the study as they were made available by wind farm operators, not according to a pre-determined sampling design. Consequently, turbines were surveyed for different periods and turbines that may have had different characteristics were added to the pool of sampled turbines over time, potentially affecting study results. Although the authors specifically assert that the results of the study cannot be extrapolated to turbines that have not

---

[2] ~~The last paragraph of p. 47 (in Chapter 3) indicates the basic sampling unit was a string of turbines. This sampling scheme imposes a structure on the data where an individual turbine is a subunit of a string. Because of this structure,~~

[3] ~~attempt to~~

[4] ~~and consequently, almost all of the tests reported in Chapter 8 are invalid.~~

been sampled, this still represents a methodological drawback that was caused it seems by the wind farm operators.

The authors assert that it is not necessary to adjust for observer ability in reporting fatalities around turbines because they know that mortality will be underestimated. The study does, however, rely heavily on comparisons of numbers of birds killed at different turbine types. Observer ability could bias fatality rates up or down and consequently could alter conclusions about different turbine types and locations if different observers disproportionately surveyed one type of turbine or landscape position.

The authors assert that it is not necessary to adjust for scavenger removal because it is already known that mortality will be underestimated. The analysis relies on numbers of fatalities that are detected; differential scavenger removal throughout the study area would affect results of all subsequent analyses. If scavenging were uniform across the study area and among turbine types this variable probably would not affect conclusions, but the report contains no information to indicate that this is true.

Finally, the report did not address the existing literature on birds colliding with tall, lighted structures at night. Although raptors and grassland birds are presumably killed most by turbines at Altamont Pass during the day, collisions with migratory birds *may* also occur at night, *especially if taller turbines are installed in the future*. The study does not record whether turbines were lighted, and does not recognize that recommending that turbines be replaced on the tallest possible towers may actually increase mortality of migratory birds.[5]

We note here that the turbine operators at APWRA significantly hindered the design of the research project. Access was only granted to subsets of turbines from the start of the project, limiting the ability of the investigators to conduct random (or stratified random) sampling (or even complete sampling). Furthermore, the investigators reported that staff for turbine operators may have buried or hidden carcasses of birds. These factors must be eliminated to improve any future studies at APWRA. Although the investigators apparently agreed to do the study even under these conditions and therefore bear some responsibility for these limitations themselves, the greater interest of increasing knowledge about this system cannot be served if the turbine operators do not cooperate in research efforts and indeed, the investigators stated that they would not undertake a similar study if full access were not granted.

---

[5] *Regarding this issue, Smallwood and Thelander responded, "Some comments were irrelevant or confused, caused by lack of familiarity. For example, Review Team 1 appeared amazed that we neglected to discuss turbine lighting as an issue in the APWRA, but wind turbines in the APWRA are not lit. The issue of nocturnal migrants colliding with tall towers on the east coast of the U.S. cannot be extrapolated to wind turbines on the west coast, especially these small ones in the APWRA. The wind turbines in the APWRA are nowhere near the heights of communication towers, so citing the literature on collisions with communication towers would be irrelevant." By way of clarification, we suggested that the issue should be discussed given that Smallwood and Thelander recommend taller turbines in the repowering program. Groups of towers more than 200 feet tall would need to be lighted, and taller towers is currently the trend in the wind industry. This research has the potential to be applied elsewhere (despite admonitions by the authors to the contrary) and we therefore thought this issue should be addressed.*

Specific Comments

Further detail about our general comments and specific questions are raised in the following responses to aspects of each Chapter. Where relevant to the discussion, we have quoted text from the BioResource Consultants report (in *italics*) or reprinted figures. Our comments and questions are presented in Roman text. *In the final report, italicized text in our comments indicates revisions since the draft review.*

Chapter 1– Introduction

Page 7, "*In March 1998, the National Renewable Energy Laboratory (NREL) initiated research to address certain complex questions: What is the full extent of bird mortality in the APWRA? What are the underlying causes of the mortality? What role do bird behaviors play in mortality at wind turbines? Is mortality predictable at wind turbines with certain suites of characteristics? If it is, then can management strategies be developed to reduce mortality?*"

The questions raised by authors in paragraph 2 are valid, but could have been expressed as sequential components that, acting in concert, result in mortality of birds. The primary components are:

(**1) physical and operational attributes of turbines** (i.e., variables of: turbine model, turbine size, rotor diameter, tip speed, window, rotor-swept area/sec., tower type, tower height, blade color scheme, perch guard, low reach of blades, high reach of blades — as used in analyses for Table 7-2),

(2) placement of individual turbines (or turbine strings) relative to topography and prevailing wind (i.e., variables of: orientation to wind, derelict turbine, whether in wind wall, position in string, position in farm, turbine congestion, elevation, slope grade, physical relief, whether in canyon, slope aspect),

(3) ecological aspects relative to avian foods that may affect behavior of birds near turbines (e.g., variables of: edge index, rock piles, rodent control, cattle pats-grass, cattle pats-turbines, cottontails-grass, cottontails-turbines, and vegetation) and

(4) behavior and seasonal abundance of avian species near (within 300 m) of turbines in "rotor zone" (e.g., variables of: season of the year, time spent flying (20 behaviors — as in Table 8-2 and Table 8-18), flight height, distance from turbines, dangerous flights, time spent perching (26 perch structures as in Table 8-2).

Inasmuch as Step 1 (susceptibility) and Step 2 (vulnerability) result in Step 3 (impacts, i.e. mortality), the first two terms mean nearly the same thing ("capable of being affected" vs. "capable of or susceptible to" some variable. *Notwithstanding the existence of literature using these terms*, the use of these two terms as meaning different things is jargon that is not familiar to most readers, *even ecologists and wildlife scientists*. The authors should either provide a more detailed explanation of the difference between susceptibility and vulnerability or avoid this usage. Furthermore, although terminology is defined throughout the document, a glossary of terms in an appendix would be useful.

Page 8, first paragraph, last sentence: "*For this report we have combined the data from both research efforts.*"

[6]The non-random addition of study sites *during these two studies* confounds various analyses. This is especially important because of the year-to-year variation in measured fatalities.

Page 9, first paragraph under 1.1.2: "*The **placement and** operation of wind turbines can make birds vulnerable to wind turbine collisions…*"

This sounds as if birds can die at wind turbines (fly into them) even if turbines are not operating (blades not turning). Are deaths in this manner minuscule compared to deaths in moving blades or is this known?

Page 11, Section 1.1.4, first paragraph: *"….then the probability of an individual being killed by a wind turbine occurring on a particular environmental element would equal the proportion of the wind turbines associated with…*"

Not sure of the wording here. Would it ***equal*** the probability or just be associated with it?

Page 12, "*At selected turbines in the APWRA, we compiled data separately for bird behaviors, wind turbine and tower characteristics, fatality searches, fatality search results, maps of rodent burrow systems, and various other physical and biological factors."*

Both here and elsewhere, the authors fail to provide a rationale for arbitrary actions. How were turbines selected? The authors do note that the turbine selection was a result of the operational constraints of APWRA, "*Only about 28% of the APWRA's total wind turbine population was included in the project initially, due to limitations placed on access to turbines*" (also p. 12). Even with such limitations, the approach (although not stated in the report) seems to have been to survey all turbines to which access was granted, rather than selecting either a random sample or a stratified random sample that would have included representation of each turbine type. [7] This limits the conclusions of the study considerably, and, indeed, the authors acknowledge later that their models cannot be extrapolated to turbine strings that were not surveyed. Bird mortality could be higher or lower at turbines never studied. Furthermore, the addition of turbines to the search effort opportunistically creates *potentially* severe problems for the analyses. Because measured fatalities varied from year to year, the addition of a large number of a specific type of turbine during a "low" fatality year would give the false impression that a certain turbine type caused less mortality when data were pooled over multiple years.

Page 19, Table 1-1.

It is of interest that for 2 of the 3 turbine types (i.e., Bonus, Micon) "percent time in operation" is lacking and these two have higher mortality rates than other types except for the Kenetech KCS-

---

[6] ~~Data collected from two different research studies are not as robust as data collected for all variables in the same time period. This circumstance will limit the evaluation of interaction effects among some (maybe even) critical variables.~~

[7] *The authors provided a more detailed explanation of the sampling scheme in response to comments. While we still believe the sampling scheme was flawed (in no small part because of the turbine operators' unwillingness to provide access), this additional detail should be included in the methods section of the report.*

56, which has the highest number of carcasses associated with it in the APWRA (see Fig. 2-6). Table 1-1 lists percent time in operation as only 39% for the Kenetech KCS-56 type. It may be that operating duration was less for the Bonus and Micon turbines, or that the Kenetech KCS-56 just kills more birds because of its unique mechanical attributes. *The failure of turbine operators to provide data such as operation time compromises the ability of researchers to provide guidance to the wind industry.*

Comment deleted.[8]

Page 20, next to last paragraph: "*Within the APWRA study area, we performed focused studies involving smaller areas or select groups of wind turbines.*"

The authors never provide a rationale for how turbines were selected for the focused studies.[9] This description also gives the mistaken impression that turbines were the sampling unit, when the sampling unit was actually the turbine string (p. 47).[10] Neither this text nor the following chapters describes the selection process as being random. If turbine strings for the more focused studies were not selected randomly (that is, not every turbine string had an equal probability of being included in the focused study), then results of the focused studies on rodent burrows and bird activities cannot be extrapolated to the non-sampled turbine strings.

Page 27, last paragraph: "*To uncover and understand the patterns of bird mortality at a wind farm one must first interpret the influences on wildlife ecology that are caused by wind turbines. They are artificial structures installed in an otherwise natural setting that can have a profound influence on how arrays of interrelated landscape components function.*"

These statements suggest that an additional component to the four we presented earlier needs to be included. That is, without before and after turbine installation studies (which authors have acknowledged are needed) some of the ecological aspects are confounded inasmuch as the act of installing the turbines changes the food base that in turn affects bird behavior and may increase exposure to effects of turbines, even if the turbines are not operating.

Chapter 2 – Cause of Death and Locations

Page 28, last paragraph: "*The statistical tests included mostly one-way analysis of variance (ANOVA) and least significant differences (LSD) between groups. All LSD tests reported below were associated with P-values < 0.05. We also estimated Pearson's correlation coefficient for the distance of the carcasses and elevation of the tower base.*"

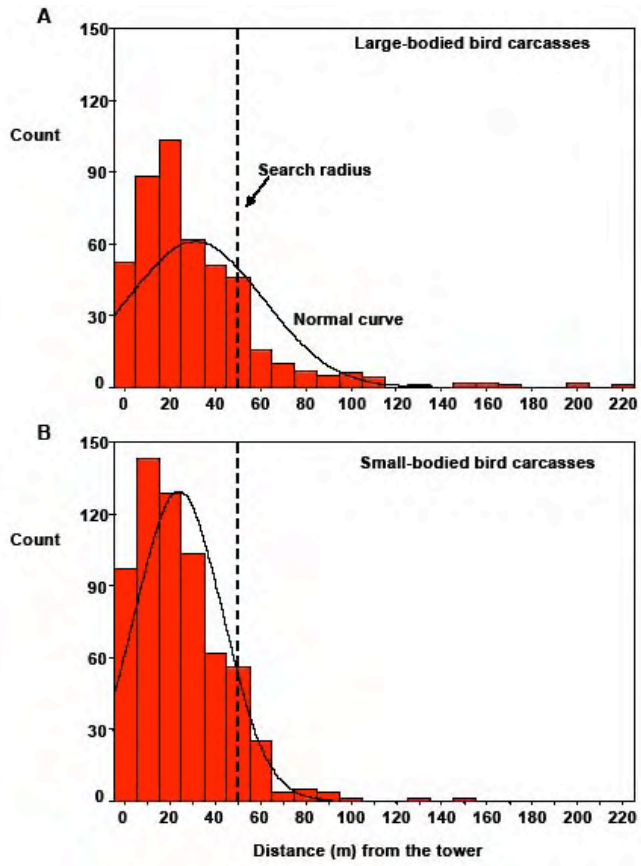
---

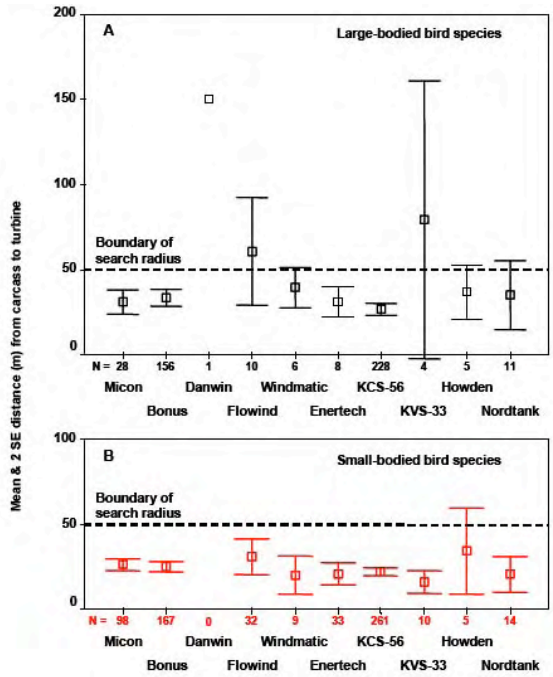
[8] ~~Page 20, first paragraph.~~
~~The authors give a general description of the study area but no vegetation map is provided. Such a map would be useful to show that either turbines are all in similar habitats, or to provide a means to visually assess the degree to which vegetation may affect survey results (e.g., lower detection of carcasses in taller vegetation or increased scavenger abundance near dense cover).~~

[9] *The authors provided more detail about the selection process in their response to these comments. Their description of a "systematic" approach needs to be fully described in the methodology sections of the report.*

[10] *In response to comments the authors claimed that both the turbine and the turbine string were the sampling unit. Both cannot be true. If turbine strings were sampled in a single visit then the turbine string is the sampling unit. This does not preclude analysis of individual turbines, but any tests must incorporate the nested nature of the data.*

The use of distances to the individual birds is pseudo-replication, which invalidates the results of the one-way ANOVA tests. The sampling unit is the string of turbines.[11] Consequently, the individual turbines are sub-samples of the strings. For this situation the appropriate analysis of the distances is a two-factor nested design. Furthermore, the authors use one-way ANOVA seemingly without regard for the underlying assumptions of the procedure, which include normality of error distribution and homogeneity of variance across variable levels. Figures 2-9 (p. 39) and 2-12 (p. 43) (reproduced below) illustrate violations of both assumptions.

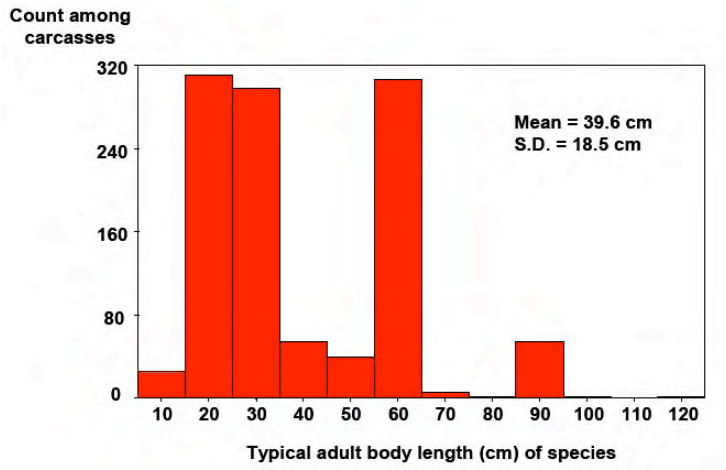




Use of LSD for *post-hoc* multiple correlations dramatically increases the chance of Type I error (i.e., labeling differences as significant when underlying population means are not). For example, in the discussion of blade tip speed (top of p. 42), with 10 categories there would be 45 possible LSD tests, which would lead to a Type I error probability of 90%.

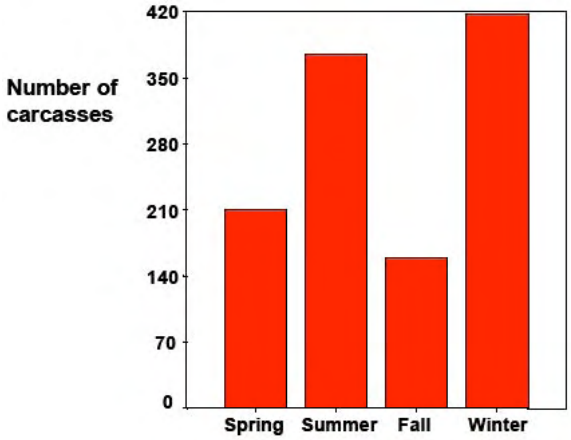
[11] *In response to this comment the authors claim that they can switch from the turbine being the sampling unit to the turbine string being the sampling unit depending on the analysis. This is not true, the data are collected in a nested format that must be accounted for in the analysis. A similar analysis can be conducted, but it cannot presume that each turbine is independently sampled.*

Page 29, Figure 2.1.



The choice of 38 cm as *a "natural beak" for dividing between* large and small body sizes seems arbitrary (see Figure 2-1 reproduced above). *To us the "natural break" occurs closer to* 50 cm. *An alternative could be* the median.[12]

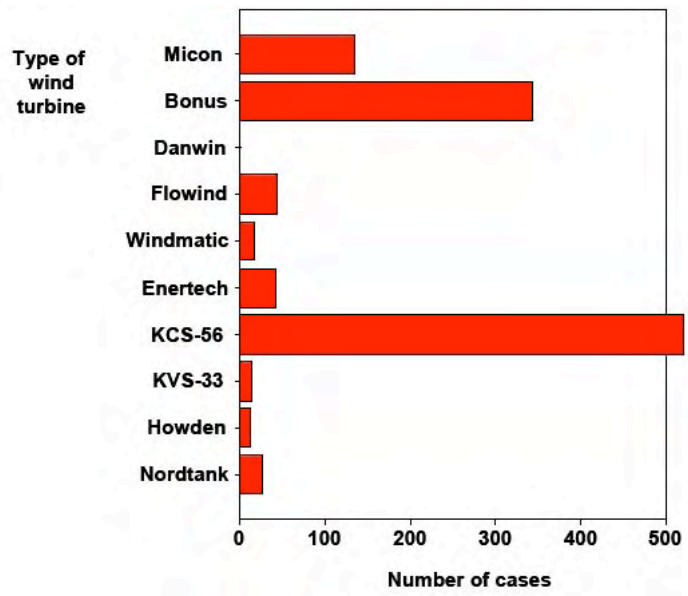
Page 36, Figure 2-5:



These results are presented as uncorrected counts. For comparability, the fatalities need to be expressed as carcasses per search effort, which needs to be clearly defined, e.g., hours or area or a combination, i.e., search effort per unit area. As raw counts, the reader does not know if the seasonal differences result from differences in search effort or seasonal changes in mortality.

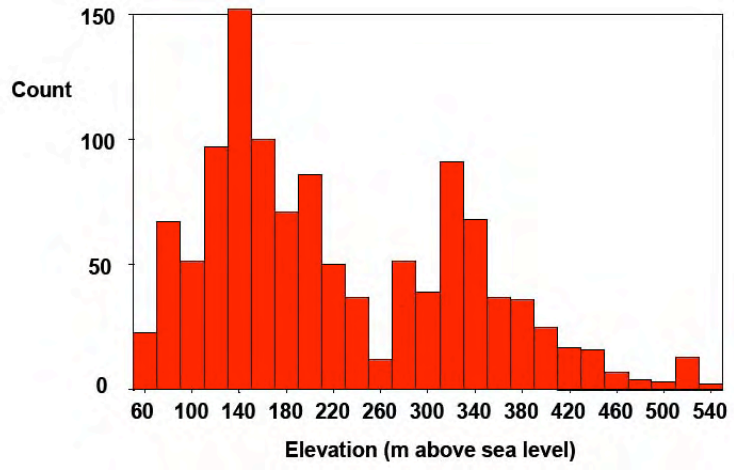
[12] *seem like more reasonable choices*

Page 35, Figure 2-6.



Results in Figure 2-6 should be expressed either as mortality per turbine type per search effort, or the graphic should express mortality as a percentage of the total mortality *and* this or another adjacent graph should depict the percentage represented by each turbine type so that readers can quickly assess whether some turbine types are associated with mortality disproportionate to their prevalence on the landscape. The figure as currently constructed could be misleading.

Page 38, Figure 2-8.

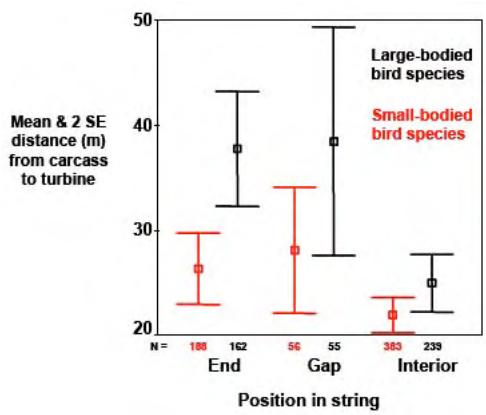


This figure and associated text should be expressed in fatalities per turbine at each altitude, or should express the mortality as a percentage of the total mortality and also should graph the percentage represented by each elevation class. The current figure does not provide much useful information because it is not clear if the pattern results from the elevational distribution of turbines or an inherent elevational pattern in mortality.

Page 41, Figure 2.11.

[13]The LSD tests described on p.38 indicate that the relationship between distance and height is not linear (i.e., the 43-m tower mean is less than the intermediate height towers.) *So the presentation of this figure, and the analysis it represents is meaningless.* In addition, the scatter plots show clear violations of the assumption of constant variation in distance across the tower heights.

Page 42, second paragraph: "*The distance of carcass locations from the wind turbines differed according to whether the wind turbine was located at the end, at a gap, or in the interior of a string of towers (ANOVA F=11.11; df = 2, 455; P < 0.001), and post-hoc LSD tests found distances to be 13 m greater on average from end and edge of gap turbines, compared to interior turbines.*" See also Figure 2-13:



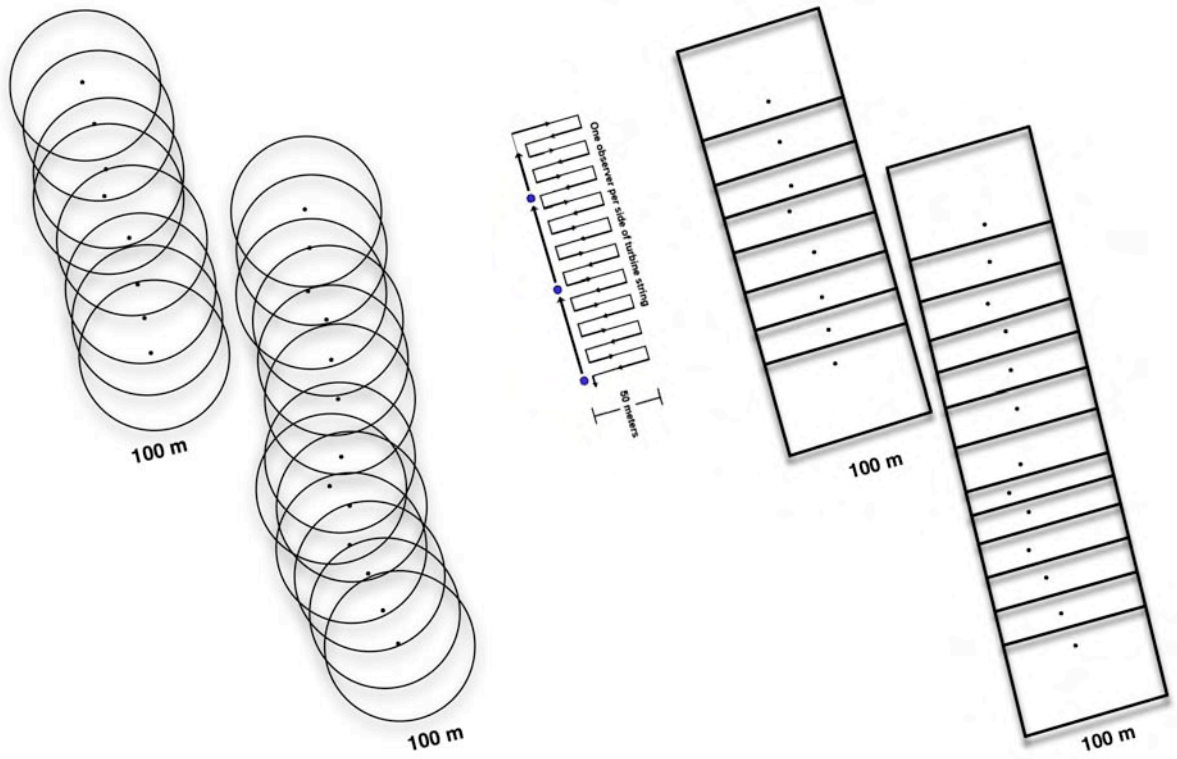
Assuming for the moment that the methodological problems (i.e., pseudo-replication resulting in inflated degrees of freedom, repeated *post-hoc* tests increasing probability of Type I errors) did not affect results, it is significant for other aspects of the study that dead birds were detected on average farther from end turbines and at gaps. This result *is probably spurious, resulting from*[14] a systematic problem allocating carcasses to turbines (and more importantly to turbine type). When a string is searched, those carcasses found along the string will be allocated to the nearest turbine (see p. 45). However, for end turbines and turbines at gaps, carcasses in a greater area would be allocated to the turbine. *This occurs* because turbines are often located *far less* than 100 m from one another. Therefore, a smaller total area allocated to turbines on the interior of strings and a greater area allocated to those at the ends and at gaps. *To illustrate this problem, we digitized the turbine locations from Figure 6-43 of the report. We used the scale to draw 50-m radius circles around each tower. This 50-m radius overlaps substantially for this set of turbines and we must conclude that the authors allocated mortalities to the closest turbine. To visualize the search area, we drew a rectangle around each turbine string that was 100 m wide and reaching 50 m beyond each end turbine. This represents the total search area for the string and is consistent with the authors' own figure (at left). We then divided up this total search area into rectangles that would be attributed to each turbine. Visual inspection of this figure leads to the conclusion that the search area for end turbines is far greater than for interior turbines*

---

[13] ~~Both halves of this figure are meaningless (i.e., R^2 =0.01) and inappropriate.~~

[14] ~~suggests that there is~~

(which would lead to both increased mortality estimates and to a greater carcass distance). Furthermore, it leads to the conclusion that because turbines are so closely spaced (at least in this example, which is not unique) attributing avian mortalities to a single turbine could result in many misattributions.



These figures are based on a figure in the original report and rely on the scale in that figure being accurate. The left side shows 50-m radius circles around two strings of turbines and the right side shows the areas presumably attributed to each turbine based on the description of the search methodology. The small inset in the middle is a copy of a figure by Smallwood and Thelander showing the survey methodology.

The extent of this problem could be ascertained by reporting the average distance between turbines in a string. If this distance is less than 100 m, then the conclusions of the turbine level analysis become difficult to justify because of misattribution of mortalities, and the observed greater mortality caused by end turbines could be the result of a larger search area. The authors acknowledged in response to this comment that the area searched at end turbines could be different than interior turbines but thought this difference was small. If our figure above correctly depicts the search areas, this difference could be as much double or triple the size, depending on inter-turbine spacing. The search area for each turbine should be calculated and reported to resolve this question.

[15]The authors *then* should adjust for the different search areas for interior vs. end and gap turbines. Adjusting for these differences may change the conclusion that end and gap turbines

[15] ~~End turbines are likely to be situated at the top of slopes (resulting in carcasses falling father away), which the authors use as an explanation for the increased distance to carcasses. However, this pattern is not likely to hold for~~

cause more mortality. It is possible that this observed relationship is merely a result of the greater search area for end and gap turbines, especially because turbines are often spaced closer than 100 m within strings (see e.g., Figure 6-41). This aspect of the methodology could jeopardize all of the turbine-level analyses in the report *because the <100 m distance between turbines will lead to some unknown degree of misattribution of mortality to individual turbines. Furthermore, our figures above show that the search area for turbines varies considerably, even on the interior of strings. As the authors prepare these data for publication in journals, a possible direction might be to use a GIS to depict the location of all avian fatalities, then describe the characteristics of turbine strings within a buffer around each fatality. This would avoid the potential problem of misattribution, especially for carcasses found equidistant from two turbines.*

Comment deleted.[16]

Page 45, first paragraph: "*We found birds beyond the 50-m search radius because the search crew members could sometimes see carcasses at these greater distances as they approached the 50-m termini of their transect segments.*"

Inclusion of these carcasses will result in a higher apparent mortality rate at those turbines where detectability is higher for any reason.[17] Because information about detectability was not gathered, it is not possible to assess the effect of this bias.

Chapter 3 - Bird Mortality

Page 46, Section 3.1 Introduction, paragraph 2.

We are not sure about mortality being expressed relative to megawatts (MW) of rated power generated per year. We can understand why authors chose to express mortality in these terms, but each type of turbine does not have the same relative effect on killing birds because of the inherent attributes of each type of turbine and we know that three different models seem to kill the most birds. Unless deaths per MW / year (or numbers of actual birds killed per year) can be clearly linked with "hours of rotating blades / year" for the particular type of turbine in question, the use of MW / year to associate with mortality may be misleading because rated power MWs do not kill birds, mechanical blades do. We therefore agree with the authors' response to this

---

~~gap turbines, and the often steep ground ("*precipices of very steep hills descending into ravines and canyons*") *could* also result in fewer carcasses being detected. The most logical explanation is that the implementation of the survey protocol, including the inclusion of carcasses located beyond 50 m, resulted in a greater effective search area for end and gap turbines.~~

[16] ~~Page 44, second paragraph: "*Carcass distances from wind turbines differed significantly by season of the year (ANOVA F=3.61; df =3, 630; P=0.013), and post hoc LSD tests revealed that fatalities in spring were significantly closer to wind turbines (mean = 19.6 m) than were fatalities during summer (mean = 24.8 m), fall (mean = 28.1 m), and winter (mean = 23.5 m)*."~~
~~Assuming that this pattern is real, it suggests that detectability of carcasses differs by season. Because carcasses greater than 50 m from turbines were included only as observed from within the 50 m search radius, their inclusion increases the average distance of carcasses from the turbine. The authors should investigate whether carcasses from > 50 m caused this pattern. That would be logical, because vegetation is usually tallest in the spring in Mediterranean grasslands. If this pattern does result from detectability differences, it would underscore the need to account for detectability in the study design and to account for seasonal variation in search effort.~~

[17] ~~(e.g., vegetation is lower, slopes are not steep, etc.).~~

comment that MW is essentially a proxy for rotor diameter and time in operation. With this additional understanding of the authors' motives, we accept this metric as a second best metric until information on time in operation can be obtained and combined with rotor sweep.

Page 47, last paragraph before Methods: "*Finally, we extrapolated our mortality estimates to the portion of the APWRA not sampled in order to characterize the range of likely project impacts per species and larger taxonomic groups*."

The non-random sampling scheme[18] *requires that* such extrapolation *be supported by evidence that the unsampled portion of the APWRA is well represented by the data that were collected*.

Page 47, Section 3.2 Methods.

Several details are omitted from the Methods section that directly affect any judgment about the validity of the data collection methods. 1) P. 47 indicates the 1,526 turbines were sampled, but gives no specifics about how the sampled turbines were selected. *If all were searched, this should be stated.* The same criticism applies to the additional comments that note that other groups of turbines were added periodically. How were these selected for inclusion? 2) No mention is made of any efforts to prevent double counting on successive visits. Found carcasses were flagged, but no mention was made whether that flagging was permanent throughout the course of the study. *Details about carcass removal or flagging to prevent double-counting should be included in the report.* 3) No discussion is presented about the search sequence. Were strings searched in the same order throughout the rotation?

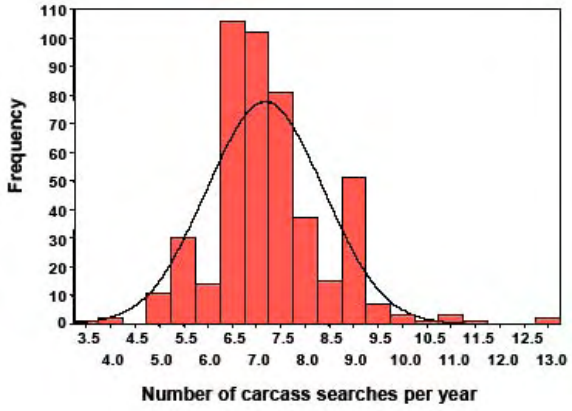
The search effort per turbine varied substantially by turbine.[19] How would this difference in effort affect the reliability of estimates of mortalities? Are these mortality estimates more conservative than if the same effort for the first group of turbines had been applied to the larger second group? And did the types of turbines in the strings differ between the two sampling periods? The sampling unit is described as the turbine string. Are turbine strings most always composed of the same type of turbine?

---

[18] ~~does not support~~

[19] ~~The authors acknowledge the disparity in searches for dead birds between the time periods (March 1998–Sept 2002 with 3 or 4 years for each month around 1,526 turbines) vs. (November 2002–May 2003 with only 1 search per month around 2,548 turbines) and they note that all turbine strings were searched every month.~~

Page 49, Figure 3-1.



Given the range of search effort per turbine per year (Figure 3-1), fatality estimates should be corrected upwards to adjust estimates for turbines searched less frequently. Authors assume that the same number of fatalities would have been found during a given year regardless of whether twelve searches or eight searches were performed. They acknowledge that fewer carcasses would be detected at turbine strings with fewer searches but do not adjust for this factor. What supports the assumption that the influence of search effort on carcass detection would not affect the subsequent analysis?

Page 49, second paragraph: "*Searcher detection and scavenger removal rates were not studied, because it had already been established that mortality in the APWRA is much greater than experienced at other wind energy generating facilities. We were unconcerned with underestimating mortality, and in fact we acknowledge that we did so. We were more concerned with learning the factors related to fatalities so that we can recommend solutions to the wind turbine-caused bird mortality problem. Thus, we put our energy into finding bird carcasses rather than into estimating how many birds we were missing due to variation in physiographic conditions, scavenging, searcher biases, or other actions that may have resulted in carcasses being removed.*"

Searcher detection and scavenger removal rates[20] *could affect* the results of the analyses.[21] Although both search detection ability and scavenger removal would result in underestimates of total mortality, these influences are not constant over space and time. Consequently, detection and scavenging rates would affect the results of all subsequent analyses if they are not constant over space and time. Both detection and scavenging rates are likely also affected by vegetation, which varies over space and time. Given that the remainder of the study involves multiple tests of the number of fatalities and the characteristics of the related turbines, the nonrandom, geographically varying effects of scavenging and detection are indeed centrally important to the study. This effect would be especially profound for turbines that were searched infrequently. Indeed, there could be massive scavenger losses, especially of small birds, even at the average 50-day period between searches.

---

[20] ~~are not inconsequential to~~
[21] ~~as implied by the authors.~~

Page 51, "*Because we did not perform trials to estimate searcher detection and scavenger removal rates, we relied on published estimates from other studies.*"

This adjustment results in simply inflating fatalities by a constant rate, but it does not incorporate the differences across space and time that *almost* certainly exist. This adjustment therefore does nothing to counteract the nonrandom influence of vegetation on detection and scavenging rates, or on observer detection ability to the extent that observers were not assigned to survey routes randomly. *If the authors believe that scavenging and detection rates are constant across the APWRA, they should provide evidence to support this assumption.*

Page 52, first paragraph: "*based on our experience with raptor carcasses in the APWRA, we did not believe that these scavenger removal rates were accurate for raptors, and we halved the removal rate estimates reported by Erickson et al. (2003).*"

What *specific evidence*[22] led *the authors* to believe that the scavenger removal rates were inaccurate for raptors?

Page 59, Figure 3-15.

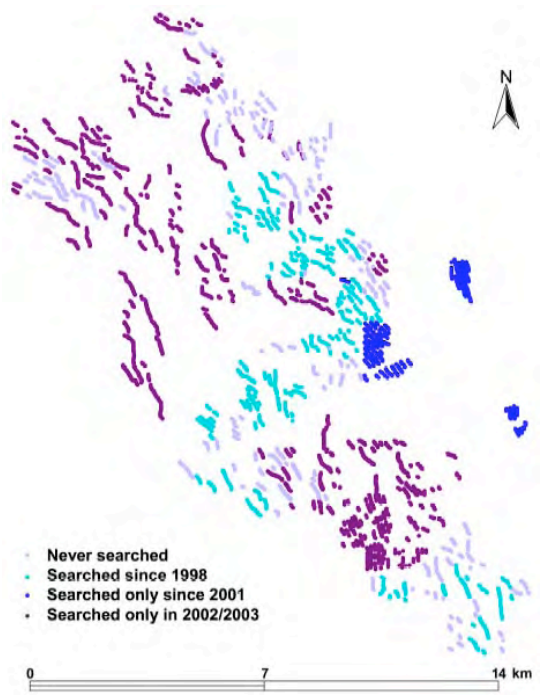


Figure 3-15 shows spatial distribution of survey effort. This figure does not appear to show a random sample. The authors should provide statistics about how the surveyed towers differ from the non-surveyed towers in key parameters (tower type, topography, elevation, turbine manufacturer, etc.). The non-random search pattern may influence other results. For example, elsewhere the authors report results for turbines that were searched four years without highlighting how the characteristics of those turbines differ from non-sampled turbines (e.g., turbine type, elevation, landscape position, etc.).

[22] ~~was the author's experience that~~

Page 64, First paragraph: It is stated that "*The mortality of all bird species combined increased steadily and significantly throughout the study, according to the comparison including all turbine strings searched for a least one year*."

Can some of this result be attributed to the increased familiarity of the investigators with the study areas, especially when areas were studied for 4 years? *If not, to what do the authors attribute this increase?*

Page 67, Table 3-3.

The right column has only turbines searched for 4 years. This is a geographically clustered sample, so it is unclear how results can be compared to the other turbines or to all other turbines at APWRA. The authors disclose that these turbines were within areas of rodent control, but do not describe the other differences from the other sampled turbines or the unsampled turbines.

Page 70, Table 3-9.

This table shows mortality per turbine string for two sets of turbines searched for different time periods. Because neither sample is random, and years of data are pooled (rather than comparing data from one year at one set to the same year at the other set), it is not obvious how the reader is to interpret this information.

It would be of interest to know how many deaths by species per year were associated with the total sum of "hours of operation / year" of all turbines and for each type of turbine in these two groups. Were there about equal proportions of each turbine type in each of these two groups? Because information like this is lacking, it is difficult to draw any conclusions from these data.

Page 75, Table 3-12.

This table provides results on a "per turbine" basis but the sampling unit was a string of turbines. *As we have illustrated above, the search methodology may have resulted in misattribution of mortality to individual turbines.*

Page 76: Regarding the nonrandom sampling scheme, the authors write "*This shortfall in our study was beyond our control, since the owners of the wind turbines allowed us access to various new groups of turbines at different times during the study.*"

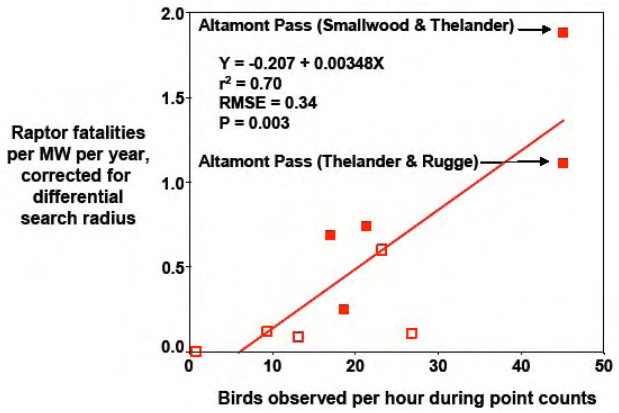
We are sympathetic in that the wind turbine operators did not allow access to turbines uniformly so that designing a random sampling scheme was difficult, if not impossible. This remains, however, a shortcoming of the study. The authors should have restricted all comparisons of mortality rates to turbines that were sampled during the same period and within a random sampling framework. *If a rationale can be presented to support the nonrandom sampling scheme, then results should be compared for similar time periods.*

Chapter 4 – Impacts from Wind Mortality

Herein the authors attempt to compare mortality rates at wind farms as determined in different studies. The authors make many assumptions about scavenging, detection, and search interval

that cannot be verified *because they did not collect information about the influence of these factors in their study.*

Page 79, Figure 4-2.



In the regressions of raptor fatalities by birds observed per hour it seems that most of the explanatory power comes from the current study and its precursor at Altamont pass. Furthermore, the two high fatality estimates constitute partial duplication of the same data, because it seems that the data from Thelander and Rugge are incorporated into Smallwood and Thelander.

Page 83 and following, Figures 4-5 through Figure 4-7.

We are not convinced that the mortality rates from the different studies can be compared. Furthermore, the use of "bird observations" as a metric is not particularly useful because it is already apparent from the data that avian species are not all equally vulnerable to collision with turbines.

Chapter 5 – Range Management Issues

This chapter has several methodological problems: 1) The turbine level analyses are pseudo-replication and the analyses using one-way ANOVAs are therefore not valid. 2) The two transect types are paired by turbine string and should be analyzed using an analysis method that accounts for the pairing, e.g. a block or repeated measures design. 3) The use of LSD tests results in a very high probability of Type I errors. The chapter does not contain information about the sampling scheme (e.g., is it randomized, is it stratified by turbine type?).

Aspects in this chapter follow our component framework #3. Even without operating the turbines, their establishment modifies the local environment by changing the food base that may affect the behavior of birds and cause some low-level mortality. The effect on behavior, in turn, may predispose birds to be hit by turbines when they are operating and cause higher levels of avian mortality.

Page 90, Section 5.2 Methods.

Unfortunately, the amounts of lateral edge and vertical edge were characterized as "some", or "lots". If we understand the layout correctly, these variables could have been quantified in terms of *x* meters of lateral or vertical edge. Also, please describe *in the report* the difference between ridge crests and ridgelines. Where these topographic classifications made with automated Geographic Information System tools or based on judgments in the field or another technique?

Page 90, Paragraph 5.

How did the authors determine that cottontail pellets were especially abundant?[23] Is there any citation or precedence that connects rabbit pellets with abundance? Fecal abundance as an index of animal abundance is not always reliable.

Page 91, Section 5.2.1, first sentence.

The text "March, 1998" is missing after the word "from".

Chapter 6 – Rodent Burrows

Page 111.

How did the authors choose the 571 turbines to map rodent burrows? This should be described in terms of turbine strings because strings are still the sampling unit. The choice of turbine stings appears to have been arbitrary, perhaps guided by an idea of a stratified random sample of turbine strings associated with different raptor mortality, physiography, and rodent control. If the sample was, indeed, a stratified random sample this should be stated clearly with a description of how many replicates of turbine strings were associated with the three criteria (i.e., range of raptor mortality, physiographic conditions, and level of rodent control). If not, then the method for choosing these turbine strings should be clearly described.

Page 124, first paragraph: "*Eleven strings of wind turbines were selected for seasonal monitoring purposes…*"

Were these strings selected randomly? The numbers (and types?) of turbines in the strings were widely variable ranging from 3–35 turbines, and 1 to 3 or 4 strings per group. How comparable were these groups?

[23] ~~Random transects?~~

Page 140, Fig. 6-25.

This photo suggests that type of tower, at least, was not uniform within groups of strings. Tower type seems important; did inclusion of different types of towers have any effect on results?

Page 149, first paragraph, last sentence.

Did the type of turbine have any measurable effect?

Pages 151 to 161, Fig. 6-34 through 6-44

Again it seems important to recognize the large disparities in numbers of turbines (and perhaps types of turbines?) among these sites. Is it possible that unadjusted mortality of species is related to number and type of turbines and not rodent control treatments? Is there no way to test for interaction effects?

Chapter 7 – Predictive Models

Page 186, Section 7.2.2, Analyses, Paragraph 3.

"Search effort" is defined as m^2 times number of years during which surveys were made.[24] *This definition should also include the number of minutes visited during a year (or the number of standardized visits during a year). Number of years times area is not a complete description of search effort unless each turbine was searched the same amount of time during each year. As it stands, search effort (quantified in terms of hours of searching per unit area) is not presented and perhaps not available. This shortcoming affects the credibility of mortality estimates, inasmuch as any differences in numbers of birds found may be related to search effort and not to differences in other variables (e.g., turbine type).*

Page 186, paragraph 4: "*Figure 7-1B illustrates the inverse power relationship between a fatality rate and search effort, which casts doubt on the reliability of a simple conversion of fatalities to fatality rates (mortality) for inter-string (or inter-site) comparisons and hypothesis testing.*"

---

[24] ~~How and when, does the amount of time spent on transects looking for carcasses (or number of visits per year) factor in?~~

Doesn't fatality rate imply deaths per unit of time? Not unit of area? And even more appropriate may be to express as deaths per hours of turbine operation *(if available)*, because flying into moving turbine blades is the primary cause of bird deaths.

Page 188.

The predictive model is flawed. The variables examined are clearly not independent and so summing the accountable mortality values across variables (p. 188) must necessarily overestimate the predicted impact. All model results are suspect because of this flaw. Furthermore, this is a complex study with many potential confounding factors, yet the development of the predictive model strikes us as simplistic and fails to account for such effects.

Page 189, Figure 7-2 through 7-4 and 7-8 through 7-18.

It is not always evident what the figure caption "count" means in these figures. It seems to be number of turbines, mostly.[25]

The words "search effort" are used in captions for measurements that really *are* the number of years during which searches were made, multiplied by a search area. This measurement ignores the number of visits (or hours) that each area was searched and assumes that there would be no variation in the number of dead birds found with greater or fewer visits during a year. *It does not seem prudent to assume that no variation exists in the number of dead birds found with greater or fewer visits during a year.*

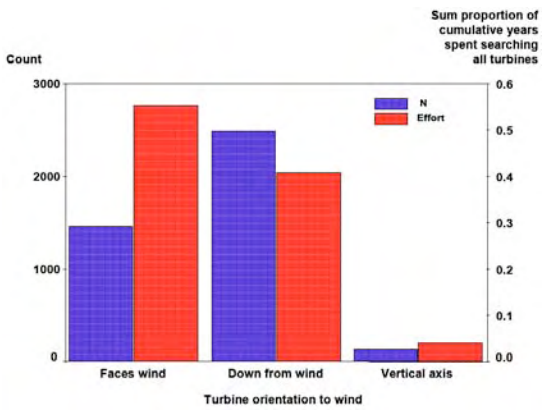
It is also peculiar that in this analysis, the authors use rotor swept area as a measurement of turbine size, rather than MW rating. We prefer the rotor swept area as a method of standardization.

Page 205, Table 7-3.

On what basis were the two groups "Hawks" and "Raptors" segregated? *The AOU checklist (7th ed.), which we consider the standard reference for bird classification, does not identify a group "Raptors". The Order Falconiformes, which is designated "Diurnal Birds of Prey", includes the Family Accipitridae, which is designated "Hawks, Kites, Eagles, and Allies". Owls in the Order Strigiformes include the Family Strigidae that includes an important study species, Burrowing Owl. As currently written the report does not provide sufficient information to know taxonomic designations of the species that comprise "Hawks" and "Raptors".*

Figures 7-2 thru 7-4 and 7-7 thru 7-13.

[25] ~~Is that correct? It also seems that~~



These figures are all misleading. The adjacent bars suggest direct comparisons, yet the opposing scales are not comparable. As an example, in Figure 7-8 (above) the left scale (count) maximum is 3,000, which is 74% (=3000/4675) of the total number of turbines, whereas the right hand scale maximum is 60%. This imbalance of scales makes the effort bars taller than they ought to be. *This information can be presented together in the same graph, but the scales should be comparable because it presents a comparison to see whether any particular turbine orientation was over- or under-sampled relative to its incidence.*

Tables 7-1 through 7-3.

The results for a large number of the Chi-squared tests in Tables 7-1 through 7-3 that are suspect because too many of the expected values for individual categories presented in Appendix C are less than 5. The authors mention this fact on p. 206 but present the tests anyway. The test ought not to have been done.

The individual turbines within the same string are not independent and just as in the ANOVAs this fact needs to be accounted for in the Chi-squared analyses. In the analysis of seasonal differences the repeat visits are not independent and that needs to be accounted for also. *This is not to say that analyses cannot be completed, but they must account for the nested nature of the data.*

Comment deleted.[26]

Page 215.

The conclusion about rock piles does not seem to be adjusted for different mortality rates in different years, and for all the other factors that differ between the samples?

Page 222, Second paragraph.

The model was not validated by withholding a subset of data then using those data to check the accuracy of the model. Such validation would be desirable. [27]

---

[26] ~~Page 210, Wind Turbine Attributes, First paragraph.~~
~~What was the rationale for excluding turbine model from the tests?~~

On p. 222, the authors ask the reader to assume "our predictive model are relatively precise" yet provide no justification for the assumption. The authors appear to be ignoring the possibility of false positive predictions. There are two aspects to a predictive model; correctly identifying as "dangerous" turbines where fatalities were found (called sensitivity) and correctly identifying "non-dangerous" turbines where fatalities are not found (called specificity). While the model for Golden Eagles has a sensitivity of 82%, the specificity is 50%. The authors argue that the model identifies a collection of "dangerous" turbines. The model's ability to correctly identify a turbine that actually has an associated fatality depends on both the sensitivity and the specificity.

A calculation, using Bayes Theorem, can be used to answer the question, what is the likelihood that more searches would "*add many more wind turbines to the pool of wind turbines documented to have actually killed members of each species*?" (p. 222, line 4). To perform the calculation, one must assume an average fatality rate. Here is a table of hypothetical fatality rates for and corresponding likelihoods that a "dangerous" turbine will be found to have killed one or more Golden Eagles.

Fatality rate	0.001	0.01	0.05	0.1	0.25	0.5
Likelihood	0.002	0.016	0.079	0.153	0.352	0.619

To interpret this table, consider this example: With an average fatality rate of 5% (.05 in the table), prior to applying the predictive model one would expect about 5% of turbine searches to produce a Golden Eagle fatality. If the searches were restricted to "dangerous" turbines (as identified by the model) then one would expect to find Golden Eagle fatalities in 8% (.079 in the table) of the searches. Thus, the model increases the chance of finding Golden Eagle fatalities from 5% to 8%. *Then* one can conclude about 92% of the turbines identified as "dangerous to Golden Eagles" will **not** have an associated fatality.[28]

Page 223, Table 7-8.

[29]Because the physical attributes of operating turbines manifest the lethal force in bird deaths, it may have been instructive to use only those variables identified in framework component #1 to develop a predictive model with AIC methods. Similarly the same approach may be applied to the other framework components as outlined at the beginning of this review to determine which variable(s) contributed to bird mortalities. From results of the four predictive models, perhaps an overall model could be developed that used the most important variable(s) from each component model.

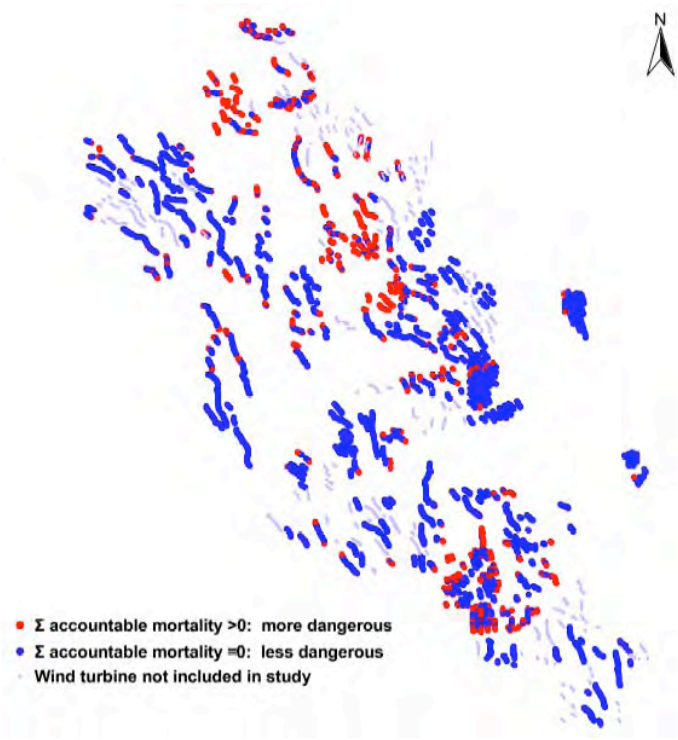
---

[27] ~~It is not clear if a subset of data was withheld from the data that was used to develop the empirical models so that they could be validated. On Page 243 (first paragraph) it is stated that "…. 472 strings that were used for developing the model." This implies that an effort was made to validate the model. Perhaps this could be confirmed and further elucidated.~~

[28] ~~and hence would not be "dangerous"~~

[29] ~~The authors appear to be selective about inclusion of "important" variables. Using the Golden Eagle as an example. The variable 'Part of wind wall' (p < .05) yet the variable 'Tower height' (p < .05) was not. The accountable mortality for 'Position in string' was reported in the table as 19 while in Appendix C it is given as 18. There were several other similar occurrences with other variables in the list.~~

Page 237, Figure 2-27.



The authors conclude that "dangerous" turbines are distributed "relatively narrowly" across the APWRA. The distribution in the maps does not seem narrow to us (see Figure 7-27 above for red-tailed hawks).

Page 244.

A typographic error seems to have resulting in a duplicate discussion of rock piles near the bottom of the page.

Chapter 8 – Bird Behaviors

Comment deleted.[30]

Page 247, paragraph 2.

Was there a random order of choosing which plot was sampled next?

Page 254.

---

[30] ~~Page 246.~~
~~What is spatial distribution of 61 study plots? It seems that they are associated with turbine strings that were chosen arbitrarily, meaning that the behavioral study plots were not selected randomly. Consequently, behaviors from these plots cannot be extrapolated to other areas within the APWRA.~~

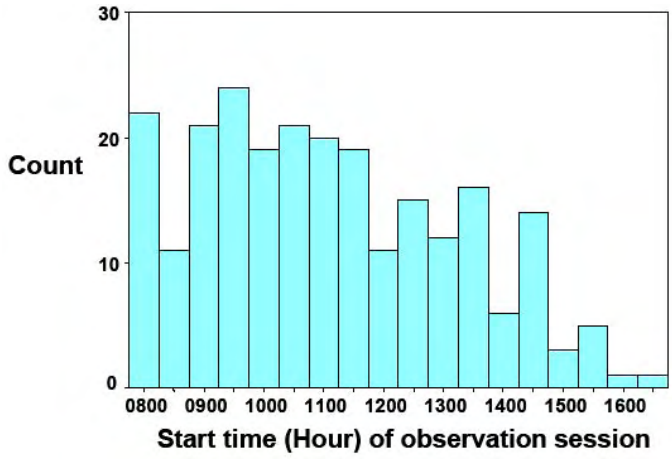
The analysis of a measured variable, such as minutes, using a Chi-squared analysis is invalid. The Chi-squared tests are not invariant to changes of scale, i.e. the results would change if the data were expressed in seconds or in hours. (Using seconds would make the tests more significant and using hours would make the tests less significant.) This invalidates almost all of the tests performed here. *The appropriate test would have been to use an Analysis of Variance.*[31]

Turbine level analysis involves pseudo-replication because turbines were sampled as strings. *A two-factor nested design could allow for investigation of turbines while recognizing their place within turbine strings.*

Page 256, paragraph 3.

Our experiences *(including over forty years conducting observational studies of birds in the field duly documented in the literature)* reinforce the authors' conclusion that the observation time for the sessions was minimal at 30 minutes. Other observational studies with which we are familiar found that 2-hour blocks, randomly assigned throughout the entire period available to observe birds, were adequate to determine reliable patterns of bird activity.

Page 257, Figure 8-3.



The behavioral surveys are biased toward morning observations and include no summer observations (Figure 8-4). This survey pattern may influence results, especially by underestimating behaviors occurring when conditions are hotter (later afternoon and in summer).

Page 331, Section 8.4.5., paragraph 5.

We agree with the authors that BACI study designs will be required to sort out effects of some variables, because the mere presence of the turbines as they are installed affects the environment and in turn affects bird behavior, which is a variable related to mortalities.

Page 332.

---

[31] ~~On p.254 the authors indicate the observed values used in the Chi-squared tests were either minutes or behavioral events. It seems that no Chi-squared tests were based on behavioral events.~~

The authors argue for taller turbines to repower at APWRA, but they seem not to consider how this will influence mortality rates for migratory songbirds. Turbines greater than 200 feet will require obstruction lighting, which is associated with increased mortality of nocturnally migrating birds. *Although mortality of night migrating songbirds at lighted towers is generally less in the West, the conclusions of this report may be applied to other areas of the country (notwithstanding the authors' best efforts to admonish readers that the study is not meant to be extrapolated to other situations).*

Chapter 9 – Recommendations

Unless and until the methodological and statistical problems described above are resolved, the conclusions reached in the report must be considered premature *from a statistical perspective. We note, however, that reanalysis of these data with other statistical methods may indeed result in similar conclusions and that the information in the report is more detailed than previous research efforts. Even as reanalysis is undertaken, the working hypotheses presented by Smallwood and Thelander can be used to implement adaptive management actions (e.g., repowering) that will test these hypotheses while potentially reducing avian mortality from collision.*

Appendix A – Measuring Impacts by MW

Page A-1, first paragraph, first sentence.

The term "confusion" may be correct but the term "complexity" also depicts the situation. It may be that it is inappropriate to try to compare mortalities between wind generating facilities because each facility has unique features for each of the four framework components, thereby preventing any reliable comparison between facilities. Conversely, the individual turbine type (and its attributes) is of utmost importance in how many birds are killed *(according to the data in tables 7-1 to 7-3)*. Variables of the other three framework components (that we outlined at the beginning) can be neutral or either increase or perhaps decrease the predisposition of birds to being killed by the turbines. But, each wind farm site is unique with specific effects of variables that cannot be fully replicated.

Comment deleted.[32]

Page A-6, paragraph 2.

Another reason to question the use of fatalities/ MW/ year is that the MW is a constant (as stated), but that the number of fatalities is variable over time and depends on amount of search effort, so that inadequate search effort in a given year will weaken the reliability of results. Authors further acknowledge (Page A-7) that this is likely that areas around wind turbines that were not searched over a long enough period will not provide a robust estimate of mortality.

---

[32] ~~Page A-2, Section 3.0 Results, paragraph 5.~~
~~This statement reinforces our earlier comments (See comments page 46, Chapter 3) that attempting to standardize by basing number of fatalities/ MW/ year instead of number of fatalities/ turbine/ year does not provide insights about effects of individual turbine types, which is the killing structure. Actually, Figure A-3b, Page A-5 depicts an even more direct metric of what kills birds — the area of rotor-swept / year, which again relates to turbine type, size and blade speed.~~

Page A-11, Section 4.0 Discussion, paragraph 1.

Indeed, it may be more convenient to express mortalities on the basis of MW / year, but information on which type of turbine and supporting structure that kills birds is not emphasized. *The authors have subsequently endorsed the use of fatalities per kWh as a metric. We are more comfortable with this, because it implicitly embodies the mechanical attributes of a turbine and duration of operation.*

Page A-12, paragraph 1.

Alternative ways to express mortalities may be by use of actual physical attributes that are involved in killing birds (i.e., those variables listed in framework component #1).

ATTACHMENT B

Review Team #2

A Statistical Review of 'Developing Methods to Reduce Bird Mortality in the Altamont Pass Wind Resource Area' by Smallwood and Thelander (August 2005, Publication #500-04-052)

Original draft: **June 26, 2006**
Revised draft: September 11, 2006

Commissioned by:

The California Institute for Energy and Environment on behalf of the California Energy Commission

Accessing the report:

The Smallwood/Thelander report can be downloaded from the following web site: www.energy.ca.gov/pier/final_project_reports/500-04-052.html

Comments regarding revision of the original June 26, 2006 peer review:

The original version of this review was submitted on June 26, 2006. On August 29, 2006, the three groups of reviewers received the Smallwood and Thelander responses to their original review. Responses to the Smallwood and Thelander comments to this review are addressed in two ways. Additions, such as this, are italicized, and deletions will be indicated in footnotes. Be aware, however, that the original review did contain a few italicized phrases. In the footnote, the deleted parts will have ~~strikethrough~~ lines through them.

Purpose of the review:

The purpose of this anonymous review is to objectively evaluate the statistical analysis used in the report 'Developing Methods to Reduce Bird Mortality in the Altamont Pass Wind Resource Area' by Smallwood and Thelander (August 2005, Publication #500-04-052) created for the California Energy Commission (CEC). More specifically, the reviewers will evaluate whether the data collection and statistical analysis methods are scientifically sound and appropriate for achieving the report goals set forth by the CEC. Policy recommendations are not to be reviewed.

This review was created by "Review Team 2". Review Team 2 was comprised of three individuals working together. One member's professional training is as a biostatistician, another as a wildlife ecologist, and the third as an environmental engineer.

Abbreviations and notation:

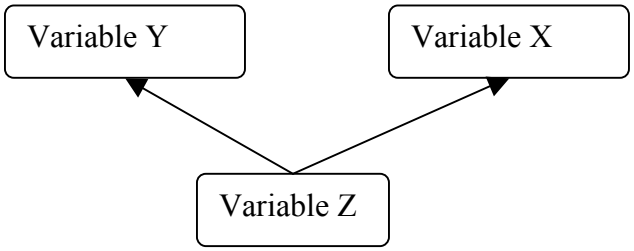
Certain abbreviations and notation will be used in this review:

- (p.73, par.2) = Page 73, paragraph 2. The first paragraph is considered the text at the top of the page, regardless of whether the text's paragraph began on that page or the previous page.
- Altamont Pass Wind Resource Area = APWRA

Terminology commonly used:

When evaluating a report based upon data analysis, understanding certain statistical terminology and ideas are necessary. We hope the following explanations will help the readers of this review better understand some basic statistical concepts.

- **Confounding:** Confounding is statistical terminology for when the researcher cannot tell if variable Y is caused by variable X or variable Z. This confusion arises because variables X and Z are related. Sometimes a dataset only has the X and Y variables and the analysis statistically indicates that Y and X are associated. It may be, however, that Y is caused by the unmeasured Z variable and that Z also causes X; X does not cause Y. Consequently the word "associated" is often used in statistics because variable X may not cause Y but it is associated with Y. (See the following figure.)



- **Null and Alternative Hypotheses:** The null hypothesis is considered to be the status quo and is to be retained by the researchers unless the data suggest strongly otherwise. (Much like the saying, "Innocent until proven guilty beyond a reasonable doubt.") The alternative hypothesis is often considered the "research hypothesis" and is in contrast to the null hypothesis. For example, a null hypothesis would be that wind turbine model A has the same bird mortality rate as model B. The alternative hypothesis, which must be stated before observing the data, could be that turbine model A has a lower bird mortality rate compared to model B. Or the alternative hypothesis may be that the mortality rates for the two models are different. If the probability of getting the collected data is considered to be suspiciously improbable assuming the null hypothesis is true and the collected data would have been more probable if the alternative hypothesis is true, the null hypothesis will be rejected in favor of the alternative hypothesis. This finding is considered "statistically significant".
- ***P*-value:** This is a probability that states the likelihood of getting the sampled data, or data even more unlikely, if the null hypothesis were true. Statistical calculations are performed assuming the null hypothesis is true. Typically, if the *P*-value is equal to or less than 0.05, there is considered to be enough doubt to believe that the data were not generated with the null hypothesis being true, but instead the data were generated with the alternative hypothesis being true. It is conventional to consider the findings to be "statistically significant" if the *P*-value is less than or equal to 0.05. Statistical significance is a measure of probability concerning the null hypothesis, not of the magnitude of the effect being investigated.
- **Type I error:** This error is often called a "false negative" or a "rejection error". This occurs when the null hypothesis is rejected in favor of the alternative hypothesis even though (unknowingly) the null hypothesis was true. If using a cut-off of 0.05 for the *P*-value, for every 20 statistical tests performed when the null hypothesis is actually true we would expect to commit one Type I error.
- **Type II error:** This error is often called a "false positive" or an "acceptance error". This occurs when the null hypothesis is retained although the alternative hypothesis is true. The probability of this event occurring depends upon many factors.
- **Power:** The probability of rejecting the null hypothesis when the alternative hypothesis is true is called the power of the test. This is the flip-side of a Type II error and can be thought of as the sensitivity of a statistical test – the likelihood of

giving a true positive. It depends upon many factors, but primarily the magnitude of the effect (if it exists) being tested for, the amount of natural and measurement variation in the variable being measured, and the sample size.

Report Overview:

The authors were challenged with a broad question, a large geographic region, limited access, and a very large number of wind turbines of various models. Their report is a quantitative exploration into the variables associated with bird mortality. Almost certainly, at least some of the many variables measured are truly linked to bird mortality – birds are certainly being killed in the APWRA. The reviewers have little confidence, however, that this report has scientifically been able to determine which of those variables are important.

Much effort went into collecting massive amounts of data; however, the authors should have focused more effort on study design and collected their data more wisely. Likewise, the data analyses could have been more thoughtful and sophisticated. The statistical analyses are applied in an automated manner that fails to fully utilize the data at hand and ignores potential confounding of variables. It seems like many of the statistics were calculated just for the purpose of producing statistical tables to the point of data dredging. Furthermore, the mathematical assumptions behind statistical tests like one-way ANOVA are ignored and thus the reported P-values should be treated as approximations. The large number of statistical tests likely resulted in many Type I errors; therefore, statistically significant findings should be treated more as an indicator of what should be explored in future studies.

Furthermore, the reviewers concur with the authors that their study does not give accurate estimates of actual bird mortality.

Broader comments addressing specific questions:

- *Was the statistical methodology used on the analysis consistent with accepted methods used in other biostatistical analyses?*

No. A very large number (>1000) of univariate chi-square tests is not common in biostatistical analyses. Interpretations of the univariate tests are clouded somewhat by shared variation among the explanatory variables (turbine attributes).

Chi-square analysis assumes that the counts are exact and not estimated counts[1]. *Adjusted counts (adjusted for scavenging and detection rates) are frequently used throughout the report and it is not always obvious whether adjusted or raw counts are being used to calculate the statistics. If adjusted counts, rather than raw counts, are used in a chi-square test, it* is not clear how *the uncertainty in adjusted counts* would influence the conclusions reached on the numerous chi-square hypothesis tests.

[1] Original review text before considering the Smallwood and Thelander response: "Chi-square analysis assumes that the counts are exact and not estimated counts ~~*as* they are in this study~~. It is not clear how ~~this~~ would influence the conclusions reached on the numerous chi-square hypothesis tests."

In estimating mortality rates for specific species due to wind turbine collisions, almost half (28) of the 60 species or groups have fewer than 5 fatalities reported in the entire project. And yet, mortality rates are still estimated and reported.

Although the study design is observational, the authors quickly jump to *confirmatory analysis methods such as* hypothesis testing and parametric analysis without exploring their datasets thoroughly. *The application of the exploratory data analysis methods such as those developed and popularized by John W. Tukey, Frederick Mosteller, and others would have been more appropriate.* What distinguishes the 20% of the turbines where fatalities were discovered from the 80% without fatalities?

Most biostatistical analyses based on analogous (but often smaller) datasets rely on multivariate models such as a general linear models, logistic or Poisson regressions, or discriminant function analyses. The authors explain that limitations in their sampling precluded these more sophisticated multivariate analyses, but this may not be true if the authors (a) carefully screen their variables to reduce the number of parameters in their models, and/or (b) clearly restrict their inferences to the turbines actually sampled.

- *Were the technical approaches used in the research appropriate for achieving stated goals?*

The stated goals for this study were to 1) quantify bird use, 2) evaluate the flying behaviors and conditions associated with flight behaviors, 3) identify the relationships between bird mortality and various explanatory variables, and 4) develop predictive, empirical models that identify areas or conditions associated with high vulnerability.

The sampling programs designed to address the first goal are not best for goals 3 and 4. (The authors used a separate approach for goal 2, which unfortunately omitted the summer.) Consequently the study design is not well crafted for achieving objectives 3 and 4.

- *Were the data collection and analysis methods and assumptions clearly stated, valid, and reliable? Were there any errors or, flaws? Were any relevant factors missing?*

The authors used *recently developed* protocols for carcass searches, *and standard* bird observations, rodent surveys, etc. to obtain the ecological data, and the technical approaches were appropriate.[2] However, the methods used to estimate bird mortality rate are suspect because (a) neither scavenging rate nor observer detection probabilities were measured empirically, values were pulled from the literature – in

[2] Original review text before considering the Smallwood and Thelander response: "The authors used ~~standard~~ protocols for carcass searches, bird observations, rodent surveys, etc. to obtain the ecological data, and ~~generally~~ the technical approaches were appropriate."

some cases based on studies in different locations; (b) a 50m search radius is insufficient to detect an adequately high percentage of carcasses, especially given the lack of rigorous data on detection rates of carcasses beyond 50m from a tower; (c) the authors adopted adjustments to published scavenging and detection rates based on assumptions that are inadequately supported with observation. For example, the following three assumed adjustments are problematic: (1) "halving" the scavenging rate for raptors, (2) elevating the scavenging rate by 10% for the 2nd set of turbines because they were checked much less frequently than those in the study from which scavenging rates were used, and (3) assuming detection rates were equally high beyond 50m, where the crews did not search rigorously. Most of these inadequacies biased mortality estimates by an unknown amount and direction. For comparative purposes of a single species' mortality rates across turbine and location attributes (Chapter 7), these biases may operate *roughly* similarly across the variables and therefore may not undermine the analysis. For examination of impact (Chapter 4), however, these biases are very problematic indeed.

- *Was the study design scientifically sound? Was there sufficient time to conduct the study (e.g., time for conducting searches, time for assessing seasonal effects)?*

The description of the sampling – how well it yielded a representative sample of all the turbines in APWRA – was inadequate, hindering our ability to rigorously assess the sufficiency of the sampling itself. Access to study the 2nd set of turbines was granted too late and the study's duration (and hence the length of their reexamination) was too short to be of maximum use to the overall project. The bird behavioral sampling did not include most of the summer season.

The sampling design is not clear. What is the sampling element? Is it the turbine or the turbine string or is it the sampling visit? How was the order selected for visiting the strings?

- *Were uncertainties described, either qualitatively or quantitatively?*

In some cases, yes; however, the very large number of univariate test significantly inflates the probability of false positive results across the entire project. The authors made no attempt to adjust, quantify, and describe this issue.

In addition, many estimates of rates were provided with no attempt to describe the associated uncertainties. For example, tables 7-4 through 7-7, 7-9, 7-11, 7-13, and 7-15, all provide estimates of *the percentage*[3] increase in mortality associated with a given variable, but no qualitative or quantitative measures of uncertainty are provided. *A percentage change is a proportion (change in number of fatalities/total fatalities) and confidence intervals for proportions are easily computed for large*

[3] Original review text before considering the Smallwood and Thelander response: "... all provide estimates of ~~rates of~~ increase in mortality..."

sample sizes and also, with somewhat more effort, for small samples. Similarly, the species or group specific mortality rates given in tables 3-11 and 3-12 are presented with no *statistical* measures of uncertainty provided*; i.e., they provide low and high estimates mortality, but do not provide the reader with a statistical measure of the quality of these estimated bounds* .

The authors do not consider any interactions, which further inflates the magnitude of the uncertainties.

- *Were findings statistically significant?*

It is likely that some number of the reported test results were statistically significant. But due to the very large number of univariate tests conducted, there is a high probability that a number of "significant" results were based on pure chance. With an accepted *P*-value of 0.05, then 5 out of every 100 tests will, on average, appear statistically significant by chance when the null hypothesis is true. No effort to account for this was made by the authors.

- *Were the conclusions supported?*

We cannot accept this analysis as one that has rigorously tested hypotheses regarding determinants of bird mortality and that could be reasonably applied in decision making. Instead, it may be more useful to consider this project an exploratory analysis that has identified a number of variables positively associated with increased mortality rates. Therefore, the product of this research is an educated list of working hypotheses. This valuable contribution can be followed by more thorough testing of said hypotheses by rigorous sampling and controlling of confounding variables via sophisticated multivariate analysis of observation data and/or controlled experimentation.

- *Other observations and comments?*

See specific comments, below.

Chapter 1: Understanding the Problem

- p.9, par.4 and 5: The authors imply that a "use vs. availability" approach to quantifying vulnerability can be effectively pursued via chi-squared tests. *A classic*[4] paper describing chi-square (goodness-of-fit) tests to examine use vs. availability of resources in a wildlife context is by Neu et al. (1974). Since that paper was published over 30 years ago, resource selection analyses involving so called use-versus-availability designs have advanced substantively (especially in the last 10 years). Now, few biologists would consider chi-square tests[5] state-of-the-art for use-versus-availability designs (see book on the subject by Manley et al. 2002 and Journal of Wildlife Management volume 2006 issue #2). Instead, most use-versus-availability designs make use some form of logistic regression functions or general linear models. *In fact, Thomas and Taylor (2006, in said volume of J. Wildlife Management), found that 35% of recent use-vs-availability studies use logistic regression; only 8% used chi-square goodness-of-fit tests.*

  Even under the context of using the chi-square test for use-and-availability analysis, the calculations require that the authors have, for each particular bird species, the number killed at turbines in a particular landscape type, the number killed in all landscape types, and the proportion of landscapes that are of that particular landscape type. The chi-square test assumes that the observed counts are accurate and any variation occurs simply from chance and not from observer error. As stated frequently in following chapters, the actual mortality counts are actually *estimated* counts and assumed to be biased low. Even assuming the mortality estimated counts are not biased low or high, this will result in inaccurate levels of statistical significance for the chi-square tests.

  And finally, chi-square tests are typically of two types: test for association/independence and test for goodness-of-fit. These chi-square tests are goodness-of-fit tests where the null hypothesis is that the counts were generated by a uniform distribution. That is, if there were no preference for the various categories of the explanatory variable, a carcass (or whatever response variable is being measured) would be equally likely to end up in any of the categories when adjusted for availability of the categories.

  Manley, B. F. J., L. L. McDonald, D. L. Thomas, T. L. McDonald, and W. P. Erickson. 2002. Resource selection by animals. Second edition. Kluwer Academic Publishers, Dordrecht, Netherlands.
  Neu et al. 1974. A technique for analysis of utilization-availability data. *Journal of Wildlife Management* 38:541-545.

- p.12, par.1: "… we are able to identify which environmental factors might have a causal relationship." After so many years of studying avian mortality associated with wind turbines prior to this work, exploratory observational studies should be

[4]Original review text before considering the Smallwood and Thelander response: "~~The "original"~~ paper ..."
[5]Original review text before considering the Smallwood and Thelander response: "...few biologists would consider chi-square tests ~~effective or~~ as state-of-the-art..."

superseded by designed experimental studies. Observational studies are not able to reveal causality. Experiments, however, can show causality. Yet there is no evidence here that any experimental design took place prior to the observations. The sample locations and times were certainly not random nor were they seemingly selected to provide contrasts in factor levels. This would have allowed them to better compare the variables of interest and help to eliminate confounding variables.

- p.14, Figure 1-1: This is a useful location map; however, a more useful map would have shown the topographic and other specific features of the APWRA. Are there distinct regions of the resource area that might be used to stratify the design?

- p.19, Table 1-1: On p.13, par.3, Table 1-1 is described as "…summarizing the wind turbine attributes of the wind turbines in our sample in the APWA." Much more information is needed here. If this is the sample, how many of each type of turbine is in the sample? How many observations (visits?) occurred at each turbine type in the first set and in the later one? What fraction of the total turbines in the APWRA does each of these types constitute? A description of the sample and the population is called for here. Are these turbines representative of the entire APWRA population?

  Some information is provided in section 7.3.1, but it focuses on sampled turbines attributes and does not provide adequate comparison to the target population (not to mention it is in Chapter 7 on page 189…a long time to wait for readers who will naturally wonder about this issue beginning in Chapter 1). In the end, the study reports data from 4,074 turbines (some with more data than others), and 1,326 turbines remained unmapped and characterized (these numbers were most easily extracted on page 352 in Chapter 9, and in our opinion should be made very prominent here in Chapter 1). But after a complete reading, the reader is still left wondering this most basic of questions – did the sampled turbines adequately represent all the turbines in APWRA? The authors need to provide a table summarizing the distribution of the sampled turbines (both sets) relative to the complete "population" of turbines. We recognize that there may be some variables that the authors cannot ascribe to turbines that were not studied (e.g., grass height surrounding the turbine), but we assume many variables are catalogued by the turbine owners (turbine model, rotor speed, etc.) and/or obtainable from GIS (elevation, slope, aspect, etc.). Figures 1-2 through 1-7 provide visuals of the distribution of sampled turbines, but they offer no information on how these distributions compare to the target population because the unstudied turbines are simply marked "unmapped", *prohibiting a visual comparison of the sampled population versus the target population*. [6]

[6] Original review text before considering the Smallwood and Thelander response: "~~This is a significant shortcoming of the report.~~"

On more minor notes, why are model numbers only given for the Kenetech turbines? The column headed "Size (kW)" should be headed "Rated Power (kW)".

- p.21, Figure 1-2 through Figure 1-7: These are the first of many colorful figures of this type in this report. Each one shows the spatial distribution of some factor. It would be useful to include an additional figure that depicts which turbines were linked to 1 carcass, 2 carcasses, etc.

Chapter 2: Cause of Death and Locations of Bird Carcasses in the APWRA

- It would seem appropriate to present the methods section, given in Chapter 3, prior to reporting the results. It is not possible to make sense out of the various results given in Chapter 2 without knowing the sampling methods used and the underlying sampling program design.

- p.28, par.5: The authors state that one-way analysis of variance (ANOVA) is commonly used and least significant differences (LSD) to compare groups. The authors should give detail as to which LSD method was used as there are several different variations, although it is doubtful this resulted in any significant changes in their calculations.

  A more important defect is the authors' excessive use of one-way ANOVA throughout this chapter and report. Many variables are tested one by one for association with mortality using one-way ANOVA. This approach makes the analyses vulnerable to confounding variables when two are more variables are highly correlated with one another, such as blade height and blade speed. The basic statistical rule that "association is not causation" can get lost in data analysis expeditions. In addition, each time a one-way ANOVA analysis is performed, the data should be graphed so that readers can see if a particular characteristic of the dataset is having heavy influence on the outcome and whether or not more subtle statistical theory violations are occurring. In light of the absence of such graphs, the *P*-values can be considered only approximate at best.

  Given the phenomenal number of univariate hypothesis tests done later in this report, it is surprising that there is no discussing of corrections for multiple comparisons here.

  It would also be helpful if the authors stated which statistical software package was used to do these analyses.

- p.29, par.2: What are the dates for season boundaries? These are not presented until Chapter 7 on page 182. Even there, the description of these dates and why they were chosen is inadequate (see later comments).

  How were days since death estimated? Were these simply guessed via personal experience? How was such experience gained?

- p.32, Table 2-1: Of the 1162 detected birds (and bats) killed by turbine collisions, almost 50% (49.5%) were restricted to 4 of the 60 species/groups reported: Red-tailed Hawk (18.3%), Rock Dove (16.9%), Western Meadowlark (8.3%), and Burrowing Owl (6.0%). Does this high concentration (i.e., 50% of deaths in 7% of the species) reflect the differences in a) abundance among these species, b) the relative risk of wind turbine collisions, or c) the probability of carcass detection?

- p.38, par.1: The methods used to search for carcasses are not described until Chapter 3. This makes the understanding of Chapter 2 material awkward for the readers unless Chapter 3 has already been read.

  The authors openly stated earlier that their search radius was 50 meters (m) and acknowledge that some "unknown proportion" of carcasses outside of the search radius went uncounted (p.28, pars.1 and 2). Yet, an unsupported statement is made here (p.38, par.1) that the "search radius included 84.7% of the carcasses of large-bodied bird species determined to be killed by wind turbines or unknown causes." How was this 84.7% calculated? In light of their search radius, it is not surprising that the majority of the carcasses were found inside the 50m radius of wind turbines. This problem is repeated later (p.42, par.5) when they note that their search radius "included 90.5% of the carcasses of small-bodied bird species." How they determine "90.5%" is left totally unclear to the reader.

  It is unclear both in this section and in Chapter 3 how the carcasses beyond 50m from the turbines were discovered. If the discoveries were accidental and not within the defined sample element, then why were they included in the analysis? If the discoveries beyond 50m were accidental, describe the circumstances of the accidents. Were the observers walking in toward or away from the turbine strings? If they were collected as part of a special study in a systematic search that extended beyond the 50m limit, then describe that study's methods and results.

- p.39, Figures 2-9 and 2-8: These figures confirm that the authors found and counted carcasses found well beyond their 50 meter search radius. That some were found as far as 200 and 220 meters distant make the idea of happenstance discovery of carcasses outside of a systematic search procedure more believable. How were these carcasses found?

If the discoveries shown in these figures beyond 50m were accidental, then, whatever the resultant pattern, it is unreliable since different sampling effort was expended within the 50m limit then beyond it. Consequently, we expect to have more discoveries within 50m then beyond it. It is no surprise that 75% of the large bodied birds were found within 42m of the tower. If we had a uniform density of birds on the ground in a 50m radius of the tower, we would expect to find 74% of the birds within 43m of the tower as shown in this simple ratio circles' areas $\frac{(\pi \times 43^2)}{(\pi \times 50^2)} = 0.74$.

Imposing a normal curve on this is unwarranted and somewhat misleading. The only patterns that are worth analyzing are within the 50 m limit. Within that limit, the distributions of discoveries with distance are similar for both large and small bodied birds. As a very minor note from the reviewers, applying the normal distribution curve to these bar graphs is not sensible considering the truncation at 0 meters and that the first bar represents only a 5 meter range while the other bars cover 10 meters. This is likely an artifact of the statistical software, but can be specified by the users. Later, the authors also put the normal curve into bar plots for non-random variables which are determined by the authors such as number of searches (Figure 3-1, p.49).

- p.40, Figure 2-10: A polar or wind rose plot would be clearer. How can the 0 and 360 degrees cells not have identical counts since they are the same direction? What is the predominant wind direction? And what about the direction the wind turbine is facing?

- p.41, Figure 2-11 and referring text p.38, par.2 and p.42, par.6: The authors use simple linear regression to show that mortality counts increase linearly with turbine tower height. The mathematical assumptions behind linear regression are not valid with this particular dataset (likely nonlinearity, non-normal distribution of errors, unequal variances) thus inadequately demonstrating statistically conclusive evidence that mortality counts are greater for taller turbines. The fact that the one-way ANOVA for wind turbine model and carcass distance was statistically insignificant (p.42, par.7) suggests the height-distance conclusion is questionable. In a confused sequence of logic, the authors state (p.42, par.6), "[the regression] predicted that for every meter increase in tower height, average distance of the carcass from the tower increased by half a meter." This clearly ignores that different wind turbine models have different tower heights, thus it may not be the height, but rather the model, that results in the carcass distance. Height and wind turbine model are confounding variables.

  The authors stated, "Distance from tower [to the carcasses] increased with tower height, according to regression analysis, although the precision was poor." The overwhelming majority of the towers were 18.5m and 24-25m tall, making this

primarily a study of these towers with a few others added in. Consequently, the observations at the lowest and highest towers had the greatest influence on the regression. [7] Even with the data for the 43m towers, the regressions only explain a trivial 1% of the variance in the distances that the carcasses were found from towers. The phrase "poor precision" is an understatement. This is the difference between "statistically significant" and biologically important.

- pp.42 and 44: A description of the tower population would be useful here. For the sampled towers and the population as a whole, how many towers of each type, what elevation distribution, what string lengths (1 to n), what spacing between towers in string, etc?

  The authors survey how carcass distance relates to multiple independent variables including tower height (continuous); blade speed (continuous); upwind vs. downwind (binomial); end, gap, or interior of string (categorical); season (categorical); whether turbine was in a canyon (categorical), slope grade (categorical); or elevation (continuous). They investigate each variable in a univariate analysis, but this may be better suited for a general linear model.

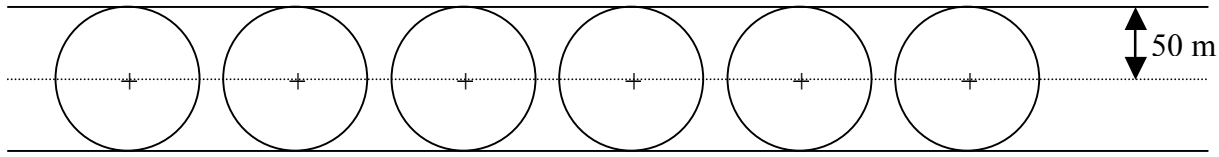
  Why are there 2 degrees of freedom (# levels – 1) in the ANOVA to test if carcass differed depending on whether the turbine was in a canyon? Either the independent variable is binomial (in a canyon or not) in which case there is 1 degrees of freedom or there were three "canyon categories" (yielding 2 degrees of freedom) that the authors did not articulate to the readers.

- p.43, Figure 2-13, p.44, Figure 2-14: The report of a strong effect of tower location within a string on the carcass distance is difficult to accept without careful analysis of the influence of the sampling method. The sampling method is described to some degree in Chapter 3, but it remains unclear how carcasses were associated with a particular tower within a string.

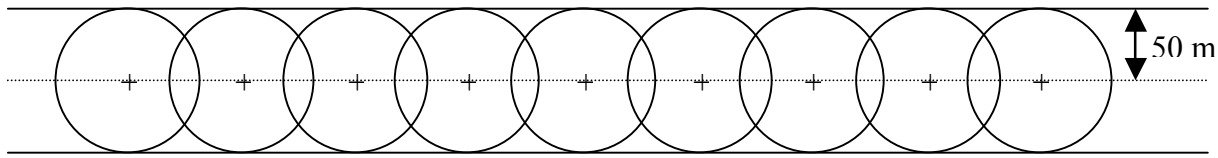
  For example, if the tower is not in a string (or if you prefer, a string of 1), then there is no confusion. Any carcass found within 50m of the tower is associated with that tower, and the search area would be $\pi \times 50^2 = 7854 m^2$. But for towers in strings, the tower spacing makes a difference. In the first sketch below, the towers are spaced more than 100m apart so that the area within 50m of each tower does not overlap with any other tower's area. (But looking forward to Figure 3-3 on page 51, will the search areas of a string then be very wide rectangles that include the spaces between the circles?) In the second sketch, the towers are less than 100m apart so there can be a lot of overlap in the 50m zone around each tower. Note that the end towers have greater area to themselves.

[7] Original review text before considering the Smallwood and Thelander response: " .. greatest influence on the regression. ~~If the 4 to 6 observations on the 43 m towers were removed, we suspect that neither of the two regressions would be statistically significant.~~ Even with the data..."

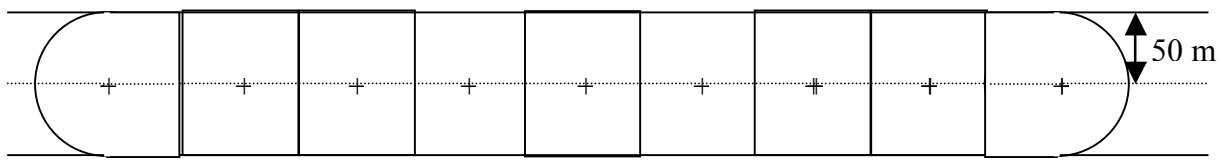
Case 1



Case 2



Based on Figure 3-3, a closer approximation to case 2 would be:



The towers not on the ends end up with rectangular search areas, where some of each rectangle is beyond 50m from a tower. On the other hand, the end towers in a string may have considerably larger sampling areas (depending on the tower spacing) and more at the further distances away from the tower.

For example, if the towers in the string were 50 m apart, then the search area for the towers in the internal string would be: height$\times$ width = (50m + 50m) $\times$ 50m = 5000m^2. For the end towers, the area would be (half of a rectangle + half of a circle) = (50m + 50m) $\times$ 25m + 0.5$\times \pi \times$ (50m)^2 = 2500m^2 + 3927m^2 = 6427m^2 which would be 29% larger than for the internal towers.

So the distance between towers in a string is important.

- p.43, Figure 2-12: The authors show standard error plots of carcass distance by the different wind turbine types. Box plots would do a more adequate job of showing the spread of the data and inform the reader of potential biases in the study with regards to various wind turbine models. Specifically, box plots would show if distance of carcass (beyond 50m) would result in reduced carcass count for a particular wind turbine model. A "mean and 2 standard error" plot is designed to show the reader the range where the true mean is likely to be. With this study, however, we are more interested in the range and general distribution

of where the carcasses are to be found rather than what would be the long term average distance of where carcasses are to be found.

In addition, for large bodied birds, 50.0% of the carcasses are associated with KCS-56 turbines, 34.1% with Bonus, and 6.1% with Micron, totally 90.2% of just 3 of the 10 turbine types. Similarly for the small bodied birds, 83.6% of the total carcasses were found at the same 3 of 10 turbine types. How many turbines of each type are there? Is this disproportion chance or pattern?

Given the happenstance data collection on carcasses beyond 50 m, the inclusion of the beyond 50 m data in the analysis is inappropriate.

- p.44, Figure 2-13: Regarding distance of carcass from wind turbine for "end", "gap", and "interior" turbines and their analysis (p.44, par.1) could suggest that carcasses tossed far from one turbine could be attributed to the turbine to which it landed closest too. This is acknowledged in the discussion (p.45, par.3). Are all wind turbines in a string alike?

- p.45, par.1: The authors state that they found 15.3% of the large bird carcasses and 9.5% of the small bird carcasses outside of their 50m search radius. It is not surprising that only small percentages of the birds were discovered beyond 50m since the search effort in that region was happenstance. It is not stated whether these carcasses were found during the observers' systematic searches or while the observers were walking to the area where a systematic search would be done.[8]

- p.45, par.3: They state that extending their search radius to 100m would include 94% of the large bird carcasses[9] *they found, but this figure does not illuminate the number of carcasses that may still be missed beyond 50, because their rigorous searches terminated at that distance.* There is a well-established body of theory for estimating density of animals (or in this case, carcasses) using the distance to each detection and modeling probability of detection as a declining function of distance. There are computer programs (e.g., DISTANCE) for this sort of thing. These programs could essentially estimate the number of carcasses that were overlooked to yield a more unbiased and accurate estimate of carcass density.

---

[8] Original review text before considering the Smallwood and Thelander response: "... where a systematic search would be done. ~~How can you discover carcasses if you do not search for them?~~"

[9] Original review text before considering the Smallwood and Thelander response: "... large bird carcasses~~, an unsupported figure~~. There is a well-established body of theory..."

- p.45, Table 2-2: [10] The effects listed for Flowind and KVS-33 turbines are based on very small sample sizes (10 and 4, respectively) and also include happenstance discoveries beyond 50 m which further distorts the intervals. The reported effect could very well be spurious.

Chapter 3: Bird Mortality in the APWRA

- p.46, par.3: The authors propose that impact of the APWRA can be measured one of two ways: (1) number of fatalities per megawatt per year or (2) number of fatalities relative to the natural mortality and recruitment rates. They choose the fatalities per megawatt because it treats a certain number of fatalities as the "cost" of producing a megawatt. The other method evaluates the long term affect on the bird population; however, some of the needed demographic variables for such a measure are logistically unreasonable to estimate and beyond the scope of this project. Other authors use fatalities per turbine per year (p.46, par.4).

  It is more an issue of policy that determines which measurement is more helpful. Although not unreasonable, fatalities per megawatt per year ignores the total number of fatalities. Total number of fatalities is an important measure that shows, at least in part, an impact on the bird populations even if you do not know the demographic conditions of the species. Fatalities per megawatt per year is a good measurement if you are trying to minimize fatalities while producing a certain amount of energy. Fatalities per wind turbine would only be helpful if you are trying to minimize the number of fatalities for a fixed number of wind turbines regardless of energy output – something only reasonable if wind turbine models all had the same energy output.

  Another issue that needs to be looked at is how often, when operating, is a wind turbine actually achieving its rated energy output? Given that a wind turbine is operating, the distribution of time operating at various output levels will likely differ among different models. So would a higher rated wind turbine be more frequently operating at sub-maximum energy output levels than a smaller wind turbine, although being a risk to birds for as many hours as the smaller turbine? This would not be represented by the megawatt per year metric. The authors do briefly mention the lack of data regarding this issue on p.347, par.1 of Chapter 9.

- p.47, par.5: The authors sampled 1,526 wind turbines (182 strings) for 4.5 years and another 2,548 wind turbines (380 strings) sampled for about 6 months (November through May) because of access issues. Although this is about 75% of the wind turbines in the APWRA, the authors do not say how they decided

[10] This original review point began with the following paragraph before considering the Smallwood and Thelander response:. "~~This table summarizes the conclusions reached in this chapter about the distances of carcasses from towers. The relationships between distance and tower height are heavily influenced by a few observations on the tallest towers and in any case, the relationships are not substantial and only statistically significant in the most narrow technical sense given the r^2 values of 1%.~~"

which turbines to survey. Did they survey every wind turbine or string for which they had access? Did they use a sample of convenience, simple random sampling, or systematic sampling? The essential question is: Can the surveyed wind turbines be considered representative of the entire population of wind turbines or at least representative of the wind turbines for which they had access?

The short duration of sampling for the second set was the result of delayed access to the turbines from the owners. Although the first set includes fewer turbines and strings, it provides the primary and superior data set because of the repeated observations, the seasons sampled, and the increased duration. The limited duration of sampling, the lack of replication, and the restricted seasons sampled greatly reduces the value of the second set. Unfortunately, the analyses do not distinguish between the two sets.

- p.48, par.4: Was there any concern about whether severed body parts from one mutilated bird (wind turbine or scavenger caused) could have indicated more than one fatality?

- p.48, par.1 and p.49, par.2: The authors write, "...we recently found that 85%-88% of the carcasses occurred within 50m of the wind towers." The absence of any described systematic method of how they searched beyond 50m makes this estimate questionable. The authors then write the following:

> "Searcher detection and scavenger removal rates were not studied, because it had already been established that mortality in the APWRA is much greater than experienced at other wind energy generating facilities. We were unconcerned with the underestimating mortality, and in fact we acknowledge that we did so. We were more concerned with learning the factors related to fatalities so we can recommend solutions to the wind turbine-caused bird mortality problem. Thus, we put our energy into finding bird carcasses rather than estimating how many birds we were missing due to variation in physiographic conditions, scavenging, searcher biases, or other actions that may have resulted in carcasses being removed." (p.49, par.2)

With this statement, readers must treat all bird mortality estimates as relative estimates and not as the exact counts or unbiased estimates. Regardless, the authors go ahead and attempt to come up with reasonable mortality estimates.

- p.49, par.1: What is the sampling element in use in this chapter? The authors "… express mortality as the number of fatalities per MW per year …" The total

number of fatalities observed on a string divided by the total rated power output from the string and divided by the total duration of sampling. This indicates that the sample size is the string, so that each string, not turbine, has an associated fatality rate. So sample sizes should be the number of strings visited, not turbines visited.

- p.51, par.1: The authors did not assess searcher detection rates in this study and selected to use literature values: 85% detection rate for raptors and 41% for non-raptors. Solely in this chapter, these detection rate values are used to correct the observed counts for deficiencies in detection. This seems reasonable, but why do the authors feel detection would be 50% less likely to discover a small raptor such as a kestrel than a similar sized non-raptor, such as a robin? (This same question applies to scavenging rates as well.)

  They estimated the number of carcasses that actually existed by dividing either by 0.85 (raptors) or 0.41 (non-raptors). These calculations were equally applied to carcasses was found within or beyond the 50m search radius. This seems unreasonable to treat the beyond-50m carcasses the same as within-50m carcasses because carcasses beyond 50m were discovered by happenstance. The fraction missed beyond 50m could be much larger than their estimate.

- p.51, par.2 – p.52, par.2: The authors used scavenger removal rates and detection rates estimated in other studies to produce bird mortality estimates (p.51, par.1). A bothersome aspect of the authors' report is that they adjust the scavenger removal rates and detection rates from the other studies to rates that they believe better describe the APWRA and the time between their surveys without giving any anecdotal or empirical evidence of why they chose the numbers they did. Adding 10% to the scavenger removal rates of Erickson et al. (2003) to account for the authors' longer interval between searches appears arbitrary (p.51, par.2). Furthermore, without any support of data or other evidence the authors add (p.52, par.1), "Based on our experiences with raptor carcasses in the APWRA, we did not believe that these scavenger removal rates were accurate for raptors, and we halved the removal rate estimates reported by Erickson et al. (2003)." Underestimating scavenger removal rate will result in underestimating mortality.

  There is an error in their calculations for "halving" of the raptor removal rate. If s is the scavenging rate, the authors estimate the pre-scavenged carcass number by dividing the number of carcasses available after scavenging by $(1-s)$. After "halving" the scavenger rate, the authors simply divided by $2\times(1-s)$ while they should have divided by $1-\frac{s}{2}$. Their method reduced the scavenging rate by more than half and results in mortality estimates that are biased downward.

For example, the scavenger removal rate for carcasses of large-bodied species is 68.6% (p.51, par.2) thus the proportion of carcasses after scavenging to be found is $1 - 0.686 = 0.414$; therefore,

$$\text{Pre-scavenged number of carcasses} \times 0.414 = \text{Number of carcasses after scavenging}.$$

So to calculate the pre-scavenged number of carcasses from the number of carcasses available to be found after scavenging, we divide by 0.414:

$$\text{Pre-scavenged number of carcasses} = \frac{\text{Number of carcasses after scavenging}}{0.414}.$$

The authors halve the scavenging rate by doubling the denominator, thus 0.414 becomes 0.828. This, however, is different than halving the 68.6% down to 34.3% which would give an estimate of the pre-scavenged number of carcasses to be:

$$= \frac{\text{Number of carcasses after scavenging}}{(1 - \frac{1}{2} \times 0.686)}$$
$$= \frac{\text{Number of carcasses after scavenging}}{(1 - 0.343)}.$$
$$= \frac{\text{Number of carcasses after scavenging}}{0.657}$$

This does not equal the authors' pre-scavenged calculation of $\frac{\text{Number of carcasses after scavenging}}{0.828}$. Consequently the authors are more than halving the scavenger rate.

The combination of these various corrections results in an estimate of overall mortality that is, at best, rough and imprecise and, at worst, seriously biased (likely downward). *Inadequate*[11] consideration is given to these ad hoc corrections in evaluating the uncertainty in the mortality rate estimates provided later in this chapter.

As a last note here, the authors should make their calculations more clear to the reader. Erickson et al. 2003 provides a good template.

- p.52, par.2: The authors are correct in stating that their "mortality estimates might be conservative" because of removal of carcasses by people not involved in the

[11] Original review text before considering the Smallwood and Thelander response: "~~No~~ consideration..."

authors' study and they provide some anecdotal evidence. The authors do not account for such carcass removal.

- p.52, par.3: The authors state that, of the 1162 carcasses whose fatality was attributed to the wind turbines, 198 were more than 90 days old. Table 3.1 on pp. 64 and 65 counts fatalities as Type A (both fresh and old) and Type B (fresh; used to estimate mortality). The difference between Type A and Type B should be the number of carcasses older than 90 days. In fact the difference is 1162 – 923 = 239 which is larger than the 198 reported on p. 52. What happened to the other 41? Bats account for some, but not all.

- p.52, par. 4 and p.53, Figure 3-4: The authors state that the frequency distributions shown in Figure 3-4 are "at the string level of analysis". The caption for Figure 3-4 should reflect that the figure shows the frequency of strings with various levels of estimated mortality rates.

  It is striking that at 270 of the 562 strings searched, or 48%, no carcasses were found. A useful analysis would have been to compare the group of strings with zero fatalities to those with observed fatalities.

  Both parts of Figure 3-4 include what appears to be a truncated normal distribution. This is inappropriate since the observed distribution is quite unlike a normal curve, more closely resembling an exponential or Poisson distribution. The normal curves should be removed.

- p.52, par.5 and p.64, par.1: The authors make statements about inter-annual mortality variation for different species and types of birds at wind turbines sampled for all four years. It is assumed, but not stated, that ANOVA and LSD are used. The multiple categories of birds species/type being tested for inter-annual mortality variation makes the chance of at least one Type I error likely.

- p.52, par.5 and p.68, Tables 3-3 and 3-4: The statement about the mortality of burrowing owls based on the strings studied for 4 years vs. just 1 year refers to the right columns of Table 3-4. We suspect this should be Table 3-3.

- p.59, Figure 3-15: Year effects on mortality rate are confounded by location, as evidenced by this figure.

- pp.54-58, Figures 3-5 through 3-14: It seems as though the 95% confidence intervals in these figures were determined based on the string-based mortality rate estimates using Student's t distribution. Then it would be appropriate to provide

the sample size for each year and not just the aggregate for all 4 years. (Or was it a sample size of 160 for the 1-year strings and 62 for the 4-year strings?)

How was the confidence interval computed for 2001-2002 in Figure 3-9? It appears that the estimate is zero and the C.I. has zero width. How is this possible? Were there no barn owls killed in the 62 strings in 2001-2002? *If so, then the point should not include a confidence interval.*

- p.70, Table 3-9: To this point in this chapter, the analysis has been string based. This table refers to 1526 turbines in the first set and the 2548 turbines in the second set. The columns give the mean and standard error among strings, not turbines. What was the sample size used for each of the mean and standard error calculations? Is it number of turbines or number of strings? Are these sample sizes taken to be the same for all species or groups

  It would be useful to compare these results to the corresponding median values. It would be interesting to know how many of the median mortality estimates would be zero? Even for the shorter duration second set, 12 of the 30 (40%) species mean mortality rates are zero.

- pp.70-75, Tables 3-9 through 3-12: The authors should better explain the calculations used to produce these tables. An example using real data would be helpful.

- p.76, par.2: The authors mention high mortality estimates in the SeaWest-owned portion of the APWRA, but the Results (Section 3.3) did not articulate about spatial or owner differences in mortality rates.

Chapter 4: Impacts to Birds Caused by Wind Energy Generation

- p.78, par.3: The authors assume a 50% miss rate outside of their 50m search radius (p.78, par.3). This statement conflicts with their Chapter 2 methods (p.51, par.1) where they said the detection rate within 50m was the same as beyond 50m. Thus in Chapter 2 they used detection rates for beyond 50m of 85% (raptors) and 41% (non-raptors). A 50% detection rate beyond 50m for non-raptors would suggest a greater detection rate beyond 50m than within 50m, obviously not sensible. More reasonable detection rates would be 42.5% (raptors) and 20.5% (non-raptors) beyond 50m (i.e., half the detection rate as within the more thoroughly searched 50m).

- p.78, par.4: The authors present findings from point count surveys although they have not yet discussed the methods with the readers.

- In general, Chapter 4 does not adequately portray that the mortality estimates at APWRA from this report are likely biased low – perhaps severely. This bias comes about because: (1) detection rates for carcasses beyond 50 m could easily be well below the values used in analyses; (2) scavenging rates could easily be higher than used in analyses (because search intervals were longer for this study than in the studies from which values were obtained); and (3) scavenging rates of raptors were arbitrarily cut in half from reported scavenge rates.

Chapter 5: Range Management and Ecological Relationships in the APWRA

- In general, the authors present the reader with a blizzard of one-way ANOVA and LSD statistical tests looking at an almost endless number of variables. Having so many variables inspected individually, leaves the study highly vulnerable to Type I errors, confounding variables and difficult to interpret findings. A multivariate approach would help the authors develop a more thoughtful, concise analysis that can help control for confounding variables.

- p.91, par.3: "Vegetation height ... was 18% greater ... where rodenticides were intermittently deployed...," the authors report with a mean difference from intense rodenticide use of 4.28cm. The magnitude of 4.28cm is more meaningful if the mean heights of the grasses are also provided. It could be 1cm vs. 5.28 or 11cm vs. 15.28 which could understandably have different ecological impacts.

- p.100: The authors indicate that the index of cottontail rabbit abundance was higher on Enertech towers, on plateau slope combinations, and on southwest slopes. Were Enertech towers especially common on southwest slopes relative to other tower types? These questions are difficult to answer because they require the reader to extract information presented for other purposes elsewhere in the report. By running multivariate analyses (which may require simplifying or reducing variables – in itself a good thing), then the association between a given predictor variable and the response variable can be measured while statistically accounting for confounding variables. This is a recurring limitation of the study.

- p.103, Table 5-20. This is an example of where the authors should interpret the meaning of the analyses while paying attention to the magnitude of differences. Furthermore, the metric "cottontail abundance" is never defined. In Table 5-20 cottontail abundance is compared between "some lateral edge" and "other edge conditions" with a statistically significant "Mean difference (cm) on grass

transect" of 0.18. What does that 0.18cm represent? Is that a small biological magnitude that ends up being statistically significant because of the very large sample size of 1327?

- p.108, par.4: The authors make quick mention that, "Some of these relationships might be confounded with other variables." This is an understatement and a recurring limitation of the study. Multivariate analyses could help control for some of these confounding variables.

Chapter 6: Distribution and Abundance of Fossorial Animal Burrows in the APWRA and the Effects of Rodent Control on Bird Mortality

- p.111, par. 4: "Most wind turbine strings were selected arbitrarily, to represent a wide range of raptor mortality recorded during our fatality searches, as well as to represent a variety of physiographic conditions and levels of rodent control," the authors write. A more rigorous method of selection should have been used, such as stratified sampling. The objectiveness and unbiasedness of "arbitrary" sampling is always questionable.

- p.112, par.4 and p.114, par.5: The method of estimating degree of clustering at wind turbines using the slope from least squares linear regression is unclear (p.112, par.4). Is "corresponding search areas" the distance from the wind turbine? It then seems that the authors disregard this "regression-slope" method (p.114, par.5) for the "observe*d*-divided-by-expected" approach. Having this "regression-slope" method discussed is confusing if it is not to be used.

- p.112, par.6: The authors mention that they learned *post hoc* about rodent control. Although likely beyond the duties of the authors, the effectiveness of rodenticides to reduce raptor mortality could be better explored in the future via a carefully planned experiment.

- p.149, par.5 and p.164, par.1 and Figures 6-45 and 6-46: The simple linear regressions used to investigate association between raptor mortality and ground squirrel burrow systems are very questionable (Figures 6-45 and 6-46). The authors discuss the significance of these scatter plots (p.149, par.5 and p.164, par.1). Some of these conclusions and "significant" *P*-values are based on sample sizes of 3 (no rodent control) and 5 (intense rodent control) – it is *outside the realm of professional practice*[12] to base inferences from just 3 or 5 data points.

[12] Original review text before considering the Smallwood and Thelander response: " ... it is ~~foolish~~ to base inferences..."

Furthermore, leverage of an individual point affects all three levels of rodent control and the assumption of homogeneous error is ignored.

- pp. 164-172, Tables 6-2 through 6-11: These tables aggregate the density of burrows into categories and then total the number of bird kills for each of the three categories. It is not clear how the authors decided to define each category and information is lost by categorizing continuous data. A dot plot or histogram of the burrow densities for where carcasses were found beside a second plot of burrow densities for where carcasses were not found would have been more informative.

- Discussion, pp.172-178: The authors make good points in the Discussion regarding the negative and/or inconsistent impacts of rodent control measures, and their case is strong, we believe. They offer the caveat that, "Intense rodent control was associated with fewer golden eagle fatalities in areas of intense rodent control, but the association is not strong enough to warrant its continued use" (p.178, par.2). We think that statement is giving the rodent control measure more causal credit than it deserves. In fact, the *P*-value for the ANOVA test of golden eagle mortality rate across the three rodent control intensity levels is statistically insignificant at 0.9 (p. 172, Table 6-12). While the mean mortality estimate is slightly lower in magnitude for the intense control category, the variance is very large, and we thus have no confidence this difference is "biologically real." One could just as easily claim that, "mortality rates among rodent control intensity were statistically indistinguishable."

Chapter 7: Bird Fatality Associations and Predictive Models for the APWRA

- p.182, par.5: The authors define four seasons, but the length of the seasons are very different: spring is 92 days, summer is 117 days, fall is only 51 days, and winter is 105 days. Summer is 2.3 times as long as the fall. What is the justification for these definitions? The authors also give no explanation of how they decide "number of days since death" when a carcass is discovered.

- p.182, par.7: Although Table 1-1 does summarize the attributes of the wind turbines in the sample, it does not state the frequency of each type in the sample and the population.

- p.183, line 7: The authors need to be careful and consistent as to how they show their mathematics. They most often, but not always, use more elementary notation such as $A \div B$ instead of $\frac{A}{B}$. On the 7th line of page 183, they define "the

window of opportunity" as $\text{Window} = \text{C} \div \text{T} \cdot \text{B}$. This is equivalent to $\frac{C \cdot B}{T}$, but the equation is more sensible as $\frac{C}{T \cdot B}$, which we believe is what the authors meant. The authors should employ the use of an equation editor, like that used in Microsoft Word.

- p.183, par.2: For purposes of computing how quickly a bird clears the rotor plane, how thick is the plane? What flight speed would be required to clear the rotor plane in the allotted time?

- p.183, par.5: The tower height is defined as the distance the rotor is above the ground. Can we assume that this is the center of the rotor?

- p.184, par.1: The incidence of rock piles was reduced to a limited number of categories. Did the authors intend the categories to be: a) none, b) less than or equal to 0.25 piles per turbine, or c) greater than 0.25 piles per turbine?

- p.184, par.2: The authors employ a 40 m radius around each turbine instead of the 50 m radius stated earlier. What is the reason to redefine the sampling zone now?

- p. 184, par.4: Did the authors test the assumptions of the statistical tests (e.g., homogeneity of variances or statistical independence and normality of residuals) applied in this or any other chapter? What objectives are the authors trying to meet in reporting "weak and non-significant correlations"? How can the measures of effect*, statistically or biologically,* be meaningful if the confidence interval for the magnitude of the effect includes zero? *A nonsignificant result would imply a confidence interval that includes zero.*

- p.184, par.5: For regressions, the authors have chosen to include the RMSE to provide a measure of the "precision of the data relative to the regression line". By RMSE, we assume that the authors mean:

$$RMSE = \sqrt{\frac{Sum - of - squared - residuals}{samplesize}}$$

A more appropriate estimator for precision*, for either simple or multiple regression,* would have been the standard error of the estimates (SEE) or:

$$SEE = \sqrt{\frac{Sum-of-squared-residuals}{samplesize-\#of\ parameters}}$$

- p.184, par.6: Although this is a non-manipulative study and the existing towers, turbines, topography, etc. as well as permission for access does limit the range of choice, it is still possible to carefully select the areas of study to provide the contrasts and comparisons of interest.

- p.185, par.1: Is the term "efficient" used here in the technical sense from statistics?

- p.185, par.2: The authors discuss the 5% significance level used in the subsequent tests and the 10% level that they interpreted as indicating "trends worthy of further research". Given the immense number of univariate hypothesis tests reported in the subsequent pages, the authors should have discussed the risks of Type I errors (false positives) associated with conducting hundreds of tests.

  The total number of chi-square tests presented just in Tables 7-1, 7-2, and 7-3 is 528 (ignoring the many more chi-square tests presented in Appendices B & C). The chief disadvantage of this approach is that Type I and Type II (false negative) error rates are inversely related, creating no clear optimization. One could argue that Bonferroni adjustments are necessary to guard against very high experiment-wise Type I error stemming from so many tests. Using Bonferroni adjustments, the experiment-wise alpha (level of significance) value "should" be set as: $\alpha_{adj} = 1-(1-\alpha)^{\frac{1}{n}}$; in this case $\alpha_{adj} = 1-(0.95)^{\frac{1}{528}} = 0.000097$ for a modified Bonferroni adjustment as proposed by Shafer (Shaffer, J. P. "Multiple Hypothesis Testing." *Ann. Rev. Psych.* **46**, 561-584, 1995.)

  But if the authors bring the experiment wise alpha value this low, the Type II error rate gets unacceptably high, especially for work designed to measure environmental impact. That is, the probability of the analysis suggesting no impact when in fact there is one becomes unacceptably high. This problem further underscores the value of a smaller number of multivariate tests, as we have suggested elsewhere.

- p.185, par.3: The uses of chi-square tests "for association" are described. The chi-square tests used by the authors are more commonly described as chi-square tests for "goodness-of-fit" where they are testing whether it is plausible that the observed counts across the categories came from a uniform distribution (each category is equally likely). Although statistically legitimate, such methods fail to control for other variables, leaving the study vulnerable to confounding variables.

Why not use a general linear model, logistic (yes/no data) or Poisson (counts data) regression, discriminant analysis, or at least a log-linear analysis?

- p.186, par.3: The authors rationalize that relative search effort can be calculated as, $N_t \times R \times Y$, where N_t is the number of wind turbines in a string, R is the mean rotor swept area in m^2, and Y is the number of years the string is searched. This decision is based on Figure 7-1. It is a loose association between the relative search effort and number of fresh bird carcasses found. From this, they assume that mean rotor swept area is proportional to the number of carcasses – a circular argument since that is what they are supposed to be investigating. Keep in mind that the swept area is proportional to the squared radius of a wind turbine (Area=$\pi \times r^2$), thus the "search effort" at a wind turbine with a 3m blade will be four times as much as at a wind turbine with a 1.5m blade (half the size) even if they physically searched the surrounding grounds equally. Thus the wind turbine with a 3m blade will have to kill four times as many birds to have the same rate of mortality as the 1.5m blade wind turbine, ignoring megawatt output. In Appendix A, the authors do show a positive relationship between megawatt output of a turbine and mortality. Perhaps the authors are trying to copy epidemiology studies which use "people years" when calculating risks for cancer; e.g., following 100 people for 5 years is equivalent to following 250 people for 2 years. Here this would correspond to "turbine years". It is a strong assumption to say that the variable "rotor swept area" is just as important as the variables "time" or "number of wind turbines" with regards to the number of expected bird carcasses.

- p.186, par.4 and p.187, Figures 7-1 A & B: Figure 7-1A presents the relationship between the number of birds recently killed at turbine strings and the measure of search effort used. [13] Which of the variables account for the observed variation in the search effort: the number of turbines in the string, the mean rotor swept area, or the number of years of searching?

  The authors suggest that Figure 7-1B illustrates an inverse power relationship between fatality rates and search effort. It would be more informative to plot the data shown on a log-log plot, which would more conveniently indicate if the relationship was in fact an inverse power relationship. It appears, however, that there may be many observations with fatality rates of exactly zero, but it is difficult to tell since the vertical axis does not show a zero.

[13]Original review text before considering the Smallwood and Thelander response: "...search effort used. ~~Of the 472 data points, only 32 or so exceed 10,000 m^2·yr of search effort and only 2 of the 472 exceeds 30,000. Consequently, these extreme values of the total dataset have the principal influence on the regression results.~~ Which of the variables account..."

Figure A4 (p. A-8) suggestions a mechanism that would produce the relationship suggested for Figure 7-1B. This indicates that the sampling approach yields stable estimates only after longer periods of search, which should be discussed here.

- *[14]*

- p.189, par.2: So now the sampling element is the wind turbine and no longer the string. What fraction of the total population of wind turbines does this sample of turbine models represent? It is important to the reader to know if these sampled wind turbines are representative of the APWRA population of wind turbines.

- p.190, Figure 7-2: The figure shows that the authors' study is essentially a study of KCS-33 and Bonus wind turbines. Furthermore, the "effort" for Bonus wind turbines is almost three times that of the number of Bonus wind turbines studied. Is that a result of the "relative effort" definition and that Bonus wind turbines' rotor sweep area is three times that of most other turbine models?

- p.189, par.8 and p.202, Figure 7-18: Based on the authors' definitions of seasons, fall is the shortest season (51 days) and so would be expected to have less sampling effort. Given the length of the seasons and assuming a uniform distribution of sampling times throughout the year, we would expect 25% of the observations in the spring, 32.1% in the summer, 14.0% in the fall, and 28.8% in the winter. Comparing this to the bar heights in Figure 7-18, the sampling effort is higher than expected in the spring, lower in the summer, higher in the fall, and on target in the winter. Is this a result of their sampling effort definition? It is not clear.

- p.192, Figure 7-4: Why is effort so many times greater for the wind turbines with 2141 rotor plane swept per second?

- *[15]*

---

[14] The original review's point was removed before considering the Smallwood and Thelander response:: "~~p.188, par.2: "Positive values express the percent of total fatalities likely killed at wind turbines due to the attribute associated with the value..." The use of the word 'due' implies causality, although at best they can only claim 'association'.~~"

[15] The original review's point was removed before considering the Smallwood and Thelander response:: "~~pp.193 and 194, Figures 7-5 and 7-6: These figures show scatter plots where an outer single point has high leverage (influence). Conclusions are essentially being determined by the one point furthest to the right.~~"

- p.199, Figure 7-14 through p.201, Figure 7-16: Why are the bin widths increased in going from graph A to graph B for each set of graphs? In graph B of each pair of graphs, the bin widths are not equal.

- p.203, Table 7-1: The dangers of multiple hypothesis testing arise in Table 7-1 when 204 chi-square tests are performed. (This is repeated again in Tables 7-2 and 7-3.) This can be kindly called "data exploration" or criticized as a "data dredging". Regardless, with 204 statistical tests, if all data were a result of a uniform distribution across each category, researcher error or biased post-hoc categorization did not cause any non-uniform distribution, and each test were independent of one another, you should expect 5% of the tests to give p-values less than 0.05. So there is a high chance of Type I errors when so many tests are performed. Also many variables may be correlated, such as "tower height" and "high reach of blades". So if a test was significant for "tower height" you should expect it to also be significant for "high reach of blades". In addition, a more clear explanation is needed as to why some variables such as "rodent control" and "Slope aspect" are tested twice.

  There are methods to help reduce the problems of multiple testing, such as Bonferonni corrections that make the p-value for declaring a "statistically significant result" much less than 0.05 for each test. This makes the overall chance of a Type I error only 5% if all tests were actually not significant. The problem with such adjustments is that the statistical power then decreases for each test opening the door for Type II errors thus making the researchers miss important variables. The authors should take a more selective and thoughtful approach to investigating the variables and use generalized linear models or multiple regression. These more advanced methods would help reduce some confounding by allowing the authors to control for other variables when testing another. The authors did, however, state that they only used the predictive model for variables that were statistically significant and showed gradients along a continuum (p.188, par.3).

  Furthermore, what are the sample sizes for each of these chi-square tests? A large sample size can produce very small p-values (very high statistical significance) even though the magnitude of difference from the uniform distribution is minimal; i.e., lacking biological significance. When the authors discuss the finding from the chi-square tests, they report something along the line of, "Wind turbines with variable X killed disproportionately more birds of species Y." What magnitude is implied by "disproportionately"? With a large enough sample size, it could be a biologically insignificant increase that is likely just a result of confounding. This issue of magnitude is addressed in Table 7-5 (p.215), but the percent magnitudes still need to be put side-by-side with real numbers to make them more meaningful.

- pp.207-209, Figures 7-19 through 7-21: There appears to be considerable spatial clustering of the golden eagle, red-tailed hawk, and burrowing owl fatalities. The variation in duration of study does not coincide with the clusters. Similar spatial clusters appear in all three figures. There is no discussion of *these figures*[16] in *this* narrative. Are these clusters the result of turbine type clustering, variation in elevation, concentration of avian habitat, or some other factors?

- pp.210-219, Tables 7-4 through 7-7: Percentage increases in mortality are listed for various species in association with 12 factors. Confidence intervals should be provided for each of these percentage values so that the precision of the estimated effect can be evaluated. How many of these confidence intervals would include zero, indicating that the magnitude of the effect might plausibly be zero?

- p.219, par.1 and pp.220-221, Figures 7-22 and 7-23: The authors note the seasons with relatively higher fatalities than expected but neglect to point out the seasons with unusually lower fatalities than expected. Specifically, the red-tailed hawk, American kestrel, and burrowing owl all show much lower fatalities than expected in the spring. Why would this be true? Similarly, there were no fatalities of mallards in the fall. Why would this be so?

p.222, par.2: "The empirical models developed were tested only against the database of the 4,074 wind turbines from which the data were obtained for model development," state the authors. Testing the quality of a statistical model on the same dataset from which it was developed is bad practice. The selected model may fit that specific dataset well, but not be robust enough to predict outcomes well from a similar but different dataset. Some statisticians, for example, will randomly set some fraction of the original data to the side (test set), fit a model on the remaining data (learning set) and then see how will it predicts the data that had been set aside. This is repeated until all data have been set aside once in the test set. Once a good model has been determined, it is fit to the entire dataset. This concept is much-addressed in the ecological statistical literature (e.g., Fielding and Bell 1997, Boyce et al. 2002, Knightes and Cyterski 2005), and there are numerous analytical approaches to minimize the circularity without requiring the collection of new independent data. The authors need to address these issues.

Fielding, A. H. and J. F. Bell. 1997. A review of methods for the assessment of prediction errors in conservation presence/absence models. Environmental Conservation 24:38-49.
Boyce, M. S., P. R. Vernier, S. E. Nielsen, and F. K. A. Schmiegelow. 2002. Evaluating resource selection functions. Ecological Modelling 157:281-300.
Knightes C. D. and M. Cyterski M. 2005. Evaluating predictive errors of a complex environmental model using a general linear model and least square means. Ecological Modelling 186:366-374.

[16] Original review text before considering the Smallwood and Thelander response: "There is no discussion of ~~this~~ in ~~the~~ narrative."

- p.222, par.3: This argument is independent of any observations made by the authors. It represents circular reasoning. It argues that if the model is correctly predicting which turbines are relatively more dangerous, then the reason no bird fatalities were found at most of these dangerous turbines is just that we did not look long enough. This might be true but this work can neither support nor refute it.

- p.223, Table 7-8: The authors have so far only conducted univariate chi-square hypothesis tests. They now seek to combine the results in an ad hoc fashion into a model which amounts to a scoring system. If the authors want to develop a multivariate model, they should apply appropriate methods such as logistic or Poisson regression.

- p.224, Table 7-10: The authors' interpretation of the results presented in this table is unusual. They group the observations by the results (e.g., 0, 1, 2, 3, etc. fatalities) and compare the fractions that were predicted to be "more dangerous" and "less dangerous". This is a backwards approach to evaluating the predictive model. The observations should be grouped by the predictions (not the results) and the percentages of each group that experienced fatalities should be compared.

  For example, using the golden eagle data, we can assemble a 2 x 2 table of relative risk:

	Predict 0 fatalities	Predict ≥1 fatalities
Observed 0 fatalities	2007	2014
Observed ≥ 1 fatalities	10	43
Total	2017	2057
% with fatalities	0.5%	2.1%

  So although the turbines predicted to be more dangerous were about 4 times more likely to experience fatalities than the turbines predicted to be less dangerous, 97.9% of those predicted to be more dangerous experienced zero fatalities. On p.222, par.3, the authors argued that this large rate of false positives is attributable to the short duration of sampling. If so, then the turbines studied for 4 or more years should show a stronger response. Is this effect stronger for the turbines studied for longer periods?

- p.226, p.229, p.231, p.235, Figures 7-24, 7-26, 7-28, 7-30: In the A part of each of these figures, the authors have again grouped the observations by the results and not the predictions. Since they are attempting to evaluate the quality of the predictions, their approach is inappropriate. Like residual plots for logistic regression, the observations should be grouped by prediction (ranges of the scores) and the fraction of turbines experiencing fatalities should be compared among the prediction groups.

- p.242, par.1: The authors state that they "… were unable to account for interactions effects between independent variables." If more appropriate tools had been applied to the model development, investigation of interaction effects would have been straightforward.

- p.242, par.2: The authors claim that elimination of 20% of the turbines might reduce the mortality by 40%[17]. How was this determined?

- p.243, par.2: The authors state that the Bonus, Micon, and KVS-33 turbines are the most dangerous. How was this determined? It is likely the authors intended to include the KCS-56 instead of the KVS-33 based on the total bird fatalities reported in Table D-3. Is it possible that there are more fatalities for these turbines because there are more of them, not that they are more dangerous per unit?

- p.245, par.2: The authors state that wind turbines that are at the end of strings or are isolated kill more birds that wind turbines on the inside of strings. It is important to keep in mind that carcasses tossed far enough by a wind turbine that is on the inside of a string can be misattributed to either its left or right neighbor. Wind turbines at the end of a string can only have their kills misattributed to another wind turbine only if it tossed towards the string. Wind turbines that are isolated will not have any chance of getting their carcasses misattributed.

Chapter 8: Bird Behavior in the APWRA

- p.246, par.4: Biologists only collected bird behavior data from mid-October through mid-May. What about mid-May through September, especially since summer is when the winds are strong? Perhaps young prey or different types of prey are available more during certain months? Also, how were the 61 observation plots selected: randomly or by convenience?

- p.247, par.2.: The observation plots had a fixed radius of 300 m, so the term *variable distance circular point observations* is not really appropriate. Variable-radius plots are more commonly used in so-called "distance based sampling" in which the distance to *each* bird observation is used to estimate probability of detection as a means of calculating bird density (which is not the intent of the authors). The authors did assign birds to one of 3 distance categories (based on distance to turbine), but the furthest category was truncated at 300 m. As

[17]Original review text before considering the Smallwood and Thelander response: "...by ~~80~~%."

Reynolds (1980) states, "With the variable circular plot method no maximum distance restrictions are placed on any observation" (p.310). "Distance-based sampling" is a large sub-discipline within wildlife ecology and boasts a sizeable literature (see Volume 119 Issue 1 [2002] of *The Auk* for several recent papers on this subject), and while Reynolds et al. (1980) is a classic citation and influential in the development of current methods, it is not up-to-date with recognized methods.

- p.247, par.3: The authors state that the 61 observation plots were sampled 4 times each or "once every three to four weeks". How can the sampling cover 210 days and at the same time be once every 21 to 28 days? With one sampling at the start and one at the end, the interval between samplings would need to be about 70 days.

- p.250, Table 8-2: More explanation is needed to distinguish the types of flight behavior in Table 8-2. Contouring and surfing sound alike.

- p.251, par.1: The authors assume that, "the number of on-the-minute observations represented the same number of continuous minutes of the same activity." This is a standard assumption with conventional wildlife behavioral sampling, and is likely valid if sample sizes are large enough. This issue has been discussed extensively in the literature (see classic book by Martin and Bateson, 1993), the authors should make use of citations on the subject and defend that the assumption is valid. Also, they should identify their sampling technique within the conventional behavioral sampling lexicon – i.e., there are very standardized differences between focal animal sampling, scan sampling, and instantaneous sampling. The authors likely did the latter, but they should review these terms and identify which best describes their approach. (Martin, P. and P. Bateson. 1993. Measuring Behavior, An Introductory Guide. Cambridge Univ. Press, London, UK.)

- p.253, par.5: Chi-square tests are performed to test for disproportionate behavior under various conditions. Observations (data points) used in a chi-square test should be independent of one another. Having a single bird provide multiple observations through time removes that independence, thus invalidating the chi-square analysis. If a bird is soaring one minute, it is more likely to be soaring during the next minute. Even if a bird only contributed one observation; it could be recounted as a new bird if it disappeared for only 30 seconds (p.247, par.5).

- p.260, par.3 and p.260, Figure 8-9: The authors state that an asymptote for some behaviors is reached by about 9 minutes and for others by 20-27 minutes. It is not clear what asymptotes they are referring to. The vertical axis on Figure 8-9A

does not include zero, which exaggerates the magnitude of the change. Why did the frequency of behaviors increase with time? Does this suggest birds took some time to habituate to human presence (as suggested by Reynold et al. 1980 and others)? Or does it mean in took 8-30 minutes for observers to begin to fully "notice" (authors' term) behaviors in the observation plots? The term *special behaviors* is inadequately defined.

- p.256, par.5: The authors absolutely did not observe 855 minutes of flying; they recorded 855 incidences of flight among 3884 observations at minute intervals. There is a difference between these two. This is a problem with equating minutes of an activity with frequency of its observation at 1-minute intervals.

- p.256, par.6 and p.262, Figure 8-11: The authors state that Figure 8-11A shows the relationship between the number of flights through the rotor zone and the total number of flights observed during a session. What is the slope, r^2 value, or standard error estimate for the relationship? Is this a chance pattern? Regardless, it makes sense that if there are more incidences of flight, there will be more incidences of flight through the rotor zone. And if birds are perching – thus not flying – there will be fewer incidences of flight through the rotor zone.

- p.264, par.2: Were any bird collisions with turbine blades observed?

- p.265, Table 8-3: The table totals for the sum of minutes of flying (855) does not match the total of the column (828). Are there other raptor results not tabulated? Similarly the total provided for the sum of minutes perching column is 3029 but the column total is 2909. And the total given for the number of flights through the rotor zone (153) does not agree with the column total of 147.

  The turkey vulture, red-tailed hawk, and American kestrel account for 87% of the minutes flying and 90% of the flights through the rotor zone, but according to Table 3-1 they only account for 22.9% of the total turbine caused fatalities and for 58.1% of the total raptor fatalities cause by collisions. Why this great disparity?

- p.266, Table 8-4: In this table there are several behaviors or groups of behaviors that have zero recorded minutes of activity for all listed species and yet three other flight behaviors listed in Table 8-2 are not included (e.g., high soaring, mating, and land). Why were these omitted?

- p.267, Table 8-5: There is a discrepancy between the minutes perching for American kestrels between this table (1065) and Table 8-3 (1103).

- p.269, par.4: Many of the environmental variables may have coincidentally been correlated with when the birds were sighted. For example, "Golden eagles and American kestrels perched more often than expected by chance during cooler temperatures, which was also more or less when they flew more often." So were Golden eagles and American kestrels mostly observed during the cooler months? Would *such confounding*[18] also cause an association with certain seasonal types of wind? And how can they be perching more and flying more at the same time? Would not one increase while the other decreases?

- pp. 270-275, Tables 8-6 through 8-11: The authors have again conducted 132 univariate hypothesis tests without correcting for multiple comparisons.

- pp. 283-307, Tables 8-12 through 8-16: This time there are 792 simultaneous tests conducted without correction for multiple comparisons.

Chapter 9: Conclusions and Recommendations

- p. 339, par.3: The authors state that birds are disproportionately killed by wind turbines mounted on tubular towers. However, because of the tubular vs. lattice towers differ in many other respects (rotor length, tip speed, blade height, etc.), without examining effects of tubular vs. lattice towers while controlling for the other confounding variables via multivariate analysis, the univariate analyses are suspect.

- p.353, par.5: The authors state:

  > "We also had little control over the application of sampling effort across the APWRA, and so the differential sampling effort we applied precluded multivariate statistical methods, which would have been useful for managing the shared variation among measured variables. These factors required us to rely on univariate tests."

  The lack of management of shared variation among variables is indeed a major limitation of this study. But unrepresentative, incomplete sampling is a problem for univariate as well as multivariate analyses. There is no reason why the authors cannot employ more state-of-the art analytical tools to try to disentangle

[18]Original review text before considering the Smallwood and Thelander response: "Would ~~that~~ also cause..."

the multiple measured variables, with the strong caveat that the sampling was likely inadequate.

ATTACHMENT C

Review Team #3

September 8, 2006

Review of 'Developing methods to reduce bird mortality in the Altamont Pass Wind Resource Area' by K. S. Smallwood and C. Thelander (2004)

This final report is a revision of an initial report that considers the work of Smallwood and Thelander (2004). I have responded to their comments regarding broader issues raised by all of the reviewers first. I then consider their responses to my specific comments in my initial report. Changes in my initial report are represented in italics. In some places, I state the authors' response and then further explain my point. In other cases, my initial statement and their response require no further comment. Clearly, the authors considered the reviewers' comments and thoughtfully explained how they would modify their methods if future work was to be performed.

Pseudoreplication was the first broad issue raised by reviewers to which the authors responded. However, the first 2 paragraphs under this section (bottom of page 5 and top of page 6) are not about pseudoreplication, rather they refer to nonrandom selection of turbine strings. Hence, I will address these paragraphs and related material first and then consider the pseudoreplication issue offered in the third and fourth paragraphs under the first subtitle (i.e., second and third paragraphs of page 6).

Nonrandom sampling of turbine strings

Smallwood and Thelander agree that random sampling initially would have been preferable, but they disagree with the premise that statistical inference depends on samples being randomly selected. Thompson (*Sampling* 2002:2) states 'Sampling is usually distinguished from observational studies, in which one has little or no control over how the observations on the population were obtained.' Subsequently, he emphasizes that random sampling can avoid many of the factors that make observational data "unrepresentative". The take home message here is that Smallwood and Thelander have performed an observational study. Observational studies are commonly performed in the biological sciences. However, as such, they do have limitations. One of these limitations is the degree to which defensible inferences can be made. In my initial review, I questioned the inferential capability of this work because of the nonrandom sample of turbine strings. I encouraged the authors not to infer to the larger APWRA because they do not have a random sample from the APWRA. While this is not their fault (they surveyed all turbines to which they were granted access), nevertheless, its implications remain. A nonrandom sample does not necessitate that their results are not representative. For instance, Cochran (*Sampling Techniques* 1977:10) states that under the right conditions, some common types of nonrandom sampling can give useful results. However, he further states that the only way of examining how good one of them may be is to compare the results to a situation in which the results are known, either for the whole population or for a probability sample. If one acknowledges that inference should only be made to those turbines which were surveyed, the question then becomes does the time period in which they surveyed reasonably represent the temporal nature of the system, for

which I have no answer. However, at least they have multiple years of surveys and multiple surveys within each year for the portion of turbines surveyed.

On page 7, Smallwood and Thelander state they could not have purposely biased selection of turbines- I have no such suspicions. I am only concerned about unknown sources of bias due to nonrandom selection. The authors selected turbines systematically to intersperse searched and unsearched turbines. This is a common approach used in biological studies. However, a systematic approach can be a random method if the initial starting point is randomly selected from within the collection of units to be sampled, thereby allowing all intervals of sampling units to have a nonzero chance of selection. A systematic random sample has the same structure as a cluster sample, and can be treated as a cluster sample of size one. Variance estimation is not straightforward for such samples, unless the systematic sample is replicated (i.e., more than one collection of spaced units is selected). I encourage the authors to read chapter 12 in Thompson's book (Thompson 2002).

Pseudoreplication

To properly discuss this issue, one must identify the sampling or experimental unit. First, an experiment has not been performed. Treatments have not purposely been applied to subjects with the notion of eliciting a response. Second, has a sample of turbines strings been examined? The answer is yes, but it is a nonrandom sample of the entire APWRA (to be discussed subsequently). What was measured? On page 47, the authors indicate that turbine strings were surveyed using transects. In this sense, the turbine strings are the sampling unit. Turbines within strings might be considered subsamples, depending on the situation. Certainly, individual turbines within strings are correlated spatially and temporally in terms of measurements made. Thus, use of individual turbines must be considered carefully so as not to inflate sample sizes in statistical tests.

Extrapolation

Extrapolation of the mortality estimates may be looked upon as acceptable by some researcher, but statisticians as a whole do not encourage such practices because extrapolation is inherently problematic. The authors state they believe the extrapolation was reasonable, and attempt to justify it based on the interspersion of turbines searched, etc. 'Trust me' is not the basis of good science. Point estimates are of little value in estimation. Suppose I was to ask the authors to develop a 90% confidence interval around their extrapolation, could they do so? In regression modeling, there exists a defensible means of doing so, but only within the range of the data collected, so why then are the authors so willing to go beyond the range of their data in this case?

The authors clarify that they did not extrapolate the results of fatality associations to wind turbines outside the measured set. My question is then, to what population are the statistical tests referring? I suspect the answer might be that their population is the collection of mortality strikes annually at the wind turbine set. If so, the important question then becomes were the turbines searched enough throughout the year to adequately represent it in its entirety, and if not, were the times within the year randomly selected so as to infer to an annual timeframe.

Confounding
Their response to this problem was basically an acknowledgment that it may have occurred. They give an example of how they could reduce some of these effects if given the opportunity to revise the report. While I agree that not all instances of confounding might have caused misinformation, I remain concerned that some of the conclusions drawn might be inappropriate because of lack of attention to the issue. My recommendation to the authors is to identify the degree to which certain variables are confounded, i.e., look at the 'treatment structure' of the turbine strings and identify situations in which confounding can be reduced or eliminated. When confounding cannot be averted, clearly state this whenever a test is performed and when interpreting the test result. For instance, they could say, 'it remains unclear whether variable B or C is actually associated with this high mortality rate'.

Multiple Comparisons, Multivariate tests, and use of Chi Square Tests
I agree with the authors that alternative analysis methods will not eliminate all of the problems associated with the data set, e.g., nonrandom sampling, lack of replication, etc. The selection of variables was under their control as were the analytical methods and it is here that more thought should be placed. I recommend the authors consider use of some form of multiple regression in combination with model selection methods. The latter may rely on information theoretic approaches, Mallow's Cp, likelihood ratio testing, etc. Specific methods should be discussed in consultation with a statistician. For instance, logisitic regression could be used with the response being two categories (deaths occurred or not) vs. several explanatory variables, including effort. Polytomous logistic regression would allow for more than 2 categories, so they could have high mortality, low mortality and none as the response variable. An alternative would be to use generalized linear models such as Poisson regression, in which effort could be an explanatory variable, or using multiple regression with a response variable of mortalities per unit effort. I understand the idea behind the authors' use of the Chi square goodness of fit testing, but believe this approach is inferior because it is only considering one variable at a time. The authors later sum over individual variables, but there is no basis for doing so under the sampling design and replication with which they had to work.

Type I error
The authors response to this criticism has identified a misunderstanding of what Type I error and P-values represent. First, probability is prospective. Second, whether or not a Type I error has been committed does not depend on the P-value (other than the test had to be rejected, or below the stated alpha). If I were to perform 100 statistical tests using an alpha level of 0.05 in each, then it would be very reasonable to have 5 rejections even when none of the null hypotheses were false. The authors' response is that many of their P-values were small. This does not change the above interpretation; notice I did not even mention the P-values for the 5 rejections. A P-value is not a measure of effect size and it is not legitimate to adjust an alpha level based on a p-value. Another viewpoint from the authors is that they are willing to live with some type I errors. For example, suppose 100 tests were performed and 15 were rejected. Again, let's assume 5 were rejections of true nulls. The important question would then be which ones are false rejections and which are not. The authors are less concerned about answering this question than I expected.

What follows is an edited version of the initial report

General Summary

Smallwood and Thelander do an excellent job of describing background information and previous literature regarding bird mortalities at wind farms. A strong case is made for the relevance of the work they performed. Their objectives were multi-faceted and involved examining bird behaviors, raptor prey availability, turbine characteristics, landscape features, and bird mortalities. Each of these components is a substantial endeavor and the authors are to be commended for examining this multitude of factors. Clearly, the authors used a 'heads up' approach that involved making observations, and then attempting to collect data and evaluate hypotheses formulated from these observations. For example, they expanded their research to examine fossorial animal burrows and effects of rodent control on bird mortality.

While I believe they have substantial information regarding all of their objectives, the details of how to extract this information and to what the information refers to become the most important considerations I have regarding the report. I believe that alternate analytical methods should be explored for examining their data, which may lead to different interpretations of their data. Even if the interpretations remain similar, I would have greater confidence if their analyses were improved upon. Furthermore, I would suggest extensive consultation with a statistician to assist in such an endeavor. *Smallwood and Thelander have indicated a willingness to do so if allowed to revise the report.* It is my belief that a great deal of thinking and interaction with a statistician is needed to ensure that the information contained in this impressive data set is properly interpreted. Given the variety of data collected, including different scale issues, multiple response and explanatory variables (they refer to as dependent and independent, respectively), various data types (categorical, continuous, ordinal), etc., I believe many hours of interaction is needed to identify the most prudent analysis approach for each of the chapters. I would anticipate spending well over 100 hours on such a project in a statistical consulting role.

My criticisms that follow are meant to be constructive, not destructive, and I hope that the authors will take them as such. I likely have misunderstood their intentions at various places in their report and I assume the authors will correct my thinking on these portions of my assessment. Thus, I suspect some of my criticisms can easily be rectified. There was an opportunity to submit questions and have them answered by the authors. However, in my experience, the importance of the nonstatistical aspects of statistical consulting cannot be understated. For example, direct verbal interaction is the best means for arriving at a consensus understanding of the material. Use of reflective listening is a very useful technique for making sure everyone is 'on the same page'. The format for this review did not allow such interaction; however, I recognize the utility of the process employed in this review.

Strength of inference is determined by the study design. Strength of evidence is determined by the data alone. In this case, the study design does not lend itself to a strong inferential setting for two reasons. First, the variables of interest are likely confounded because turbine string placement was not designed with their study objectives in mind. Hence, there are many factors that potentially affect the response variables of interest that are not separately estimable because all combinations of

explanatory variables are not represented on the landscape. For example, if one was only interested in the effects of aspect (north, east, south and west) and tower height (say 4 categories), then one would need 16 tower-aspect combinations represented, with replication, for sufficient estimability of main and interactive effects. Thus, this study lacks replication of the set of 'treatment' combinations of explanatory variables being examined (see Johnson 2002 for a discussion on the importance of replication in wildlife research).

Smallwood and Thelander indicate some misunderstanding of this issue in their response, in which they state the above problem is the reason they tested one variable at a time. There may be 'replication' of most of the variables, but there is substantial confounding of these variables. When variables are confounded, their effects are not separately estimable, therefore one-variable-at-a-time analyses can be misleading. Tests can be conducted one variable at a time, but such tests do not escape the problem of confounded effects. For example, if explanatory variable A affects the response variable C, but explanatory variable B does not, a test of B's effect on C is likely to be detected if A and B are confounded, even though it has no effect, because it is confounded with A, which does have an effect. Ultimately, their approach may have led to several spurious effects, upon which management recommendations might be made. Confounding is not something that can easily be addressed, it must be seen as a problem that limits inference. In addition, there are good reasons why multiple regression is commonly used (two or more explanatory variables in one model) as opposed to multiple versions of simple linear regression using one explanatory variable at a time. Multiple regression enables one to assess the impact of one explanatory variable in the presence of other explanatory variables. More complex models typically are more helpful in providing sufficiently precise predictions. When using multiple regression, one must be mindful of potential multicolinearity problems when two or more explanatory variables are highly correlated. Multiple regression is not an example of multivariate methods, the latter refers to the situation in which there are multiple response variables, not multiple explanatory variables. Their approach of univariate Chi-square goodness of fit tests to identify individual variables of interest, and the associated metric of the ratio of observed to expected as a measure of effect (which they later sum over for certain variables), is more of an ad hoc approach that does not readily draw upon formal statistical procedures for model selection and testing. As stated earlier, I believe a substantial amount of communication should take place with a statistical consultant to best arrive at an understanding of some important statistical issues, which I can only assume are poorly understood at this point.

Second, because the sample of sites studied were not randomly sampled any inferences based on statistical tests are not statistically defensible. The collection of sites studied represents those locations to which the researchers were allowed access or which were accessible due to environmental conditions. I do not fault the researchers for this shortcoming; they surveyed what was made available to them. However, it does weaken the strength of inference they are able to impart and I suggest they reconsider the context of many of their statements of inference. For instance, hundreds of tests (Chi-square, analysis of variance, tests of correlation, etc.) are performed in the report, but the population(s) to which inferences are made is at best, unclear, and at worst, inappropriate.

In response to this criticism, the authors clearly state that they had no investigator bias. Randomization is important to avoid any source of bias, many of which are unknown to the investigator (I never meant to imply there was an investigator bias). For instance, the collection of turbines surveyed may underrepresent or overrepresent a particular landscape attribute, which may affect likelihood of strikes. In my initial report, I acknowledged that the nonrandom sampling was partly unavoidable and does not necessitate that their sample of turbine strings is poor. They surveyed all of the turbines in the first set of turbines that were made accessible. However, in the second set, they selected a portion systematically. Systematic sampling is commonly in biological studies because it forces a spatial evenness to the sampling. Before sampling has begun, one can envision a set of sampling units, called a primary unit, separated by a specific interval. As long as the initial selection of the primary unit is randomized, a random sample has been achieved. Often, researchers use their judgment to select the primary unit, which needlessly results in a nonrandom sample. It is unclear how the authors selected the turbine strings in the second set. In any case, if the population being inferred to is the entire APWRA, a nonrandom sample of this area has been taken. One approach the authors could take would be to simply describe the results from their sample. There would be no uncertainty in these descriptive statistics, although there are detectability issues that would imply the estimates presented are below the true mortality number. Such an approach would eliminate the extrapolation of mortality estimates to the entire APWRA that I criticize later. If the authors choose to infer to the entire APWRA, then such estimates cannot be defended statistically, i.e., acceptance depends on the judgment of their audience.

In some instances, the authors did have the opportunity to randomly sample from the collection of turbines to which they had access. Granted, this target population is not the entire APWRA, but random sampling would have justified inferences to a larger population. For example, when studying rodent burrow locations, the authors have not indicated that a random sample of turbines or turbine strings was selected from those available; they stated turbine strings were 'arbitrarily selected'. Thus, I am not clear as to what population the inferences are being made.

Hypothesis testing is one form of statistical inference. Estimation of parameters is another form of inference the authors employed. Limitations due to nonrandom sampling must also be considered in this context. For example, in estimating bird and raptor mortality (e.g., see executive summary) for the Altamont Pass Wind Resource Area (APWRA), a nonrandom sample of turbine locations was surveyed from the entire APWRA. While I do not blame the researchers for this circumstance, it does remove the defensibility of statistical inferences. Random sampling does not necessarily ensure a representative sample, it only ensures that on average, in repeated sampling, the population will be well-represented (a conceptual property). That the authors have surveyed a majority of turbines provides some support for the notion they have a good representation of the population, however, only approximately 28% of the turbines were measured at least 3 years. If the authors can make the case that their sites are representative of the larger APWRA, then perhaps the scientific inference (not statistical) being made will be acceptable to all. I realize that many of the environmental attributes might be unknown for those sites not visited, but I would think that the turbine/tower features would be known for all turbines in the APWRA and could be considered within

the context of those that were sampled. In their executive summary, they estimate the number of raptors (and all birds combined) killed annually in the APWRA. I suggest they *(changed the wording here from Personally, I would*) refrain from making such inferences and limit estimates to the area specifically surveyed, given the strength-of-inference limitations due to nonrandom sampling and the uncertainty in the 'adjustments' they make in their estimation process. Their projections for all wind-generating facilities in the United Stated (page 86) should also be considered as extrapolations without much credence.

They further state that the risk to birds has increased substantially over the past 15 years, indicating a formal trends analysis. This support for this statement is not satisfactory unless more information is given on consistency of detection rates over this time period. Note there are many factors that can cause inconsistencies in observed counts over time, including surveyor differences, environmental differences, and animal behavior differences. The authors did suggest that birds may have altered their behavior in response to the presence of turbines in the area. I suspect that there are many survey methodological differences over this time span as well. *The authors indicated a willingness to replace chapter 4 with a more rigorous treatment of the issue if given the opportunity.*

There is also a temporal component of their sampling which is clear in the general sense that turbines were surveyed a different number of occasions, etc., but the details of how often/when specific turbines were surveyed and how this might affect interpretations are not explicit. For instance, if tubular towers were surveyed more often and at times of higher bird abundance than other tower types, then greater mortalities observed may merely be a function not of the tower type, but greater effort and timing of greater bird abundance. Some sites (1526 turbines) were surveyed over a 3-year period, with a staggered entry of turbines over that period, so turbines had differing numbers of surveys executed within that period. For example, on the middle of page 48 they state they were limited to 685 turbines through 1999. Another set of turbines (2548) were surveyed over a 6-month period. If I understand their methodology, the authors computed estimates of availability that account for survey effort by placing landscape and turbine features on a relative (proportional) basis. However, in most use-versus-availability studies I have seen, 'availability' is assumed known, when it is almost always estimated. What is 'available' from the human perspective of what is on the map, is not necessarily available to the animals of interest, even if they are highly mobile, because they may not have such a map in mind when making decisions. Alldredge et al. (1998) provide a nice overview of statistical approaches to resource selection studies that nicely clarifies the set of assumptions underlying such analyses. If the authors do not modify their analytical methods, which I strongly recommend, then at the least they could more explicitly state the assumptions underlying their approach, the likelihood that the assumptions are valid, and the ramifications if not valid. *The authors have indicated a willingness to do so if given the opportunity to revise the report.*

Another important design component includes consideration of what the multiple competing hypotheses are and how best to discriminate among them. When possible, readers of this report should be informed of the hypotheses under consideration and how the sampling scheme used can discriminate among these hypotheses. For example, there is an entire section of work in chapter 2 that examines the distances of 'small' versus

'large' birds from the turbine, yet there is no explanation of why the data are being partitioned as such, i.e., what the hypothesis is, and how this partitioning relates to assessing the efficiency of their search radius. Another example of the importance of considering one's hypotheses is demonstrated in chapter 6. The objectives are clearly stated, but the *a priori* hypotheses regarding the effects of rodent control are not stated. They allude to the notion of ineffectiveness, but I would like to see explicit hypothesis statements. Lacking the benefit of observations, my scientific hypothesis would be that increased intensity of rodent control results in fewer raptors, thus lowering susceptibility to strikes and thus lowering observed fatalities. *The authors stated that they would strive to clarify their hypotheses if given the opportunity for revision.*

How best to discriminate among hypotheses is an important design consideration. For example, in studying the effects of rodent control, the authors did select sites with a wide range of observed raptor mortality and rodent control intensity. This approach enhanced the ability to distinguish among competing hypotheses. However, they did not random sample turbine strings according to these features (e.g., a stratified design), thus limiting the defensibility of inferences made.

Care must be taken when attempting to demonstrate 'treatment' effects. For instance, the treatments must be effective in their application. My understanding is that rodent control was aimed specifically at eliminating ground squirrels, but not other species (e.g. pocket gophers). Clearly, if one species is targeted, that does not necessarily imply a significant reduction in overall prey availability, in fact, it may increase it. *On the other hand, elimination of one species might focus raptor presence in other locally defined areas with other prey species. The important question would then be are these areas more or less susceptible to bird strikes?* Thus, I question if the rodent control 'treatments' were substantial enough to observe an effect. *In their response, the authors assured me that the treatments were effective in killing ground squirrels*. Rodent control might be effective in reducing raptor susceptibility if prey availability (in total, not just ground squirrels) is substantially reduced and this reduction is discerned by raptors. How best to isolate treatment effects is also an important component in identifying treatment effects. In applying the rodent control treatments, were other variables that influence raptor mortality controlled? If not, not, then confounding effects may make it difficult to isolate the effect of the rodent control treatment. In summary, I suggest that the efficacy of the management actions taken be considered before discarding their usefulness. Similar arguments apply to topics such as benefits of perch guards, etc. Smallwood and Thelander have made decisions and recommendations based on observations from considerable survey effort. Again, I believe there is tremendous value in their data, interpretations of which must be carefully considered for that value to be realized. *In responding to this statement, the authors state all of their conclusions were based on evidence. While I agree that they are using evidence (data), the details of corrected interpretation of that data become important in determining validity. For instance, the presence of confounding effects may indicate a variable has an effect when in fact it does not, or vice versa (one variable may counteract another). I was simply stating that a rigorous experiment should be performed that allows isolate of specific treatments to identify treatment effects. I do not believe that such an experiment has been correctly executed as of yet.*

Regarding strength of evidence, the authors have completed a notable amount of work. Having surveyed the majority of turbine locations, some of which were surveyed multiple years, I believe there is substantial information to be discerned from the collected data. The key to harvesting that information is proper context and hard thinking about what metrics make sense, appropriate use of statistical tests, placing outcomes of statistical tests in the context of biological relevance, etc. In attempting to determine causal factors of bird mortalities, the authors surveyed locations around turbines in the APWRA that were accessible. They identified all known bird strike mortalities (approximately 1045 if the number due to unknown causes is removed) and examined associations of various turbine/tower, landscape, and environmental factors with these mortality counts. Thus, a retrospective observational study has been performed with the intent of determining causal relationships. I believe most statisticians would agree that establishing causation requires a more rigorous approach in study design. Romesburg (1981) stated that causation requires more than correlative evidence; one must eliminate other possible causes and must demonstrate similar associations that are plausible over a wide range of circumstances. Many conclusions stated in the report are plausible, but I am not convinced that causation has been established anywhere in the report.

Their analysis approach was generally that of examining associations one variable at a time. Thus, hundreds of computations resulting in hundreds of statistical tests were performed. One of the problems with such an approach is that there are likely to be spurious effects. The probability of falsely rejecting a true hypothesis at least once is essentially one when more than 50 tests are performed. Another problem with examining one variable at a time is that there are likely to be confounding effects unless all combinations of factors (34 variables) are represented. For example, if certain turbine models tend to be at higher elevations, or appear on canyons more often, then greater than expected raptor mortalities may be due elevation or landscape considerations rather than turbine model itself. Alternatively, it may be that an interactive effect exists, such that the combination of various features increases the potential for strikes. The one-variable-at-a-time approach to analysis can mask such outcomes. The authors recognize the potential for confounding effects in other studies (see appendix B, page 2), and even admit a small portion of their study (rodent control in chapter 6) is prone to confounding, yet they fail to see the possibility of confounding effects in the majority of their one-variable-at-a-time analyses. I am confident that there are several reasonable approaches that can be made to analyze their data considering multiple explanatory variables simultaneously. For example, I suggest that logistic regression or Poisson regression be considered for some types of analyses. These modeling approaches will allow for multiple explanatory variables and will not necessitate the data reduction that has been used to categorize certain continuous explanatory variables.

In summary, I believe there are specific design flaws which limit the validity of inferences to a larger population (which in some cases is not clearly defined). Reliance on descriptive information may not be seen as 'scientific' but I disagree with the notion one must make inferences to have useful information. Second, I suggest the authors rethink their analytical approaches. More appropriate methods of analysis would strengthen my belief in their stated outcomes and recommendations. Extensive consultation with a statistician is recommended.

(I deleted a section here that began as 'If I were considering this for publication…' because the point was moot.)

References:

Alldredge, J.R., D.L. Thomas, and L.L. McDonald. Survey and comparison of methods for study of resource selection. Journal of Agricultural, Biological, and Environmental Statistics. 3:237-253.

Johnson, D.H. 2002. The importance of replication in wildlife research. Journal of Wildlife Management 66:919-932.

Romesburg, H.C. 1981. Wildlife science: gaining reliable knowledge. Journal of Wildlife Management 45: 293-313.

Executive Summary

Specific Comments

In describing their approach, the authors state they presented mortality estimates as ranges, where the lower end was adjusted for likely outside of their search area, and the upper end was adjusted for fatalities missed due to undetected carcass removal. I would consider both of these to be upper-end adjustments, actually using both simultaneously would provide a higher upper end. The lower end would be represented by unadjusted values *(Smallwood and Thelander acknowledge this in their response)*. The upper-end estimates must be interpreted with caution. The adjustments made are based on detectability estimates from other studies that are likely to be study-specific for a variety of reasons. In addition, the range for the estimated number of raptors (and all bird combined) killed annually is given for the entire APWRA. I suggest the authors consider the accuracy of their estimator given the design weaknesses at the larger scale of sampling the APWRA and at the smaller scale of bird carcasses detected. They further state that the risk to birds has increased substantially over the past 15 years, however, I am not convinced that the data have been collected in a manner that allows for trend estimation due to likely inconsistencies in survey methods, personnel, effort, animal and environmental differences.

Justification for their defined metric of mortality as mortalities per megawatt (MW) per year *was not convincing (I modified this ending from 'is not properly stated'). As with any metric, the variable of interest must be clearly defined.* They give the reason of 'to avoid the false appearance that larger turbines kill more birds'. If total number of fatalities at a site is the variable of interest, then *their metric* is *in*appropriate. To compare deaths as a function of turbine size, then fatalities per turbine *(e.g., grouped by size categories)* is an appropriate metric and does not give a 'false appearance'. *I now understand that by using their metric that includes megawatt production, they attempted to adjust for time of operation, i.e., it is used as a surrogate. I agree that this factor should be considered in one is trying to identify which turbine types are more dangerous, but use of MW production is only a partial surrogate for operation time (which they admit in their response). So, for example, if 100 birds are killed by 100 turbines of each size class operated the same length of time, I would conclude that bird mortality is equivalent between the 2 turbine sizes. However, if one accounts for MW production, the larger turbines would have a smaller metric of mortalities per megawatt per year*

assuming that larger turbines generate more power. Based on this scenario, I can only conclude that on a cost-benefit basis, the larger turbines are better, but in an absolute sense of fatalities, larger turbines are no better than smaller ones. By incorporating the MW produced by each turbine, they have simply factored in the benefit of generated power in this cost representation. *Wind speed also underlies in the MW production number, thus it is difficult to isolate tower size effects with any metric unless one can control the other factors (operation time, wind speed, etc.). I hope this clarifies my initial point that the metric to be used depends on the objective of the measurement itself.* There are advantages of using this metric, many of which are stated in appendix A, but the advantage depends on how the metric is to be used.

The authors state that at least 3 years of carcass searches are needed before stabilization of the percentage of non-zero mortality values. Are they saying that if a sample of 100 turbines is surveyed, at least 3 years are needed to estimate the percentage of those 100 that kill at least one bird? I am not convinced this is a useful metric. Rather than focusing on the turbines where zero, or even an occasional mortality occurs, should not the focus be on those characteristics at turbines where numerous mortalities occur (e.g., see figure 3-4). They proceed to interpret this result by stating that one must survey at least 3 years before getting a 'good' estimate of mortality rate. Mortality rate (expressed by fatalities per turbine per year) is not the same metric as percentage of turbines with at least one fatality. While I agree that more data is better for estimation in general, they have not demonstrated 3 years of data are necessary for a 'good' estimate of the mortality rate. The term 'good' in the context of bias would require knowledge of true mortality rate. The term 'good' in the context of precision would require some definition of what precision is needed for the estimates to be useful. *The latter can be examined with their data for those turbines searched multiple years. In other words, they could report the variation in estimates from one year to the next to estimate how many years of data are needed to achieve estimates within a certain margin of error with a stated level of confidence. In other words, given variance estimates, sample size estimation is possible.*

I disagree with the general statement that their test results for associations were statistically and biologically sound. Reasons for my contrary opinion are provided throughout my review. The predictive power of their 'models' reported in this section and later is likely a poor indicator of model validation given that the same data used to construct the models has been used to test them.

Chapter 1 Understanding the Problem

General comments

The first chapter gives an overview of the project, the objectives and their general approach toward meeting those objectives. Important definitions are made regarding their usage of terms like susceptibility, vulnerability, etc. For instance, vulnerability is measured here on a relative, not absolute basis. *The clarity provided with these definitions is helpful to the reader. However, many other terms are not as clear.* For instance, how is *bird* use measured *(I previously used the term 'habitat use')*? -this is a very important question when interpreting results. How close does a bird have to be to the reference point (e.g., rotor) to count as use? The authors use the word 'nearby' wind

turbines, but I am uncertain what that implies. Is flying over an area for a few seconds treated the same as when a bird perches or hunts in the same area over several minutes or hours? Their phrasing suggests they consider the proportion of sampling periods in which use was detected, but this does not indicate duration of use per se. How do they treat observations of multiple birds at the same time? Are pairs treated as one observation? *Smallwood and Thelander responded by stating this chapter was meant to be conceptual and as such, perhaps these comments are premature.* Many of these questions are explicitly answered in later chapters, but some are not.

Specific comments

The authors correctly state that a preferred study design would have been to use a before-after control impact design but that such was not possible given the prior existence of the turbines. On the bottom of page 9, the authors present a 'model' for vulnerability as the ratio of observed and expected use. I suggest they restate this as a metric, not a model. It is not clear why the Chi-square symbols are in the numerator and denominator of this expression. A Chi-square statistic can be computed based on the sum of squared differences of observed and expected values, divided by the expected values. The authors should clearly state this is a 'goodness of fit' approach to testing. Section 1.1.3 is nice section on the difficulty of measuring impact. I would like to know, however, how the number of mortalities per year in the APWRA compares to other hazards, such as collisions with vehicles or airplanes, or deaths due to poaching or contaminants. This would give the reader some perspective on the magnitude of impacts of strike mortalities in the APRWA. I realize that for some species, e.g., the golden eagle, car collisions are unlikely, but what about other human-induced sources of mortality? *I only mention this in case the value of wind farms was being questioned (relative to its costs). Similar considerations have been ongoing, for instance, regarding the utility of dams versus the impacts they have on river systems.* Section 1.1.3, introduces the notion that by comparing observed and expected frequencies, one is able to identify which environmental factors might have a causal relationship (see p. 12, 4th and 5th sentences of first paragraph). The term 'might have' is important, because this is merely an observational study and thus causation cannot be established. The objectives are then described; the sampled population (initially) is identified as approximately 28% of the APWRA's turbine population due to limitations placed on access, and a brief description of 'midcourse' corrections is stated. Section 1.1.4 introduces the idea of 'use versus availability' in terms of assessing mortalities and associations with turbine location by considering what percent of mortalities one would expect given random use of the sampled area versus the number actually observed. This reasoning is the basis for much of the statistical testing (Chi-square goodness of fit tests) presented later in the report. I question to what population is the statistical inference being made with these tests. *Smallwood and Thelander state the population is the sample used in the chi-square test. I am confused by this statement. Statistical testing consists of using sample data to infer about a larger population. No testing is needed if one's interest is only the sample. They further state that for the highly significant tests, inference can be drawn to birds using the APWRA. First, the authors are misinterpreting P-values as effect size. Use of the term 'highly significant' demonstrates this misunderstanding (common to many scientists). Second, inference to the entire APWRA is not supported statistically as discussed earlier*

and certainly does not depend on the P-value from a hypothesis test. I agree that by examining the observed/expected ratios, one can describe places where more or fewer mortalities occurred than expected with random use of the sampled area. However, is it reasonable to assume that birds use landscapes randomly? *Smallwood and Thelander admit this is unreasonable to assume, but is the null condition. However, all of their statistical tests are based on this false assumption, which calls into question the value of the P-values that are reported. These are all computed assuming the random use, which is clearly false. Anderson et al. (2000) and others have referred to these as 'silly nulls'.* On page 20, the authors mention a focused study on bird behavior involving about 1500 wind turbines. Did they randomly sample these turbines from the collection of all turbines they studied? If so, then they could make inferences to the larger collection of turbines they surveyed, but again, I would suggest they resist the temptation to infer to the entire APWRA.

References

Anderson, D.R., K.P. Burnham and W.L. Thompson. 2000. Null hypothesis testing: problems, prevalence and an alternative. Journal of Wildlife Management 64:912-923.

Chapter 2 Cause and Death and Locations of Bird Carcasses in the APWRA

General comments

This chapter basically describes the mortalities found by species, cause, season, turbine model, etc. These detected mortalities form the basis for their development of associations, causal factor identification and recommendations based on their general conclusions. Thus, how the data were collected (chapter 3) become very important in interpreting what the data represent. Obviously, the fact that they searched around wind towers suggests they are predisposed to finding mortalities due to collisions than, say, due to natural predation or other factors (e.g., disease). Statistics referring to most mortalities being found near KCS-56 turbines and Bonus turbines or during certain seasons are useful in a descriptive sense, but are less useful in inferring associations unless placed into the context of search effort, availability of turbine types, local raptor abundance, and other factors likely to influence bird strike mortality numbers. My perspective is that the most useful data from this chapter are reported as the percentage of mortalities within their search radius, based on the relative number of birds found outside the search radius. Their recognition that end towers may require a search radius larger than 50m to find 90% or more of carcasses in the 'world of the turbine' is valuable for future survey efforts. Unfortunately, I did not see any attempt to estimate their detectability within their search radius. I see little value of the statistical testing in this chapter (see specific comments below).

Specific comments

At the top of page 30, they state a total of 1162 fatalities caused by collisions and by unknown causes were found. Table 2-1 identifies all 1162 fatalities as wind turbine collisions. Why are the unknowns folded into this column of the table? *Smallwood and Thelander responded by saying they assumed the unknown mortalities were due to brid strikes.* Note that 3 of these observations are bats and 42 are unknown species or group.

By assessing the 'efficiency' of their search radius, I assume the authors are referring to what percentage of bird strike mortalities are contained within their search area. Thus, they are actually referencing detectability and attempting to determine the validity of 100% detection rates. In most biological studies, detectability is less than 100% and estimates of detection rates are necessary for actual abundance estimation. Clearly, the observation that carcasses were found beyond a 50-m radius indicates that their mortality estimates are underestimates of true mortality. I am curious as to why the authors did not expand their search radius to lessen this bias; however, I can appreciate that a larger search area would mean considerably more search effort that logistically may not have been possible. I am curious as to what the detection rate may have been within their defined 50-m radius. The researchers could have directly estimated detection rates using various techniques (double observer, capture-recapture, removal approaches, or distance sampling if actual distances of carcasses were measured for each carcass). Instead, they related distance to carcass as a function of bird body size, wind turbine attributes, season, etc. They later state (page 49 bottom) that they were unconcerned with underestimating mortality, yet they spend much of chapter 2 examining carcass distances to assess 'efficiency of search radius'. What *a priori* hypotheses did they have regarding bird body size and distance from turbine? Given a clear association, how is that useful for determining detection rate? I can understand how certain information such as how carcass distances are associated with turbine height or rotor speed might be used to aid in the design of future survey efforts. For instance, if taller wind turbines displace strikes over a 100-m distance, then a suggestion would be to use a larger search area when quantifying bird strikes. This purpose is stated at the bottom of the second paragraph on page 28.

The purpose of the arbitrary distinction of small and large body lengths in section 2.2 is unclear to me, as is any age classification. The analyses that followed was size-specific, but I do not understand the reason for such a partitioning. *Further, the grouping used is arbitrary (they used a natural break in the histogram of body lengths).* I also do not understand the statement that they lacked sufficient funding to factor in the slope of the hills from each wind turbine. Are they saying they could not afford a clinometer? *The authors clarify in their response that they did not measure inclination during their data collection effort, so to return to make these measurements would require more funding.* I am curious as to how Pearson's correlation coefficient (p. 28 bottom) was calculated for assessing the linear association of carcass distances and elevation of tower base. A given tower base may have had multiple carcasses with multiple distances. Did they treat each of the carcasses as independent observations or did they compute an average distance for all carcasses at a given tower base?

In section 2.3, the authors state on page 29 bottom that most carcasses were discovered during summer and winter. Is that because more surveys were performed then, or a greater abundance of birds were present, or a lesser number of birds were present, but they used the area over a much longer duration than passing migrants do in spring and fall? The number of bird carcasses next to KCS-56 and Bonus turbines is drastically higher than all other turbine types. My question is 'is it the turbine type that predisposes it toward more bird strikes, or are there simply more of these turbines or that they were surveyed more often or are these turbines in places were birds are more abundant as a result of some other attribute, e.g., landscape feature? *Their response to*

this question indicated there were more of these turbines and they were searched the longest. Thus, based on these data, one should not conclude these are the most dangerous types of turbines.

The ANOVAs reported in this section demonstrate statistical detectability of differences among means of carcass distances by tower height. I am not convinced that any of these analyses are useful given the purpose of this data collection. It was my understanding that the purpose of examining bird carcass distance was to 'assess the efficiency of their search radius'. Testing for differences in mean distances in not an effective approach to determining how large search radii ought to be at different tower locations. One needs to look at the distribution of the distances and/or actually estimate detectability of bird carcasses due to strikes with wind turbines. There are 2 levels of detectability here. The first level of detectability concerns what proportion of strike victims are beyond a prescribed search radius? The authors have collected information for estimating this proportion. The sentences that state 'Our search radius included 84.7% of the carcasses of large-bodied birds (90.5% for small-bodied birds) determined to be killed by wind turbines or unknown causes' are the most informative in this section, although I would eliminate the distinction of large and small bodied birds and eliminate the unknown cause counts. Figure 2-12 is also useful here in demonstrating that the 50-m radius contained approximately 95% of carcasses in most cases (the large variances for KCS is curious, the lone (extreme) observation for the Danwin turbine is also notable). The second level of detectability pertains to within their search radius, what percentage of carcasses is found? Later in chapter 3, page 51, they state they adopted a searcher detection rate of 85% for raptors based on Orloff and Flannery (1992) and 41% for nonraptors based on two other studies. How valid these estimates are for the current study is unclear *due to potential methodological differences (personnel, survey method, time frame, etc.).*

Other concerns are 1) why was ANOVA used for the continuous explanatory variables? Regression modeling seems to be more appropriate (which they also report, but do not emphasize). *Reduction of numerical data to categorical results in information loss, loss which is avoidable.* 2) As mentioned previously, there is a potential for confounding effects given only one variable (e.g., tower height) is being examined in each analysis. For example, in the large-bodied bird section, if the 32-m towers were placed more often on sites with greater slopes, then the comparatively large 57m average distance may be a function of slope, not tower height (they recognize this shortfall on page 45 bottom, but fail to see that there are many factors, like blade speed, that should also be jointly considered in analysis). I would suggest the authors consider regression analyses that include several pertinent explanatory variables, rather than a one-variable-at-a-time analysis approach. *If only continuous variables exist, then multiple regression can be used. If both continuous and categorical explanatory variables exist,* such as rotor direction (upwind or downwind) or wind turbine location (end, gap or interior) *then analysis of covariance is appropriate. The main point here is that by only considering one variable at a time, they are blind to problems of confounded variables, and are not able to best identify which variables are most important (when considered collectively). In their response, the authors indicated they believed the one-variable-at-a-time approach forced them to examine the data more carefully than using multiple variable (they use the term multivariate) methods. This logic would indicate that more*

sophisticated models are less useful, when in fact, they are more useful, e.g., have more statistical power. The authors do need to be aware of possible multicolinearity problems when using multiple explanatory variables and thus examining associations of the explanatory variables is recommended. I would also suggest that the authors consider the precision of the model as well as the assumptions underlying its use. If the precision is poor, or if the underlying assumptions are not met, I would not rely on the model and its estimates or predictions. To quote Michael E. Soulé, "Models are tools for thinkers, not crutches for the thoughtless". 3) The differences detected with the LSD tests that are reported are not biologically important in my opinion (e.g., the means ranged from 26m to 33m for large-bodied birds), nor are correlations meaningful when the sample correlation coefficient itself is near zero (see page 44, bottom). If the authors disagree, they need to make a case for why the differences are meaningful, that is they must identify what is a meaningful effect size.

Finally, I do not understand the last sentence in the first paragraph of the discussion (page 45). Clearly, they do have an unknown proportion of actual carcasses given that carcasses were located beyond their search radius *(they agree with this statement), yet their sentence is 'We assume that we did not find an unknown proportion of ...carcasses' Perhaps they intended to say 'We assume that we did not find all carcasses'. Clearly, my misinterpretation of this sentence remains.* Thus, their observed counts of mortality likely do not represent all strike mortalities over the defined spatial and temporal sampling period. They later clearly state this on page 49, bottom. Perhaps they did not intend to have the word 'not' in this sentence. The authors appropriately recognize alternative reasons why bird distances may be identified farther away for turbines at the ends of strings and on hills.

Chapter 3 Bird Mortality in the Altamont Pass Wind Resource Area

General comments

In this section, the sampling methods for counting bird carcasses are described. A staggered entry of turbines was used as access to turbines became available. A second set of wind turbines was added in November of 2002, and were surveyed until May 2003. According to a statement in the executive summary, mortality estimates should not be deemed reliable until 3 years of surveying has been conducted, use of this portion of data would seem to be inconsistent with this statement. *At many places in the report, the authors have qualified results with preceding statements such as 'we have low confidence in the mortality estimates' or 'note that the correlation was low'. But having done so, they then proceed to draw conclusions as if their results are valid and/or meaningful. In their response, the authors claim their suspicions about bias in turbine access issues was warranted.- A conclusion based on low-confidence data. This approach can be commonly found in the scientific literature, but that does not make it tenable. The authors state the assumptions upon which their results are based are invalid, and then later proceed to make conclusions as if the assumption violations did not exists and the data were confirmatory. I believe the data should be reported, but suggest the authors refrain from drawing any conclusions.* For those turbines searched one year or more, temporal coverage is stated to be approximately 7 times per year. Two people searched for carcasses within 50m of each turbine and 50m beyond the end turbine. The authors

state they did not estimate searcher detection and scavenger removal rates because they were unconcerned with underestimating mortality, yet later, they adopt corrective measures for these processes from other studies in estimating mortality *(I deleted the end of this sentence stating this was contrary to their prior stated indifference).* They conclude by stating their mortality estimates might be conservative. However, I would suggest they may also be overestimates if the set of adjustments *for scavenger removal and searcher detection they used from other studies* were not applicable to their surveys. *The authors claim their adjustments were conservative, but I am not convinced that estimates from other studies are applicable to their methods. For instance, if they surveyed more frequently than other studies, scavenger removal rates would be lower. If they were more diligent in surveying for carcasses, then detectability might be higher in their work.* An abundance of statistical tests were performed, testing for time variation in mortality (Tables 3-3 through 3-8). In conducting several hypothesis tests using a Type I error rate of $\alpha = 0.05$ (comparison-wise error rate), the authors are likely to have some null hypotheses rejected due to type I error. That is to say that some null hypotheses will be rejected even though they are true because the type I error rate for the entire collection of tests (experiment-wise error rate) will be much greater than 0.05. *If the authors are concerned about false rejections, they could use a Bonferroni adjustment for the alpha level for their test (which lowers the comparison-wise alpha level as a function of the number of tests performed). In doing so, the probability of at least one type I error is reduced, but the observed P-value must still be lower than the comparison wise alpha level to reject Ho. In essence, a more restrictive criterion is being used to reject Ho. Simply having the P-values listed does not take care of the problem unless they clearly state what the comparison-wise alpha level is.*

Specific comments

Their metric for reporting bird mortality is clearly the number of fatalities per megawatt of power per year. The authors give previous mortality estimates from other authors, but these were reported in deaths per turbine per year. Hence the numbers from this study cannot be directly compared unless one knows the megawatts per turbine from other studies. So, I am a bit confused by the statement (page 47) that their purpose was to estimate mortality so that comparisons could be made to other sites. (I see later in table 3-12, their use of fatalities/turbine/year for these comparisons.) *The authors responded to this by stating that almost all reports now use their metric. I hope that its limitations are obvious to all (see earlier discussion regarding this metric on page 8).* I would also be interested in knowing if their survey methodology differed from previous work. If so, then they should be cautious in making comparisons of observed mortality rates. For example, if their search methods were more thorough, then observed mortality differences may be due to detection differences, rather than actual mortality differences. *The authors agree that caution is warranted, but despite that, I wonder if they proceed without this caution in mind.*

In this same section, the authors state that they extrapolate to the portion of the APWRA not sampled in order to characterize the range of likely project impacts…. I would caution the authors that because the sample of sites they have studied is not a random sample, such extrapolation is not supported from a statistical inference perspective. The authors are aware of problems associated with nonrandom sampling of

other studies (see page 179, second paragraph), but have overlooked application to their study. As stated before, if the authors can justify that the areas they studied are similar to those not studied, then perhaps one might accept some plausibility in their projections for the entire APWRA. However, they authors later state that they do not know the attributes of the tower locations not surveyed, which compels me to suggest that the authors refrain from broader inference and report their results for the observed sample of sites. Their descriptive information for the sites sampled is clearly valuable and represents the majority of the study area, thus they should refrain from making broader inferences that are not defendable. In their conclusions in chapter 9, they

On page 47, they state they were unable to search all turbine strings throughout the study or equally in frequency, so that time spans and seasonal representation varied at turbine strings. Again, I cannot blame them for logistical constraints, but they must take care in analyzing and interpreting patterns in data in light of the fact that they do not have a well-designed study in which all combinations of factors are represented with replication. *I acknowledge that the authors provided cautionary statements, but as previously stated, wonder if these concerns are forgotten when drawing conclusions.* On the bottom of page 48, in determining time since death, how much do weather conditions affect these estimates? On page 51, the authors describe adjustments they made to their observed mortality counts. For instance, they state they adopted a searcher detection rate of 85% for raptors based on Orloff and Flannery (1992) and 41% for nonraptors based on two other studies. They further assumed scavenger removal rates (differential by small and large birds) based on Erickson et al. (2003). They added 10% to these rates for the second set of wind turbines. How valid any of these estimates are for the current study is unclear. Thus, I am skeptical of their estimated number of fatalities for the APWRA given in the executive summary and presented in tables 3-10 through 3-12. Clearly, differences in personnel, search effort and design, habitat, weather conditions, predator density, etc., would lead to differences of these rates from other studies. I suggest the authors concentrate on the observed mortalities in their study and consider optimal strategies for harnessing the information contained therein. As an example, Figures 3-5 through 3-14 present means and standard errors of mortality estimates (per MW). The use of standard errors implies the authors are inferring to a larger population mean. Therefore, it is important to clarify to what population their interval estimates refer. If they are estimating the mean for the entire APWRA, then the same limitations of inference apply as stated before, given the nonrandom collection of sites that have been surveyed. *In their response, the authors state the unsampled turbines did not differ from the turbines they searched. Is that true for all aspects, e.g., landscape features, bird abundance and use of those areas? I have no reason to believe that they purposely biased the selection of turbines, but I would be interested to know how they can guarantee that these unsurveyed locations were similar in all aspects to the measured turbines.*

If, on the other hand, they are estimating mortalities at surveyed sites for the entire year based on a sample over time, then how they sampled over time becomes important in considering the validity of the inference being made. I suggest the authors consider using descriptive statistics of mean and standard deviation (not standard error), which are informative in terms of describing their sample of sites surveyed. This data shows that higher mortality occurred for red-tailed hawks and barn owls during the year

1999/2000, although the reasons for this are not stated. There was also more variability in observed mortality rates during this year. In some cases, consideration of the variability is just as interesting as a measure of center, so they might consider why there was more variation in mortalities this year. By relying on descriptive statistics for their sample of sites, the hypothesis testing in tables 3-3 through 3-8 is not needed. *To their credit, the authors seem very willing to make many of the suggested changes if allowed the opportunity to revise the report.*

In their discussion in section 3.4, the authors reference higher mortality rates at Sea-West-owned turbines than other portions of the APWRA. I was not able to find the supporting evidence for this statement. In addition, I fail to find how ownership would affect mortality per se. Perhaps the ownership issue is tied to some other attribute of which I am unaware. I would like to see more clarification from the authors about the point being made. Further clarification regarding the biological significance of any raptor mortality (last sentence of page 76) would be beneficial. *The authors provided further clarification in their response.*

Chapter 4 Impacts to Birds Caused by Wind Energy Generation

General and specific comments

Examination of bird mortality in relation to bird abundance is a worthwhile endeavor; however, actual bird abundances were unknown. The authors have relied upon measures of relative abundance, determined from point counts. Usefulness of such indices relies on the assumption that the detection rates of birds are similar across time and space. The assumptions should be explicitly stated. There are a host of reasons why this assumption is likely to fail, including observer differences, animal differences and environmental differences (see Anderson 2001, Ellingson and Lukacs 2003). Logically, I would anticipate some sort of positive association between abundance and mortalities, however the positive associations estimated in figures 4-1 and 4-2 are likely not valid for the above detailed reasons. Figure 4-5 B actually demonstrates a trend counter to this assumed association, so an explanation would be helpful for the reader.

Comparison of mortality rates reported from several previous studies is unwise given the vast differences in survey designs, methodologies, site-specific and surveyor-specific differences. I appreciate the attempt to synthesize information from several studies, thus placing their results in context. However, almost all of the assumptions/adjustments made in this section are likely to be false and the extent of bias due to this failure unknown.

In the results, the authors state that bird mortality did not correlate with radius of search around the wind turbine. I am likely misinterpreting this statement, but if one increases a search area, one can only find more carcasses, not less, so I find this confusing. The data presented on mortality at APWRA from 1988 to 2000 has not been described in terms of consistency of methodology, surveyor ability, effort, consistency of environmental conditions that might affect detection rates of carcasses, wind turbine numbers, etc. Thus, I find it difficult to accept the reported trends as meaningful. The authors recognize the importance of standardized methods (see page 86), yet they have not made it clear they have met this requirement in their analyses and in fact, they explicitly state some differences among these studies.

References:
Anderson 2001. The need to get the basics right in wildlife field studies. Wildlife Society Bulletin 29:1294-1297.

Ellingson, A.R. and P.M. Lukacs. 2003. Improving methods for regional landbird monitoring: a reply to Hutto and Young. Wildlife Society Bulletin 31: 896-902.

Chapter 5 Range Management and Ecological Relationships in the APWRA
General comments
Methods described here are not detailed enough to fully evaluate this section. For instance, did they randomly place their transects (string and grass) within a defined area around the turbine string, or were these haphazardly or judgmentally placed? How was average vegetation height measured along a transect, based on every plant or at specific points (i.e., a point transect sampling approach)? How might detectability of cattle pats, rabbit pellets, lizards, mammals, etc., differ in different locations? As before, to what population is inference being made with the statistical tests? *The authors' response to this is that it depends on the strength of the test or how small the P-value was. Identification of the population to which a test or inference is being made does not depend on a P-value. This clearly is a misconception of what a P-value represents. There has been a tremendous amount of discussion on the utility of hypothesis testing, what P-values represent, etc., in the biological literature in recent years (see, for example, Cherry 1998, Johnson 1999, Anderson et al. 2000). I suggest the authors read some of these articles as well as counter viewpoints (Eberhardt 2003).* Use of the word 'significant' needs to be clarified as statistically detectable rather than biologically meaningful. I recommend the authors reserve the word significant only when referring to a biologically meaningful observation. Their one-variable-at-a-time approach may confound the observed associations. For instance, the authors give many examples of how vegetation height differed according to aspect, physical relief, etc. Although I am not convinced these are meaningful differences, they do indicate numerous variables are being considered, and a one-variable-at-a-time analysis procedure has inferential limitations which have been discussed previously. What is the rationale for associating turbine or tower type to lizard counts?

Specific comments
Use of the phrase, 'tended to be significant' on the middle of page 91 is either improper interpretation of P-value as related to effect size or from a decision making perspective of null hypothesis testing a way of circumventing a yes-no answer in the formal test. First, a P-value is not an indication of effect size, after all, its value can be changed simply by changing the sample size; the same estimate can yield differing P-values. A P-value can be interpreted as the probability of observing a result as or more extreme than that observed given the null hypothesis (Ho) is true.

The biological significance of the observations has not been justified, conclusions have been based on statistical detectability. For instance, the mean difference of vegetation height comparing heavy versus intermittently employed rodenticides was 4.28

cm). Is this a meaningful difference, i.e., *do prey species and/or raptors perceive this difference? The authors responded to this question by stating that published literature indicates changes in grassland use by animals according to vegetation height, but I still have no idea if a difference a few centimeters falls within that umbrella statement.* The summary table 5-25 summarizes their findings, but underlying all of these associations are questionable interpretations of biological importance.

The mean differences in Table 5-1 are confusing, for instance, when comparing plateau to plateau (these are the same variable) and when comparing plateau to peak and slope (one cannot perform LSD on 2 *different* variables), they must use another multiple comparison procedure which allows combinations of means. Note also that it is improper to follow nonrejection of an ANOVA with LSD multiple comparisons because there is no control for type I error in such a case. *Apparently, I did not note that their alpha level was 0.10 in this chapter. It is unclear why they used a more liberal test in this chapter, which results in a higher Type I error rate.*

Table 5-3 gives several correlation coefficients, most of which indicate a weak linear association at best, yet they highlight the statistical detectability of these cases. For example, the last sentence in the discussion on page 54 restates that vegetation height correlated positively with number of cattle pats, but the sample correlation coefficient was only 0.19, a weak association at best. *Smallwood and Thelander state that they did not characterize them as significant. By simply reporting these values in the results section, I agree. However, when they reiterate a positive relationship in the discussion, they are then interpreting the result as significant. By not having the qualifier 'weakly' positively correlated in their discussion, the reader may be misled to think the correlation is relevant.* One correlation coefficient given for vegetation height and 'percent in canyon' is moderate (r = 0.46), but I am not certain this is meaningful. The population of turbines being measured is clearly stated as the 1526 wind turbines measured through August 2002. Thus, statistical inferences might be made to this collection of turbines if sampled appropriately *(they agree)*. The important consideration then becomes how the turbines were sampled. Were transects randomly placed within a defined area around each turbine or the turbine string? If a multistage random sampling design were implemented, estimates would be possible at both the individual turbine and turbine string level. My impression is that random placement was not made at either scale, thus limiting their inferential capability. In my comments on this section, I have provided a few examples in which I question the value in the perceived associations. Rather than repeat the same concerns for each table/response variable combination, consider my concerns to apply to all of the variables and associated analyses performed in this chapter (cattle pat counts, cottontail index, lizard index).

References

Anderson, D.R., K.P. Burnham and W.L. Thompson. 2000. Null hypothesis testing: problems, prevalence and an alternative. Journal of Wildlife Management 64:912-923.

Cherry, S. 1998 Statistical tests in publications of The Wildlife Society. Wildlife Society Bulletin. 26:947-953.

Eberhardt, L.L. 2003. What should we do about hypothesis testing? Journal of Wildlife Management 67: 241-247.

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

Chapter 6 Distribution and Abundance of Fossorial Animal Burrows in the APWRA and the Effects of Rodent Control on Bird Mortality.

General comments

The idea that reducing raptor prey populations in the APWRA may discourage raptors from visiting the APWRA makes intuitive sense, although I am not sure if this effect can be achieved at such a large scale (as opposed to more locally near wind turbines). I am not convinced that elimination of one component of prey, specifically ground squirrels, would achieve the desired result either. I commend the authors for making observations regarding gopher and squirrel burrows and their proximity to turbines, and developing research regarding the variables. Their objectives are clearly stated as relating ground squirrel and pocket gopher distribution and abundance to rodent control intensity, physiographic and turbine attributes, and comparing raptor mortality to densities and contagion of burrow systems, but a priori hypotheses are not explicitly stated. Burrow densities are implicitly being used as indices to abundance. I have no knowledge of whether or not this assumption is reasonable. For example, the burrows may represent a population size that existed several years prior to the current observed raptor morality or numbers of burrow per animal may differ depending on landscape or predator abundance features. How ephemeral are these burrow systems? The seasonal effects reported in setion 6.3.2 indicate tremendous burrow variability within a year, but are numbers of individuals fluctuating that much? *The authors provided many clear answers to these questions in their response.* Changes in the numbers shown in various figures make me question the relevance of burrows as an index to prey abundance, let alone prey availability. See my discussion on chapter 4 regarding the utility of indices. *Here, I was referring to the tremendous fluctuations in burrow systems as seen in their figures. Smallwood and Thelander state in their response that these fluctuations are real (well established in the literature).* Two metrics of contagion used in their analyses. Given the observed variation in burrow numbers throughout a year, how did they relate observed bird mortality over a year or several years to burrow contagion? *Here, I was asking what measure of contagion did they use for a given year if the measure changed throughout the year.*

Specific comments

Wind turbine strings studied were selected arbitrarily (not randomly), hence limiting statistical inferences *(the authors agree and clarify that this was exploratory work))*.

Figure 6-4 (and later 6-45 and 6-46) presents results comparing burrow system densities between areas with rodent control and areas lacking rodent control, however, for the latter, only 3 observations were available. The strength of evidence here is very weak *(the authors agree)*. In figure 6-5, they discarded an outlier without explanation. If a poor measurement was made so that this result was unreliable, then state this clearly. If,

however, it does not fit their predefined opinion on what should occur, then this is not a viable reason to omit it from the analysis.

The authors resort to transformations at various places without explanation or consideration of what is then actually being compared. For example, what does Figure 6-6A tell us? The variable based on a log-log regression in this figure and Figure 6-7 has not been explained. *A description of these figures is provided in the text, but the authors did not state why a log-log plot is used rather than working with the data on the original scale.* Why are normal curves shown in all frequency distributions? *There was no meaningful reason for doing so.*

I find it curious that ground squirrel avoidance was stated to differ between summer and other seasons (see page 124), given the degree of overlap of *summer with fall and spring* (Figure 6-27B). *The authors clarify in their response the difference is detectable between summer and winter.*

The term 'tended to differ significantly' or 'tended to correlate' is used at various places (e.g., see pages 124, 146, 166) to reflect a P-value less than 0.10 but greater than their stated alpha level of 0.05. As previously stated, this is either improper interpretation of P-value as related to effect size or from a decision making perspective of null hypothesis testing, a way of circumventing a yes-no answer in the formal test. Note again that it is improper to follow nonrejection of an ANOVA with LSD multiple comparisons (see page 124, first paragraph of section 6.3.3) because there is no control for type I error in such a case.

The reported correlations that follow are weak to moderate at best and thus I would not read too much into the statistical detectability.

From a nontechnical perspective, I find the observed relationship between rodent control intensity and pocket gopher burrow density interesting in that its highest level is at a moderate level of rodent control. Cottontail burrows demonstrated an opposite pattern. Why would gopher clustering differ by aspect for control areas, but not for nontreated areas? I could not find explanations for several reported effects.

Comparisons of raptor mortality and small mammal burrow distributions were executed only considering burrow density and thus are prone to many confounding effects as previously stated. At various times, the authors recognize the potential complexity of what they are attempting to measure, but they then put this consideration aside and proceed to analyze, interpret and conclude. *They respond by saying what else are we to do? My suggestion would be to eliminate material where treatment effects cannot be isolated, or at the very least, only report descriptive information and do not proceed to interpret it. In other words, if one cannot isolate a treatment effect, or cannot meet important assumptions, than no modeling is better than proceeding as if everything was valid.*

The finding that mallard mortality was related to rodent control intensity was dismissed as a spurious effect. I agree with this conclusion, but it illustrates an important point. When one can develop an ecological explanation for an observed result *a posteriori*, the result is more heavily weighted as 'truth'. I believe this approach to science is ubiquitous, but not ideal and has lead to many spurious results. *Their response to this issue was commendable.*

Chapter 7 Bird Fatality Associations and Predictive Models for the APWRA

General comments

The authors begin this chapter by stating the importance of identifying causal factors of bird fatalities. They then state that collisions are rarely observed and that inferences must be drawn from carcass locations. Such inferences are merely associative, not causative. I suggest the authors present their observations in the former context as I believe the latter is not attainable within the context of the current study. Causation is best established through experimentation and requires reasoning beyond statistical analysis. Establishing causation with observational studies requires meta-replication of the system of interest, consistency of results, and a plausible explanation for the observed behavior. Preferably, potential hypotheses should be formulated *a priori* and the evidence for each hypothesis should be documented with the observed data. Far too often, ecological explanations are developed *a posteriori*, which leads to spurious results. The authors are aware of possible spurious results (see page 218 discussion of mallard fatalities), but they fail to see this potential when explanations can be developed *a posteriori*. *Here, I was referencing their willingness to describe the mallard fatalities as spurious, but they did not warn the reader (to my knowledge) that other detectable associations might also be spurious.* More importantly, they fail to see this potential in their overall approach to analysis. For instance, 34 explanatory variables have been measured at each turbine site, 12 bird species examined leading to hundreds of single variable tests of associations (Tables 7-1, 7-2, 7-3). There is great potential for Type I error (rejection of the null hypothesis of no association when there is no association) when considering all of the tests being performed. A total of 408 tests were performed in Tables 7-1 and 7-2. The probability of not making a Type I error (using the stated $\alpha = 0.05$ level) is $(0.95)^{408}$. Thus, the probability of making at least one type I error is $1-(0.95)^{408}$, which is essentially one. *In their response, Smallwood and Thelander seemed to have confused the alpha level of a test and a P-value; it is as if they want to vary the alpha level according to the observed P-value. The alpha level is a preset measure that is decided prior to data analysis based on the level of Type I error that is acceptable to the researcher. I have not overestimated the probability of at least one Type I error. The P-value is entirely data-dependent. Note also that probability stated above is for at least one type I error, therefore this does not mean that out of 408 tests, only one will be a Type I error. The authors would like to act as if only one of the 408 tests was a false rejection; this is a misinterpretation of statistical hypothesis testing and surprises me.*

A simple approach to reducing overall (experiment-wise) type I error rate (α_e) is to lower the comparison-wise error rate (α_c), such that $\alpha_c = \alpha_e/2k$ where k is the number of tests being performed, i.e., a Bonferroni procedure. *Thus, to maintain and experiment-wise error of 0.05, the authors should use a comparison-wise alpha level of 0.000006, in which case the authors will be surprised to find that even their results with P-values of 0.001 would not be rejected. Such a small alpha illustrates the severity of the situation; it is not intended for actual use. The authors need to reduce the total number of tests performed (which reiterates why the one-variable-at-a-time approach is not recommended) and then use a Bonferroni correction. Even then, the comparison-wise error rate will be small, and thus, many reported detectable results will likely not be detectable.*

The authors have counted fatalities at all locations surveyed and measured variables (both environmental and turbine-specific) in an attempt to reveal 'robust' patterns. It is not clear to what the intended patterns are 'robust'. Based on the observed patterns, predictive models were developed, yet the models they developed were never clear to me. For instance, one could develop a model such that the expected number of fatalities Y (per MW per year) is a function of explanatory variables X, W and Z ($Y = \beta o + \beta_1 X + \beta_2 W + \beta_3 Z$). Different modeling approaches have different link functions, for example, Poisson regression typically uses a log link function (which is different than a log transformation). Alternatively, one could use logistic regression which uses a logistic link function to model presence/absence of mortality as a function of the explanatory variables. Model selection, that is identifying the best or set of better models in the set of possible models would be an important component of such an endeavor as well as assessing the goodness of fit of the most general model in the model set. Burnham and Anderson (2002) is one of many pertinent references on the subject. The main point here is that I do not really understand what their 'model' is. *The authors' response is that their models were summations of accountable mortality across the variables selected for the model. First, the one-variable-at-a-time approach is wrought with problems as previously discussed, including confounding. Therefore, anything based on a sum of these variables has inherent limitations. Second, the assumption of additivity of these variables is questionable, in part because their one-variable-at-a-time approaches have problems of confounding and also because variables may have interactive effects. Third, I am not convinced they are describing a model; rather it sounds more like a metric of sorts. A predictive model might look like the following example using multiple regression:* $E(Y) = \alpha + \beta_1 canyon + \beta_2 towerheight + \beta_3 turbinedensity + \ldots + error$ *where Y represents the number of bird strikes. Such a model considers several explanatory variables simultaneously in the estimation process as opposed to one-variable-at-a-time approaches. Note that the variable 'canyon' is an indicator variable, so this is really an example of analysis of covariance.* Is the assumption that their predictive 'models' are relatively precise appropriate (page 222)? Testing of their 'models' appears to have been performed using the data to develop the model, which is an inappropriate means of evaluating models (see Olden et al. 2002). The authors should consider using a portion of their data for model development, reserving the other portion for model evaluation.

Specific comments

On pages 179 to 182, the authors describe an abundance of previous studies which have presented conflicting conclusions regarding causal factors of collisions. Such disparity suggests to me, as stated earlier, that demonstrating causation is not a simple task and the actual mechanism underlying bird strikes may be very complex, e.g., a combination of many environmental and turbine-based factors. Again, the overall analysis approach has been to look at associations one explanatory variable at a time, thus leading to potential confounding effects and spurious results. *For clarification, I am not suggesting multivariate statistics be used (for example, MANOVA or factor analysis), I am stating that multiple explanatory variables be simultaneously used in modeling, in which case model selection methods will play an important role. Use of multiple explanatory variables is still considering a univariate modeling approach, i.e., there is one response variable (mortalities per unit of search effort).*

On page 182, last full paragraph, the last sentence is unclear. I believe the authors meant to imply that some turbine attributes were collinear or highly correlated, thus similar associations with bird mortality were observed when looking at each variable separately.

On page 184, the first three paragraphs of section 7.2.2 are wrought with lack of information and misstatements. The authors state that the assumptions of the corresponding hypothesis tests were satisfied, but they do not state what those assumptions are and how they were assessed. For example, Pearson's correlation assumes bivariate normality. Did they assess normality of each variable? How? *(Their response is essentially 'we did not, but other biologists do not either'.) Such reasoning is for lack of a better word, ridiculous. First, normality is easily assessed and if violated, would suggest use of Spearman's rank correlation. This is a basic concept that is taught in introductory statistics classes.* The least squares regression models they used assume the errors are independently and identically distributed as a normal distribution with mean zero and constant variance. How did they assess normality of the residuals? How did they test equal variance? Analysis of variance (ANOVA; which is really regression with a categorical explanatory variable) has the same assumptions. Were these also assessed for these analyses? *Their lack of attention to important assumptions leads me to conclude that extensive consultation with a statistician is needed.*

Misstatements include 'Correlation analyses are summarized by the coefficient of determination, R^2, when prediction is the ultimate objective. I believe they meant to say 'Regression analyses….' The coefficient of determination measures the percent of variation in the response variable that is explained by the regression model (i.e., the collection of explanatory variables). Correlation analysis is merely descriptive and summarizes the degree of the linear association between variables. Thus, one can have a strong association (curvilinear) but Pearson's coefficient will be small (close to zero). The statement 'We report weak and nonsignificant correlations when doing so meets our objectives.' -the latter part sounds dubious and confuses several issues. Statistical detectability is directly affected by sample size, n. Thus, with large n, it is possible to have a sample correlation coefficient of $r = 0.1$, and yet have an associated P-value = 0.001 for the test of Ho: $\rho = 0$ vs. Ha: $\rho \neq 1$. In this case, the linear relationship is very weak and is not notable. *The authors clarify that nondetectable results can be interesting, particularly when they contradict previous conceptions, to which I agree.* Alternatively, you can have small sample sizes that result in nonrejection of Ho even though $r = 0.75$ indicates a fairly strong positive linear association. I agree that in the latter case, such results may be reported as long as they are taken as suggestive, not confirmatory.

Given the collection of several explanatory variables, the authors should consider using partial correlation in which one or more variables are controlled when considering the association between two variables of interest. *(I have made similar statements referring to use of multiple regression, yet in this context, the authors agree with the suggestion.)* They also incorrectly state that the coefficient of determination (R^2) is based on the steepness of the regression slope. This is only true within a specific context in which one considers a specific data set and several lines that are being fit simultaneously to the data. In general, the coefficient of determination is defined as the portion of variation in the response variable that is explained by the explanatory variable. For

example, if all observations fall on the fitted regression line, then r^2 is one, and this is true regardless of the slope of the line unless the slope is zero, in which case there is no variation in the response variable. *In their response, the authors clarify they were referring to the sample value, r^2. This does not change my original point.*

They state that several key assumptions of ANOVA were not met due the absence of a block design. A block design is not necessary for ANOVA, blocks are sometimes useful for reducing variability, but their absence does not preclude assumptions from being met. Equal treatment replication (balanced design) does not preclude successful analysis via ANOVA or other techniques. However, when all treatment combinations are not represented, e.g., fractional factorial designs, then considerable thought must go in to analytical approaches for meaningful comparisons that isolate treatment effects to be made.

I do not understand what the numbers in Table 7-8 on page 223 represent. The description is that the numbers represent the largest accountable mortality values calculated from the chi square tests…. Similarly, I do not understand how a specific wind turbine attribute can be reliably associated with X% of mortalities (pages 224-241), given the combination of variables at any given turbine and the inability to control all other factors in their analysis. It follows that I am not confident that their form of model assessment (described as the percentage of correctly predicted dangerous turbines where species-specific fatalities occurred) is a useful metric for assessing model performance.

References:

Burnham, K.P., and D.R. Anderson. 2002. Model Selection and Multimodel Inference: A Practical Information Theoretic Approach. 2nd ed. Springer Verlag, NY.

Olden, J.D., D.A. Jackson, and P.R. Peres-Neto. 2002. Predictive models of fish species distributions: A note on proper validation and chance predictions. Transactions of the American Fisheries Society; 131: 329-336

Chapter 8 Bird Behaviors in the Altamont Pass Wind Resource Area

General Comments

Quantifying bird behavior or intensity of use is an enormous task that is very complex in its actual measurement and analysis. My understanding of their approach is briefly given below. The researchers surveyed 61 plots encompassing between 6 and 52 wind turbines. A total of 1500 wind turbines were surveyed, although these are not independent in the sense that they are located in strings. These plots were surveyed on 4 separate occasions over a 7-month period. The plots were delineated with a 300-m buffer around the focal wind turbines. Two observers scanned the plot area over a 30 minute period, and recorded bird locations, behaviors, etc., at the turn of each minute. Basically, this means there are 30 points in time for each session. Behavioral events were also recorded, such as flight through a string of wind turbines, etc. A total of 120.6 hours of sighting was executed. Again, I commend the authors for their efforts.

From an inferential perspective, I would ask the authors to clarify how these specific sites were selected for observations. It looks as though a variety of wind turbine types were selected purposefully (Table 8-1) that cover a previously referenced area (see page 246 bottom), but a stratified sampling approach could be used to ensure random

selection and representation of all turbine types as well. *The authors clarify the sample is all Set 1 turbines, thus, inference is limited to this set and should be made to APWRA.* In reference to their analysis, I am concerned that their approach is inadequate to identify meaningful relationships for reasons similar to that stated for other chapters.

First, the approach of examining one variable at a time oversimplifies what is a very complex situation. For example, by only considering minutes of perching by temperature levels, the observed difference of observed and expectations may be the result of another variable, such as wind speed, which often is associated with time of day and temperature. That is to say that such an approach does not identify causal mechanisms. To their credit, the authors do mention another source of complexity, the notion that birds may adapt their behavior in response to the presence of wind turbines (page 246). But, I am not at all confident in *m*any of their findings in this entire chapter because of their one-variable-at-a-time goodness of fit approach to analysis. *The authors have responded by stating some associations are irrefutable. For instance, they found golden eagles perched disproportionately more in canyons. I will again clarify my concern that one-variable-at-a-time approaches can be misleading. While they may have found greater than expected perching in canyons (by considering all perching observations), it may be , for example, that specific turbine types were located more often in canyon areas, and that eagles prefer to perch on these turbines. Thus, it is not the presence of a canyon that is driving the response, rather it is another variable confounded with presence of a canyon. Their conclusion that perching occurred more often in canyons is not incorrect, but it may be misleading, leading one to believe that canyons themselves induce more perching rather than another possible attribute of this study area. I am not stating that turbine types were disproportionate in canyons, but by only considering one variable at a time, they have eliminated the means for controlling for other explanatory variables.*

Second, I question whether many of the stated 'significant differences' are biologically meaningful. For example, in Table 8-6, the chi-square test for time of day effects on perching minutes of all raptors resulted in a P-value less than 0.005, yet the percent deviation from the expected value is less than 3 percent for all categories. In the same table using temperature as the explanatory variable for golden eagles, there is a 20 percent negative measure for temperatures of 60-69 degrees and a 21 percent positive measure for temperatures between 70 and 79 degrees. Do the authors believe that golden eagles perch less due to temperature in the 60s and more in the 70s and then less in the 80s? *Their response is that they are simply reporting all of their results. I am suggesting they be more responsible in interpreting their relevance.*

I recommend that the analysis approach in this chapter (and several others) be changed. I would need to know more to specifically advise on how they should proceed, but I will make the following statements and suggestions. First, birds (or animals in general) do not use habitat randomly. Any assumption of random use is a 'silly null' hypothesis which is certainly false (see Anderson et al. 2001). (*see previous discussion of this issue).* Second, while I agree that understanding how birds use APWRA would be useful in putting bird fatalities in context, I fail to see how associations with variables such as temperature would be useful. *(Here, I was referring to the inability of managers to control environmental variables, not the biological associations. Smallwood and*

Thelander have erroneously jumped to the conclusion that this statistician does not understand any biology.)

Third, I suggest the authors condense the 30-minute level information to percentages for that survey period and consider relating the percentage of time perching to the set of meaningful explanatory variables collectively. This approach treats each 30 minute period as the measure made on each sampling unit (plot). Such an approach eliminates concerns about the covariance structure between successive minute-by-minute observations. Whether or not one needs to consider the relationship between survey periods on the same site is another issue (perhaps repeated measures structure should be used?). Once a reasonable modeling approach is identified, they must develop appropriate models for consideration and use a well-defined model selection process.

Specific Comments

For me to completely comprehend the intricacies of the data collected would require considerable face-to-face interaction with the researchers. Examples of my uncertainties are stated below. I am not sure whether the focal set of wind turbines was always a complete string Also, based on figure 8-2, it looks as though a 300-m buffer may include additional sets of wind turbines; thus, they too provide opportunities for perching, 'dangerous flights', etc. How is their presence handled in terms of analysis at the wind turbine level even though they are not the focal set? I failed to see the distinction between plot level and string level of analysis mentioned on the middle of page 247. *Their response indicates all wind turbines in the observation plot were considered.* How do they count number of minutes perched if 2 birds are perched in the study area for the first 5 minutes of the study period. Is that 10 bird-perching minutes? If so, then the number of bird-minutes is the metric, rather than minutes alone. They state on page 247 that for each record, they recorded the species, … predominant flight behavior, flight direction, distance to nearest turbine, number of passes by a turbine, and flight height relative to the rotor zone. How does one record flight direction if they traveled a flight was multidirectional, circular, etc? Is distance to nearest turbine the smallest distance observed during the entire flight? How does one define a pass by a turbine? If turbines are in a string a bird flies 100 m above the string, are all of these turbines counted? How much error is there in measuring flight height relative to rotor zone when birds are considerable distance from the observers? *The authors offer useful responses to these questions. Note the importance of many of these details. Two birds perched for 5 minutes is treated the same as 1 bird perched for 10 minutes, yet the former situation has more birds susceptible to strikes than the latter.*

Birds may have exited the area for more than 30 seconds, only to return again and be considered as a new bird. Pseudoreplication is a concern here, but clearly the researchers cannot be expected to recognize individual birds per se. Several quantitative or ratio level variables, such as temperature and wind force, were reduced to ordinal categories. Information loss occurs in such a process and is unnecessary. I suggest the authors use these variables as continuous explanatory variables in a model effort other than chi-square goodness of fit tests. I do not understand why they compared the correlation of flight frequency in the rotor zone with flight time to that of perching time (page 256). I would assume that when a bird is perching, it is not flying at all, and thus is not flying in the rotor zone. I also do not understand the interpretation of frequency of

behavioral observations during a 30-minute session on page 256. If the initial presence of observers is modifying bird behavior, then this is an important observation, which suggests that an initial 'settling' period should occur prior to the actual observation period. However, I am not sure this was the point they were trying to make. Finally, it is hard to evaluate the plausibility of their discussion points and recommendations at the chapter's end given the lack of enthusiasm I have for their analysis methods. I will say again that they do have valuable information, some of which was presented descriptively, that should be considered as meaningful.

References:

Anderson, D.R., W.A. Link, D.H. Johnson, and K.P. Burnham. 2001. Suggestions for presenting the results of data analyses. Journal of Wildlife Management 64:912-923.

Chapter 9 Conclusions and Recommendations

General comments

This section of the report was clearly written and suggests several management alternatives. It is my belief that some form of adaptive management (Walters 1986, Walters and Holling 1990) should take place given the amount of data already collected by this project and others cited therein. My opinion is there is still uncertainty regarding the causal mechanisms behind bird strike mortalities. However, I think it would be feasible to evaluate the effects of many of the recommendations made in this chapter in a cost-effective manner. For example, I assume it would cost very little to paint blades for a sample of turbines. Assuming that experiments can be conducted that control for other potential factors affecting mortality, one can isolate the effect of the management action.

Specific comments

Their first recommendation, at least on the surface, seems reasonable. If sampling in the WRRS program is haphazard and/or voluntary-response based, then from a scientific monitoring perspective, the data collected is of little value. That is not to say that there is nothing to be gained by observations of maintenance workers in the area, because the cost is presumably nothing. I am unsure that the comparisons of observed fatalities are fair (same time period, same locations surveyed, etc.), but if they are, and a consistent relationship between the 2 methods could be established, then such a system would be similar in worth for trend estimation. However, consistency is something not easily obtained in any index-based study (Anderson 2001).

Their conclusion regarding rodent control is counterintuitive to me, but that does not make it wrong. However, I question whether the conclusion is correct given the potential weaknesses associated with assessing the effects of rodent control treatments (see chapter 6 evaluation). They further state that even if rodent control were effective, displacing raptors would result in a net loss of raptors from the remaining habitat. *Smallwood and Thelander state that populations would be reduced through displacement because these species cannot be crowded into smaller spaces.* That is only true if populations elsewhere are at carrying capacity, which I doubted to be the case. *They state that carrying capacity is difficult to define, yet this assumption is implicit in their above statement regarding crowding into smaller spaces. Secondly, would not displacing birds from the APWRA into other suitable habitats reduce the potential for bird strikes, a likely*

goal of everyone involved? (I deleted a sentence here that stated if birds were at carrying capacity there might be less concern about the observed mortalities)

Their third recommendation (and its subcomponents) is similar in the underlying idea to the second recommendation: reducing prey availability may reduce raptor susceptibility and thus fatalities. Thus, I find it interesting that they advocate ceasing the rodent control program, but encouraging fossorial animals to be farther from wind turbines. *Note that I referred to similarity in the underlying idea of reducing prey availability, not the spatial differences of these approaches that they refer to in their response.* However, I agree that one might want to eliminate the rodent control program for other reasons (e.g., adverse impacts on other species of importance).

In their test of perch guard effectiveness (recommendation #4), did they control for other factors that may affect mortality? Similar to the results for rodent control, it may not be the method itself that is lacking effect; it may be the implementation, for instance, if the chicken wire readily falls apart. *I would think that any effort to keep birds away from the towers at all times might reduce their susceptibility to strikes. So, the question remains as to whether effective perch guards could reduce strikes. The authors readily admitted that the perch guards implemented thus far were ineffective, e.g., chicken wire falling apart.*

Their conclusion that wind turbines at the end of strings are edges of clusters kill disproportionately more birds is very plausible, and hence their suggestion of adding pole structures is worthy of consideration for experimentation. However, it may be that by adding pole structures more birds will collide with the turbine because of the visual impedance mentioned in recommendation number 9. *I misinterpreted this recommendation. However, pole placement at the end of turbine strings might offer additional perching opportunities, thereby bringing birds in closer proximity to the turbines and increasing susceptibility.*

Most of the remaining recommendations are yet to be proven as effective management options. I suggest that pilot studies be used in which the efficacy of a small set of management actions (a subset of their listing) can be evaluated without the confounding effects of other possible mortality factors. *The authors respond by saying that pilot studies have been tried before and were inconclusive. Why then do the authors believe enough is known to implement mitigation measures at a very broad level?* Based on my limited knowledge from this report, I am not convinced that enough is known to warrant universal implementation of certain mitigation measures (see page 348 bottom). *I make this statement because I do not believe causal mechanisms have been identified and because of confounding issues from their one-variable-at-a-time analyses.* I also question the degree of error one would have in the estimated number of bird mortalities over 10 years as the input to Smallwood's estimator of are for support described on the top of page 348. *To clarify, the estimated number of bird mortalities over 10 years is being based on what, the number of mortalities found during their study? Extrapolating any such estimates to other periods is untenable, and as such*, unreliable input into a model almost certainly results in unreliable output.

In section 9.3, Thelander and Smallwood state they were unable to extend their model predictions to the turbines not characterized. Given that an appropriate predictive model exists (which I have previously questioned), I would suggest that to properly evaluate the model, these sites would provide an independent means of evaluating the

model *via future data collection*. Their use of the same data for model development and testing violates the independence necessary for proper model evaluation (see Olden et al. 2002). I agree with much of their description of limitations in this section. Many of the concerns they state are equivalent to what I have stated in my review and provide the basis for my concern regarding the validity of stated conclusions. I assume that when they are referring to multivariate statistical methods on page 353, they are referring to multiple response variables, not multiple explanatory variables, but their one-variable-at-a-time approach to analysis indicates there may be confusion regarding this terminology. Multiple regression and analysis of covariance are considered univariate methods because one response variable is being considered. I suggest that these methods could be employed to better examine mortality as a function of several explanatory variables. Multivariate methods generally refer to analysis in which multiple response variables are being considered simultaneously, which requires consideration of the covariance structure among these variables.

I am not sure how they arrived at their estimated mortality reductions on page 354. These are likely purely speculative.

References:

Anderson 2001. The need to get the basics right in wildlife field studies. Wildlife Society Bulletin 29:1294-1297.

Olden, J.D., D.A. Jackson, and P.R. Peres-Neto. 2002. Predictive models of fish species distributions: A note on proper validation and chance predictions. Transactions of the American Fisheries Society; 131: 329-336

Walters, C.J. 1986. Adaptive Management of Renewable Resources. MacMillan, New York.

Walters, C.J. and C.S. Holling. 1990. Large-scale management experiments and learning by doing. Ecology 71: 2060-2068.

Appendix A

This section is intended to explain the rationale for adoption of fatalities/MW/year as the metric of choice as opposed to fatalities/turbine/year. First, the authors criticize the metric of fatalities/turbine/year by stating it can be misleading. Their example on page A-2 demonstrates how fatalities rates using number of turbines as a reference can appear to differ at 2 locations even when the same number of deaths occurs at these locations. The reason for this is the sites have differing numbers of turbines. The same differences can be illustrated using their metric of fatalities/MW/year. For example, if farm A generates 40MW/year and farm B generates only 4MW/year and 100 fatalities occur at both places, then the rate is 2.5 fatalities/MW/year for farm A, but is 25 fatalities/MW/year at farm B. This might mislead someone to believe that more fatalities occurred at farm B. *This issue has been discussed previously.* Later, they compare regression sums of squares relating MW to bird deaths and turbine numbers to support their proposed metric compared to the turbine per year metric. This approach did not compel me to see the advantages of their metric.

The main difference between their metric and the one that uses turbines is that theirs incorporates MW produced per turbine, thus the cost (mortalities) is stated within more of a context of the benefit (MW production). *I have now modified this viewpoint in an earlier description of their defined metric.* Initially, I questioned the purpose of comparing mortality rates among different wind energy generating facilities. If one is only considering management of a particular wind facility, then knowledge of how one place compares to another is not useful. After all, different facilities have many different environmental factors likely to be important in the process that leads to bird strike fatalities. Later, in the discussion, the authors mention replacement of older turbines with larger turbines capable of greater MW generation. Thus, if one wanted to compare fatality rates between time periods at a given site, it may be advantageous to use their metric if considering the cost-benefit aspects of the power generation. They have presented better arguments for using their metric here than previously.

Also in the results section, the authors refer to the relationship between time span surveyed and number of turbines with non-zero mortality (Figure A6). It seems obvious to me that the more you search, the more likely you are to find at least one mortality at a turbine, so I do not understand why they are emphasizing this point as being important. The proportion of turbines where at least one bird has been killed is not a metric that I find particularly useful and does not translate directly to fatalities per turbine or MW per year. I do not agree that this relationship demonstrates that most of their turbines were not sampled long enough to robustly estimate mortality. I am not sure what is meant by 'robustly' here, but in section 4.4.2 on page 86, the authors use the word robust to imply reliability based on high precision. Precision, or repeatability, is only one component of accuracy. Bias, or the deviation of an excepted value of an estimator from the true parameter is equally, or perhaps more, important.

Appendix B

Appendix B attempts to explain the differences in their study from those reported in Kerlinger and Curry (2003). I agree that the incidental counting/voluntary response of bird carcasses is not a rigorous approach to estimating mortality rates if that is what the Wildlife Reporting and Response system relies upon. In various places, the authors refer to 'less robust' estimates of other researchers and then proceed to compare estimates from various studies. Robustness implies a resiliency to, for example, an assumption violation. It is not clear to me what their use of the term implies. In comparing estimates, they report Kerlinger and Curry (2003) underestimated mortality relative to their mortality estimates. Such comparisons are unwise if they constitute different spatial or time periods. Of course, one must know the true mortality to say which estimate is 'better'.

Appendices C and D

These sections consist of details of the chi-square goodness of fit approach to examining use versus availability of various landscape and turbine features. I have previously commented on the efficacy of this approach to analysis.

ATTACHMENT D

RESPONSE TO THIRD REVIEW OF SMALLWOOD AND THELANDER (2004)

K. SHAWN SMALLWOOD and CARL THELANDER

28 August 2006

Herein we respond to the third review of Smallwood and Thelander (2004), entitled "Developing methods to reduce bird mortality in the Altamont Pass Wind Resource Area." Prior to its release by the California Energy Commission (CEC), our report was peer-reviewed by two scientists expert with the issue of bird collisions with wind turbines: Dave Sterner and Sue Orloff. Our revised report was reviewed by the wind turbine owners and their consultants, and by California Department of Fish and Game and U.S. Fish and Wildlife Service biologists. We revised the report again before the CEC released it. A year later the CEC had our report reviewed again after wind turbine owners and industry trade organizations/lobbyists complained about CEC staff conclusions in a white paper on the wildlife impacts of wind turbines. This white paper was prepared in support of the 2005 Integrated Energy Policy Report (IEPR) process.

The CEC administered a second review by Drs. Michael Morrison, Christine Schonewald, and Jan Beyea. The California Wind Energy Association (CALWEA) claimed this second peer review was invalidated, in part, by the reviewers' acquaintance with Smallwood. However, it is unlikely qualified peer reviewers could be found who are unfamiliar with our work, because the pool of scientists working on this problem is small.[1] The second peer review was as valid as any performed at scientific journals, except we were not given the opportunity to revise the report.

Smallwood responded to the second peer review, but the reviewers apparently did not see his responses. He also worked with CEC staff to respond to lobbyist and consultant comments on our 2004 report in reaction to the CEC staff white paper for the IEPR. The staff responses composed 205 pages, so a lot of careful thought and explanation was provided in response to the industry comments. Regardless, CALWEA urged the CEC to conduct a third review.

The methods and many of the same results and interpretations were also reviewed by three scientific peers prior to National Renewable Energy Lab's (NREL) release of our 2003 progress report. Our final report to NREL was also reviewed by three scientists prior to its 2005 release. A favorable review by four scientific peers was also recently completed on a paper submitted to a scientific journal, and minor revisions to the paper are underway. Thus, a number of reviews were already completed on our methods and results prior to this one.

In this review, three teams of statisticians were asked by the California Energy Commission to respond to specific questions posed by the CEC. Carl Thelander and I were then given three weeks to respond to the review comments. We will provide general responses to common or important comments, followed by a list of methods we would use in a future, similar study, based on what we have learned and on review comments with which we agree, and we then respond to specific comments.

---

[1] Furthermore, peer reviewers of scientific journal submissions often are acquaintances or colleagues of the authors because the pool of available qualified scientists tends to be small in most fields of study, especially in fields such as wildlife biology, conservation biology, or ecology.

CONTENTS

Review process……... 3
Organization of responses……. 4
Overview of review comments……. 4
Broader issues raised by the reviewers……. 5
Pseudoreplication…… 5
Nonrandom sampling of turbines…… 6
Extrapolation or results…… 7
Multiple comparisons with inter-correlated variables without appropriate corrections 9
Use of multivariate or multiple response tests……. 10
Inappropriate use of chi-square tests on measured variables……... 11
Differences in observer ability were not incorporated…… 12
Differences in scavenger removal rates were not incorporated…….. 13
Confounding……... 14
Type I error……. 14
Improving research on bird and bat collisions with wind turbines……... 15
Site utilization……. 17
Behavior research methods…….. 17
Sample size and spatial and temporal scope…….. 18
Using human observers…….. 18
Advanced Integrated Radar and Camera System (AIRCAMS)……. 19
Data analysis…… 19
Abundance research methods…… 20
Diurnal raptors…… 20
Raptor nest surveys……. 20
Nocturnal raptors……. 21
Grassland songbirds……. 21
Bats…….. 21
Fatality Searches…….. 21
Mortality metric…… 22
Mortality estimator…….. 22
Scavenger removal…… 23
Searcher detection……. 24
Background mortality…… 25
Crippling bias…… 25
Search radius bias…….. 25
Left-censoring of data…… 25
Ecological relationships…….. 26
Responses to specific comments…… 26
Review Team 1…… 26
Review Team 2…… 49
Reviewer 3…….. 97
References……. 136
Attachments...……. 139

REVIEW PROCESS

The CEC embarked on a third review of our report, but this time applied conditions that were unusual for peer review at scientific journals and the National Academy of Sciences. We were told the third review would follow the National Academy of Sciences model, but we disagree that this review has followed that model. This review has deviated from the National Academy's model of peer review for the following reasons.

(1) As was the case after the second peer review, the CEC again disallowed our revision of the report in response to this review. We are mystified about the purpose of a review which does not lead to the report's revision and improvement. The purpose of scientific peer review is to improve reports of scientific research. The conclusions of peer reviewers are not necessarily right while the authors' conclusions are wrong. Peer review comments contribute to the scientific process by adding additional perspectives that the authors can use to improve their product. Receiving and responding to review comments without improving the report makes little sense to scientists.

(2) The third review was performed only by statisticians. Employing only statisticians prevents achieving a balanced review, which is a hallmark of scientific peer review.

(3) The statisticians were directed to examine certain aspects of the report, which could have biased the review. Normally, journal editors and National Academy referees of the review process do not direct the reviewers' efforts to certain aspects of reports under review.

(4) The National Academy's process usually allows reviewers to contact the authors and ask questions to clarify their understanding of the report, but in our case it appears the reviewers were not given the opportunity to contact us. None did. It would have helped had the reviewers asked us questions because it was evident in their comments they were naïve about the history of bird and bat collision research and the methodology typically applied to the type of research we performed at wind turbines. The reviewers appeared naïve about what we could and could not have done with regard to study design.

(5) After submitting our response to comments, the reviewers will have the opportunity to comment further on the report and to our responses. The National Academy normally appoints a referee to oversee the authors' responses to comments and to make sure they are reasonable and comprehensive, but the reviewers do not get the chance to see the authors' responses, or to respond to them. We do not understand why the reviewers get the unusual opportunity to comment on our responses to comments.

(6) The National Academy of Sciences normally identifies the reviewers once the process is completed, but in our case it appears the reviewers will remain anonymous. We feel strongly that the reviewers be identified. An increasing trend among scientific journals is for reviewers to sign their names to their reviews, because doing so encourages constructive reviews.

Despite our view that this review process as inconsistent with either National Academy of Sciences review or conventional review used by scientific journals, many of the comments were constructive. Normally, we would have used these comments to revise our report. Instead, we will use them to improve our manuscripts under preparation for submission to scientific journals.[2] However, we agreed to respond to these comments so that we may also recommend how future studies can be improved to avoid or minimize the methodological shortfalls identified by the third review and to which we agree exist.

ORGANIZATION OF RESPONSES

The reviewers provided 80 pages of review comments, which were unnumbered. For the sake of efficiency, we responded first to comments on broader topics, and followed this section with comment-specific responses. Our conventions to responding to comments are detailed below.

Symbol	Meaning
R#	Reviewer number, so R1 represents Review team 1.
P#	Page number where a particular comment can be found.

A comment followed by the citation, (R1:P3), means the comment can be found on page 3 of Review team 1's comment letter. Most of our responses symbolize only the page number (e,g,, P3) because the corresponding review team was identified as a heading. Some comments lack a page number because it was from the same paragraph as the preceding comment that was identified by a page number. Comments appear in italics, and our responses follow in normal font. We skipped over some comments that were addressed specifically in other locations.

OVERVIEW OF REVIEW COMMENTS

Many useful comments were provided by the reviewers, and we agree with many of them. However, as stated, the review could have been more constructive had the reviewers been familiar with the issue. Some comments were irrelevant or confused, caused by lack of familiarity. For example, Review Team 1 appeared amazed that we neglected to discuss turbine lighting as an issue in the APWRA, but wind turbines in the APWRA are not lit. The issue of nocturnal migrants colliding with tall towers on the east coast of the U.S. cannot be extrapolated to wind turbines on the west coast, especially these small ones in the APWRA. The wind turbines in the APWRA are nowhere near the heights of communication towers, so citing the literature on collisions with communication towers would be irrelevant.

Communication between the reviewers and us could have lessened the impact of the reviewers' lack of familiarity with the wind turbine collision issue. The CEC originally told us that we would be responding to queries. We could have informed the reviewers about certain aspects of our study, just as we informed two of the three reviewers of our report during the second peer review. However, we received no queries from the reviewers.

---

[2] It is important for non-scientific readers of this document to be forewarned that results in scientific journal papers do not always correspond with the results published in a preceding agency report. Continued research and analysis often lead to differences in methods used, results obtained, and interpretations of results. Since our report was released in 2004, we have made many advances in our analysis of the data and our presentation of the results.

The review would have been more helpful had it been balanced. The CEC hired us to perform the study because we are biologists, yet the CEC had our report reviewed the third time solely by statisticians. The biological elements of the report are of paramount interest and the statistical information, when applied, is intended to support the biological elements. After reading the reviewers' comments, we agree it would be prudent to consult with statisticians during the study design and analysis phases, but biologists are needed to conduct studies of this nature because they understand the biology and tend to be more pragmatic with respect to study design and analysis.

Reading through comments made by Review Teams 1 and 2, we found many contradicting and redundant comments. For example, reviewers of Chapter 2 expected to see methods and results from Chapters 3 and 7 to be presented in Chapter 2, as if the reviewers of Chapter 2 did not realize the methods and results they expected to see in Chapter 2 appeared later in the report where they were appropriate.

Review Teams 1 and 2 raised issues early in their reviews which were resolved later in their reviews, but the original comments remained. In some cases, these reviewers disagreed with particular conclusions we made, but later agreed with these same conclusions. One example includes early comments disagreeing with the mortality metric we used, but later agreeing we used the best of the available metrics. By leaving comments on issues that the review teams later resolved, any reader of these reviews would need to read through the entirety of the comments before realizing that many of the issues raised by these reviewers turned out to be non-issues.

Finally, we must comment on the tone of the reviews from Review Team 2. Their reviews would have been better received had they made many fewer misleading and incorrect comments. Also, they could have found a more helpful way to express their disagreement with one of our methods other than to declare it "foolish." This is just not the type of language that ought to be used by scientists in a review, anonymous or otherwise.

BROADER ISSUES RAISED BY THE REVIEWERS

Pseudoreplication (R1:P1)

According to Review Team 1, statistical inference depends on samples being randomly selected and measured. We disagree with this premise. Random samples are not always desirable (Hurlbert 1974), nor are they always possible. It is more important to achieve treatment replication and interspersion within the study area than it is to sample study units randomly. Often in biological field investigations, investigators sample systematically to achieve interspersion of treatments, which minimizes gradient effects. Field investigators sometimes attempt to incorporate random sampling into studies of rarely occurring events or units, referred to as adaptive cluster sampling. Taking a simple random sample of statistically rare events can doom a study to insufficient sample sizes.

As an example of sampling randomly from wind turbines to perform fatality searches, Anderson et al. (2004, 2005) divided their Tehachapi and San Gorgonio wind farm study areas into plots, and selected randomly from among these plots. From wind farms composed of thousands of wind turbines, their random plot selections resulted in fatality searches at 201 wind turbines at one wind farm, and a similar number at the other. As a consequence, the logistics of their study were cumbersome, and their sample sizes of fatalities were insufficient for drawing reliable inferences of the factors related to wind turbine collisions.

Review Team 1 criticized the report when they pointed out (R1:P1) that turbine strings were treated as a unit, and then individual turbines were treated as independent samples. On page 47, we wrote "*...we chose the string of turbines as one of our study units because searches were efficiently performed on them.*" Note that we identified the turbine string as ***one*** of our study units. On page 331, we suggested birds possibly perceive wind turbine strings as a unit, and perhaps sometimes attempt to fly around the turbine string, which might partly explain why more birds are killed by end-of-row turbines. However, we did not conclude the wind turbine string was the only possible study unit, and we obviously considered both the turbine string and the individual turbines to be reasonable study units. Whereas birds might perceive turbine strings as units, they also likely see individual turbines as units, and the latter is the actual unit that kills birds. Because we regarded either study unit as reasonable, we provided results at two levels of analysis – at the individual turbine level and at the string level. We did this as a service to the readers, who can then more easily come to their own conclusions about which study unit is more appropriate, and how much stock they want to put into our results associated with each unit. We disagree with the reviewer's conclusion that our hypothesis tests at the string level caused the hypothesis tests at the turbine level to be pseudoreplicated. The latter were treated and presented independently of the former.

Some of the hypotheses we tested at the individual turbine level of analysis could not be tested at the string level. For example, the position of the turbine in the string could not be tested for association with collisions at the string level. Prior researchers in the APWRA had proposed the position of the turbine in the string as a possible key factor, beginning with some of the earliest research efforts. This factor was tested using chi-square tests by Howell and Noone (1992), Howell et al. (1991) Orloff and Flannery (1992), and more recently by Kerlinger et al. (2006) at the High Winds project in Solano County, California. Hypotheses can be tested on either the turbine string or the individual turbine as the study unit so long as the measured set is clearly defined and there is adequate representation of each category or group included in the measured set.

Nonrandom sampling of turbine strings

We agree random sampling from the APWRA's hundreds of turbine strings at the start of our study would have been preferable. However, most of these turbine strings were not available to us at the start of the study. Instead, we were incrementally granted access to groups of turbine strings throughout the study. Until we were granted access to the last set of 3,800 turbines, we selected all the turbines we were allowed to search. The last set was the largest, and we clearly could not search all the turbines in this set within the remaining time and budget. Having to select turbines out of this last set, we adopted the approach advocated by Hurlbert (1974). We

opted to select these turbines systematically in order to intersperse searched and unsearched wind turbines, and in order to more fully represent the north-south and east-west gradients of land use practices, wind turbine models, and environmental conditions across the APWRA.

An important point to consider regarding our systematic selection of turbine strings from the last group made available to us was that at the time of our selection we had no knowledge of which turbines were more dangerous to birds than any other turbines. Our selection of the last set of wind turbine strings to be searched could have been considered random for all practical purposes because our selection was naïve about which turbines killed more birds. And this point brings us to the crux of our response to comments on this issue.

Generally, randomization is used in field experiments to prevent or minimize investigator bias. Being naïve to which turbines were more dangerous to birds, and having selected turbines systematically when we clearly could not search the entire set (i.e., Set 2), we fail to see how our turbine selection could have biased the study. We agree with Krebs (1989), who pointed out that so far no good evidence exists to support the notion that systematic sampling in complex systems "leads to biased estimates or unreliable comparisons."

Randomization would have helped to minimize biases unrelated to the investigators, in our opinion. Where it would have helped is in preventing confounding caused by spatial and temporal differences among groups of turbines as they were added incrementally to the study (beyond our control). For example, the last set of turbines added to the study was searched over a 6-month period after searches ceased among the other turbines, so this last set was temporally separated and therefore prone to confounding due to conditions that may have differed between this period and the preceding 4 years. This is why we presented the mortality estimates separately between wind turbine Sets 1 and 2, as well as presenting them combined.

Extrapolation of results

The reviewers were concerned that nonrandom sampling produced results that could not be extrapolated to the other 25% of the APWRA wind turbines that were not sampled. Whereas we extrapolated our mortality estimates from the sampled 4,000 wind turbines to the 1,300 wind turbines we did not search, we in fact did not extrapolate our results of fatality associations to any wind turbines outside our measured set of wind turbines. We were satisfied reporting results of tests performed on our measured set because it was a large measured set, including the majority of the wind turbines in the APWRA. Our predictive models, as well as the mitigation recommendations in follow-up reports by Smallwood and Spiegel (2005a-c), were directed only to the wind turbines included in our fatality searches, and not to the unsearched turbines. We even cautioned readers about extrapolating our results to other wind farms.

Fatality associations were reported along with information readers could use to decide whether and to what degree they felt comfortable extrapolating our results. We explained how we included wind turbines in our study and how we measured variables. We listed our assumptions, although not as clearly as we could have. We also provided both the observed and expected chi-square cell values in the appendices, as well as guidance on how to assess the reliability of the tests. We provided notification of tests with P-values < 0.005, as well as those with P-values <

0.05 and < 0.10. On page 185, we summarized our presentation of chi-square test as follows: "*The observed/expected values derived from χ^2 tests are used as measures of effect, and need to be interpreted based on the P-value of the test, whether the expected number of observations was larger than 5 (smaller than 5 is generally regarded as unreliable), and the magnitude of the ratio. These latter considerations for assessing the significance of particular observed/expected values we leave to the reader*."

To summarize, we provided the information readers would need to decide whether and to what degree to extrapolate our results, but the only extrapolation we performed was on the mortality estimates. The only extrapolation we made was extrapolating mortality estimates from 75% of the APWRA's wind turbines to the 25% of the other wind turbines scattered within the APWRA. We believe this extrapolation was reasonable. Figure 1 shows the turbines to which mortality estimates were extrapolated, and from which the extrapolations were made. The unsearched turbines were well interspersed among the searched turbines. They were also the same models and sizes as the searched turbines: 1,021 KCS56 100 kW turbines, 5 Bonus 120-kW turbines, 6 Howden 330-kW turbines, 57 Nordtank 65-kW turbines, and nearly 200 unknown turbines since removed for the Buena Vista repowering project. Other investigators routinely extrapolate their mortality estimates to unsearched turbines, and even incorporate an extrapolation term for this purpose in their mortality estimators. However, they usually do not achieve the level of interspersion of unsearched and searched turbines we achieved.

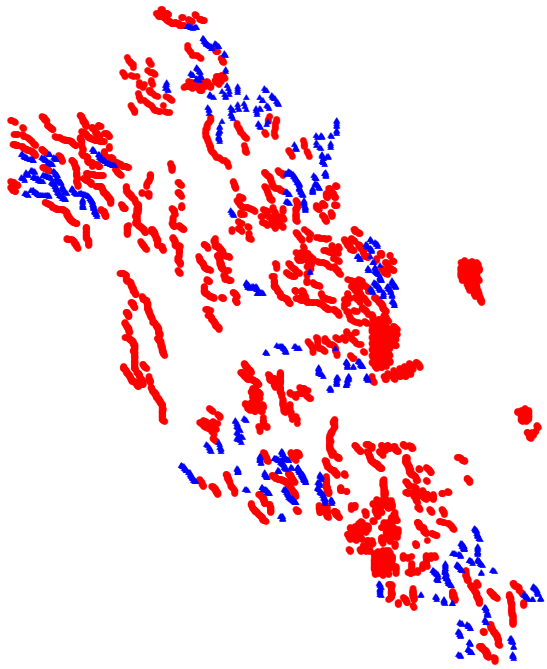


Figure 1. Wind turbines searched (red circles) and unsearched (blue triangles) for fatalities.

Multiple comparisons with inter-correlated variables without appropriate corrections

The reviewers felt we compared many inter-correlated variables without correcting for their shared variation. The reviewers appeared to have misunderstood our approach. We decided to use chi-square tests one variable at a time because our fatality data were derived from differential sampling effort. Some turbine strings were searched twice, whereas others were searched up to 34 times. We could have pretended to account for this differential sampling effort by transforming the fatality data into rates, such as fatalities per search, or fatalities per MW per year. This approach would have enabled our use of multivariate analysis, stepwise multiple regression analysis, or other methods to reduce the effects of multicollinearity. We chose not to convert our data to rates because we felt their underlying search effort varied too greatly.

We presented our methodically performed goodness-of-fit tests as a service to the reader, in case any reader had an interest in a particular variable. Given this was an agency report, and not a scientific journal paper, we felt we should take advantage of the opportunity to present everything we did. But we did not present our methodical test results with the intention of comparing them all at the same time to draw inferences. We presented them as our first of several steps to reducing the number of variables used in synthesis and in developing predictive models.

As we stated above, we screened these test results for inclusion in the next step of our analysis. We selected from this large set of test results those with small P-values, interpretable gradients in fatalities versus categories or levels in the association variable, relatively fewer expected cell values <5, and large measures of effect. We also tried to represent each larger factor with only one or a couple of representative variables in our analysis. Although we did not use currently popular statistical methods for data reduction and variable selection, we sought to achieve the same ends using our more laborious approach. We believe our approach required a lot more thinking about shared variation, confounding, and other data issues than is commonly applied by biologists utilizing more sophisticated data reduction and variable selection methods. Contrary to claims we ignored possible confounding and multicollinearity, we were more aware than anyone of these possibilities, because of the approach we took. We tempered our inferences accordingly.

The proof of our variable screening is in the syntheses presented in our discussion sections. We did not attach biological significance to every goodness-of-fit test we performed. Far from it, we carefully selected which results to discuss as significant. Furthermore, our predictive models are not composed of all the variables we presented earlier in the chapter on fatality associations. Instead, these models were composed of small subsets of the variables tested for goodness-of-fit earlier in the chapter. It is in these models and in the syntheses where variables ought to be examined for confounding, multicollinearity, and likelihood of type I errors. Focusing on possible multicollinearity, confounding, and Type I errors among the goodness-of-fit tests preceding our screening of the results is misdirected.

Use of multivariate or multiple response tests

Whether we used discriminant function analysis, Poisson regression, logistic regression, or a general linear model, we still would have faced some of the same problems discussed by the reviewers. When it came to selecting the statistical tests to use, there was no opportunity to change the experimental design, which was largely dictated to us by the wind turbine owners and existing circumstances. As we stated on page 353, "…*our study design was constrained by its post-hoc nature. We had little to no control over the replication and interspersion of treatments, including control treatments. Thus, our results were prone to inflation of measured effects and to confounding*." This is not to concede our design was flawed and should not have been implemented, but it does emphasize that we recognized and properly documented the limitations of our design and the results it generated.

The selection of variables *was* under our control, however. Smallwood (1990) faced this same problem using discriminant function analysis to predict success or failure of exotic species invasions based on many measured variables. Smallwood (1990) used principal components analysis (PCA) to group predictor variables by their degrees of shared variance, and then he selected only one variable from each PC for inclusion in the discriminant function analysis. He did this to minimize multicollinearity.

We recognized from the outset we could have used Smallwood's (1990) approach, but we decided it was not needed because we could easily identify the groups of predictor variables that shared variation. And when it came to synthesizing our results at the ends of Chapters 7 and 8, we selected only one or two variables from each group to discuss. And when it came time for model development, we selected only one or two variables from each group for inclusion. Thus, each of our predictive models only included one or two variables representing wind turbine attributes, because the wind turbine attributes were highly inter-correlated.

Another problem with using multivariate or multiple response methods is the limitation of our sample size. Even though we accumulated the largest number of fatalities among the studies of bird collisions with wind turbines to date, our sample size was still relatively small for use with multivariate or multiple response tests. Each additional predictor variable and each additional interaction term further partitions the underlying data set into smaller subsets, and some of these subsets might be insufficient. Multivariate and multiple response methods might be more practical for larger sets of fatality data collected over longer time spans using equal numbers of fatality searches among wind turbines.

Next we will discuss some specific issues with the multivariate and multi-response analyses proposed by the reviewers.

Discriminant function analysis uses only continuous predictor variables, so it was unsuitable for the many categorical variables we measured. We could have used categorical variables by treating them as dummy variables, but dummy variables are difficult to interpret when used in discriminant function analysis.

Logistic regression analysis yields a dichotomous outcome, which was more limiting than we originally preferred. In the end, whereas our models yielded continuous output from -1 through 1, we interpreted the output dichotomously, which was the level of interpretation we felt comfortable making. Since our report, Smallwood and Spiegel (2005a,b,c) improved on these models considerably, but they still refrained from using logistic regression. Smallwood has steered away from using logistic regression with these data because the dependent variables were obtained from differential sampling efforts. However, in hindsight, we could have selected the portion of our data set including relatively equal sampling effort, and developed logistic regression models from those data. We predict our results would have been very similar, but it would be worth trying this test. One advantage of this approach would be the rest of the fatality data would be held aside for use in validation. We might try using logistic regression with our data when we prepare manuscripts for submission to scientific journals.

Multicollinearity remains a problem for logistic regression, and would need to be dealt with in similar fashion to the way we dealt with it. Also, outliers could pose a problem for logistic regression and other general linear models, whereas our ratings approach quashed the large effects of outliers. Again, however, we are willing to try logistic regression.

Poisson regression would have worked had we a restricted predictor variable selection to only those variables with 20% of expected cell values >5, which is the same rule of thumb we applied to inclusion of predictor variables in our simple models. Also, it would have worked had we no outliers, an adequate sample size, and independence of observations. Furthermore, we would have had to categorize continuous variables, which would have given up information, but we ended up doing this anyway using our ratings approach. Overall, defending our use of Poisson regression would have been more problematic than defending the approach we used.

Inappropriate use of chi-square tests on measured variables

The reviewers were split on whether we appropriately used chi-square tests on animal behavior data. Two reviewers pointed out that behavior or activity data collected sequentially from birds were likely not independent. For example, a red-tailed hawk observed soaring one minute will likely be observed soaring the next minute because it is the same bird responding to similar environmental stimuli. But the test assumption of independence of observations does not apply so much to the momentum in the bird's behavior as it does to the increased likelihood the observer will record the bird again as a direct result of seeing the bird a minute ago. We would agree the behavior of the bird can affect independence of observations in certain cases, such as tracking individual birds as they fly around, or by returning attention each minute to locations where a bird had been seen perching, but in most other cases the observers performing 360˚ visual scans are slowly turning to change their viewsheds. We had the observers do this in order to approach independence of observations, though we would agree we likely did not truly achieve it.

Lacking complete independence of observations, the question then becomes how significant is this to the chi-square tests we used? We disagree with the reviewers' conclusion that many or all test results related to behavior should be dismissed because a test assumption may have been violated. We believe many of the results were too strong to dismiss. For example, we observed

burrowing owls flying within 50 m of wind turbines 10 times other than expected, and the corresponding chi-square test was significant with a very small P-value. Burrowing owls are obviously flying disproportionately more often close to wind turbines, and this pattern is biologically significant because we found a relatively large number of burrowing owls killed by wind turbines. We cannot dismiss this result just because we may have violated a test assumption.

Differences in observer ability were not incorporated

We believe Review Team 1 overestimated the variation in searcher detection across the seasons and across environmental conditions in the APWRA. A review of the photos in our report illustrates the general lack of variation in vegetation cover types and heights. Our search areas were almost entirely in annual grassland grazed by cattle. There was little spatial variation in grass height, and spring was the only season of the year when grass height was greater than a few inches (Photos 1 through 3). It is always possible we missed a few more small bird and bat carcasses during spring, but we do not believe we missed enough to matter substantively.

Photo 1. This barn owl feather pile was easily visible on the short-stature vegetation in the APWRA. It is relatively easy to find bird carcasses present within our search areas during fatality searches.

Photo 2. Golden eagles were easy to find anywhere in the APWRA during any season, as long as we were out there looking for them.

Photo 3. Great horned owl carcass found during a search in spring, when the grasses tend to be a little taller. Even with taller grasses, bird carcasses are relatively easy to spot.

Differences in scavenger removal rates were not incorporated

In the absence of evidence, Review Team 1 concluded scavenging is distributed unevenly across landscapes, which confound patterns in fatality detections among wind turbines. In fact, nobody knows whether scavenging varies significantly across the APWRA. We agree, however, that it might vary spatially. But so could any number of other factors affecting the fate of carcasses deposited under APWRA wind turbines. We reported our discoveries of raptor carcasses under rocks (e.g., Photo 4), as well as those picked up by personnel administering the WRRS, who neglected to inform us about some of the carcasses removed.[3] The rates of carcass removal and illegal carcass hiding also might have varied across the APWRA. We have no idea how many carcasses were removed or hidden, for example, by maintenance workers and ranchers. This likely occurred due to the contentious nature of this research and its financial impacts on those people. We acknowledge these types of potential error, and yet we stated our simple assumption of equal scavenger removal across the APWRA for the purpose of performing our tests. We assumed our assumption can be weighed by the reader when deciding how much trust to place in the results.

[3] The following is a note Smallwood made during the study: "Golden eagle carcass (BRC No. 1244) had been picked up by Tara Dinman on 9/16/02, but we found the feathers on 10/15/02. Not until after we found the feather trail did we learn of Tara's removal of the carcass." Tara Dinman worked for Greenridge Services, LLC., and routinely picked up birds reported by maintenance personnel and by us.

Photo 4. Red-tailed hawk carcass discovered under a rock pile. We did not routinely pick up rocks to look for bird carcasses, so we do not know how many we missed because they were buried.

Confounding

We agree with the reviewers many or our test results may have been confounded or spurious, and we pointed some of these out in our report. We simply reported what we found, in a methodical, laborious manner due to the differential sampling effort in our study. We never ignored potential confounding, and we took steps to minimize it. Had we the opportunity to revise the report, however, we would take additional steps to minimize confounding while testing hypotheses. For example, we could test whether fatalities related to tower type by selecting only those fatalities occurring at towers on ridge crests in order to prevent any confounding that may have resulted from a particular tower type occurring more often in canyons.

Not all instances of confounding would have invalidated our results, and not all spurious relationships were a problem, either, as they related to management recommendations. For example, even though associations between raptor behaviors and month of the year may have been confounded with temperature, the most pragmatic mitigation recommendations were directed toward month of the year.

Type I error

The reviewers pointed out that a statistical significance threshold of 0.05 likely resulted in Type I errors. They pointed out this threshold of 0.05 should average about 1 Type I error in 20 tests, and because we produced hundreds of tests we must have committed multiple Type I errors. However, the reviewers did not acknowledge many of our test results were non-significant at this threshold, and many were associated with P-values much smaller than 0.05. Type I errors would not have occurred among the many non-significant results, and they would have been rarer than 1 in 20 tests associated with P-values much smaller than 0.05. Test results associated with P-values < 0.005 would have produced about 1 Type I error in 200 tests.

Furthermore, this comment was directed toward the many tests presented as the foundation or first step in the development of syntheses and predictive models. We used several criteria to select the test results for use in synthesis and development of predictive models. The important question would be how many of these selected test results would be products of Type I error? For red-tailed hawk, as an example, 2 of the test results were significant at the 0.05 level, and the other 10 were significant at the 0.005 level. Given our selection criteria, we think it unlikely Type I errors were committed more than one or two of the test results used in synthesis or development of predictive models.

IMPROVING RESEARCH ON BIRD AND BAT COLLISIONS WITH WIND TURBINES

Smallwood and Thelander's (2004) report to the CEC has now been reviewed three times. As part of our response to the third review we agreed to suggest how future, similar research could be improved based on our experience and on those review comments with which we agreed. We agree with the reviewers' suggestion that manipulative experimentation would contribute more to our knowledge of the problem than continued mensurative study, though we disagree with the reviewers' claim that little can be learned from a mensurative study. Manipulative experimentation requires sufficient understanding of the factors potentially contributing to direct and indirect impacts, as well as cooperation from the wind turbine owners in experimentally designing the wind farm or manipulating the design of an existing wind farm. At the start of our study, our knowledge of bird and bat collisions was insufficient to justify experimental manipulations. Since our study, the wind turbine owners have resisted our every suggestion of experimental manipulation of treatments in the APWRA. Manipulative experimentation was never an option during our study, and will remain infrequent in wind farms for financial reasons.

If we were asked to repeat research of the collision problem in a wind farm, we would agree to perform the research only if we were able to improve on our methods. We would probably perform another mensurative study. Wind turbines have become too expensive to install in experimental designs to suit researchers of bird and bat collisions. Once installed, it is too expensive to move wind turbines to achieve experimental research objectives. The only measures left to the experimental researchers would be painting schemes, acoustical devices, lighting, and land use practices. But we concluded the most important factors include siting, wind farm configuration, and height of the rotor plane above the ground. These are factors not easily manipulated to serve experimental research objectives.

Were we to agree to perform manipulative research, we would do so only if we believed the manipulation would reduce direct or indirect impacts; we would not agree to implement any treatments thought more dangerous to birds or bats. From the researchers' standpoint, experimentation of wind farm attributes to test treatment effects on bird and bat collisions poses much the same problem faced by the wind turbine owners in operating a business while knowingly taking birds and bats protected by environmental laws. Performing a mensurative experiment, researchers are more likely to be free of this burden because they are not active participants in killing protected species. Performing a manipulative experiment, the researchers are actively taking part in killing birds and bats in order to learn how and why they are getting

killed. Whereas we understand the need for manipulative experimentation, we would be unwilling to participate with it unless we are reasonably sure the treatments will be effective.[4]

One example of a manipulative experiment we would be willing to implement would be the experimental addition or integration of the VMA, Inc. vertical axis wind turbine into a wind farm, especially replacing end-of-row horizontal axis turbines with this one. Based on what we learned about bird responses to wind turbines of various designs and arrangements in the APWRA, we are reasonably confident VMA Inc.'s wind turbine would not kill birds. Another treatment we would endorse would be end-of-row pylons or other barriers intended to encourage birds to fly farther around the last operating wind turbine in a string.

In this section of our responses to comments, we will suggest how future research can be improved, but we will largely restrict our suggestions to mensurative research methods. We restrict ourselves to the scope of research we performed or we would have liked to have performed, and we will not suggest research or research methods that could be conducted at sites without wind turbines. We will not suggest the types of research that ought to be performed in advance of decisions about where wind farms will be constructed. We will not suggest research methods to test the effectiveness of potential mitigation measures, although the methods we will describe can be extended to tests of mitigation effectiveness. Our focus will be on performing research at locations where wind turbines are planned or already operational. We will also recommend study methods implemented in compliance with permit conditions, because pre- and post-construction utilization and mortality surveys can and should contribute more effectively to knowledge of the factors related to wind turbine collisions and indirect impacts.

Smallwood and Thelander (2004) contributed much more to our knowledge of factors associated with wind turbine-caused bird collisions, but its design and scope could not answer all the questions that remain about how birds are killed by wind turbines. More research is needed to identify causes of collisions and what measures could be taken to effectively reduce mortality caused by wind turbines. Much of this research can be directed to particular questions raised by the Smallwood and Thelander (2004) study, as well as by other studies.

Directed research efforts are needed, but a great deal also could be learned from ongoing and future programs to monitor wind turbines for fatalities and to characterize bird and bat utilization of the wind farm. Whereas many of the fatality monitoring and bird utilization studies at wind farms are performed to satisfy mitigation requirements, and are not research efforts in the sense of testing hypotheses, they can still be performed in a manner that improves our knowledge about wind turbine collisions. Our recommendations will target research studies, but will at times apply to perfunctory pre- and post-construction studies at wind farms pursuant to permit conditions.

---

[4] Some participants with the bird collision issue in the APWRA have increasingly argued that experimentation is warranted for those mitigation measures we recommended that are more uncertain in their effectiveness. We would agree with this argument so long as it is applied to modifying the existing APWRA, but we argue that any experimentation of measures in new projects ought to be directed to the measures with greater certainty in their effectiveness. Those situations we identified as potentially dangerous to birds, but for which we are uncertain, ought to be avoided in new projects whenever feasible, rather than included experimentally.

SITE UTILIZATION

Bird and bat utilization research has been performed among wind farms to achieve two objectives: (1) characterization of behaviors that might relate to collisions, and (2) estimation of relative abundance. Flight and perching patterns in environmental settings associated with wind turbines can be related to collision rates, and these relationships can be used to forecast relatively safer locations and heights above ground to install wind turbines for minimizing collisions. Relative abundance estimates can be compared within and between wind farms to quantify direct and indirect impacts, as well as changes in impacts following the implementation of mitigation measures.

Increasingly the methods used to achieve objective (1) have been used to achieve objective (2), but inappropriately (Smallwood 2006). The principal research method used to achieve objective (1) has been the 360˚ visual scan, which can be useful for objective (2) so long as it consists of instantaneous counts and so long as the search radius is reasonably close to the observers, and methods are similar among project sites or through time. It is also useful only for particular bird species, and not for grassland songbirds, migrants, and many other species. Characterizing relative or absolute abundance requires the application of specific, well-accepted methods, such as applying distance estimators to data obtained from grassland songbirds flushed by observers walking along strip transects.

Both of the general objectives of utilization research need to be pursued both pre- and post-construction of wind turbines. More needs to be known about how birds and bats change behavior patterns in response to the installation and operation of wind turbines (some behavior patterns might change in response to installation, whereas others might respond specifically to turbine operations). Also, research is needed to quantify the species-specific changes in relative abundance and spatial distribution following wind turbine installations.

Appropriate experimental designs would be BACI and Impact-Gradient designs, as well as a combination of these designs.

Behavior Research Methods

Variables that need to be quantified include the following.

- Species
- Height above ground
- Flight direction
- Specific behavior, e.g., soaring, contour flight, hovering, powered flight, attack on prey
- Perch, e.g., ground, fence, wind turbine
- Spatial coordinates of the observed bird or bat
- Association with vegetation, landscape feature, slope, elevation
- Proximity to wind turbine, if data are collected post-construction

Additionally, the analyst will need to know the environmental conditions associated with each observation, including weather and wind conditions in the observation area during the

observation session (and any changes during the session). Peripheral data should include the following.

- Wind direction at OP, at least every 15 minutes during the session
- Wind speed at OP, at least every 15 minutes during the session
- Visibility including weather conditions affecting visibility during session
- If post-construction, then operating wind turbines should be identified
- Name of observer(s)
- Observation session start time
- Temperature at start time

Sample size and spatial and temporal scope

Large sample sizes are needed to adequately characterize flight and perch patterns in a proposed or existing wind farm. Minimum sample sizes have yet to be estimated, partly because both field and analytical methods have been under development. Adequate sample sizes will depend on the size of the project and the area involved, so a per-ha sample size requirement will likely be necessary.

Behavior observations need to be made during each season of the year, because species assemblies and behaviors change seasonally. Also, wind directions and wind speeds can change seasonally, and can affect flight patterns.

Behavior observations also need to span the entirety of the proposed or existing wind project area. Contiguous plots can be arranged throughout the project area, or if the project area is very large, then potential plots can be sampled randomly.

The need for sufficient sample size and spatial coverage of the project area will be challenging for rarely occurring species, as well as for nocturnal observations. In some cases, radio-telemetry might be the only efficient means to gather sufficient information on movement patterns of rarely occurring or nocturnal species.

Using human observers

Visual scans have emerged as the most common bird observation method in wind farm project areas. Visual scans are usually performed from vantage points, and typically last between 10 and 30 minutes. In the future visual scans need to be performed long enough for birds in the area to acclimate to the observers and to behave more naturally. Ten and 20 minute scans need to be replaced with hour-long scans.

Observers need maps onto which to record animal locations, and these maps need topographic data and search area boundaries to assist with estimating the locations of observed animals. A GIS analyst should be a collaborator on every future research project. The areas in the observation plot actually visible to the observers from the observation point need to be delineated so that only these areas are used in the subsequent analysis. GIS can be used to perform this step, which can be exemplified on a web site established by Lawrence-Livermore National Lab

(http://eed.llnl.gov/renewable/). To date researchers collecting bird utilization data have assumed they can see the entirety of the area within the maximum observation distance, but this is not the case, and the proportion of the area not visible varies by site and by maximum distance used. Future comparisons of utilization data need to account for this variation in visible areas.

If we were to conduct another similar study, we would limit observations of large-bodied birds to within 300 m away from the OP, and of small-bodied bird species to within 100 m (see Attachment B).

Advanced Integrated Radar and Camera System (AIRCAMS)

After twice using human observers on the ground, we would pursue the development and implementation of remote detection systems to perform bird and bat utilization research. Human observers are expensive, record only a fraction of the bird locations and behaviors in the project area, operate only during the daytime, and are prone to bias and measurement error in recording the position of the target in space. We would pursue the development of AIRCAMS.

AIRCAMS would consist of GPS linked to computer, anemometer, wind vain, thermometer, strip cameras, heat cameras, and controlling software to detect and track bird and bat targets at whichever heights above ground they are detected. Observers would initially calibrate the controlling neural network to identify behaviors by flight patterns and species by image pixel patterns, but after calibration the system might operate with minimal human oversight. AIRCAMS would capitalize on the advantages provided by other technologies. For example, the heat camera would be used only to verify targets are alive, but imagery would not need to be collected in long-term computer memory. Also, relative to human observers, radar would detect many more targets, and due to the integration of other system components, such detections would be made much more accurately and frequently.

Data analysis

No matter which research method is used to accumulate data of bird or bat behaviors in the field, the problem of autocorrelation will need to be addressed because most statistical tests assume independence of observations. Observations of birds and bats will not be independent when observed within a project site over time. Independence of observations can increase by increasing the time intervals between observations, but the logistical efficiency of the research will diminish as the duration of these intervals is increased.

Avian Risk of Collision.--Any and all utilization data should be shared collaboratively with Richard Podolsky (or someone equally qualified) so that he can further develop the predictive power of Avian Risk of Collision (ARC). ARC's accuracy and utility at proposed new wind farms, or changes to existing wind farms, can only improve with improved understanding of bird flight patterns in areas where wind farms are likely to be developed. Each research project focused on bird and bat collisions should include a budget for coordinating with Podolsky to improve ARC. ARC would especially benefit from data on avoidance behaviors as birds and bats encounter wind turbines, as well as the ranges of flight patterns and flight speeds used by birds and bats in various environmental conditions.

Map-based forecasting of safer wind turbine locations in the wind farm.—Any and all utilization data should be shared collaboratively with one or more GIS analysts (e.g., Lee Neher) for spatial analysis and development of map-based indicators of locations relatively safer and more dangerous to birds or bats. These hazards maps could be color-coded to indicate where occurrences of particular species are more likely, where particular behaviors are more likely, and where collisions will be more likely, after relating utilization data to collision data in existing wind farms.

Abundance Research Methods

Estimates of abundance or relative abundance are useful for comparing abundance to collisions to identify which species are more susceptible to wind turbine collisions. Also, and perhaps more importantly, these estimates enable crude estimates of biological impact by relating the numbers of animals killed to the numbers occurring in the local environment. A steady influx of individuals into vacated territories, however, could confound such a simple risk assessment, so additional demographic data and abundance estimates outside the project boundary would help with interpretation. Population Viability Analysis (PVA) might be ideal for use in wind farm settings, but PVA is costly and debate continues over which dependent variable to measure and how to interpret the results.

Diurnal raptors

If we were unable to deploy AIRCAMS, we would perform instantaneous counts while doing visual scans. These counts would be divided by the number of hectares visible to the observer, so the metric would be the number of individuals per hectare.

Raptor nest surveys

We did not survey for raptor nests, but if we were to conduct another related study we would search for nests in and around the project site. For most species nest distribution will be difficult to relate to turbine-caused mortality due to pseudoreplication and other reasons. For most species the value in nest surveys will be in testing whether nest occupancy and productivity changes after the wind farm is constructed. No distance has yet been established at which nesting raptors might be affected by wind turbines, so until such a minimum distance is established we would assume searching the area within 4 km of wind turbines would suffice. To decide whether raptor nest occupancy responded to wind turbine proximity, we would rely on a simple non-parametric test, or if we obtained a sufficient sample size we would use logistic regression.

We would map the locations of burrowing owl burrows, distinguished between nest burrows and refuge burrows. Evidence indicates burrowing owl burrow density near wind turbines relates to burrowing owl collision rates, so mapping of burrowing owl burrows in wind farms is warranted. After sufficient data are collected, analysts should be capable of forecasting burrowing owl impacts based on the proposed wind turbine model and siting locations.

Nocturnal raptors

We performed no nocturnal surveys for raptors, so our utilization data poorly represented nocturnal owls and other nocturnal volant species. If we were to perform additional related research, we would use night-vision or thermal imaging systems to detect nocturnal species. We would also pursue development of AIRCAMS so that we can remotely collect animal detections during both day and night. By tracking multiple targets in the sampled area at once, AIRCAMS would also be useful for instantaneous counts and relative abundance estimation.

Grassland songbirds

We did not measure grassland songbird utilization during the CEC-funded portion of our study, but we did search for these species while funded by NREL. If we were to perform an additional, similar study, we would more aggressively survey for grassland songbirds.

Whereas visual scans are intended to identify all birds in a project area, the spatial distributions and relative abundances of many bird species cannot be characterized reliably without using methods long-ago developed specifically for these species. For example, abundance estimators were developed for grassland songbirds, which are most efficiently sampled by walking line transects and measuring distances and directions to flushed birds. Appropriate sampling methods implemented in an impact-gradient design would be useful for comparing grassland songbird abundance (1) before and after a wind project is development, (2) at various distances away from the wind turbines, and (3) between wind project sites.

Bats

We made no effort to characterize bat utilization of the APWRA, but if we were to perform additional, similar research we would survey for bats. We would pursue the development of AIRCAMS, as well as the integration of AIRCAMS with SonoBat, which uses a neural network to recognize bat species by echolocation (J. Szewczak, pers. comm.).

FATALITY SEARCHES

One of three approaches should be used for selecting turbines to be searched:

(1) Select all turbines;

(2) Random or stratified random sample;

(3) Systematic sample selection.

Selecting all turbines should be the preferred approach every time, but sampling will be necessary when available budgets are too small for the size of the project. Whether the sampling should be random or systematic will depend on sample size, turbine layout, and logistical issues related to the fatality searches. The important objective of sampling is to intersperse treatments

and to minimize investigator bias. We would not begin another similar study if given incremental access to the wind turbines, as we were during our last study.

We would not begin another similar study unless we were given a sufficient time span to search for fatalities. In other words, we would not begin fatality searches if we were given less than a year to complete the searches, and we would prefer that 4 or 5 years of searches be committed in funding and access.

We would use shorter fatality search intervals, perhaps randomly selecting some subset of turbines or turbine strings to be searched every day.

The search radius should encompass most of the area carcasses are reasonably expected to fall, based on past experience at wind farms, wind turbine rotor diameter, tower height, and the steepness of slope. We would use a search radius greater than 50 m, perhaps 60 m for the older wind turbines in the APWRA. The search radius for larger, new-generation wind turbines should probably be 75 m or greater, depending on the height above ground and diameter of the rotor.

We would try using trained dogs to find bird and bat carcasses.

MORTALITY METRIC

All future studies should express mortality as the number of fatalities per kWh since the last fatality search. This unit of power generation expresses the actual power generated over the period of time since the last fatality search. Wind turbines do not operate at their rated capacities, and the actual power generation varies due to location, season, and maintenance issues associated with particular wind turbines. The term kWh expresses both the wind turbine's rotor diameter and its operation time since the last fatality search. If one turbine operates less than another but kills equal numbers of birds over some time frame, then the former is more lethal than the latter. We believe use of this metric would dramatically improve our ability to explain the variation in fatalities, and to develop an understanding of causal factors. This metric would require cooperation from the wind turbine operators to the extent they make available the electrical output data of each turbine.

MORTALITY ESTIMATOR

Smallwood (2006) identified two basic estimators used by researchers of bird and bat collisions among wind farms. The most recent mortality estimator used by WEST, Inc. (2006) was the following:

$$M_A = \frac{\bar{c}}{\left(\frac{\bar{t} \times p}{I}\right) \cdot \left(\frac{e^{I/\bar{t}} - 1}{e^{I/\bar{t}} - 1 + p}\right)}, \qquad \text{eqn. 1}$$

where $\bar{c}$ is the average number of carcasses observed per year, $\bar{t}$ is the mean number of days until carcass removal, p is the observer efficiency rate, and I is the search interval in days. This version of the WEST, Inc. estimator was revised from a previous version after Shoenfeld (2004) concluded it biased mortality estimates about 23% low.

Most other investigators estimating wind turbine-caused mortality have used the following formula to adjust for the fatalities not found due to scavenger removal and searcher detection error:

$$M_A = \frac{M_U}{R \times p}, \qquad \text{eqn. 2}$$

where M_U is unadjusted mortality expressed as either number of fatalities per wind turbine per year or number of fatalities per MW of rated capacity per year, R is the proportion of carcasses remaining since the last fatality search,[5] and is estimated by scavenger removal trials, and p is the proportion of carcasses found by fatality searchers during searcher detection trials. Additional adjustments could be incorporated into eqn. 2, such as background mortality (M_B), crippling bias (M_C), and search radius bias (M_S):

$$M_A = \frac{M_U}{R \times p} - M_B + M_C + M_S \ . \qquad \text{eqn. 2b}$$

Scavenger Removal

Based on the adjustment terms used to date, eqn. 2 is most sensitive to variation in the proportion of carcasses remaining after a set number of days corresponding with both the duration of the scavenger removal trial and the interval between fatality searches in wind farms.

We would strive to obtain species-specific estimates of scavenger removal rates. Surrogates might be useful in future scavenger removal trials once scavenger removal rates have been characterized for each species. To date, the opposite sequence has been employed – using surrogates such as rock doves and chickens without even knowing whether they are removed by scavengers at the same or similar rates as the species of interest.

Conducting another similar study, we would try to use fresh carcasses rather than frozen carcasses. Most scavenger removal trials have utilized frozen carcasses, but evidence is mounting that vertebrate scavengers more readily remove fresh carcasses, or carcasses that were never frozen.

We would deploy body parts and carcasses with injuries consistent with turbine collisions. Whole carcasses may not transmit the same types or levels of odors as damaged carcasses, and vertebrate scavengers may not respond at the same rates. Furthermore, dismembered birds, especially large-bodied birds, might be scavenged at higher rates simply because they are easier to pick up and carry off.

We would deposit carcasses at rates consistent with the rates of carcass deposition caused by the wind turbines in order to avoid scavenger swamping. Vertebrate scavengers are typically self-

[5] The proportion of carcasses remaining can be converted to the percentage remaining simply by multiplying the proportion by 100%, and the percentage remaining can be converted to the proportion remaining by dividing by 100.

limiting in number and distribution due to home range and territory maintenance. These animals need time to forage over their home ranges, and once a carcass is found they need to remove it, consume it, and digest it. Faced with 10, 20, 30 carcasses at once, local vertebrate scavengers cannot process this many carcasses at once. Scavenger removal trials based on these pulses of carcasses are prone to bias, which we would avoid.

We would deposit carcasses where scavengers might have established foraging routes, i.e., by wind turbines. During our research in the APWRA we noticed foxes and coyotes patrolling wind turbine strings, as well as common ravens. To more truly characterize scavenger removal rates, we would place bird and bat carcasses near wind turbines, rather than on plots far from wind turbines.

We would attempt to identify scavengers responsible for collecting carcasses by using cameras with event detectors. Capturing the scavenging event on camera would inform us of which of the species of scavenger are scavenging which species killed by the wind turbines, and would give us exact dates of removal. Identifying the scavenger species would provide the means for researchers at wind farms elsewhere to forecast scavenging rates. For example, if common ravens typically take carcasses of species A at the rate of X per common raven per carcass-day, then the impact of scavenger removal might be estimated after estimating the number of common ravens in the project area.

If eqn. 1 is ever going to be useful, we would improve the accuracy in estimating mean days to carcass removal by deploying cameras with event detectors.

Use carcasses deposited by wind turbines (most potentially accurate approach) if exact date of the collision can be established: (1) AIRCAMS; (2) event detectors. However, this approach could make use of only those collisions attributed to species, meaning the researchers would need to either have the means to identify the species remotely (AIRCAMS) or by visiting the carcass before it disappears. Sufficient sample sizes of carcasses detected using these means would eventually eliminate the need for carcass removal trials.

Searcher Detection

Using carcasses collected from previous fatality searches, we would deposit carcasses at random locations extending from the wind turbines to the maximum search distance. We would deposit both carcasses and carcass parts to simulate the types of evidence actually found by fatality searchers. We would also deposit only 1 or 2 carcasses per day, so that fatality searchers do not get wise to the trial.

To the extent possible, we would throw carcasses from the wind turbine string toward locations we intended to place them. Throwing bird carcasses would sometimes leave feather trails, such as discovered by fatality searchers. Also, thrown carcasses will land in the vegetation similarly to how they land after being killed by wind turbines. Throwing the carcasses also would minimize walking over the area, leaving shine on depressed grass and revealing signs of footfalls that could be noticed by the fatality searchers. The actual locations of thrown carcasses can be offset into a GPS using laser rangefinders, compass, and clinometer.

Background Mortality

We encountered considerable uncertainty in attributing cause of death to many of the bird carcasses we found near wind turbines. Unless the bird was cut in half or dismembered due to blunt-force trauma, we could not be certain the bird had been killed by the wind turbine. Nevertheless, it was reasonable to assume the wind turbine caused the deaths of most of the birds and bats found within 50 m of wind turbines because birds and bats do not normally drop dead in such concentrations. However, if we were to perform additional, similar research, we would prefer to have sufficient funding to support necropsies. Necropsies would especially be helpful now that West Nile Virus (WNV) has spread throughout much of California.

Additionally, we would conduct background mortality surveys, or searches for bird and bat carcasses at locations lacking wind turbines. These background mortality searches would be performed in similar habitat conditions, but at least hundreds of meters from wind turbines to minimize contamination of these sites by birds injured by wind turbines or killed at wind turbines and carried to these sites by vertebrate scavengers.

Crippling Bias

During our work in the APWRA we found a number of birds still alive, but injured severely enough to not survive long in the wild. Undoubtedly, some unknown number of birds were similarly injured but never found by us because they were able to move away from our search areas. If we were to perform another related study, we would pursue development and deployment of AIRCAMS so that we could detect the frequency of collisions followed by the animal's movement away from the search area.

Search Radius Bias

We would search a randomly selected subset of turbines to much greater distances from the turbines than the standard search in order to estimate the proportion of carcasses ending up outside the standard search radius.

Left-Censoring of Data

As far as we are aware, adjustments to mortality estimates can only be made on those wind turbines or turbine strings where at least one fatality was found. Turbines or turbine strings with 0-values cannot be adjusted for searcher detection error or scavenger removal rate, as examples, because 0 divided by either of these adjustment terms equals 0. Therefore, wind turbines need to be searched for fatalities long enough to reasonably find a fatality, and we found in the APWRA that at least 3 years is needed before 90% of the wind turbine strings yield at least one bird fatality each. With greater search frequencies, this time period might be reduced.

ECOLOGICAL RELATIONSHIPS

After we started our study in the APWRA, we initiated some exploratory research efforts into ecological relationships that might help explain some of the variation in fatalities among wind turbines. These research trials were not originally funded by NREL or CEC; we initiated them after noticing intriguing patterns. We noticed increased clustering of pocket gopher burrow systems around wind turbines tended to correspond with greater numbers of red-tailed hawk fatalities. We noticed burrowing owl burrow density associated with more burrowing owl fatalities at wind turbines. We noticed cattle often graze and lounge around wind turbines, and turbines with more cattle pats nearby tended to be associated with more burrowing owl collisions. These and other patterns prompted us to initiate research trials to better understand how these patterns might relate to wind turbine collisions.

If we were to perform another, related study, we would pursue ecological investigations similar to those we performed before. We would expand on certain investigations we performed before, using random or stratified random selection of study plots. We would use a similar plot selection approach for any new investigations, as well, but we would also consider systematic or arbitrary plot selection if the investigation was sufficiently exploratory.

Where feasible, we would also replace index expression of certain variables with quantitative measurements, such as grass height index with sample of grass measured in cm. We would add measures of percent ground cover, and perhaps one or two additional measures of vegetation biomass and cover.

RESPONSES TO SPECIFIC COMMENTS

REVIEW TEAM 1

P3: *Finally, the report did not address the existing literature on birds colliding with tall, lighted structures at night*.

None of the wind turbines were lit, so turbine lighting was not a factor we could address. This is the first of many examples of apparent naivety on the part of the reviewers with regard to the issue of bird collisions with wind turbines in the Altamont Pass. We raise this point not to criticize the reviewers but only to note that these types of comments would not have been made by reviewers more familiar with the issue or who had corresponded with us, and the overall length of the reviews could have been either shortened or redirected to substantial issues.

P3: *The study…does not recognize that recommending that turbines be replaced on the tallest possible towers may actually increase mortality of migratory birds*.

It may be true that turbines placed on taller towers will kill more migratory birds, but nocturnal migrants have so far been much less a concern in the APWRA. Concerns over large numbers of night migrants are typically associated with areas where such migrations are extensive such as in the Great Lakes region and the eastern US. This concern is not as prevalent on the west coast of the US. We concede, however, that more nocturnal migrants may be killed than we are aware of,

and that more may be killed by turbines on taller towers. Still, our predicted changes in raptor mortality following repowering on taller towers are bearing out after the first year of operations by Diablo Winds. On page 332 we predicted an 80% reduction of raptor mortality following repowering to turbines placed at 29 m above aground, but Diablo Winds was placed on towers extending to 26.5 m above ground, so it is no surprise to us that the repowering only reduced raptor mortality by almost 70%. On page 354 our species-specific predictions also faired reasonably well. We did not expect red-tailed hawk mortality to increase like it did, but we were highly accurate in our predictions of mortality reductions for burrowing owl and American kestrel. So far, we are unaware of whether Diablo Winds killed nocturnal migrants, but none were reported after the first year.

P3: *The questions raised by authors in paragraph 2* [page 7] *are valid, but could have been expressed as sequential components that, acting in concert, result in mortality of birds*.

We agree it would have been helpful to present the introductory text in the manner recommended. In fact, the sequence of components recommended is the sequence in which we presented the results of chi-square tests in Chapters 7 and 8.

P4: *Inasmuch as Step 1 (susceptibility) and Step 2 (vulnerability) result in Step 3 (impacts, i.e. mortality), the first two terms mean nearly the same thing ("capable of being affected" vs. "capable of or susceptible to" some variable. The use of these two terms as meaning different things is jargon that is not familiar to most readers. The authors should either provide a more detailed explanation of the difference between susceptibility and vulnerability or avoid this usage.*

We defined the terms, susceptibility, vulnerability and impacts on the first page of the report's Introduction. We explained these terms were from the ecological indicators literature, and we cited that literature as "*(Rapport et al. 1985; Cairns and McCormick 1992; O'Neill et al. 1994; Rotmans et al. 1994; Schulze et al. 1994; USDA 1994; Battaglin and Goolsby 1995; Wilcox et al. 2003; for examples see Zhang et al. 1998, 2003)*." This literature is well established, and international conferences have been held on the ecological indicators approach. The terms are no more jargon than are coefficient of determination, Type I error, or significance in the field of statistics. We assume that because the reviewers are primarily statisticians, they may be unfamiliar with these commonly used ecological terms.

P4: *Data collected from two different research studies are not as robust as data collected for all variables in the same time period. This circumstance will limit the evaluation of interaction effects among some (maybe even) critical variables. The non-random addition of study sites confounds various analyses. This is especially important because of the year-to-year variation in measured fatalities*.

Data were not collected from two different research studies. It was the same study funded by two different organizations and on a continuous time schedule. The NREL research ended and the CEC funding continued that exact same research and study design with respect to collecting fatality information and other data associated with each found fatality.

P4: "*The* ***placement and*** *operation of wind turbines can make birds vulnerable to wind turbine*

collisions…"

This sounds as if birds can die at wind turbines (fly into them) even if turbines are not operating (blades not turning). Are deaths in this manner minuscule compared to deaths in moving blades or is this known?

Bird collisions with turbines that are not operating are possible, but much less likely than with operating wind turbines. What we meant by the phrase "placement and operation" is that one has to put a turbine at a location flown by birds before birds can run into it, whether or not the turbine operates. We were distinguishing the term vulnerability from susceptibility. Even birds that are susceptible to colliding with wind turbines will not collide with them until the turbines are up and running, and when the turbines are up and running, then the susceptible birds are also vulnerable to collision.

P4: "*….then the probability of an individual being killed by a wind turbine occurring on a particular environmental element would equal the proportion of the wind turbines associated with…*"

Not sure of the wording here. Would it ***equal*** *the probability or just be associated with it?*

We are not sure of the wording, either. We are attempting to characterize the expected null condition in a chi-square test. There may be a better way of wording this.

P5: *Furthermore, the addition of turbines to the search effort opportunistically creates severe problems for the analyses. Because measured fatalities varied from year to year, the addition of a large number of a specific type of turbine during a "low" fatality year would give the false impression that a certain turbine type caused less mortality when data were pooled over multiple years*.

We agree the addition of turbines as the study progressed could have confounded the results of tests for association, but we disagree this potential confounding created a "severe" problem because we do not know the magnitude of inter-annual variation in mortality.

P5: *It is of interest that for 2 of the 3 turbine types (i.e., Bonus, Micon) "percent time in operation" is lacking and these two have higher mortality rates than other types except for the Kenetech KCS-56, which has the highest number of carcasses associated with it in the APWRA (see Fig. 2-6). Table 1-1 lists percent time in operation as only 39% for the Kenetech KCS-56 type. It may be that operating duration was less for the Bonus and Micon turbines, or that the Kenetech KCS-56 just kills more birds because of its unique mechanical attributes.*

We agree. It would have been better to have known the operating times of the turbines. We regularly requested those data from the turbine owners, but were denied access to them. We were forced to design our studies around this problem.

P5: *The authors give a general description of the study area but no vegetation map is provided*.

We saw no need to provide one. The vegetation is almost entirely annual grassland throughout the APWRA. The readers can see what this looks like in Photos 1-1 through 1-8 in the report, as well as in all the other photos depicting landscape scenes in the report. Providing a vegetation map would have been pointless.

P5: *The authors never provide a rationale for how turbines were selected for the focused studies. This description also gives the mistaken impression that turbines were the sampling unit, when the sampling unit was actually the turbine string (p. 47).*

The first sentence is incorrect. Rationale was provided in each case, as explained in the methods sections of Chapters 5, 6, and 8.

The second sentence is incorrect, although we could have worded this paragraph better. Sampling units were both the turbine and the turbine string, depending on the test and the research objective.

P5: *These statements suggest that an additional component to the four we presented earlier needs to be included. That is, without before and after turbine installation studies (which authors have acknowledged are needed) some of the ecological aspects are confounded inasmuch as the act of installing the turbines changes the food base that in turn affects bird behavior and may increase exposure to effects of turbines, even if the turbines are not operating.*

Correct, except we need to caveat our agreement with the last sentence. We think the existence of the wind turbines created perching opportunities that changed the ecology of the APWRA, regardless of whether the turbines are operating, but we wouldn't agree that the mere existence of non-operating turbines would increase the exposure of birds to collision, at least not substantially. We understand that last year WEST, Inc. searched for fatalities among wind turbines that were turned off during the winter. We would be interested to know whether they found bird fatalities at non-operational turbines, although we would still need to know whether the blades of those turbines were locked in place.

P6: *The use of distances to the individual birds is pseudo-replication, which invalidates the results of the one-way ANOVA tests. The sampling unit is the string of turbines. Consequently, the individual turbines are sub-samples of the strings.*

The premise of the comment is incorrect. We measured bird carcasses as distances from individual turbines, not from turbine strings. Our sampling unit was the turbine string only in the case of making mortality estimates, and for some selected tests for association with measured variables. Our sampling unit in Chapter two was the individual turbine.

P6: *Furthermore, the authors use one-way ANOVA seemingly without regard for the underlying assumptions of the procedure, which include normality of error distribution and homogeneity of variance across variable levels. Figures 2-9 (p. 39) and 2-12 (p. 43) (reproduced below) illustrate violations of both assumptions.*

These are good points, suggesting that other tests might be more appropriate in the cited cases. In our experience, however, the tests used often obtain the same results when P-values are small.

In the case of Figure 2-12, we doubt a chi-square test or any other test would have been significant, just as the ANOVA test was not. In the future, we will explore the use of other tests for these hypotheses.

P6: *Use of LSD for post-hoc multiple correlations dramatically increases the chance of Type I error (i.e., labeling differences as significant when underlying population means are not). For example, in the discussion of blade tip speed (top of p. 42), with 10 categories there would be 45 possible LSD tests, which would lead to a Type I error probability of 90%.*

So, we likely committed a Type I error in one of the 45 post-hoc tests. To a statistician, this likelihood might seem unacceptable, but our experience with field biology gives us a more positive outlook on it. What makes this criticism all the stranger to us is that we used LSD tests in this example to downplay the significant test result we obtained with the one-way ANOVA. We wrote, "*We found that carcass distances from wind turbines differed significantly, based on blade tip speed (ANOVA $F = 3.72$; $df = 9, 455$; $P < 0.001$), although LSD tests revealed that the differences were only due to two turbine models operating at intermediate-fast speeds and otherwise there was no gradient from slow to fast speeds*." In other words, we examined the test result with LSD tests and a visual inspection of the means for a gradient among the possible tip speeds. After the extra care we took to examine this relationship, we cast doubt on its significance.

P6: *The choice of 38 cm as the dividing threshold between large and small body sizes seems arbitrary (see Figure 2-1 reproduced above). Fifty centimeters or the median seem like more reasonable choices*.

On page 28 we wrote, "*Bird species were represented by typical body length (cm) as reported in National Geographic Society (1987), and were categorized as small (< 38 cm) or large (> 38 cm), the cutoff based on a natural break in a histogram of body length (Figure 2-1)*." In our opinion, a natural break in the histogram makes more sense than an arbitrary summary statistic of central tendency. Furthermore, the mean body length was 39.6 cm (see Figure 2-1), which was a lot closer to our cutoff of 38 cm than was 50 cm.

P6: *These results are presented as uncorrected counts. For comparability, the fatalities need to be expressed as carcasses per search effort, which needs to be clearly defined, e.g., hours or area or a combination, i.e., search effort per unit area. As raw counts, the reader does not know if the seasonal differences result from differences in search effort or seasonal changes in mortality*.

We recognized this, which is one reason why we presented Figure 7-2. The counts in Figure 2-6 can be compared to the proportions in Figure 7-2 to get an understanding of which turbine models associated with more fatalities. Simpler yet, search-adjusted counts can be compared to search-adjusted, expected counts in the appendices reporting chi-square test results.

P6: *Results in Figure 2-6 should be expressed either as mortality per turbine type per search effort, or the graphic should express mortality as a percentage of the total mortality and this or another adjacent graph should depict the percentage represented by each turbine type so that readers can quickly assess whether some turbine types are associated with mortality*

disproportionate to their prevalence on the landscape. The figure as currently constructed could be misleading.

The premise of this comment is incorrect. This chapter was not intended to compare mortality estimates or to identify factors associated with fatalities. On page 28, we identified the purpose of this chapter, "*In order to assess the efficiency of our search radius, we tested whether the distance of the carcass from the wind turbines related to the body size of the bird species, wind turbine attributes, season, and physiographic conditions*." There was nothing more of substance to this chapter.

P7: *This figure and associated text should be expressed in fatalities per turbine at each altitude, or should express the mortality as a percentage of the total mortality and also should graph the percentage represented by each elevation class. The current figure does not provide much useful information because it is not clear if the pattern results from the elevational distribution of turbines or an inherent elevational pattern in mortality*.

This graph was only meant to be descriptive. For a fuller understanding of how elevation affected fatalities (assuming no confounding, which we discovered and discussed later), see Figure 7-14, Table 1, and the appendices reporting the chi-square tests.

P7: *Both halves of this figure* [2-11] *are meaningless (i.e., R^2 =0.01) and inappropriate. The LSD tests described on p.38 indicate that the relationship between distance and height is not linear (i.e., the 43-m tower mean is less than the intermediate height towers.) In addition, the scatter plots show clear violations of the assumption of constant variation in distance across the tower heights*.

The turbines on 43-m towers were few and they had shorter rotor diameters than did turbines on 29- and 32-m towers, so we would not put a lot of stock in the reduced distance from the turbines on 43-m towers. Whereas the reviewer believes these relationships shown in the figures are "meaningless," we disagree, but we do recognize the relationships as weak. Note that we included multiple statistics along with the scatterplots so that each reader can decide for himself whether the relationships are meaningless or indicative of trends.

Page 7: *…it is significant for other aspects of the study that dead birds were detected on average farther from end turbines and at gaps. This result suggests that there is a systematic problem allocating carcasses to turbines (and more importantly to turbine type)*.

We disagree with this comments' premise. It, in our view, suggests turbines at ends of rows and at edges of gaps killed more birds. It seems premature to conclude there is a systematic problem with the study only because the reviewer saw consistent results indicating more birds are killed by end-of-row turbines.

Because turbines are often located fewer than 100 m from one another, this results in a smaller total area allocated to turbines on the interior of strings and a greater area allocated to those at the ends and at gaps.

We acknowledge there is likely a small difference in area.

End turbines are likely to be situated at the top of slopes (resulting in carcasses falling father away), which the authors use as an explanation for the increased distance to carcasses.

This is usually not true. Turbine strings tend to straddle hills while oriented along ridgelines, extending from the valley bottoms to the ridge crests and peaks, and sometimes down the other side, as well. End turbines tend to occur on steeper slopes lower down the hill relative to interior turbines. See the tables below.

	End of string		Edge of gap		String interior	
Topography	Obs	Exp	Obs	Exp	Obs	Exp
Ridge crest	146	224.4	82	55.1	749	697.5
Peak	26	18.8	3	4.6	53	58.5
Plateau	30	45.2	1	11.1	166	140.6
Ridgeline	544	574.5	112	141.0	1845	1785.5
Slope	354	238.7	62	58.6	623	741.8
Saddle	56	57.9	22	14.2	174	179.9
Ravine	9	5.5	4	1.4	11	17.1
Totals	1165	---	286	---	3621	---

	End of string		Edge of gap		String interior	
Slope grade (%)	Obs	Exp	Obs	Exp	Obs	Exp
0-1	216	294.6	78	84.7	1017	931.7
2-5	166	198.9	37	57.2	682	628.9
6-14	314	266.7	86	76.7	787	843.5
15-58	215	150.8	61	43.4	395	476.9

However, this pattern is not likely to hold for gap turbines, and the often steep ground ("precipices of very steep hills descending into ravines and canyons") will also result in fewer carcasses being detected.

There is no evidence carcasses were harder to detect on steeper terrain. We believe this comment is incorrect.

The most logical explanation is that the implementation of the survey protocol, including the inclusion of carcasses located beyond 50 m, resulted in a greater effective search area for end and gap turbines.

We disagree. The most logical explanation is that more birds are killed by end-of-row turbines. Orloff and Flannery (1992) found the same results, and so did Kerlinger et al. (2006) at the High Winds project in Solano County, California. Howell and Noone (1992) and Howell et al. (1991) also found the same result.

It is possible that this observed relationship is merely a result of the greater search area for end and gap turbines, especially because turbines are often spaced closer than 100 m within strings (see e.g., Figure 6-41).

It is possible the differential search area affected the results. We can certainly check, but we think the reviewer overestimates the effect.

This aspect of the methodology could jeopardize all of the turbine-level analyses in the report.

We fail to see the chain of logic leading to this statement. This aspect of the methodology could affect the tests for association between fatalities and turbine position in the string, but there is no basis for concluding the rest of the turbine-level analyses in the report are jeopardized by it.

P8: *Assuming that this pattern is real* [that carcass distance from wind turbines differed by season of the year, and were significantly shorter in spring]*, it suggests that detectability of carcasses differs by season. Because carcasses greater than 50 m from turbines were included only as observed from within the 50 m search radius, their inclusion increases the average distance of carcasses from the turbine. The authors should investigate whether carcasses from > 50 m caused this pattern. That would be logical, because vegetation is usually tallest in the spring in Mediterranean grasslands. If this pattern does result from detectability differences, it would underscore the need to account for detectability in the study design and to account for seasonal variation in search effort*.

It is true that grasses are taller during spring, but carcasses are found almost as easily during spring as during other times of the year. First, many carcasses fall on the dirt roads and dirt or gravel areas maintained around the turbines (Photo 5), as well as on tower pads (Photo 6), where grass height is not a factor. Second, carcasses falling on the grass create depressions that are easy to see (see Photo 7), and often there are feathers or other body parts that assist with carcass discovery. The exceptions would be in stands of mustard or thistle, but these stands are relatively few in the APWRA. Furthermore, whenever searchers encountered taller or thicker vegetation, they searched the area more intensively. This searcher adjustment to the environmental conditions is routine.

During the spring of 2003 we measured grass height between 20 and 40 m of 1,526 wind turbines. We found grass height was lower than 40 cm at 91% of the turbines, the mean was 25 cm, and the maximum recorded was 100 cm. These heights did not challenge the searchers' ability to detect carcasses. Furthermore, we checked our data to test whether the number of carcasses found beyond 50 m were homogenous to those found within 50 m as each related to season of the year. We found the two samples of carcasses were homogenous in their association with season ($\chi^2 = 2.23$, d.f. = 3, $P > 0.10$). We did not find evidence of a bias in finding carcasses farther away as a result of differences in grass height between seasons.

Photo 5. Golden eagle carcass found on dirt during late spring.

Photo 6. A western meadowlark carcass was found on the tower pad during spring.

Photo 7. A gull carcass found during spring depresses the grass.

P8: *Inclusion of these carcasses* [beyond 50 m from the turbines] *will result in a higher apparent mortality rate at those turbines where detectability is higher (e.g., vegetation is lower, slopes are not steep, etc.). Because information about detectability was not gathered, it is not possible to assess the effect of this bias*.

We found no evidence of a bias. The reviewer jumped to the conclusion there is a bias between carcasses found >50 m from turbines and vegetation height. There is no evidence to support this conclusion, and our check of the data (see previous response) refutes it.

P8: *We are not sure about mortality being expressed relative to megawatts (MW) of rated power generated per year. We can understand why authors chose to express mortality in these terms, but each type of turbine does not have the same relative effect on killing birds because of the inherent attributes of each type of turbine and we know that three different models seem to kill the most birds. Unless deaths per MW / year (or numbers of actual birds killed per year) can be clearly linked with "hours of rotating blades / year" for the particular type of turbine in question, the use of MW / year to associate with mortality may be misleading because rated power MWs do not kill birds, mechanical blades do*.

We agree the blades kill the birds, but unmoving blades do not. We believe the most resolute mortality metric would be the number of birds killed per kWh because kWh is the product of rotor diameter and the time the turbine actually operated. However, we were not given data on the operation time or power output of the turbines, so we used fatalities per MW per year. Even if we were given the opportunity to revise our report, we could not measure mortality in terms of kWh unless the wind turbine owners gave us data of power output per turbine. Until the turbine owners give researchers their power output data, the wind turbine's rated power output will remain as close as researchers will get to representing both the rotor diameter and the operation time of the turbine in expressing turbine-caused mortality.

Our invention of this metric is widely accepted by our peers as an improvement over simply stating mortality as birds killed per turbine, and over using rotor swept area as the metric. Over time, new and better metrics will be developed after wind turbine owners share their power output data, but for our purpose for making comparisons, the metric we used is applicable and useful.

P8: *The non-random sampling scheme does not support such extrapolation* [to the turbines that were not sampled].

We disagree. The 25% of the wind turbines that were not sampled were interspersed among the turbines that were sampled, and they consisted of the same turbine models and environmental conditions.

P8: *Several details are omitted from the Methods section that directly affect any judgment about the validity of the data collection methods. 1) P. 47 indicates the 1,526 turbines were sampled, but gives no specifics about how the sampled turbine were selected. The same criticism applies to the additional comments that note that other groups of turbines were added periodically. How were these selected for inclusion*?

All 1,526 wind turbines were searched for fatalities. We probably should not have used the word *sampled* in this case. Groups of turbines were added to Set 1 as they were made available to us, and when they were made available to us, we searched all of them.

2) *No mention is made of any efforts to prevent double counting on successive visits. Found carcasses were flagged, but no mention was made whether that flagging was permanent*

throughout the course of the study.

Double counting carcasses was never an issue. First, bird carcasses were removed by Greenridge Services, LLC. Second, bird carcasses left in the field were easily identified as carcasses previously found, and there were some found a second time. We could tell whether the carcass had been seen before based on location, decomposition level, body parts and carcass condition described on the data sheet, photos, and flags left near the carcasses for this very reason.

3) *No discussion is presented about the search sequence. Were strings searched in the same order throughout the rotation*?

Not true. On page 48 we wrote, "*With two to three people searching 120–150 wind turbines per week, 685 turbines could be sampled once every five to six weeks, thus completing approximately eight fatality search cycles in 12 months during 1998 through 1999, when we were limited to 685 turbines. Not all turbine strings were searched every month due to changes in field strategies or for reasons out of our control, such as fire hazards and flooded roads. As we were allowed to search around additional wind turbines, our search rotations took longer and our frequency of searches per year declined*." Maybe our description of the search interval could have been more thorough. Generally, the turbines were searched in the same sequence each search interval.

P9: *The authors acknowledge the disparity in searches for dead birds between the time periods (March 1998–Sept 2002 with 3 or 4 years for each month around 1,526 turbines) vs. (November 2002–May 2003 with only 1 search per month around 2,548 turbines) and they note that all turbine strings were searched every month*.

This description of our search effort is inaccurate.

How would this difference in effort affect the reliability of estimates of mortalities?

We do not know, but almost certainly not to an extent that would have fundamentally changed any of our conclusions or recommendations on how to reduce mortality.

Are these mortality estimates more conservative than if the same effort for the first group of turbines had been applied to the larger second group?

The question implies a hypothetical situation that we cannot answer or speculate about.

And did the types of turbines in the strings differ between the two sampling periods?

Mostly no, but there were some minor differences.

The sampling unit is described as the turbine string. Are turbine strings most always composed of the same type of turbine?

Yes.

P9: *Given the range of search effort per turbine per year (Figure 3-1), fatality estimates should be corrected upwards to adjust estimates for turbines searched less frequently. Authors assume*

that the same number of fatalities would have been found during a given year regardless of whether twelve searches or eight searches were performed. They acknowledge that fewer carcasses would be detected at turbine strings with fewer searches but do not adjust for this factor. What supports the assumption that the influence of search effort on carcass detection would not affect the subsequent analysis?

The answer to the last question was because we were focused on raptors, of which the carcasses of large-bodied species typically remain in the environment for months. For other birds, and even for large-bodied raptors, we agree with the reviewer that turbine strings searched fewer times during a year will likely yield lower mortality estimates. At the time we wrote the report, we had no means available to adjust our estimates accordingly, but Attachment B includes a tool kit we can use to make the adjustment, if needed.

P9: *Searcher detection and scavenger removal rates are not inconsequential to the results of the analyses as implied by the authors*.

The premise to the comment is misleading/wrong. We never said nor implied they were inconsequential. However, we do not believe searcher detection error is high in the APWRA, at least not during our study. Scavenger removal rates can be substantial, but we directed our research funds to finding carcasses, as we stated. We knew the funding required to perform an adequate scavenger removal study would rival our overall budget for the reasons given in Attachment B. We were unwilling to perform a scavenger removal study in the same manner others had been performing them. We need directed research on this issue.

P9: *Indeed, there could be massive scavenger losses, especially of small birds, even at the average 50-day period between searches*.

We agree. This is one reason our uncertainty ranges were so large around our mortality estimates of small-bodied bird species. Our uncertainty ranges were our scientific statements of low confidence in the mortality estimates, which is standard practice.

P10: *This adjustment* [of mortality estimates by scavenger removal rates] *results in simply inflating fatalities by a constant rate, but it does not incorporate the differences across space and time that certainly exist*.

It is our practice to withhold conclusions about how the world works until some evidence can be used to support the conclusion. The reviewer cannot know for certain that scavenging rates vary across the APWRA, at least not until defensible scavenging trials have been performed.

P10: *This adjustment therefore does nothing to counteract the nonrandom influence of vegetation on detection and scavenging rates, or on observer detection ability to the extent that observers were not assigned to survey routes randomly*.

As stated elsewhere, the APWRA is almost uniform in its annual grass cover. The reviewer appears to be guessing about the environmental conditions in the APWRA, and how they relate to scavenger removal rates and searcher detection error.

P10: *What was the author's experience that led them to believe that the scavenger removal rates*

were inaccurate for raptors?

Our 4.5 years of experience in the APWRA, following several years of experience in the APWRA by Orloff and Flannery (1992, 1996). Also, see Attachment B.

P10: *Figure 3-15 shows spatial distribution of survey effort. This figure does not appear to show a random sample. The authors should provide statistics about how the surveyed towers differ from the non-surveyed towers in key parameters (tower type, topography, elevation, turbine manufacturer, etc.). The non-random search pattern may influence other results. For example, elsewhere the authors report results for turbines that were searched four years without highlighting how the characteristics of those turbines differ from non-sampled turbines (e.g., turbine type, elevation, landscape position, etc.).*

We did not collect any data on the wind turbines that we did not search for fatalities, but they were generally like those we did search.

P10: *Can some of this result be attributed to the increased familiarity of the investigators with the study areas, especially when areas were studied for 4 years*?

By the second search, we were pretty familiar with each search area. They look similar from place to place. We do not believe an increase in familiarity affected fatality counts.

P10: *The right column has only turbines searched for 4 years. This is a geographically clustered sample, so it is unclear how results can be compared to the other turbines or to all other turbines at APWRA. The authors disclose that these turbines were within areas of rodent control, but do not describe the other differences from the other sampled turbines or the unsampled turbines*.

We agree this sample is geographically clustered, but it is also spread throughout much of the APWRA. It includes the same or similar mix of wind turbine models as found elsewhere in the APWRA. We only pointed out the rodent control efforts among these turbines because it was the one condition that was common to this entire sample of turbines, and like we pointed out, could be a confounding factor.

P10: *This table* [3-9] *shows mortality per turbine string for two sets of turbines searched for different time periods. Because neither sample is random, and years of data are pooled (rather than comparing data from one year at one set to the same year at the other set), it is not obvious how the reader is to interpret this information*.

We left it up to the reader to interpret the table. Sufficient information is provided to form a conclusion. The reader can examine the uncertainty ranges and decide how big a range is too big, and how small a range qualifies the estimate as sufficiently reliable. We simply presented as much of the information that we could, and that we thought was relevant, and we left it up to the reader to decide how far along our sequence of steps to agree with us. That neither sample was random makes little difference to us because one sample included all the turbines made available to us, and the other included 72% of the available sample, which was selected systematically in the absence of forehand knowledge of which strings would kill more birds. There were 3 differences between the samples we cared more about, including different periods of time the

two sets were searched, different periods of time between the searches at each set, and the fact the second set was only searched through half a year.

P10: *It would be of interest to know how many deaths by species per year were associated with the total sum of "hours of operation / year" of all turbines and for each type of turbine in these two groups. Were there about equal proportions of each turbine type in each of these two groups?*

We agree with the statement about the need for data on turbine operations, but we were not given these data. Most of the turbines were of common types between the two groups, but there were some differences, as well. However, like we reported, environmental variables appeared to be more significant to fatalities than turbine type.

P11: *Because information like this* (continued from previous comment) *is lacking, it is difficult to draw any conclusions from these data*.

Whereas we would have preferred the turbine operations data, we still believe our mortality estimates are useful for multiple reasons. One reason our estimates are useful is because they help identify shortfalls in current research practices, as well as means to surmount those shortfalls. Also, some of the estimates are more reliable than others.

P11: *This table* [3-12] *provides results on a "per turbine" basis but the sampling unit was a string of turbines*.

Turbine strings are composed of wind turbines, so expressing the number of fatalities per turbine is simply a matter of dividing the number of fatalities by the number of turbines in the string.

P11: *We are sympathetic in that the wind turbine operators did not allow access to turbines uniformly so that designing a random sampling scheme was difficult, if not impossible. This remains, however, a shortcoming of the study. The authors should have restricted all comparisons of mortality rates to turbines that were sampled during the same period and within a random sampling framework*.

We disagree. Again, the purpose of random sampling is to avoid investigator bias. We sampled all the turbines in one set, and 72% of the turbines in a second set, the latter of which was sampled systematically and naïvely with respect to likelihood of turbines to kill birds. Sampling frameworks do not have to always be random to be valid, and they often are not random in field studies.

P11: *Herein the authors attempt to compare mortality rates at wind farms as determined in different studies. The authors make many assumptions about scavenging, detection, and search interval that cannot be verified*.

Verification is simply a matter of collecting the reported results from available research reports and comparing them. If you get the same or similar results as us, then our results are verified.

P11: *In the regressions of raptor fatalities by birds observed per hour it seems that most of the explanatory power comes from the current study and its precursor at Altamont pass.*

Furthermore, the two high fatality estimates constitute partial duplication of the same data, because it seems that the data from Thelander and Rugge are incorporated into Smallwood and Thelander.

The only duplication of data in the regression would be between Thelander and Rugge (2000) and Smallwood and Thelander (2004).

P11: *We are not convinced that the mortality rates from the different studies can be compared. Furthermore, the use of "bird observations" as a metric is not particularly useful because it is already apparent from the data that avian species are not all equally vulnerable to collision with turbines*.

We, too, are not convinced that fatality rates from different studies can be compared. See Attachment B for a discussion on the biases caused by comparing utilization estimates between studies using different search areas. There are other biases, as well. Our use of bird observations as a metric was more for its use as an indicator. We believe now, however, that we can do better.

P12: *Aspects in this chapter follow our component framework #3. Even without operating the turbines, their establishment modifies the local environment by changing the food base that may affect the behavior of birds and cause some low-level mortality. The effect on behavior, in turn, may predispose birds to be hit by turbines when they are operating and cause higher levels of avian mortality*.

We agree with this statement, and we are able to agree because our research in the APWRA led us to this conclusion.

P12: *Unfortunately, the amounts of lateral edge and vertical edge were characterized as "some", or "lots". If we understand the layout correctly, these variables could have been quantified in terms of x meters of lateral or vertical edge.*

We expressed edge as an index, which is why it was called the "edge index." We disagree it should have been measured precisely because measurement of such variables for comparison to a statistically rare event can result in false precision.

P12: *Also, please describe the difference between ridge crests and ridgelines. Where these topographic classifications made with automated Geographic Information System tools or based on judgments in the field or another technique*?

Ridgelines are ridge features that continue downslope toward the valley or stream, whereas ridge crest is the ridge feature at the prominence of the hill. Ridgelines and ridge crests were designated in the field by either Smallwood or Brian Karas. Since the report was completed, GIS tools have been developed by Lee Neher to identify ridge crests and ridgelines in the APWRA.

P12: *How did the authors determine that cottontail pellets were especially abundant? Random transects? Is there any citation or precedence that connects rabbit pellets with abundance? Fecal abundance as an index of animal abundance is not always reliable*.

On page 90 we explained how the transect was arranged, and we explained that every wind

turbine available to us in Set 1 was indexed for cottontail pellet abundance. In other words, we did not sample randomly because we visited *all* the turbines that were available. For this effort, a single observer coordinated with Smallwood to subjectively apply this Index. Whereas presence of pellets might consist of one or two pellets, and exceptional number could be pellets scattered all over the strip transect the entire extent of the transect segment corresponding with the turbine. It was just an index. And we assumed if a reader did not like it for some reason, he can ignore the results of the subsequent association test.

P12: *How did the authors choose the 571 turbines to map rodent burrows? This should be described in terms of turbine strings because strings are still the sampling unit. The choice of turbine stings appears to have been arbitrary, perhaps guided by an idea of a stratified random sample of turbine strings associated with different raptor mortality, physiography, and rodent control. If the sample was, indeed, a stratified random sample this should be stated clearly with a description of how many replicates of turbine strings were associated with the three criteria (i.e., range of raptor mortality, physiographic conditions, and level of rodent control). If not, then the method for choosing these turbine strings should be clearly described*.

In the first paragraph under 6.2 Methods (page 111), we wrote: "*We mapped rodent burrows near 571 wind turbines, composing 70 strings of wind turbines in the APWRA. Most wind turbine strings were selected arbitrarily, to represent a wide range of raptor mortality recorded during our fatality searches, as well as to represent a variety of physiographic conditions and levels of rodent control*." A random sample likely would not have obtained the same range of conditions we felt we needed to learn how rodent burrow distributions are affected by rodent control efforts, and about how they relate to raptor fatalities. A stratified random sample would have been superior to a random sample, and had we been granted access to all the turbines and turbine strings at the beginning of our study, and had we known in advance about the rodent control program, then we would have selected turbines for rodent burrow mapping as a stratified sample.

P12: *Were these strings selected randomly? The numbers (and types?) of turbines in the strings were widely variable ranging from 3–35 turbines, and 1 to 3 or 4 strings per group. How comparable were these groups*?

We recall discussing whether we should select these strings randomly, but we cannot verify that we did select them randomly. They may have been selected randomly from the available pool of turbine strings at the time, i.e., from those we were given access. We did not compare rodent burrow distributions among the groups of turbines, so the question about how comparable they were is irrelevant. We were in interested in seasonal differences. On page 124 we wrote, "*Eleven strings of wind turbines were selected for seasonal monitoring purposes…*"

P13: *This photo* [Figure 6-25] *suggests that type of tower, at least, was not uniform within groups of strings. Tower type seems important; did inclusion of different types of towers have any effect on results*?

The 3 large wind turbines in this photo never operated during our study. They were derelicts. Most strings of wind turbines were uniform in tower type, but a few were not.

P13: *Did the type of turbine have any measurable effect*?

Not in this case.

P13: *Again it seems important to recognize the large disparities in numbers of turbines (and perhaps types of turbines?) among these sites. Is it possible that unadjusted mortality of species is related to number and type of turbines and not rodent control treatments? Is there no way to test for interaction effects*?

Actually, all the turbines in all but one group were either Micon or Bunus turbines on tubular towers of the same height to the rotor hub, so they were almost all the same type of turbine and tower. The exception was the group of Vestas 100-kW turbines on lattice towers in Figure 6-44. However, why would rodents care about the turbine type? We do not think variation in turbine type confounded our results in this case because tower type and turbine model are irrelevant to ground squirrels and pocket gophers. The tower bases are more important, and in this case all the tower bases were the same type and size of concrete pad.

P13: *"Search effort" is defined as m^2 times number of years during which surveys were made. How and when, does the amount of time spent on transects looking for carcasses (or number of visits per year) factor in*?

It did not factor in. Search effort was defined on pages 185 and 186.

P13: *Doesn't fatality rate imply deaths per unit of time? Not unit of area*?

It can be either, but in our application we defined it as per unit of MW per unit time.

P13: *And even more appropriate may be to express as deaths per hours of turbine operation, because flying into moving turbine blades is the primary cause of bird deaths*.

We agree, but we did not have the turbine operation data. We asked for them, but we did not get them.

P13: *The predictive model is flawed. The variables examined are clearly not independent and so summing the accountable mortality values across variables (p. 188) must necessarily overestimate the predicted impact. All model results are suspect because of this flaw. Furthermore, this is a complex study with many potential confounding factors, yet the development of the predictive model strikes us as simplistic and fails to account for such effects*.

Whereas we agree many of the variables initially tested for association with fatalities were "clearly not independent," we disagree the predictive models were flawed for that reason. The predictive models included one variable, and occasionally two or three variables, from each factor, which was identified by this reviewer as "primary components" earlier in the review. On page 222 we wrote, "*Turbine size (i.e., power output) was not used because it correlated strongly with other turbine attributes that were already used*." We should have elaborated on this point, because our additional elaboration would have helped explain the combinations of variables appearing in the models. Table 7-9, for example, includes only one wind turbine attribute, even though six turbine attributes were significantly associated with golden eagle fatalities. We made

an effort to choose variables that were highly significant in their associations with fatalities, showed clear gradients of fatalities relative to the variable, and were as orthogonal as possible from other variables.

P13: *It is not always evident what the figure caption "count" means in these figures. It seems to be number of turbines, mostly. Is that correct*?

Correct. What was counted was specified in the caption of each figure.

P13: *It also seems that the words "search effort" are used in captions for measurements that is really the number of years during which searches were made, multiplied by a search area. This measurement ignores the number of visits (or hours) that each area was searched and assumes that there would be no variation in the number of dead birds found with greater or fewer visits during a year*.

Correct.

P13: *It is also peculiar that in this analysis, the authors use rotor swept area as a measurement of turbine size, rather than MW rating. We prefer the rotor swept area as a method of standardization*.

We do not understand why our use of rotor swept area is peculiar in this analysis. Its use better fit the objectives of this analysis. Its use in estimating mortality was a problem because we would not under stand how many times the rotor swept area was actually swept. MW of rated capacity did not inform the metric of actual operation time, either, but at least it took us all one step closer to using the metric that is clearly superior, and that is the number of fatalities per kWh.

P14: *On what basis were the two groups "Hawks" and "Raptors" segregated*?

Hawks is a subgroup of raptors.

P14: *These figures are all misleading. The adjacent bars suggest direct comparisons, yet the opposing scales are not comparable. As an example, in Figure 7-8 (above) the left scale (count) maximum is 3,000, which is 74% (=3000/4675) of the total number of turbines, whereas the right hand scale maximum is 60%. This imbalance of scales makes the effort bars taller than they ought to be*.

We disagree, and this is the first time we have encountered this complaint about these bar charts. The bars are color coded and labeled, and the reader need only compare the red bars to the red bars, and the blue bars to the blue bars. The combination of the bars was intended to save space.

P14: *The results for a large number of the Chi-squared tests in Tables 7-1 through 7-3 that are suspect because too many of the expected values for individual categories presented in Appendix C are less than 5. The authors mention this fact on p. 206 but present the tests anyway. The test ought not to have been done*.

We are mystified by this comment. We provided guidance on how to interpret the chi-square

tests, including the rule of thumb to use caution when interpreting tests with a threshold percentage of expected cell values <5. Then we summarized the test results in Tables 7-1 through 7-3, and provided all details of each test in the appendices so that the readers can decide for themselves how to interpret the test results. We did not force anyone to accept our results, and we went out of our way to provide the readers the means to assess the tests for themselves. In the meantime, we were attentive to the number of expected cell values <5 when we synthesized our test results. Table 7-8 identifies significant test results composed mostly of expected cell values >5, and these test results were our candidates for development of the predictive models. We provided each of our steps in developing our models, but the reviewer appears to have mistakenly concluded our step one resulted in the models. This was not the case. We provided step one as a service to the reader, and nothing more.

P14: *The individual turbines within the same string are not independent and just as in the ANOVAs this fact needs to be accounted for in the Chi-squared analyses. In the analysis of seasonal differences the repeat visits are not independent and that needs to be accounted for also*.

If we follow the logic of these two criticisms, then no events we might count in our universe are independent. And at certain levels of thinking about independence of observations, we can agree that no observations are truly independent. However, scientists are more practical than this ideal view of the universe, and are willing to overlook small degrees of non-independence. We were expected to analyze our data, which is what we did. Claiming non-independence between turbines and within seasons should justify no tests would also mean that no tests should be performed at all, which we believe unreasonable.

P14: *What was the rationale for excluding turbine model from the tests*?

Turbine attributes are more interesting to the bird collision issue than turbine model because (1) the models incorporate suites of attributes; (2) some turbines in the APWRA were hybrids, or combinations of parts from different models; and (3) turbine models keep changing in the APWRA whereas turbine attributes merely change values as turbines are replaced. Furthermore, turbine models often shared turbine attributes, so relying only on attributes could reduce the number of attribute values used in a test for association. Also, some turbine models were so scarce that their inclusion in the tests for association were less useful than was lumping their attributes in with the attributes of other turbine models, at least whenever their attributes were shared by other models.

P14: *The conclusion about rock piles does not seem to be adjusted for different mortality rates in different years, and for all the other factors that differ between the samples*?

The main difference between the samples was fewer artificial rock piles and more natural rock piles among the Set 2 turbines. On page 244 we wrote, "*The presence of rock piles was only significant for the original set of 1,526 wind turbines we sampled. Wind turbines with these rock piles nearby killed more raptors, and disproportionately more western meadowlarks and horned larks. The addition of 2,548 wind turbines in 2002-03 to our sample changed the association test results involving rock piles. We noted during field studies that the areas where we added wind*

turbines included many natural rock piles and rocky outcrops, which likely provided many opportunities for raptor prey species to find refuge and for bird species to perch upon."

P14: *It is not clear if a subset of data was withheld from the data that was used to develop the empirical models so that they could be validated.*

We did not withhold data for model validation.

P14: *On p. 222, the authors ask the reader to assume "our predictive model are relatively precise" yet provide no justification for the assumption. The authors appear to be ignoring the possibility of false positive predictions.*

This conclusion makes no sense because we provided no indication we ignored the possibility of false positive predictions. As to our assumption the models are relatively precise, that statement was not a request of the readers to accept our assumption the models are precise. We were simply conditioning the conclusion that followed. In other words, we were not telling the reader that future fatality searches absolutely will add additional turbines to the "more dangerous" category, but rather they likely would if our models are any good. Real fatality searches can test whether more turbines will be identified as dangerous much more effectively than applying hypothetical rates to Bayes Theorem.

P15: *A calculation, using Bayes Theorem, can be used to answer the question, what is the likelihood that more searches would "add many more wind turbines to the pool of wind turbines documented to have actually killed members of each species?" (p. 222, line 4). To perform the calculation, one must assume an average fatality rate. Here is a table of hypothetical fatality rates for and corresponding likelihoods that a "dangerous" turbine will be found to have killed one or more Golden Eagles.*

Fatality rate	*0.001*	*0.01*	*0.05*	*0.1*	*0.25*	*0.5*
Likelihood	*0.002*	*0.016*	*0.079*	*0.153*	*0.352*	*0.619*

To interpret this table, consider this example: With an average fatality rate of 5% (.05 in the table), prior to applying the predictive model one would expect about 5% of turbine searches to produce a Golden Eagle fatality. If the searches were restricted to "dangerous" turbines (as identified by the model) then one would expect to find Golden Eagle fatalities in 8% (.079 in the table) of the searches. Thus, the model increases the chance of finding Golden Eagle fatalities from 5% to 8%. Thus, one can conclude about 92% of the turbines identified as "dangerous to Golden Eagles" will **not** *have an associated fatality and hence would not be "dangerous".*

At face value there is a problem with the reviewer's suggestion that more searches would barely matter to the number of turbines identified as "more dangerous." More than 2,500 turbines were searched only twice, so there is much potential to find additional golden eagles at wind turbines where we did not find them before. Finding these additional golden eagles would likely increase our understanding of the factors related to golden eagle collisions, and the predictive models would change accordingly.

We performed 20,804 fatality searches among turbines predicted by our model to be less threatening to golden eagle. This effort produced 4 golden eagles used in our mortality

estimates, for a rate of 0.190 golden eagles per 1,000 fatality searches. We performed 11,678 fatality searches among turbines predicted by our model to be more threatening to golden eagle. This effort produced 21 golden eagles used in our mortality estimates, for a rate of 1.798 golden eagles per 1,000 fatality searches. Searches at the turbines predicted to be more threatening to eagles turned up 9.45 times more eagle carcasses than those at turbines predicted to be less threatening to eagles. We conclude, therefore, that future searches restricted to the turbines predicted to be more dangerous would generate about 8 to 9 times more eagles per 1,000 searches, after allowing for some inflation of the model's performance due to the post-hoc nature of our model assessment.

Whereas we agree we would not actually find golden eagle carcasses at all, or even most, of the turbines predicted by our model to be more dangerous, it is preposterous to conclude turbines are not dangerous to golden eagles simply because we did not find their carcasses there. This is like saying freeways are not dangerous to pedestrians simply because a few surveys for dead pedestrians were negative.

P15: *The authors appear to be selective about inclusion of "important" variables. Using the Golden Eagle as an example. The variable 'Part of wind wall' ($p < .05$) yet the variable 'Tower height' ($p < .05$) was not. The accountable mortality for 'Position in string' was reported in the table as 19 while in Appendix C it is given as -18. There were several other similar occurrences with other variables in the list.*

Yes, we were selective, which is a point that the reviewer appears to have misunderstood when alleging our models were hopelessly confounded and flawed by multicollinearity. On page 222 we wrote, "*Table 7-8 summarizes the associations between variables and species that were most reliable for use in model development. Some variables were not used for model development because doing so would be nonsensical from an ecological standpoint. For example, season of the year did not fit into models built around all of the data, including from all four seasons, considered together. Other variables not entered into the models included perch deterrent and blade color schemes. Turbine size (i.e., power output) was not used because it correlated strongly with other turbine attributes that were already used*." In other words, we screened the association test results for inclusion in the model. The variables we used had to be highly significant in their associations with fatalities, show clear gradients in their relation to fatalities, include few expected cell values <5m, and had to make sense.

P15: *Because the physical attributes of operating turbines manifest the lethal force in bird deaths, it may have been instructive to use only those variables identified in framework component #1 to develop a predictive model with AIC methods. Similarly the same approach may be applied to the other framework components as outlined at the beginning of this review to determine which variable(s) contributed to bird mortalities. From results of the four predictive models, perhaps an overall model could be developed that used the most important variable(s) from each component model.*

Essentially, the reviewer described the screening approach we used, except we did not use AIC. We believe AIC was only just emerging as a tool among wildlife biologists in 2003, which is when we wrote most of the report. However, we are unsure we would have used AIC, even had we been familiar with it at the time.

P15: *The authors conclude that "dangerous" turbines are distributed "relatively narrowly" across the APWRA. The distribution in the maps does not seem narrow to us (see Figure 7-27 above for red-tailed hawks).*

The "dangerous" turbines to red-tailed hawk appear relatively narrow to us. Almost all of these turbines are at the ends of strings and the edges of turbine fields. Unfortunately, the map does not depict the locations of canyons, but our knowledge of where they are also contributes to our conclusion the distribution of dangerous turbines was relatively narrow. Perhaps the reviewer is being arbitrarily narrow in interpreting our use of the word *narrow*.

P16: *What is spatial distribution of 61 study plots? It seems that they are associated with turbine strings that were chosen arbitrarily, meaning that the behavioral study plots were not selected randomly. Consequently, behaviors from these plots cannot be extrapolated to other areas within the APWRA.*

On page 246 we wrote, "*These 61 plots covered all of the area studied during the behavior research performed under funding from the National Renewable Energy Laboratory (Smallwood and Thelander, in review)*." In other words, no random selection was warranted because our behavior study encompassed the entirety of the Set 1 wind turbine study area. Neither was the selection arbitrary; we included in each plot the area and the turbines we could see from each observation point, but we achieved complete coverage of the study area. Also, our objective was not to extrapolate our behavior observations to the rest of the APWRA.

P16: *The analysis of a measured variable, such as minutes, using a Chi-squared analysis is invalid. The Chi-squared tests are not invariant to changes of scale, i.e. the results would change if the data were expressed in seconds or in hours. (Using seconds would make the tests more significant and using hours would make the tests less significant.) This invalidates almost all of the tests performed here.*

These are good points, though we disagree with the ultimate conclusion. We used on-the-minute instantaneous sampling, instead of on-the-second sampling, in order to reduce dependence between observations, and using on-the-hour observations was entirely impractical. We were aware of the problem of dependence between observations, but also cognizant of the choice each bird can make while flying or perching or performing any behavior. Each animal chooses its behavior, and has this choice all the time. A bird seen hovering one minute may or may not be hovering the next minute, and it probably will not be hovering every minute the observer is conducting a sampling session. We were interested in learning whether certain species hover disproportionately more often under certain conditions, and we believe we learned where they do despite lack of complete independence between observations.

We noticed the reviewer suggested no alternative to the approach we used.

P16: *On p.254 the authors indicate the observed values used in the Chi-squared tests were either minutes or behavioral events. It seems that no Chi-squared tests were based on behavioral events.*

Not true. There were many chi-square tests of events and of flight attributes that were not numbers of minutes, summarized in Table 8-8, 8-9, 8-11, 8-14, and 8-15.

P16: *Turbine level analysis involves pseudo-replication because turbines were sampled as strings.*

This is a fallacious argument. Just because we searched for fatalities one string at a time does not mean we could not relate our behavior observations to plots, to strings, or to individual turbines. The reviewer applied a false data structure to our analysis.

P16: *Our field experiences reinforce the author's conclusion that the observation time for the sessions was minimal at 30 minutes. Other observational studies with which we are familiar found that 2-hour blocks, randomly assigned throughout the entire period available to observe birds, were adequate to determine reliable patterns of bird activity.*

It would be helpful if the reviewers would cite their sources. We need to be able to assess the validity of the reviewers' conclusions, and besides, citing sources contributes to the constructive intent of reviews.

P16: *We agree with the authors that BACI study designs will be required to sort out effects of some variables, because the mere presence of the turbines as they are installed affects the environment and in turn affects bird behavior, which is a variable related to mortalities.*

And how would the reviewers come to this conclusion if it was not for our studies of range conditions, fossorial mammal distributions, bird behaviors, and fatalities? We point this out because the review started out by stating our conclusions are premature, and it restated this conclusion on the very next page.

P17: *The authors argue for taller turbines to repower at APWRA, but they seem not to consider how this will influence mortality rates for migratory songbirds. Turbines greater than 200 feet will require obstruction lighting, which is associated with increased mortality of nocturnally migrating birds.*

Lighting has been an issue on communication towers, but not conclusively on wind turbines, and not on the west coast. Furthermore, we acknowledged in our report there could be surprises following repoweirng. Following the first year of fatality monitoring at Diablo Winds in the APWRA, however, it appears our recommendation for repowering was sound. The Diablo Winds repowering project appears to have reduced avian mortality 70% (Smallwood 2006, Attachment B).

P17: *The term "confusion" may be correct but the term "complexity" also depicts the situation. It may be that it is inappropriate to try to compare mortalities between wind generating facilities because each facility has unique features for each of the four framework components, thereby preventing any reliable comparison between facilities. Conversely, the individual turbine type (and its attributes) is of utmost importance in how many birds are killed. Variables of the other three framework components (that we outlined at the beginning) can be neutral or either increase or perhaps decrease the predisposition of birds to being killed by the turbines. But, each wind farm site is unique with specific effects of variables that cannot be fully replicated.*

We agree the reviewers might ultimately be correct that mortality ought not be compared among wind farms, but the conclusions used to make the argument lack evidence. The reviewer appears to have decided turbine type is the most important factor related to bird deaths, but we wonder how that conclusion was reached.

P17: *This statement reinforces our earlier comments (See comments page 46, Chapter 3) that attempting to standardize by basing number of fatalities/ MW/ year instead of number of fatalities/ turbine/ year does not provide insights about effects of individual turbine types, which is the killing structure. Actually, Figure A-3b, Page A-5 depicts an even more direct metric of what kills birds — the area of rotor-swept / year, which again relates to turbine type, size and blade speed*.

We strongly disagree that the metric, fatalities per turbine per year, is more instructive than fatalities per MW per year, and the other researchers of bird collisions agree with us on this point. It is senseless to compare fatalities per turbine per year when the turbines can vary from 40 kW to 2.5 MW. Also, this is not the suggestion the reviewer made earlier. The reviewer recommended fatalities per rotor-swept area per year. We looked at this latter metric, but we did not like the non-linear relationship between the metric and rotor-swept area. We believe a superior metric will be fatalities per kWh.

P17: *Another reason to question the use of fatalities/ MW/ year is that the MW is a constant (as stated), but that the number of fatalities is variable over time and depends on amount of search effort, so that inadequate search effort in a given year will weaken the reliability of results. Authors further acknowledge (Page A-7) that this is likely that areas around wind turbines that were not searched over a long enough period will not provide a robust estimate of mortality*.

We agree, which is why we recommended using fatalities per kWh in the future.

P17: *Indeed, it may be more convenient to express mortalities on the basis of MW / year, but information on which type of turbine and supporting structure that kills birds is not emphasized*.

Now the reviewer <u>agrees</u> with us about the metric we used.

REVIEWER TEAM 2

P4: *Almost certainly, at least some of the many variables measured are truly linked to bird mortality – birds are certainly being killed in the APWRA. The reviewers have little confidence, however, that this report has scientifically been able to determine which of those variables are important.*

We never claimed to "determine" causes of fatalities, so this comment is misleading and unfair. Right in the introduction we state our aim to identify "possible" relationships between bird mortality and bird behaviors, tower designs and environmental variables. In our report we described how we carefully processed the many tests that we reported, leading to the strongest associations. Nowhere in the report did we claim our study was definitive or that additional research was unneeded.

P4: *The statistical analyses are applied in an automated manner that fails to fully utilize the*

data at hand and ignores potential confounding of variables.

We warned of possible confounding on pages 67, 108, 183, 246, and 353. On page 353 we wrote, "*We had little to no control over the replication and interspersion of treatments, including control treatments. Thus, our results were prone to inflation of measured effects and to confounding*." On page 2 of App. B we wrote "*Our study also was prone to confounding due to a gradient in rodent control intensity across the APWRA, but much less so than was Kerlinger and Curry's study. Our samples within areas of no rodent control were more interspersed within the other rodent control treatment intensities, and our samples within areas of intense control were also more interspersed with the other treatments. Had we been granted access to all the wind turbines earlier during the study, we could have achieved a much greater degree of treatment interspersion*." We did not ignore potential confounding.

P4: *It seems like many of the statistics were calculated just for the purpose of producing statistical tables to the point of data dredging*.

We were contracted by the CEC to explore the data for patterns and relationships useful to recommending mitigation measures. Had we selectively performed tests, we would have been accused of being selective. We were as thorough as we could be, but we also screened our tests and pared them down to those most significant, most orthogonal, and most meaningful. We disagree this approach should be characterized as data dredging.

P4: *...the mathematical assumptions behind statistical tests like one-way ANOVA are ignored and thus the reported P-values should be treated as approximations*.

We disagree we ignored the assumptions of ANOVA, but we agree the P-values should be treated as approximations. We do not believe we gave any indication in the report that we regarded the P-values as anything other than approximations. This is why the reader will find our use of the term "indicated" in association with many of the ANOVA tests, and supporting LSD tests.

P4: *The large number of statistical tests likely resulted in many Type I errors; therefore, statistically significant findings should be treated more as an indicator of what should be explored in future studies*.

Whereas we agree with the recommendation, we believe the reviewer overestimates the number of Type I errors likely committed. On page 1 of this review, the reviewer explained how a cut-off level of 0.05 would result in 1 out of 20 tests being the result of a Type I error. We agree with the example, but many of our tests resulted in P-values much smaller than 0.05, so many fewer than 1 out of 20 significant tests resulted from Type I error. Among our many tests significant at the 0.005-level, we can expect about 1 out of 200 tests to have been the result of Type I error, meaning about 199 out of 200 tests likely were not the result of Type I error. We can accept these odds.

- P4: *Was the statistical methodology used on the analysis consistent with accepted methods used in other biostatistical analyses?*

No. A very large number (>1000) of univariate chi-square tests is not common in biostatistical analyses. Interpretations of the univariate tests are clouded somewhat by shared variation among the explanatory variables (turbine attributes).

We agree the reporting of so many univariate tests may be considered by some to be somewhat unusual in biostatistical analysis, but the tests we used are the oldest and most widely accepted tests available. The chi-square tests we used predate ANOVA by at least 30 years, and were developed in the nineteenth century.

P4: *Chi-square analysis assumes that the counts are exact and not estimated counts as they are in this study. It is not clear how this would influence the conclusions reached on the numerous chi-square hypothesis tests.*

We cannot understand this statement. Which of our counts were estimated?

P4: *In estimating mortality rates for specific species due to wind turbine collisions, almost half (28) of the 60 species or groups have fewer than 5 fatalities reported in the entire project. And yet, mortality rates are still estimated and reported.*

Yes, and with appropriately large uncertainty ranges.

P4: *Although the study design is observational, the authors quickly jump to hypothesis testing and parametric analysis without exploring their datasets thoroughly. What distinguishes the 20% of the turbines where fatalities were discovered from the 80% without fatalities?*

Answering the reviewer's question was the very point of our hypothesis-testing. Earlier on this same page of the comment letter, the reviewer accuses us of data dredging, and now we are accused of not exploring our data thoroughly. Which statement applies?

P5: *The authors explain that limitations in their sampling precluded these more sophisticated multivariate analyses, but this may not be true if the authors (a) carefully screen their variables to reduce the number of parameters in their models, and/or (b) clearly restrict their inferences to the turbines actually sampled.*

Not only did we carefully screen our variables for model development, but we also restricted our inferences to the turbines sampled, which we referred to in the report as the "measured set." However, we chose not use multivariate analysis for the reasons given. Were we to revise the report, we would try logistic regression or a general linear model, but not Poisson regression or discriminant function analysis.

P5: *The authors used standard protocols for carcass searches, bird observations, rodent surveys, etc. to obtain the ecological data, and generally the technical approaches were appropriate.*

Whereas we agree our technical approaches were generally appropriate, we disagree we used standard protocols for bird observations and rodent surveys. **There are no sampling protocols for rodent surveys and bird observations in wind farms.** In fact, published guidelines suggest

that site specific conditions be used to determine such protocols. Much more research will be needed on these types of methodologies applied to wind farms before protocols are appropriate, and we believe we provided some of this needed research.

P5: *However, the methods used to estimate bird mortality rate are suspect because (a) neither scavenging rate nor observer detection probabilities were measured empirically, values were pulled from the literature – in some cases based on studies in different locations*;

After reviewing the available reports of scavenger removal and searcher detection trials, we are all the more relieved to have not performed these trials ourselves in the APWRA. Investigators had been copying each others' methods and making little real progress. Smallwood's recent review revealed biases that need to be addressed with directed research (Attachment B)

(b) *a 50m search radius is insufficient to detect an adequately high percentage of carcasses, especially given the lack of rigorous data on detection rates of carcasses beyond 50m from a tower*;

How did the reviewer come to this conclusion? Earlier in this same paragraph this reviewer stated we followed standard protocols for carcass searches. In fact, we did.

(c) *the authors adopted adjustments to published scavenging and detection rates based on assumptions that are inadequately supported with observation. For example, the following three assumed adjustments are problematic: (1) "halving" the scavenging rate for raptors, (2) elevating the scavenging rate by 10% for the 2nd set of turbines because they were checked much less frequently than those in the study from which scavenging rates were used, and (3) assuming detection rates were equally high beyond 50m, where the crews did not search rigorously. Most of these inadequacies biased mortality estimates by an unknown amount and direction*.

Why are these adjustments deemed inadequate? We explained our rationale for these adjustments, which, except for (3), were based on experience. Furthermore, it is not true our adjustments biased mortality estimates by an unknown amount and direction. To come to such a conclusion, one first needs to identify the bias, not speculate there is bias. In fact, halving the scavenging rate for raptors would have biased the mortality estimate low, if in fact we were wrong to halve the rate, and it would have halved the mortality estimate.

For comparative purposes of a single species' mortality rates across turbine and location attributes (Chapter 7), these biases may operate roughly similarly across the variables and therefore may not undermine the analysis. For examination of impact (Chapter 4), however, these biases are very problematic indeed.

The reviewer states an opinion, but this opinion is naïve of the other sources of error and bias not addressed in our report. See Smallwood (2006, Attachment B) for a more thorough review of these errors and biases. Our assumptions and our adjustments were likely inaccurate in various ways, but these inaccuracies are rather trivial compared to other sources of error and bias that have yet to be addressed in any study of bird collisions with wind turbines. We can say with confidence, however, that the low ends of our uncertainty ranges are minimal levels of impacts

measured by our study. The impacts are larger than we reported, and they also include indirect and cumulative impacts.

P6: *Were uncertainties described, either qualitatively or quantitatively?*

In some cases, yes; however, the very large number of univariate test significantly inflates the probability of false positive results across the entire project. The authors made no attempt to adjust, quantify, and describe this issue.

As explained earlier, the reviewer exaggerates the likely frequency of Type I errors. Many of our test results had P-values <0.005. It is not true we made no attempt to adjust this issue because we screened the test results for use in the model-building phase of our project.

In addition, many estimates of rates were provided with no attempt to describe the associated uncertainties. For example, tables 7-4 through 7-7, 7-9, 7-11, 7-13, and 7-15, all provide estimates of rates of increase in mortality associated with a given variable, but no qualitative or quantitative measures of uncertainty are provided.

The values in these tables were not rates, so the comment is incorrect. These tables presented percentages of increases in the number of fatalities associated with particular levels or categories of association variables. They were not mortality estimates. Also, these percentages were presented as indicators. Assigning uncertainty ranges to indicators would result in the presentation of false precision.

Similarly, the species or group specific mortality rates given in tables 3-11 and 3-12 are presented with no measures of uncertainty provided.

This statement is incorrect. We stated right in the legends of Tables 3-11 and 3-12, "*We regard the mortality estimates in the left and right columns as the low and high values of the uncertainty range for each species or group*."

P6: *It is likely that some number of the reported test results were statistically significant. But due to the very large number of univariate tests conducted, there is a high probability that a number of "significant" results were based on pure chance. With an accepted P-value of 0.05, then 5 out of every 100 tests will, on average, appear statistically significant by chance when the null hypothesis is true. No effort to account for this was made by the authors*.

We responded to this comment already. Many of our tests resulted in P-values <0.005, yielding the likelihood 199 out of 200 tests were not products of Type I error. The last sentence it the comment is wrong because one of our screening criteria for including variables in model development was low P-value.

P7: *We cannot accept this analysis as one that has rigorously tested hypotheses regarding determinants of bird mortality and that could be reasonably applied in decision making. Instead, it may be more useful to consider this project an exploratory analysis that has identified a number of variables positively associated with increased mortality rates. Therefore, the product*

of this research is an educated list of working hypotheses. This valuable contribution can be followed by more thorough testing of said hypotheses by rigorous sampling and controlling of confounding variables via sophisticated multivariate analysis of observation data and/or controlled experimentation.

We agree the research resulted in an educated list of working hypotheses, but we maintain this list is much more educated than existed prior to our study (see Attachment A). Prior to our study speculation and anecdotes founded most notions of what factors contribute to bird collisions with wind turbines, and now we've explored data from a mensurative study. Certainly more research is needed to add to our understanding of factors contributing to bird collisions, but it would be irresponsible of decision-makers to ignore the patterns we reported and the mitigation measures we recommended. Decision-makers should use the best information available, and ours is some of the best information available.

P7: *The authors imply that a "use vs. availability" approach to quantifying vulnerability can be effectively pursued via chi-squared tests. The "original" paper describing chi-square (goodness-of-fit) tests to examine use vs. availability of resources in a wildlife context is by Neu et al. (1974).*

This is not exactly true. Pearson (1900) and Fisher (1924, 1950) provided the theoretical and mathematical framework for this approach, and then Larsen (1936), Greze (1939), Shorygin (1939), and Ivlev (1961) utilized components of chi-square tests for indexing resource selection. Jacobs (1974) also contributed the same year as Neu et al. (1974). This history of the development of this approach was summarized in Smallwood (1993).

Now, few biologists would consider chi-square tests effective or state-of-the-art for use-versus-availability designs (see book on the subject by Manley et al. 2002 and Journal of Wildlife Management volume 2006 issue #2). Instead, most use-versus-availability designs make use some form of logistic regression functions or general linear models.

It is presumptuous and a gross mischaracterization to claim that most wildlife biologists would consider chi-square tests ineffective or outmoded. The percent of papers using of chi-square tests in the Journal of Wildlife Management increased 30% over 20 years through 1996 (Stauffer 2002), and in 1996 about 3.5% of papers in this journal used log-linear models or logistic regression while 48% used chi-square tests. We are aware that the use of logistic regression and log-linear models has increased in use-and-availability analysis, but neither has chi-square tests gone away. As an Associate Editor of the Journal of Wildlife Management, Smallwood has administered the reviews of many recent papers, and he has also reviewed papers for other journals. He's certainly seen these more recent modeling approaches used, but he continues to see chi-square tests as a mainstay of use-and-availability analysis. But all this said, the appropriateness of a test should not be decided by the winner of a popularity contest.

P7: *The chi-square test assumes that the observed counts are accurate and any variation occurs simply from chance and not from observer error. As stated frequently in following chapters, the actual mortality counts are actually estimated counts and assumed to be biased low. Even*

assuming the mortality estimated counts are not biased low or high, this will result in inaccurate levels of statistical significance for the chi-square tests.

The reviewer's expectation of count accuracy, though enviable, would disqualify the use of chi-square tests and just about every other test, as well, from the majority of biological field studies. Observer error lessens the accuracy of the majority of wildlife studies. This is probably why none of the 15 biologists who reviewed our reports on our Altamont Pass research identified observer error as an issue related to our use of chi-square tests (3 reviewers of our NREL progress report, 3 reviewers of our NREL final report, 2 reviewers of our CEC final report, another 3 reviewers of our CEC final report one year after release, 4 reviewers of our burrowing owl manuscript submitted to a symposium proceedings).

P7: *And finally, chi-square tests are typically of two types: test for association/independence and test for goodness-of-fit. These chi-square tests are goodness-of-fit tests where the null hypothesis is that the counts were generated by a uniform distribution. That is, if there were no preference for the various categories of the explanatory variable, a carcass (or whatever response variable is being measured) would be equally likely to end up in any of the categories when adjusted for availability of the categories.*

We are aware of these chi-square basics, and in the report we described the null condition as it related to our chi-square tests. Is there another point to this comment we are missing?

P8: *"… we are able to identify which environmental factors might have a causal relationship." After so many years of studying avian mortality associated with wind turbines prior to this work, exploratory observational studies should be superseded by designed experimental studies. Observational studies are not able to reveal causality. Experiments, however, can show causality. Yet there is no evidence here that any experimental design took place prior to the observations. The sample locations and times were certainly not random nor were they seemingly selected to provide contrasts in factor levels. This would have allowed them to better compare the variables of interest and help to eliminate confounding variables*.

When the reviewer says, "After so many years of studying avian mortality associated with wind turbines prior to this work…," we wonder what studies the reviewer is referring to. The industry is in its infancy and few modern wind farms are monitored at all for fatalities. Orloff and Flannery (1992) was the only major research study performed prior to ours; other bird collision studies at wind farms were small in scope or consisted of regulatory compliance monitoring. We are unaware of any opportunities to date in which researchers have had the luxury of designing a manipulative experiment at wind farms, other than painting blades of existing turbines. We certainly did not have that luxury, and so we feel this criticism is unfair and misdirected. If we are going to conduct manipulative experiments, then the wind turbine owners need to allow us the opportunity to work with them on wind turbine siting during the wind farm planning stage. Until they are willing to allow us that opportunity, we will continue to learn as much as we can from mensurative studies.

P8: *This is a useful location map; however, a more useful map would have shown the topographic and other specific features of the APWRA. Are there distinct regions of the resource area that might be used to stratify the design?*

We tried providing topographic information in our maps, but the maps get so busy that they are useless. However, if we were able to revise the report, we would work with a GIS specialist to overlay the wind turbine map onto a lightly shaded grayscale coverage of the geographic relief.

P8: *Table 1-1 is described as "…summarizing the wind turbine attributes of the wind turbines in our sample in the APWA." Much more information is needed here. If this is the sample, how many of each type of turbine is in the sample? How many observations (visits?) occurred at each turbine type in the first set and in the later one? What fraction of the total turbines in the APWRA does each of these types constitute? A description of the sample and the population is called for here. Are these turbines representative of the entire APWRA population?*

Had we the opportunity to revise the report, we would add the information suggested by the reviewer. Figure 7-2 provides some of this information, but we agree more would be better.

P8: *But after a complete reading, the reader is still left wondering this most basic of questions – did the sampled turbines adequately represent all the turbines in APWRA*?

After explaining that we searched for dead birds at 75% of the 5,300 wind turbines in the APWRA, and after providing a map of the turbines searched and not searched, we assumed the reader would assume, yes, the sampled wind turbines adequately represented the APWRA. We suspect that most people familiar with this topic would agree.

P9: *We recognize that there may be some variables that the authors cannot ascribe to turbines that were not studied (e.g., grass height surrounding the turbine), but we assume many variables are catalogued by the turbine owners (turbine model, rotor speed, etc.) and/or obtainable from GIS (elevation, slope, aspect, etc.)*.

The reviewer's assumption may be correct, but he over-estimates the willingness of the wind turbine owners to cooperate with us. After our study, WEST, Inc. catalogued the wind turbines we did not search. They were the same types of turbines.

P9: *Figures 1-2 through 1-7 provide visuals of the distribution of sampled turbines, but they offer no information on how these distributions compare to the target population because the unstudied turbines are simply marked "unmapped." This is a significant shortcoming of the report*.

We do not understand the comment, nor do we understand how the perceived shortcoming is significant. Our target population was the APWRA, and we searched every turbine we could search with the access granted and the budget available.

P9: *On more minor notes, why are model numbers only given for the Kenetech turbines?*

We provided the information provided to us by the wind turbine owners.

P9: *It would be useful to include an additional figure that depicts which turbines were linked to 1 carcass, 2 carcasses, etc.*

Figures 7-19 to 7-21 show fatalities on maps. We hope the reviewer will agree it is difficult to see the fatalities at this scale. This is why we depicted numbers of fatalities per string in Figures 3-16 to 3-19. However, we never felt it was very helpful to show maps of where we found dead birds, because the maps would be influenced heavily by search effort, which varied across the APWRA from 2 searches per turbine to 34 searches. The maps of model predictions in Chapter 7 are more useful.

P9: *It would seem appropriate to present the methods section, given in Chapter 3, prior to reporting the results. It is not possible to make sense out of the various results given in Chapter 2 without knowing the sampling methods used and the underlying sampling program design.*

That is why the first sentence of the Methods section in Chapter 2 reads, "*The field methods used to find and record fatalities are described in Chapter 3*."

P9: *The authors state that one-way analysis of variance (ANOVA) is commonly used and least significant differences (LSD) to compare groups. The authors should give detail as to which LSD method was used as there are several different variations, although it is doubtful this resulted in any significant changes in their calculations*.

Had we the opportunity to revise the report, we would add this detail, although we agree it would make little or no difference to the outcome of the tests.

P9: *A more important defect is the authors' excessive use of one-way ANOVA throughout this chapter and report. Many variables are tested one by one for association with mortality using one-way ANOVA. This approach makes the analyses vulnerable to confounding variables when two are more variables are highly correlated with one another, such as blade height and blade speed.*

We would agree with the reviewer had we attempted to combine these variables tested by ANOVA in a predictive model without screening the variables for multicollinearity, but we made no such models in Chapter 2. The reviewer's argument over our use of ANOVA in Chapter 2 is a red herring.

P9: *The basic statistical rule that "association is not causation" can get lost in data analysis expeditions*.

This is a condescending and unproductive, rhetorical comment to which we cannot reply.

P9: *In addition, each time a one-way ANOVA analysis is performed, the data should be graphed so that readers can see if a particular characteristic of the dataset is having heavy influence on*

the outcome and whether or not more subtle statistical theory violations are occurring. In light of the absence of such graphs, the P-values can be considered only approximate at best.

This reviewer accused us earlier in the review of data dredging, but now recommends we graph out all the details of our ANOVA results. We are inclined to provide the reader as much information as possible, however, and if we had the opportunity to revise our report we would add the graphs.

P10: *Given the phenomenal number of univariate hypothesis tests done later in this report, it is surprising that there is no discussing of corrections for multiple comparisons here*.

This is because we did not make multiple comparisons. We methodically tested whether fatalities associated with each and every variable we measured, then we screened the associations for inclusion in a predictive model. Our screening method was in the report. The variables we used had to be highly significant in their associations with fatalities, show clear gradients in their relation to fatalities, include few expected cell values <5m, and had to make sense. We also selected variables to be more orthogonal relative to the others in the model.

P10: *It would also be helpful if the authors stated which statistical software package was used to do these analyses*.

We used SPSS. There is no explanation why this information is helpful, and it should not matter for this review.

P10: *What are the dates for season boundaries? These are not presented until Chapter 7 on page 182. Even there, the description of these dates and why they were chosen is inadequate (see later comments)*.

Many of these comments are addressing inconsequential details. Why ask about the dates of season boundaries when you know they appear on page 182?

P10: *How were days since death estimated? Were these simply guessed via personal experience? How was such experience gained?*

On page 48, we wrote, "*Each fatality was classified as a "fresh kill" or as "old remains," depending on the estimated time since death. Fatalities were considered fresh when carcasses and small remains were estimated < 90 days since death. Old remains included highly decomposed and dismembered carcasses with weathered and discolored feathers, missing flesh, and bleached, exposed bones. These carcass characteristics led observers to believe that the time since death was before the initiation of search rotations at the particular wind turbines*." And yes, we relied on experience as field biologists. We also placed bird carcasses in screens and set them in the APWRA for monitoring. We have multiple sequences of photos of bird carcasses monitored through time, and which helped guide our estimates of days since death. For example, Photos 8 through 12 depicts one of these sequences, but note every other photo we took in this sequence is not shown here. Undoubtedly, our guesses were not 100% accurate, but they were sufficient for our needs and within our budget constraints.

Photo 8. European starling found freshly killed in the APWRA and placed in a mesh cage to protect it from vertebrate scavengers and monitor the condition of its carcass.

Photo 9. The condition of the European Starling carcass the day it was found and placed in cage.

Photo 10. The condition of the European Starling carcass one week after placement in cage.

Photo 11. The condition of the European Starling carcass 28 days after placement in cage.

Photo 12. The condition of the European Starling carcass 78 days after placement in cage.

P10: *Of the 1162 detected birds (and bats) killed by turbine collisions, almost 50% (49.5%) were restricted to 4 of the 60 species/groups reported: Red-tailed Hawk (18.3%), Rock Dove (16.9%), Western Meadowlark (8.3%), and Burrowing Owl (6.0%). Does this high concentration (i.e., 50% of deaths in 7% of the species) reflect the differences in a) abundance among these species, b) the relative risk of wind turbine collisions, or c) the probability of carcass detection*?

The answer is "all of the above." Chapter 8 presents the information needed to assess alternative hypotheses (a) and (b), and our Attachment B addresses hypothesis (c).

P10: *The authors openly stated earlier that their search radius was 50 meters (m) and acknowledge that some "unknown proportion" of carcasses outside of the search radius went uncounted (p.28, pars.1 and 2). Yet, an unsupported statement is made here (p.38, par.1) that the "search radius included 84.7% of the carcasses of large-bodied bird species determined to be killed by wind turbines or unknown causes." How was this 84.7% calculated? In light of*

their search radius, it is not surprising that the majority of the carcasses were found inside the 50m radius of wind turbines. This problem is repeated later (p.42, par.5) when they note that their search radius "included 90.5% of the carcasses of small-bodied bird species." How they determine "90.5%" is left totally unclear to the reader.

We calculated the proportion of carcasses found within 50 m as the following:

$$\frac{N_{50}}{N_T},$$

where N_{50} was the number of carcasses we found within 50 m of the turbines we searched, and N_T was the total number of carcasses we found while performing fatality searches, including those found within 50 m and farther than 50 m.

P10: *It is unclear both in this section and in Chapter 3 how the carcasses beyond 50m from the turbines were discovered. If the discoveries were accidental and not within the defined sample element, then why were they included in the analysis? If the discoveries beyond 50m were accidental, describe the circumstances of the accidents. Were the observers walking in toward or away from the turbine strings? If they were collected as part of a special study in a systematic search that extended beyond the 50m limit, then describe that study's methods and results.*

As depicted in the photos in our report, the grass is short. It is easy to see carcasses, including those outside the search areas. On page 45, the first sentence of our Discussion section read, "*We found birds beyond the 50-m search radius because the search crew members could sometimes see carcasses at these greater distances as they approached the 50-m termini of their transect segments*." Our fatality searchers were instructed to record all the carcasses they found, and sometimes they found carcasses farther away. Smallwood, while mapping rodent burrows and wind turbines, would find carcasses at other turbine strings, and on his first day in the APWRA he spotted a golden eagle carcass across a ravine (Photo 13). It had been electrocuted.

Photo 13. This golden eagle had been electrocuted on a distribution pole. It was spotted by Smallwood from across a ravine, perhaps 250 m away, even though he was not looking for carcasses. Bird carcasses often stood out on the grasslands of the APWRA.

Photo 14 shows a golden eagle we found 87 m from a wind turbine. We determined, however, this eagle carcass had been dragged more than 30 m, based on a trail of feathers. Tracking feather trails to the bird carcass was not unusual during our study. On the other hand, some carcasses were found outside the search radius, but right next to it. Photo 15 shows a burrowing owl recorded at 51 m, just 1 m from the boundary of our search area. In our opinion, omitting this owl from the analysis would have been unjustified. In another case, we recorded a red-tailed hawk at 105 m from the wind turbine, but the bird was alive. For analytical purposes we regarded the bird as dead because its wing was broken and it would not have survived in the wild. (The bird hopped away and likely died somewhere else.) Yet another red-tailed hawk was found alive 150 m from the wind turbine, but its head injury later proved fatal after it was taken to a rehabilitation center.

Photo 14. Golden eagle carcass outside our search radius but, spotted from within our search radius.

Photo 15. Decapitated burrowing owl found 51 m from a wind turbine.

Reports of collision monitoring at wind farms refer to carcasses found between searches as *incidental* finds. Carcasses found outside the search radius are sometimes called incidental finds, as well, and sometimes they are simply counted as part of the search. Among 20 reports of collision monitoring at wind farms, 5 excluded incidental finds, 5 included them, 1 estimated mortality with and without incidental finds, and 9 did not explain how they handled them.

Researchers working at wind farms can be reluctant to exclude carcasses found incidentally or beyond the search radius because they are acutely aware of their sample size limitations. Mortality estimates are routinely made with sample sizes much smaller than the sample we obtained. Future monitoring and research efforts can reduce the magnitude of this problem by shortening the fatality search interval so that the searcher can find more of these carcasses during scheduled searches and before scavengers carry them off.

We decided to include the carcasses beyond 50 m because many had obviously been killed by the wind turbines. Our search radius of 50 m was relatively far for fatality monitoring at wind turbines at the time, but we found the radius to be too short because about 13% of the carcasses we found were outside our search areas. Since then search areas have expanded throughout the US.

P11: *If the discoveries shown in these figures beyond 50m were accidental, then, whatever the resultant pattern, it is unreliable since different sampling effort was expended within the 50m limit then beyond it. Consequently, we expect to have more discoveries within 50m then beyond it. It is no surprise that 75% of the large bodied birds were found within 42m of the tower. If we had a uniform density of birds on the ground in a 50m radius of the tower, we would expect to find 74% of the birds within 43m of the tower as shown in this simple ratio circles' areas*

$$\frac{\left(\pi \times 43^2\right)}{\left(\pi \times 50^2\right)} = 0.74.$$

Imposing a normal curve on this is unwarranted and somewhat misleading. The only patterns that are worth analyzing are within the 50 m limit.

We agree more carcasses were likely to be found within 50 m, and we are not claiming any sort of surprise that most of the carcasses were found within 50 m. The point of this comment is lost on us because we were not attempting to express mortality as the number of fatalities per square meter. We were interested in how many fatalities were caused by the wind turbine. Therefore, we disagree the patterns worth analyzing were only those within 50 m.

P11: *A polar or wind rose plot would be clearer. How can the 0 and 360 degrees cells not have identical counts since they are the same direction? What is the predominant wind direction? And what about the direction the wind turbine is facing?*

We agree a wind rose plot would be clearer, and 0 and 360 degrees should have been combined. On page 13 we wrote, "*Steady winds from the southwest blow across Altamont Pass during about April to October. Differential air temperatures form as the warmer Central Valley east of Altamont Pass draws in cooler, marine air from San Francisco Bay to the west. Winds are more erratic at other times of the year. They can originate from any direction*." Most of the turbines rotated to face the wind or away from it, depending on their design, and vertical axis turbines responded to wind from any direction.

P11: *The authors use simple linear regression to show that mortality counts increase linearly with turbine tower height. The mathematical assumptions behind linear regression are not valid with this particular dataset (likely nonlinearity, non-normal distribution of errors, unequal*

variances) thus inadequately demonstrating statistically conclusive evidence that mortality counts are greater for taller turbines. The fact that the one-way ANOVA for wind turbine model and carcass distance was statistically insignificant (p.42, par.7) suggests the height-distance conclusion is questionable. In a confused sequence of logic, the authors state (p.42, par.6), "[the regression] predicted that for every meter increase in tower height, average distance of the carcass from the tower increased by half a meter." This clearly ignores that different wind turbine models have different tower heights, thus it may not be the height, but rather the model, that results in the carcass distance. Height and wind turbine model are confounding variables.

The reviewers mischaracterized our report. Whereas we provided an interpretation of the regression models, we also tempered that interpretation with cautionary notes. On page 38 we wrote, "*Distance from tower increased with tower height, according to linear regression analysis, although the precision of the model was poor…*" On page 42 we wrote, "*A linear regression slope was significant but imprecise (Figure 2-11B), and it predicted that for every meter increase in tower height, average distance of the carcass from the tower increased by half a meter*." Other readers in addition to these reviewers may care to disregard these regression results, but we provided them the means to make up their own minds.

Whereas the reviewers characterized a simple statement we made as "a confused sequence of logic," we still feel this statement is factual and straightforward. Perhaps turbine model influenced the result, as the reviewer believes in the absence of evidence, but the simple fact remains that we found bird carcasses increasingly farther from the wind turbine the taller its supporting tower. Of course, tower heights are intimately linked to turbine models because the latter determines the former to a large extent. For example, we are not going to find KVS-33 turbines on 14-m towers. But in this case, we were more interested in whether tower heights affected how far carcasses traveled before hitting the ground, and the reason for our interest is because we were aware the APWRA will be repowered with new-generation turbines mounted on taller towers.

P12: *The authors stated, "Distance from tower [to the carcasses] increased with tower height, according to regression analysis, although the precision was poor." The overwhelming majority of the towers were 18.5m and 24-25m tall, making this primarily a study of these towers with a few others added in. Consequently, the observations at the lowest and highest towers had the greatest influence on the regression. If the 4 to 6 observations on the 43 m towers were removed, we suspect that neither of the two regressions would be statistically significant.*

With the fatalities at 43-m towers removed from the analysis we obtained the following results.

Small birds: $Y = 11.70 + 0.53X$, $r^2 = 0.01$, RMSE = 19.34, P = 0.015

Large birds: $Y = 2.13 + 1.30X$, $r^2 = 0.02$, RMSE = 29.86, P = 0.002

For the large birds, removing the birds killed at 43-m towers from the analysis steepened the regression slope and increased the significance of the regression, which was opposite of the reviewers' prediction. For the small birds, the regression model did not change. The reviewers

were incorrect in their prediction the regression models would not be significant after excluding the fatalities at 43-m towers.

P12: *A description of the tower population would be useful here. For the sampled towers and the population as a whole, how many towers of each type, what elevation distribution, what string lengths (1 to n), what spacing between towers in string, etc*?

We agree, and we would be happy to add this information to a revised report if given the opportunity. However, we point out that our report was 531 pages, so every addition of this sort of information lengthens it all the more.

P12: *The authors survey how carcass distance relates to multiple independent variables including tower height (continuous); blade speed (continuous); upwind vs. downwind (binomial); end, gap, or interior of string (categorical); season (categorical); whether turbine was in a canyon (categorical), slope grade (categorical); or elevation (continuous). They investigate each variable in a univariate analysis, but this may be better suited for a general linear model*.

We agree, and if we were given the opportunity, then we would revise the report accordingly.

P12: *Why are there 2 degrees of freedom (# levels – 1) in the ANOVA to test if carcass differed depending on whether the turbine was in a canyon*?

That was probably a typographical error. The degrees of freedom should have been 1.

P12: *The report of a strong effect of tower location within a string on the carcass distance is difficult to accept without careful analysis of the influence of the sampling method. The sampling method is described to some degree in Chapter 3, but it remains unclear how carcasses were associated with a particular tower within a string*.

On page 45 we wrote, "*Although the position of the wind turbine in the string related significantly to the distance of carcass from the tower, the effect should be expected, simply because there is greater opportunity for carcasses to be located farther from the end tower. That is, if a bird is killed by an interior turbine, its carcass is likely to fall to either side and to be associated with the neighbor tower; whereas, the end tower only has one neighbor for such a mistaken association to be made. Still, the percentage of carcasses of large-bodied bird species found within 50 m of end turbines was 79%, which was 6% fewer than all the towers considered together and 11.4% fewer than the interior turbines alone. A greater search effort is needed for large-bodied bird species at end turbines; 100 m would include 94% of the carcasses we found*." We will add that end-of-row and edge-of-gap turbines also tended to be located closer to the ravines and canyons, and the slopes were steeper. Gaps in turbine strings often resulted from ravines, so birds struck by edge-of-gap or end-of-row turbines have more airspace to fall through before hitting the ground.

P13: *The towers not on the ends end up with rectangular search areas, where some of each rectangle is beyond 50m from a tower. On the other hand, the end towers in a string may have*

considerably larger sampling areas (depending on the tower spacing) and more at the further distances away from the tower.

For example, if the towers in the string were 50 m apart, then the search area for the towers in the internal string would be: height× width = (50m + 50m) × 50m = 5000m^2. For the end towers, the area would be (half of a rectangle + half of a circle) = (50m + 50m) × 25m + 0.5× π× (50m)^2 = 2500m^2 + 3927m^2 = 6427m^2
which would be 29% larger than for the internal towers.

So the distance between towers in a string is important.

We agree with the conclusion that larger search areas were devoted to end-of-row turbines, but we do not understand how this difference in search area would result in a larger average distance of carcasses from turbines. Increasing the odds that we find more carcasses at end-of-row turbines does not mean their distances from the turbines will be greater. The searches were performed within the same search radius, after all.

P14: *The authors show standard error plots of carcass distance by the different wind turbine types. Box plots would do a more adequate job of showing the spread of the data and inform the reader of potential biases in the study with regards to various wind turbine models. Specifically, box plots would show if distance of carcass (beyond 50m) would result in reduced carcass count for a particular wind turbine model. A "mean and 2 standard error" plot is designed to show the reader the range where the true mean is likely to be. With this study, however, we are more interested in the range and general distribution of where the carcasses are to be found rather than what would be the long term average distance of where carcasses are to be found.*
We agree, and if given the opportunity to revise the report, we would use box plots instead of error bars in these instances.

P14: *In addition, for large bodied birds, 50.0% of the carcasses are associated with KCS-56 turbines, 34.1% with Bonus, and 6.1% with Micron, totally 90.2% of just 3 of the 10 turbine types. Similarly for the small bodied birds, 83.6% of the total carcasses were found at the same 3 of 10 turbine types. How many turbines of each type are there? Is this disproportion chance or pattern*?

The turbine models in our study are described in Table 1-1, and on page 42 we reported that distance from towers did not relate significantly to turbine model. Fatality associations with turbine model are reported in Chapter 7.

P14: *Given the happenstance data collection on carcasses beyond 50 m, the inclusion of the beyond 50 m data in the analysis is inappropriate*.

We disagree. There was an equal likelihood of observers seeing carcasses beyond the 50 m search boundary across the APWRA. However, if given the opportunity to revise the report, we would be happy to perform the analysis again with and without carcasses found beyond 50 m.

P14: *Regarding distance of carcass from wind turbine for "end", "gap", and "interior" turbines and their analysis (p.44, par.1) could suggest that carcasses tossed far from one turbine could be attributed to the turbine to which it landed closest too. This is acknowledged in the discussion (p.45, par.3). Are all wind turbines in a string alike*?

Usually, wind turbines are alike in the string, but not always.

P14: *The authors state that they found 15.3% of the large bird carcasses and 9.5% of the small bird carcasses outside of their 50m search radius. It is not surprising that only small percentages of the birds were discovered beyond 50m since the search effort in that region was happenstance. It is not stated whether these carcasses were found during the observers' systematic searches or while the observers were walking to the area where a systematic search would be done. How can you discover carcasses if you do not search for them*?

We did not claim any surprise is warranted on the smaller rate of discovery of birds found beyond 50 m. We simply reported what we found. How could we have discovered these carcasses? By seeing them.

P14: *They state that extending their search radius to 100m would include 94% of the large bird carcasses, an unsupported figure. There is a well-established body of theory for estimating density of animals (or in this case, carcasses) using the distance to each detection and modeling probability of detection as a declining function of distance. There are computer programs (e.g., DISTANCE) for this sort of thing. These programs could essentially estimate the number of carcasses that were overlooked to yield a more unbiased and accurate estimate of carcass density*.

The following is what we actually wrote on page 45, "*A greater search effort is needed for large-bodied bird species at end turbines; 100 m would include 94% of the carcasses we found*." We simply reported what we found, and contrary to the reviewer's claim, we did not estimate the percentage of carcasses that would be found out to 100 m.

P15: *This table* [Table 2-2] *summarizes the conclusions reached in this chapter about the distances of carcasses from towers. The relationships between distance and tower height are heavily influenced by a few observations on the tallest towers and in any case, the relationships are not substantial and only statistically significant in the most narrow technical sense given the* r^2 *values of 1%.*

Earlier we showed the reviewers were incorrect in their prediction a few of the tallest towers strongly influenced the pattern we reported, and we also showed they misrepresented our report by ignoring the fact we reported these results with cautionary statements.

P15: *The effects listed for Flowind and KVS-33 turbines are based on very small sample sizes (10 and 4, respectively) and also include happenstance discoveries beyond 50 m which further distorts the intervals. The reported effect could very well be spurious*.

The effects referred to were small, and the corresponding test results non-significant. The reviewers give the impression we attempted to report these effects as significant. In fact, we did not.

P15: *The authors propose that impact of the APWRA can be measured one of two ways: (1) number of fatalities per megawatt per year or (2) number of fatalities relative to the natural mortality and recruitment rates. They choose the fatalities per megawatt because it treats a certain number of fatalities as the "cost" of producing a megawatt.*

This comment mischaracterizes our report. We did not state we chose metric (1) as the cost of producing a megawatt. A reader can obviously interpret it that way, but that is not how we presented it.

Other authors use fatalities per turbine per year (p.46, par.4).

Expressing mortality this way is rare anymore. Most investigators adopted our recommended metric.

P15: *It is more an issue of policy that determines which measurement is more helpful. Although not unreasonable, fatalities per megawatt per year ignores the total number of fatalities. Total number of fatalities is an important measure that shows, at least in part, an impact on the bird populations even if you do not know the demographic conditions of the species. Fatalities per megawatt per year is a good measurement if you are trying to minimize fatalities while producing a certain amount of energy. Fatalities per wind turbine would only be helpful if you are trying to minimize the number of fatalities for a fixed number of wind turbines regardless of energy output – something only reasonable if wind turbine models all had the same energy output.*
It is not just a matter of policy determining which metric is more helpful. Researchers need to compare their results to the extent reasonable, and the number of fatalities per MW is serving this capacity as a surrogate to the obviously superior metric of fatalities per kWh. Wind turbine operators need to cooperate with bird collision researchers and monitors so that we can consistently obtain wind turbine output data. As for reporting the total number of fatalities in the wind farm, we agree with the reviewers, which is why these estimates appear in our Table 3-11.

P16: *The authors sampled 1,526 wind turbines (182 strings) for 4.5 years and another 2,548 wind turbines (380 strings) sampled for about 6 months (November through May) because of access issues. Although this is about 75% of the wind turbines in the APWRA, the authors do not say how they decided which turbines to survey.*

Not entirely true. On page 47 we wrote, "*During the course of the project, we periodically added groups of wind turbines into this set as access to these turbines became available*." We searched every turbine to which we were given access. Set 2 turbines were sampled systematically, but some of our initial turbine string selections proved infeasible because the turbines were either gone or were derelict. In each case we skipped these and selected the next string to the north or south in the sequence. This is why Figure 3-15 depicts some groups of turbines that were not sampled.

P16: *The short duration of sampling for the second set was the result of delayed access to the turbines from the owners. Although the first set includes fewer turbines and strings, it provides the primary and superior data set because of the repeated observations, the seasons sampled, and the increased duration. The limited duration of sampling, the lack of replication, and the restricted seasons sampled greatly reduces the value of the second set. Unfortunately, the analyses do not distinguish between the two sets*.

We do not understand this comment. Of course we distinguished between the two data sets. We described the two data sets, and we used a different set of scavenger removal rates for Set 2. We presented our mortality estimates separately, and we also combined them. We left it to the readers to decide whether they wanted to keep the estimates separate or combined.

P16: *Was there any concern about whether severed body parts from one mutilated bird (wind turbine or scavenger caused) could have indicated more than one fatality*?

No.

P16: *The authors write, "...we recently found that 85%-88% of the carcasses occurred within 50m of the wind towers." The absence of any described systematic method of how they searched beyond 50m makes this estimate questionable*.

This reviewer repeatedly raises this same issue, but we already addressed it earlier.

P16: *The authors then write the following:*

> *"Searcher detection and scavenger removal rates were not studied, because it had already been established that mortality in the APWRA is much greater than experienced at other wind energy generating facilities. We were unconcerned with the underestimating mortality, and in fact we acknowledge that we did so. We were more concerned with learning the factors related to fatalities so we can recommend solutions to the wind turbine-caused bird mortality problem. Thus, we put our energy into finding bird carcasses rather than estimating how many birds we were missing due to variation in physiographic conditions, scavenging, searcher biases, or other actions that may have resulted in carcasses being removed." (p.49, par.2)*

With this statement, readers must treat all bird mortality estimates as relative estimates and not as the exact counts or unbiased estimates. Regardless, the authors go ahead and attempt to come up with reasonable mortality estimates.

We agree with the reviewers' conclusion about how the reader should treat the mortality estimates. They should be considered relative, and yes, we did our best to present the most

reasonable estimates. Since the report, we have taken additional steps to improve the estimates (Attachment B).

P17: *What is the sampling element in use in this chapter? The authors "... express mortality as the number of fatalities per MW per year ..." The total number of fatalities observed on a string divided by the total rated power output from the string and divided by the total duration of sampling. This indicates that the sample size is the string, so that each string, not turbine, has an associated fatality rate. So sample sizes should be the number of strings visited, not turbines visited.*

Correct. On page 49 we wrote, "*We expressed mortality as the number of fatalities per MW per year (see Appendix A), where the MW were the sum of the rated power output of the wind turbines composing the string, and the number of years or fractions of a year were the time spans over which searches were performed at that string of wind turbines*."

P17: *The authors did not assess searcher detection rates in this study and selected to use literature values: 85% detection rate for raptors and 41% for non-raptors. Solely in this chapter, these detection rate values are used to correct the observed counts for deficiencies in detection. This seems reasonable, but why do the authors feel detection would be 50% less likely to discover a small raptor such as a kestrel than a similar sized non-raptor, such as a robin? (This same question applies to scavenging rates as well.)*

We used the literature-based estimates of searcher detection and scavenger removal rates available to us at the time. Even since then, investigators have been reporting the results of their trials for groups of birds, rather than for individual species. We agree the searcher detection term should have been lower for American kestrel, but we just did not have reasonable estimates available to use. The consequence of using a searcher detection term for American Kestrel that was too high was to produce a mortality estimate that was too low. Recently, Smallwood went to some length to get more reliable estimates of searcher detection and scavenger removal rates, and the results are attached as Attachment B.

P17: *They estimated the number of carcasses that actually existed by dividing either by 0.85 (raptors) or 0.41 (non-raptors). These calculations were equally applied to carcasses was found within or beyond the 50m search radius. This seems unreasonable to treat the beyond-50m carcasses the same as within-50m carcasses because carcasses beyond 50m were discovered by happenstance. The fraction missed beyond 50m could be much larger than their estimate.*

We agree the number of missed carcasses beyond 50 m could have been much larger than we estimated, but we disagree it was unreasonable to include those we did find in our mortality estimates. We add that we likely missed many birds that were injured by the turbines and later died someplace else, and we likely missed many for other reasons, as well.

P17: *The authors used scavenger removal rates and detection rates estimated in other studies to produce bird mortality estimates (p.51, par.1). A bothersome aspect of the authors' report is that they adjust the scavenger removal rates and detection rates from the other studies to rates*

that they believe better describe the APWRA and the time between their surveys without giving any anecdotal or empirical evidence of why they chose the numbers they did.

Why is it bothersome for biologists to exercise professional judgment?

Adding 10% to the scavenger removal rates of Erickson et al. (2003) to account for the authors' longer interval between searches appears arbitrary (p.51, par.2). Furthermore, without any support of data or other evidence the authors add (p.52, par.1), "Based on our experiences with raptor carcasses in the APWRA, we did not believe that these scavenger removal rates were accurate for raptors, and we halved the removal rate estimates reported by Erickson et al. (2003)."

The key phrase is "Based on our experiences with raptor carcasses in the APWRA." Since the preparation of our report, we followed up with additional investigations, and we remain at the same conclusion. See Attachment B for more details.

Underestimating scavenger removal rate will result in underestimating mortality.

We agree.

P17: *There is an error in their calculations for "halving" of the raptor removal rate. If s is the scavenging rate, the authors estimate the pre-scavenged carcass number by dividing the number of carcasses available after scavenging by* $(1-s)$. *After "halving" the scavenger rate, the authors simply divided by* $2\times(1-s)$ *while they should have divided by* $1-\frac{s}{2}$. *Their method reduced the scavenging rate by more than half and results in mortality estimates that are biased downward.*

We agree we made a math mistake, and biased our small raptor mortality estimate low. Since then we improved our estimation method and would use a different approach (see Attachment B) if given the opportunity to revise the report.

P18: *The combination of these various corrections results in an estimate of overall mortality that is, at best, rough and imprecise and, at worst, seriously biased (likely downward). No consideration is given to these ad hoc corrections in evaluating the uncertainty in the mortality rate estimates provided later in this chapter.*

We agree the mortality estimates are crude, but we disagree with the claim we gave no consideration to the mortality adjustments we used or the uncertainty created. The uncertainty ranges we used are large, and as such they state our level of confidence in the estimates. The very fact we used the estimates adjusted by scavenger removal as our upper-end estimate is another statement of our low confidence.

P19: *The authors are correct in stating that their "mortality estimates might be conservative" because of removal of carcasses by people not involved in the authors' study and they provide some anecdotal evidence. The authors do not account for such carcass removal.*

How would the reviewers suggest we account for this type of carcass removal?

P19: *The authors state that, of the 1162 carcasses whose fatality was attributed to the wind turbines, 198 were more than 90 days old. Table 3.1 on pp. 64 and 65 counts fatalities as Type A (both fresh and old) and Type B (fresh; used to estimate mortality). The difference between Type A and Type B should be the number of carcasses older than 90 days. In fact the difference is 1162 – 923 = 239 which is larger than the 198 reported on p. 52. What happened to the other 41? Bats account for some, but not all.*

Actually, bats do not account for any of the other 41 carcasses. In fact, 41 data sheets did not include an estimated number of days since death, and we excluded these from the mortality estimates. We neglected to mention this in our report.

P19: *The authors state that the frequency distributions shown in Figure 3-4 are "at the string level of analysis". The caption for Figure 3-4 should reflect that the figure shows the frequency of strings with various levels of estimated mortality rates.*

We did not understand this comment.

P19: *It is striking that at 270 of the 562 strings searched, or 48%, no carcasses were found. A useful analysis would have been to compare the group of strings with zero fatalities to those with observed fatalities.*

We believe this is what we did.

P19: *Both parts of Figure 3-4 include what appears to be a truncated normal distribution. This is inappropriate since the observed distribution is quite unlike a normal curve, more closely resembling an exponential or Poisson distribution. The normal curves should be removed.*

We agree, and we would remove the normal curves if we could revise the report.

P19: *The authors make statements about inter-annual mortality variation for different species and types of birds at wind turbines sampled for all four years. It is assumed, but not stated, that ANOVA and LSD are used. The multiple categories of birds species/type being tested for inter-annual mortality variation makes the chance of at least one Type I error likely.*

We used ANOVA, the results of which are summarized in Table 3-3 and depicted in the graphs showing means and error bars. We reported on 10 tests for differences in inter-annual mortality across 4 years, including 5 significant results with P-values of 0.009, 0.007, 0.004, 0.002, and 0.001. It is conceivable we committed a Type I error with one of these tests, but given the P-values we conclude the possibility was low. We think the reviewers mischaracterized our report again.

P19: *The statement about the mortality of burrowing owls based on the strings studied for 4 years vs. just 1 year refers to the right columns of Table 3-4. We suspect this should be Table 3-3.*

Correct. This is a typographical error.

P19: *Year effects on mortality rate are confounded by location, as evidenced by this figure* [Figure 3-15].

In the first paragraph of the Discussion section to Chapter 3 we wrote, "*Whereas we standardized our estimates of mortality by dividing the number of fatalities per MW and by the years spanning the search effort, our estimates of mortality might have been influenced by variable search efforts expressed as the number of years spanning the search period. For example, if few fatalities happened during a particular year, and we searched a group of wind turbines only during that year, then our mortality estimate from those wind turbines will be less than from other wind turbines and the comparison compromised. This shortfall in our study was beyond our control, since the owners of the wind turbines allowed us access to various new groups of turbines at different times during the study. For example, we did not gain access to our last addition of 2,548 wind turbines until late in 2002, after we completed our searches at all other wind turbines. However, this shortfall exists and needs to be divulged herein*."

Figure 3-15 does not reveal a terrible secret. We openly discussed the confounding issue. However, each test can be considered on its own, without comparison to other tests. That is, tests for inter-annual variation in mortality across 4 years will be non-confounded by restricting interpretation of the results to only those turbines searched all 4 years. We would run into a problem if we compared this test across 4 years to a test of data collected across 2 years, because the comparison would be confounded by location.

P20: *It seems as though the 95% confidence intervals in these figures were determined based on the string-based mortality rate estimates using Student's t distribution. Then it would be appropriate to provide the sample size for each year and not just the aggregate for all 4 years. (Or was it a sample size of 160 for the 1-year strings and 62 for the 4-year strings?).*

As an example, Figure 3-10 showed mortality means and confidence intervals for 'all hawks' among wind turbines searched all 4 years. In the upper right corner of the graph, we wrote "*N = 62 turbine strings*." This means 62 turbine strings were searched all four years, and these were the turbine strings included in the analysis. There was no variable sample size between years, only the same 62 turbine strings.

P20: *How was the confidence interval computed for 2001-2002 in Figure 3-9? It appears that the estimate is zero and the C.I. has zero width. How is this possible? Were there no barn owls killed in the 62 strings in 2001-2002*?

Incorrect. The answer to how it was possible to obtain a mean of zero would be no barn owls were found by us during 2001-2002 at these 62 turbine strings. We did not know how many were not found due to crippling bias, scavenger removal, or human removal of the carcasses. We cannot say how many were killed by these turbine strings.

P20: *To this point in this chapter* [Table 3-9], *the analysis has been string based. This table refers to 1526 turbines in the first set and the 2548 turbines in the second set. The columns give the mean and standard error among strings, not turbines. What was the sample size used for each of the mean and standard error calculations? Is it number of turbines or number of strings? Are these sample sizes taken to be the same for all species or groups*.

To this point in the chapter it should be apparent that 1526 turbines refers to Set 1, and 2548 turbines refers to Set 2. The sample size underlying the estimates in the table were the numbers of turbine strings, which were provided in the Methods section.

P20: *It would be useful to compare these results to the corresponding median values. It would be interesting to know how many of the median mortality estimates would be zero? Even for the shorter duration second set, 12 of the 30 (40%) species mean mortality rates are zero*.

We agree. We also point out Figure A6 (page A-9), which illustrates the percentage of turbine strings with 0 fatalities decrease through the period of fatality monitoring. Had we kept searching the Set 2 turbines, the frequency of 0-values among species and among turbine strings would have decreased.

P20: *The authors should better explain the calculations used to produce these tables. An example using real data would be helpful* [Tables 3-9 through 3-12].

An example probably would have helped, but we did provide all the information the reader needs to carry forward the calculations from Table 3-9 to Table 3-12 (see Methods section and footnotes to tables). We provided the means for the reader to apply his own assumptions and decisions about how to estimate mortality.

P20: *The authors assume a 50% miss rate outside of their 50m search radius (p.78, par.3). This statement conflicts with their Chapter 2 methods (p.51, par.1) where they said the detection rate within 50m was the same as beyond 50m. Thus in Chapter 2 they used detection rates for beyond 50m of 85% (raptors) and 41% (non-raptors). A 50% detection rate beyond 50m for non-raptors would suggest a greater detection rate beyond 50m than within 50m, obviously not sensible. More reasonable detection rates would be 42.5% (raptors) and 20.5% (non-raptors) beyond 50m (i.e., half the detection rate as within the more thoroughly searched 50m)*.

The reviewers are confusing the issue. Our Methods were worded precisely on this issue and need to be read carefully. On page 78 we wrote, "*For each reported search radius equal or larger to 50 m, we identified the proportion of bird carcasses we found beyond that radius and multiplied it by two (again, assuming a 50% miss rate).*" The proportion of carcasses found beyond 50 m is obviously a variable measured on a continuous scale, though bounded at 0 and 1. It is expressed for each species separately. The 50% miss rate in the parentheses refers to the last sentences of the preceding paragraph, and is an arbitrary expression of the proportion of carcasses occurring beyond 50 m at the time of the fatality search, and which we assume we did not detect. We are saying that of those carcasses occurring beyond 50 m, we assume we missed half of them. The reviewers already commented on this mortality adjustment term, and

concluded the miss rate could have been much higher at distances greater than 50 m from the turbines (and we agree).

P21: *The authors present findings from point count surveys although they have not yet discussed the methods with the readers*.

We did not perform these point count surveys ourselves. We had just explained in the preceding paragraphs (pages 77 and 78) that the data were from multiple studies conducted at multiple wind farms. The point count data were collected in various ways from the various studies.

P21: *In general, Chapter 4 does not adequately portray that the mortality estimates at APWRA from this report are likely biased low – perhaps severely. This bias comes about because: (1) detection rates for carcasses beyond 50 m could easily be well below the values used in analyses; (2) scavenging rates could easily be higher than used in analyses (because search intervals were longer for this study than in the studies from which values were obtained); and (3) scavenging rates of raptors were arbitrarily cut in half from reported scavenge rates*.

We agree with this conclusion. Had we the opportunity to revise the report we would incorporate Attachment B.
P21: *In general, the authors present the reader with a blizzard of one-way ANOVA and LSD statistical tests looking at an almost endless number of variables. Having so many variables inspected individually, leaves the study highly vulnerable to Type I errors, confounding variables and difficult to interpret findings. A multivariate approach would help the authors develop a more thoughtful, concise analysis that can help control for confounding variables*.

We disagree multivariate analysis would have led to a more thoughtful analysis. The approach we used was thoughtful, and in fact required closer examination of the data than tends to occur when investigators rely on multivariate tests. We agree, however, that the use of multivariate tests or multiple response tests would have produced more concise results. Had we the opportunity to revise the report we would be interested in using multivariate tests with these data.

We also disagree with the reviewers' portrayal of Chapter 5 as presenting an "endless number of variables." The number of variables we used in Chapter 5 was actually relatively small. We tested for relationships between each dependent variable and about 10 independent variables.

P21: *"Vegetation height ... was 18% greater ... where rodenticides were intermittently deployed...," the authors report with a mean difference from intense rodenticide use of 4.28cm. The magnitude of 4.28cm is more meaningful if the mean heights of the grasses are also provided. It could be 1cm vs. 5.28 or 11cm vs. 15.28 which could understandably have different ecological impacts*.

We agree it would have helped to have provided mean values. In the example used by the reviewers, however, the means are readily obtained by recognizing that the 4.28-cm difference is also the 18% difference noted in the same sentence. Dividing 4.28 by 0.18 yields 23.8 cm, which was the mean grass height on the areas where rodent control was intermittent. The mean grass height in the intense rodent control area was thus 19.5 cm.

P21: *The authors indicate that the index of cottontail rabbit abundance was higher on Enertech towers, on plateau slope combinations, and on southwest slopes. Were Enertech towers especially common on southwest slopes relative to other tower types?*

One (0.6%) of the 164 Enertech turbines was on a southwest slope, whereas 80 (2%) of the other 3910 turbines were on southwest slopes.

These questions are difficult to answer because they require the reader to extract information presented for other purposes elsewhere in the report. By running multivariate analyses (which may require simplifying or reducing variables – in itself a good thing), then the association between a given predictor variable and the response variable can be measured while statistically accounting for confounding variables. This is a recurring limitation of the study.

We disagree simplifying variables is necessarily a good thing; it is not "good" to sacrifice information unless it is to reduce the degree of false precision in the data. The use of multivariate or multiple response tests would result in division of data into small sample sizes representing each combination of variables, which poses its own significant problem. Whereas we do not believe multivariate analysis would be the fix to the confounding issues raised by the reviewer, we would be interested in applying it to the data if given the opportunity to revise the report.

P22: *p.103, Table 5-20. This is an example of where the authors should interpret the meaning of the analyses while paying attention to the magnitude of differences. Furthermore, the metric "cottontail abundance" is never defined.*

On page 90 we wrote, "*The cottontail abundance index was recorded along the string transect and grass transect. We recorded the presence or absence of cottontail fecal pellets along 40-m transects and within 5 m of the observer (the same 5 m strip transects used for cattle pats, as well as a 5 m strip transect along the turbine string). We also noted whether or not cottontail fecal pellets were especially abundant*." Thus, the index was measured as pellets absent, present, or abundant.

In Table 5-20 cottontail abundance is compared between "some lateral edge" and "other edge conditions" with a statistically significant "Mean difference (cm) on grass transect" of 0.18. What does that 0.18cm represent? Is that a small biological magnitude that ends up being statistically significant because of the very large sample size of 1327?

We mistakenly compared means of the cottontail pellet abundance index among levels of independent variables. These comparisons should have been made in contingency tables. Also, the headings in the tables summarizing cattle pat abundance should not have included cm as the units. We would fix these mistakes if given the opportunity.

P22: *p.111, par. 4: "Most wind turbine strings were selected arbitrarily, to represent a wide range of raptor mortality recorded during our fatality searches, as well as to represent a variety of physiographic conditions and levels of rodent control," the authors write. A more rigorous*

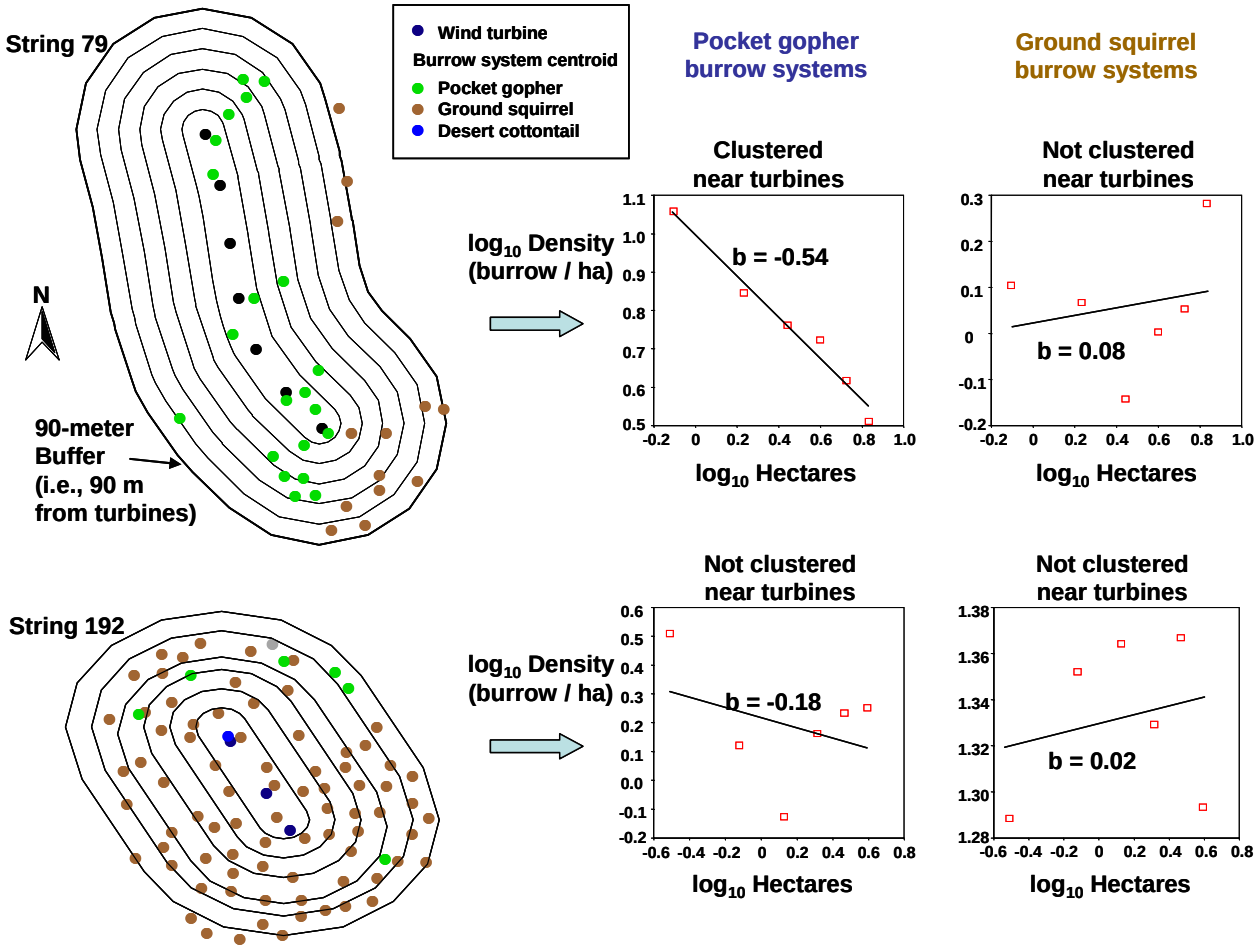
method of selection should have been used, such as stratified sampling. The objectiveness and unbiasedness of "arbitrary" sampling is always questionable.

We agree with the comment. We used an arbitrary sampling approach to maximize the range of mortality estimates in the subsequent tests, which we think was suitable for exploratory research. We agree that a more rigorous approach would be to use stratified random sampling.

P22: *The method of estimating degree of clustering at wind turbines using the slope from least squares linear regression is unclear (p.112, par.4). Is "corresponding search areas" the distance from the wind turbine? It then seems that the authors disregard this "regression-slope" method (p.114, par.5) for the "observe-divided-by-expected" approach. Having this "regression-slope" method discussed is confusing if it is not to be used.*

We discussed the regression slope method because we tried it and we think it is honest and thorough to explain what we tried and what we found to extent practical. The methods of how we applied the approach to search areas could have been clearer, as the reviewer stated. Below is a graph depicting search areas, which are the areas within each concentric boundary line.

Examples of pocket gopher and ground squirrel distributions around wind turbines where rodent control was and was not implemented.



P22: *The authors mention that they learned post hoc about rodent control. Although likely beyond the duties of the authors, the effectiveness of rodenticides to reduce raptor mortality could be better explored in the future via a carefully planned experiment.*

We concur with the comment.

P22: *The simple linear regressions used to investigate association between raptor mortality and ground squirrel burrow systems are very questionable (Figures 6-45 and 6-46). The authors discuss the significance of these scatter plots (p.149, par.5 and p.164, par.1). Some of these conclusions and "significant" P-values are based on sample sizes of 3 (no rodent control) and 5 (intense rodent control) – it is foolish to base inferences from just 3 or 5 data points. Furthermore, leverage of an individual point affects all three levels of rodent control and the assumption of homogeneous error is ignored.*

We would characterize our interpretation of regressions with 3 or 5 data points as aggressive, but not foolish. We had the data, the biological significance of which was important. We presented them in the scatter-plots so the reader can decide how much significance to give the regression results. Of course we would have rather had more data points to use in the analysis, but we were not prepared to ignore what we had.

The assumption of homogenous error was not ignored. The data were presented with the regression slopes so the reader can decide to what extent the relationships were homoskedastic. In the case of 3 data points, however, we agree we cannot assume the error in the data were homogenous.

P23: *pp. 164-172, Tables 6-2 through 6-11: These tables aggregate the density of burrows into categories and then total the number of bird kills for each of the three categories. It is not clear how the authors decided to define each category and information is lost by categorizing continuous data. A dot plot or histogram of the burrow densities for where carcasses were found beside a second plot of burrow densities for where carcasses were not found would have been more informative*.

Losing information by categorizing continuous data is exactly what this reviewer recommended on page 21 of the review, in order to utilize multivariate analysis. But to clarify our burrow clustering categories, we delineated categories by natural breaks in the histograms of burrow system clustering for each species.

P23: *Discussion, pp.172-178: The authors make good points in the Discussion regarding the negative and/or inconsistent impacts of rodent control measures, and their case is strong, we believe. They offer the caveat that, "Intense rodent control was associated with fewer golden eagle fatalities in areas of intense rodent control, but the association is not strong enough to warrant its continued use" (p.178, par.2). We think that statement is giving the rodent control measure more causal credit than it deserves. In fact, the P-value for the ANOVA test of golden eagle mortality rate across the three rodent control intensity levels is statistically insignificant at 0.9 (p. 172, Table 6-12). While the mean mortality estimate is slightly lower in magnitude for the intense control category, the variance is very large, and we thus have no confidence this difference is "biologically real." One could just as easily claim that, "mortality rates among rodent control intensity were statistically indistinguishable."*

We agree with the comment. The P-value we reported for the ANOVA test involving golden eagle mortality was 0.90. There was no difference in golden eagle mortality among rodent control treatments.

P23: *The authors define four seasons, but the length of the seasons are very different: spring is 92 days, summer is 117 days, fall is only 51 days, and winter is 105 days. Summer is 2.3 times as long as the fall. What is the justification for these definitions*?

The seasons were defined according to the biology. In general, the behaviors and long-range movements of the birds we studied change according to the seasons we defined. We based our definitions on years of research we have performed on raptors.

P23: *The authors need to be careful and consistent as to how they show their mathematics. They most often, but not always, use more elementary notation such as* $A \div B$ *instead of* $\frac{A}{B}$. *On the 7th line of page 183, they define "the window of opportunity" as* $\text{Window} = \text{C} \div \text{T} \cdot \text{B}$. *This is equivalent to* $\frac{C \cdot B}{T}$, *but the equation is more sensible as* $\frac{C}{T \cdot B}$, *which we believe is what the authors meant. The authors should employ the use of an equation editor, like that used in Microsoft Word*.

We agree, and we would use the equation editor if we were to revise the report.

P24: *For purposes of computing how quickly a bird clears the rotor plane, how thick is the plane? What flight speed would be required to clear the rotor plane in the allotted time*?

We were not attempting to measure the speed needed for the bird to clear the rotor plane. We only calculated the time the bird would have available. To better understand the flight speed needed to clear the rotor plane, refer to Tucker (1996a,b) or Richard Podolsky, who uses his model, Avian Risk of Collision, to calculate such speeds. Note, however, that birds rarely fly at their top speeds, and they do not always approach the rotor plane from a perpendicular angle.

P24: *The tower height is defined as the distance the rotor is above the ground. Can we assume that this is the center of the rotor*?

Yes.

P24: *The incidence of rock piles was reduced to a limited number of categories. Did the authors intend the categories to be: a) none, b) less than or equal to 0.25 piles per turbine, or c) greater than 0.25 piles per turbine*?

Yes. On page 184 we wrote, "*The incidence of rock piles at each turbine string was characterized as none, less, or equal to 0.25 piles per turbine, and > 0.25 piles per turbine*."

P24: *p.184, par.2: The authors employ a 40 m radius around each turbine instead of the 50 m radius stated earlier. What is the reason to redefine the sampling zone now*?

Because this time we were indexing soil/vegetation edge conditions, whereas the 50-m radius was used for fatality searches. Furthermore, we used a 300-m radius for making behavior observations, and 90 m for mapping rodent burrows. The sampling radius needs to fit the study unit being sampled.

P24: *p. 184, par.4: Did the authors test the assumptions of the statistical tests (e.g., homogeneity of variances or statistical independence and normality of residuals) applied in this or any other chapter*?

Smallwood routinely checks the residuals visually for homoscedasticity, and we reported root-mean square error as a measure of precision of the data around the regression slope.

What objectives are the authors trying to meet in reporting "weak and non-significant correlations"?

An example would be when the prevailing view was that there would have been a significant correlation. Non-significance from a statistical standpoint is often significant from a biological standpoint.

How can the measures of effect be meaningful if the confidence interval for the magnitude of the effect includes zero?

They would not be significant, which may or may not be significant to us as biologists. But where in our report did the reviewer see measures of effect with confidence intervals overlapping zero? We are not alleging they did not appear in our report, but we wonder where the reviewer saw these.

P24: *For regressions, the authors have chosen to include the RMSE to provide a measure of the "precision of the data relative to the regression line". By RMSE, we assume that the authors mean*:

$$RMSE = \sqrt{\frac{Sum - of - squared - residuals}{samplesize}}$$

A more appropriate estimator for precision would have been the standard error of the estimates (SEE) or:

$$SEE = \sqrt{\frac{Sum - of - squared - residuals}{samplesize - \#of\ parameters}}$$

Would not SEE be more appropriate for multiple regression analysis rather than for simple linear regression?

P25: *Although this is a non-manipulative study and the existing towers, turbines, topography, etc. as well as permission for access does limit the range of choice, it is still possible to carefully select the areas of study to provide the contrasts and comparisons of interest*.

There are cases where we agree with the reviewer on this point, but we largely did not have any latitude in designing the study in terms of wind turbine selection. The reviewer can say we should have, but we could not. Until the fall of 2002, we searched for fatalities at every turbine we were granted access.

P25: *p.185, par.1: Is the term "efficient" used here in the technical sense from statistics*?

It is used to summarize the arguments made in the papers that were referenced in the sentence, i.e., (Smallwood 1993, 2002).

P25: *p.185, par.2: The authors discuss the 5% significance level used in the subsequent tests and the 10% level that they interpreted as indicating "trends worthy of further research". Given the immense number of univariate hypothesis tests reported in the subsequent pages, the authors should have discussed the risks of Type I errors (false positives) associated with conducting hundreds of tests*.

If we were to revise the report, we would add a cautionary statement that some of the results could have been significant due to Type I error.

P25: *The total number of chi-square tests presented just in Tables 7-1, 7-2, and 7-3 is 528 (ignoring the many more chi-square tests presented in Appendices B & C). The chief disadvantage of this approach is that Type I and Type II (false negative) error rates are inversely related, creating no clear optimization. One could argue that Bonferroni adjustments are necessary to guard against very high experiment-wise Type I error stemming from so many tests. Using Bonferroni adjustments, the experiment-wise alpha (level of significance) value "should" be set as:*

$\alpha_{adj} = 1-(1-\alpha)^{\frac{1}{n}}$ *; in this case* $\alpha_{adj} = 1-(0.95)^{\frac{1}{528}} = 0.000097$ *for a modified Bonferroni adjustment as proposed by Shafer (Shaffer, J. P. "Multiple Hypothesis Testing." Ann. Rev. Psych.* ***46****, 561-584, 1995.)*

But if the authors bring the experiment wise alpha value this low, the Type II error rate gets unacceptably high, especially for work designed to measure environmental impact. That is, the probability of the analysis suggesting no impact when in fact there is one becomes unacceptably high. This problem further underscores the value of a smaller number of multivariate tests, as we have suggested elsewhere.

We understand and appreciate the comment, but we also point out that we did not rely on all our significant test results for developing our models. We screened them for low P-value, biologically significant gradients of fatalities, and an effort to minimize shared variation in the models. In the end we reduced the variables used in the model. However, if we were to revise the report we would attempt to use multivariate analysis as recommended. If we were to perform another study of bird collisions at wind turbines, we would again attempt to achieve equal and sufficient sampling effort among wind turbines so that multivariate analysis can be performed.

P25: *The uses of chi-square tests "for association" are described. The chi-square tests used by the authors are more commonly described as chi-square tests for "goodness-of-fit" where they are testing whether it is plausible that the observed counts across the categories came from a uniform distribution (each category is equally likely). Although statistically legitimate, such methods fail to control for other variables, leaving the study vulnerable to confounding variables. Why not use a general linear model, logistic (yes/no data) or Poisson (counts data) regression, discriminant analysis, or at least a log-linear analysis*?

If given the opportunity to revise the report we would try using logistic regression or log-linear analysis, but not Poisson regression. We agree confounding was possible, and indeed likely. We

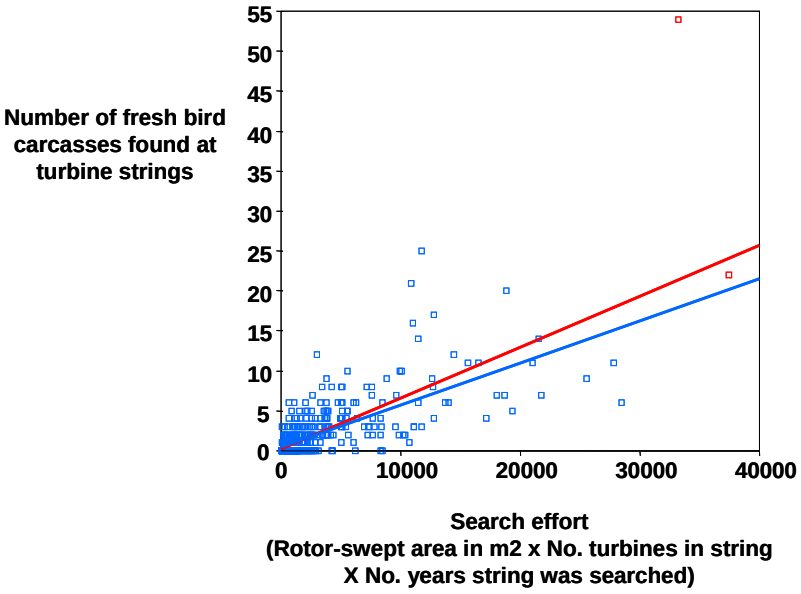
do not believe multivariate analysis would entirely remove or control for the effect of confounding, but it would certainly be helpful to minimize it.

P26: *p.186, par.3: The authors rationalize that relative search effort can be calculated as,* $N_t \times R \times Y$*, where* N_t *is the number of wind turbines in a string,* R *is the mean rotor swept area in* m^2*, and* Y *is the number of years the string is searched. This decision is based on Figure 7-1. It is a loose association between the relative search effort and number of fresh bird carcasses found. From this, they assume that mean rotor swept area is proportional to the number of carcasses – a circular argument since that is what they are supposed to be investigating. Keep in mind that the swept area is proportional to the squared radius of a wind turbine (Area=*$\pi \times r^2$*), thus the "search effort" at a wind turbine with a 3m blade will be four times as much as at a wind turbine with a 1.5m blade (half the size) even if they physically searched the surrounding grounds equally. Thus the wind turbine with a 3m blade will have to kill four times as many birds to have the same rate of mortality as the 1.5m blade wind turbine, ignoring megawatt output. In Appendix A, the authors do show a positive relationship between megawatt output of a turbine and mortality. Perhaps the authors are trying to copy epidemiology studies which use "people years" when calculating risks for cancer; e.g., following 100 people for 5 years is equivalent to following 250 people for 2 years. Here this would correspond to "turbine years". It is a strong assumption to say that the variable "rotor swept area" is just as important as the variables "time" or "number of wind turbines" with regards to the number of expected bird carcasses*.

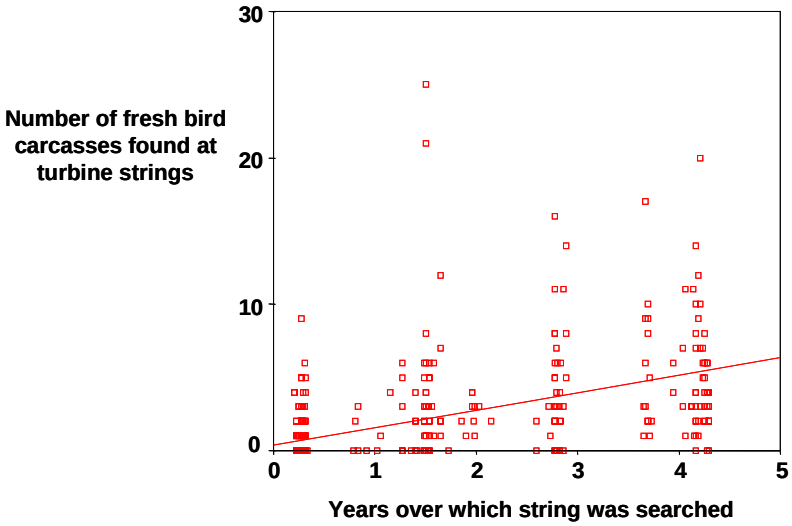
We do not understand the comment, especially the statement that we used a circular argument to arrive at our measure of search effort. We did not test for associations between fatalities and rotor-swept area at the string-level of analysis. On page 186 we explained that we used the string-level of analysis for tests of association with rodent burrow distributions. Rotor-swept area was not incorporated into our measure of search effort at the turbine-level of analysis, which is the level at which we tested for association of fatalities with rotor diameter.

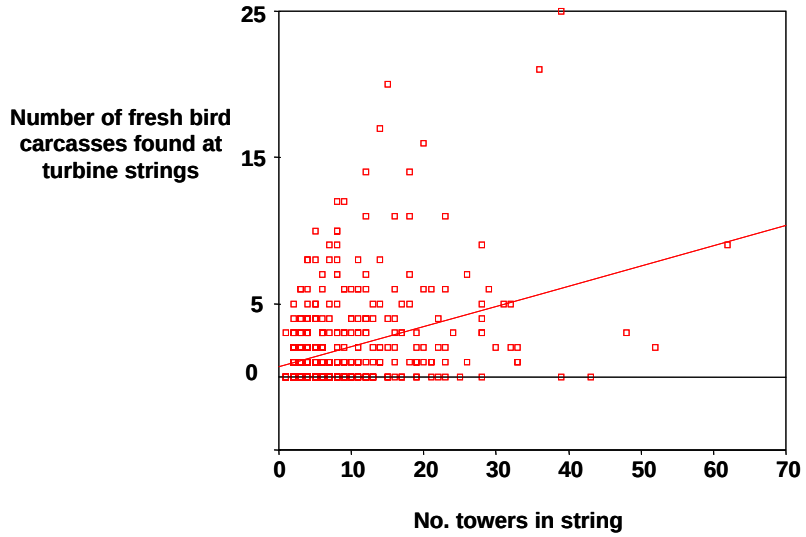
P26: *p.186, par.4 and p.187, Figures 7-1 A & B: Figure 7-1A presents the relationship between the number of birds recently killed at turbine strings and the measure of search effort used. Of the 472 data points, only 32 or so exceed 10,000* $m^2 \cdot yr$ *of search effort and only 2 of the 472 exceeds 30,000. Consequently, these extreme values of the total dataset have the principal influence on the regression results. Which of the variables account for the observed variation in the search effort: the number of turbines in the string, the mean rotor swept area, or the number of years of searching*?

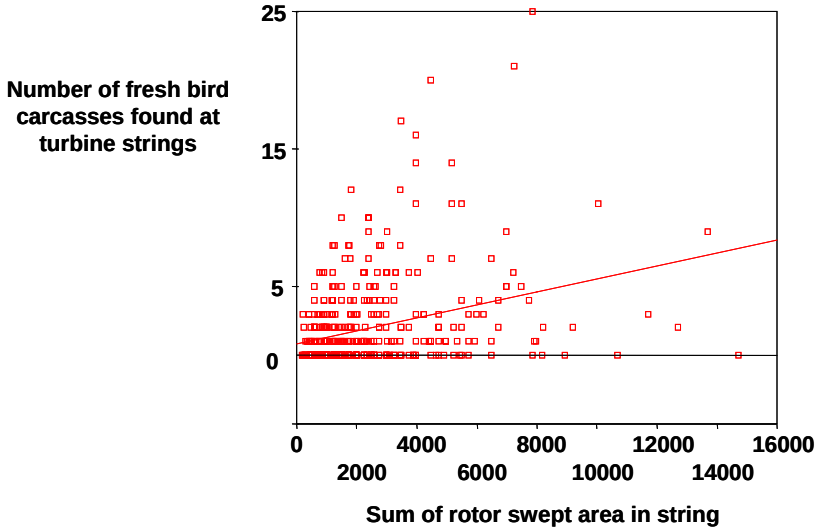
Below is a comparison of the regression slopes produced with (red line) and without (blue line) the search effort >30,000 (red squares). It appears the reviewer's prediction was inaccurate. The regression slope was influenced by more data than the two large values of search effort.



To answer the last question in the comment, we provide the plots below. Separating out rotor-swept area from number of turbines makes no sense because a string of turbines can range from 1 to about 65, so we presented the relationship between number of bird fatalities and the sum of rotorswept areas in the string (third scatter-plot down). It appears all three variables contributed to the pattern in the scatter-plot above, but the above plot was the most precise and made the most sense to us.

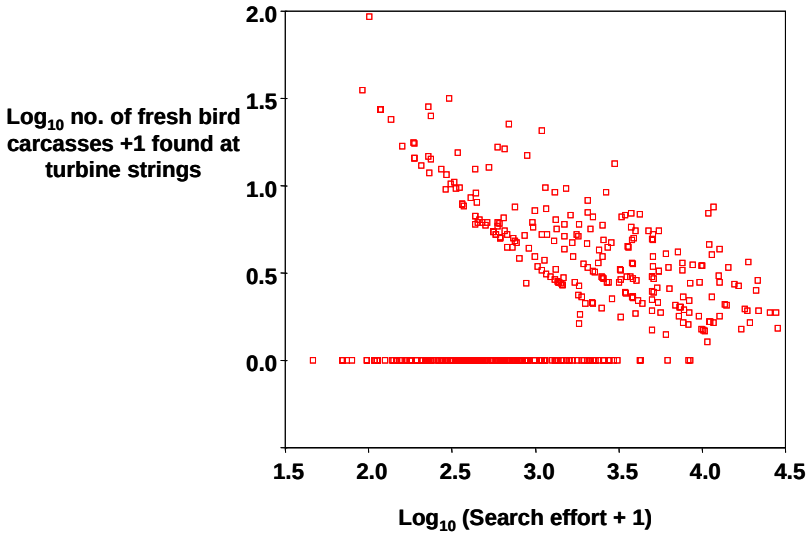






P26: *The authors suggest that Figure 7-1B illustrates an inverse power relationship between fatality rates and search effort. It would be more informative to plot the data shown on a log-log plot, which would more conveniently indicate if the relationship was in fact an inverse power relationship. It appears, however, that there may be many observations with fatality rates of exactly zero, but it is difficult to tell since the vertical axis does not show a zero.*

The reviewer's suggestion is depicted in the plot below. This plot resembles Figure A-7 on page A10 of our report. We added 1 to both variables in order to include the 0-values. The plot below, however, does not depict a linear pattern among the non-0 values, so bird mortality is not a strong inverse power function of search effort. We fit an inverse function to the data, though the r^2 value was still only 0.61. As we explained in Appendix A, continued fatality searches would eventually eliminate the 0-mortality values.



P27: *Figure A4 (p. A-8) suggestions a mechanism that would produce the relationship suggested for Figure 7-1B. This indicates that the sampling approach yields stable estimates only after longer periods of search, which should be discussed here.*

The mechanism is discussed in the Appendix, as the reviewer indicates, but that is where the CEC wanted us to discuss it. Had we the opportunity to revise the report, we would add a discussion in this chapter of the report, as well.

P27: *p.188, par.2: "Positive values express the percent of total fatalities likely killed at wind turbines due to the attribute associated with the value..." The use of the word 'due' implies causality, although at best they can only claim 'association'.*

This comment is misleading. On page 188 we wrote, "*Positive values express the percent of the total fatalities likely killed at wind turbines due to the attribute associated with the value, and negative values express the percent of the total that were expected to have been killed if fatalities were random, but that were not killed.*" Note the word *likely* preceding the word *due*. We were speculating, and while at it we used a term conveying uncertainty. We were not concluding causality, but rather suggesting it.

P27: *p.189, par.2: So now the sampling element is the wind turbine and no longer the string. What fraction of the total population of wind turbines does this sample of turbine models represent? It is important to the reader to know if these sampled wind turbines are representative of the APWRA population of wind turbines.*

We searched for fatalities at 75% of the wind turbines. The unsearched wind turbines were of the same models we searched, and on the same types of towers. They were interspersed among the turbines we searched. We provided a map of the turbines we searched and did not search (Figure 3-15).

P27: *p.190, Figure 7-2: The figure shows that the authors' study is essentially a study of KCS-33 and Bonus wind turbines. Furthermore, the "effort" for Bonus wind turbines is almost three*

times that of the number of Bonus wind turbines studied. Is that a result of the "relative effort" definition and that Bonus wind turbines' rotor sweep area is three times that of most other turbine models?

No, Bonus turbines were some of the first turbines we were given access to, and were the turbines we searched the longest.

P27: *p.189, par.8 and p.202, Figure 7-18: Based on the authors' definitions of seasons, fall is the shortest season (51 days) and so would be expected to have less sampling effort. Given the length of the seasons and assuming a uniform distribution of sampling times throughout the year, we would expect 25% of the observations in the spring, 32.1% in the summer, 14.0% in the fall, and 28.8% in the winter. Comparing this to the bar heights in Figure 7-18, the sampling effort is higher than expected in the spring, lower in the summer, higher in the fall, and on target in the winter. Is this a result of their sampling effort definition? It is not clear.*

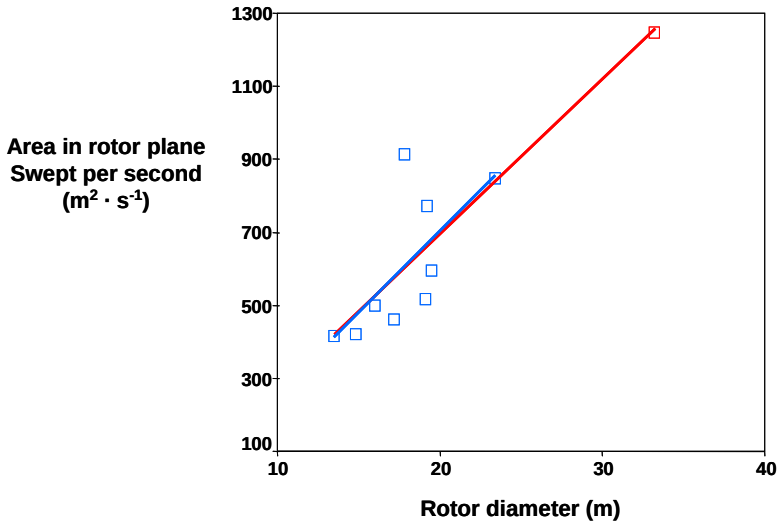
It was a result of funding. Our funding repeatedly lapsed in fall. During several years we were unable to search for carcasses during part or all of September and October.

P27: *p.192, Figure 7-4: Why is effort so many times greater for the wind turbines with 2141 rotor plane swept per second*?

For the same reason the Bonus turbines were given more effort – they were the Bonus 120-kW turbines, and were searched the longest because they were the first turbines we were granted access to search.

P27: *pp.193 and 194, Figures 7-5 and 7-6: These figures show scatter plots where an outer single point has high leverage (influence). Conclusions are essentially being determined by the one point furthest to the right*.

We disagree with this assessment. Essentially the same regression models would be obtained with or without the points to the right for Figures 7-5B and 7-6B, and the regressions for Figures 7-5A and 7-6A were not significant. (However, the Figure legend should not have stated that rotor swept area swept per second was a linear function of blade tip speed, because the regressions were non-significant.) The point we were making was that the area swept per second and the seconds between sweeps are both more responsive to rotor diameter than to tip speed, and the point remains valid with or without the data points on the right aspect of the plots. See the plot below, where the blue regression line was fit to blue symbols, and red regression line was fit to all data, including the red data point identified by the reviewer. The blue line is difficult to see because it so closely matches the red line, meaning the data point to the right did not highly influence the conclusion.



P27: *p.199, Figure 7-14 through p.201, Figure 7-16: Why are the bin widths increased in going from graph A to graph B for each set of graphs? In graph B of each pair of graphs, the bin widths are not equal.*

The bin widths are different between A and B because the variables being graphed are different. Regarding the second question, the bin widths differ because, again, the variables being graphed are different variables.

P28: *p.203, Table 7-1: The dangers of multiple hypothesis testing arise in Table 7-1 when 204 chi-square tests are performed. (This is repeated again in Tables 7-2 and 7-3.) This can be kindly called "data exploration" or criticized as a "data dredging". Regardless, with 204 statistical tests, if all data were a result of a uniform distribution across each category, researcher error or biased post-hoc categorization did not cause any non-uniform distribution, and each test were independent of one another, you should expect 5% of the tests to give p-values less than 0.05. So there is a high chance of Type I errors when so many tests are performed. Also many variables may be correlated, such as "tower height" and "high reach of blades". So if a test was significant for "tower height" you should expect it to also be significant for "high reach of blades". In addition, a more clear explanation is needed as to why some variables such as "rodent control" and "Slope aspect" are tested twice.*

The reviewer neglected to consider that many of our test results were significant with P-values < 0.005. The probability of committing a Type I error is much less among tests with these small P-values, so the reviewer's estimate of a 5% rate is misapplied.

We agree with the reviewer's point about shared variation, and that we should expect similar results across multiple variables when those variables express the same larger factor. However, we did not use all these variables in model development because we screened them and narrowed down the number of variables used. We made an effort to avoid multicollinearity when we synthesized our results in the Discussion section and when we developed our models.

P28: *There are methods to help reduce the problems of multiple testing, such as Bonferonni corrections that make the p-value for declaring a "statistically significant result" much less than 0.05 for each test. This makes the overall chance of a Type I error only 5% if all tests were actually not significant. The problem with such adjustments is that the statistical power then decreases for each test opening the door for Type II errors thus making the researchers miss important variables. The authors should take a more selective and thoughtful approach to investigating the variables and use generalized linear models or multiple regression. These more advanced methods would help reduce some confounding by allowing the authors to control for other variables when testing another. The authors did, however, state that they only used the predictive model for variables that were statistically significant and showed gradients along a continuum (p.188, par.3)*.

This comment was made earlier about the trade-off of Bonferonni corrections in their implications for Type I versus Type II errors. The last sentence of the comment acknowledges we screened our variables for inclusion in the predictive models. And we agree that generalized linear models might help sort out the variables and account for some confounding and multicollinearity, and we would use these if given the opportunity to revise the report.

P28: *Furthermore, what are the sample sizes for each of these chi-square tests? A large sample size can produce very small p-values (very high statistical significance) even though the magnitude of difference from the uniform distribution is minimal; i.e., lacking biological significance. When the authors discuss the finding from the chi-square tests, they report something along the line of, "Wind turbines with variable X killed disproportionately more birds of species Y." What magnitude is implied by "disproportionately"? With a large enough sample size, it could be a biologically insignificant increase that is likely just a result of confounding. This issue of magnitude is addressed in Table 7-5 (p.215), but the percent magnitudes still need to be put side-by-side with real numbers to make them more meaningful*.

Were the reviewers provided our full report? We reported all sample sizes used, along with corresponding measures of effect. All these data can be found in our Appendices C and D.

P28: *pp.207-209, Figures 7-19 through 7-21: There appears to be considerable spatial clustering of the golden eagle, red-tailed hawk, and burrowing owl fatalities. The variation in duration of study does not coincide with the clusters. Similar spatial clusters appear in all three figures. There is no discussion of this in the narrative. Are these clusters the result of turbine type clustering, variation in elevation, concentration of avian habitat, or some other factors*?

The tables and text following these figures summarize the factors underlying the spatial clustering of the fatalities. Each species was associated with a different set of variables, and some variables were common among species. We do not understand how the reviewer can claim there is no discussion of this clustering in our report. Numerous times we reported particular species killed disproportionately by wind turbines in canyons, by wind turbines at the edges of local turbine field, by wind turbines in sparse turbine fields, by wind turbines in areas of intermittent rodent control, and in other situations that resulted in spatial clustering of fatalities.

P29: *pp.210-219, Tables 7-4 through 7-7: Percentage increases in mortality are listed for various species in association with 12 factors. Confidence intervals should be provided for each of these percentage values so that the precision of the estimated effect can be evaluated. How many of these confidence intervals would include zero, indicating that the magnitude of the effect might plausibly be zero*?

The percentages in the table were only presented as measures of effect, and we treat them as indicators. However, we agree it would be more informative to include confidence intervals, and we would try to do that if given the opportunity to revise the report.

P29: *p.219, par.1 and pp.220-221, Figures 7-22 and 7-23: The authors note the seasons with relatively higher fatalities than expected but neglect to point out the seasons with unusually lower fatalities than expected. Specifically, the red-tailed hawk, American kestrel, and burrowing owl all show much lower fatalities than expected in the spring. Why would this be true? Similarly, there were no fatalities of mallards in the fall. Why would this be so*?

Bird species in the APWRA vary in abundance and behaviors seasonally, and the wind turbine operations also vary seasonally because the winds vary seasonally. Red-tailed hawk, American kestrel and burrowing owl nest during spring, so they are spread out all over their geographic ranges defending nest territories. American kestrels and red-tailed hawks congregate in the APWRA during winter, and burrowing owls are fledging out of the APWRA during summer and fall. We simply reported what we found, and we agree we pointed out the seasons when more fatalities were found and neglected to mention the seasons when fewer fatalities were found, but it is relatively easy for the readers to figure out which seasons fewer fatalities occurred.

P29: *222, par.2: "The empirical models developed were tested only against the database of the 4,074 wind turbines from which the data were obtained for model development," state the authors. Testing the quality of a statistical model on the same dataset from which it was developed is bad practice…*

It is common practice to test the quality of a model on the data set from which it was derived. Bootstrap analysis does this, and so do jackknife methods and other variance-exhaustion methods. Discriminant function analysis is often assessed by how well the model correctly classifies the data used to derive the model. We agree that a better assessment of models is to test them against additional data, but we disagree using the same data is "bad practice." It is certainly a better practice than not bothering to assess the performance of the model. If we had the opportunity to revise the report, not only would be strive to generate superior models – and we know we can because Smallwood and Spiegel (2005a,b,c) did – but we would also apply the better post-hoc assessment methods suggested by the reviewer.

P29: *p.222, par.3: This argument is independent of any observations made by the authors. It represents circular reasoning. It argues that if the model is correctly predicting which turbines are relatively more dangerous, then the reason no bird fatalities were found at most of these dangerous turbines is just that we did not look long enough. This might be true but this work can neither support nor refute it*.

See our response to R1:P15 regarding golden eagle mortality at turbines predicted to be more dangerous compared to those predicted less dangerous. Among the turbines we predicted to be more dangerous 9.45 times more golden eagles were found per turbine search. Many of these turbines were searched only twice during half of one year. Continuing to search these turbines will turn up more golden eagle carcasses, and more of the turbines in this group will have been documented to have killed golden eagles. Such is our prediction. It is not circular reasoning, but rather a prediction. Our prediction can, of course, be tested by actually performing more searches at the turbines.

P30: *p.223, Table 7-8: The authors have so far only conducted univariate chi-square hypothesis tests. They now seek to combine the results in an ad hoc fashion into a model which amounts to a scoring system. If the authors want to develop a multivariate model, they should apply appropriate methods such as logistic or Poisson regression*.

Our combination of results was not ad hoc. The first step was methodical testing of all the variables, and the next step was a screening of the variables for use in developing the predictive models. We decided to keep the model simple because, which is why we adopted a scoring system. However, whereas we believe Poisson regression would not be an appropriate method, we would be happy to try logistic regression if given the opportunity to revise the report.

P30: *p.224, Table 7-10: The authors' interpretation of the results presented in this table is unusual. They group the observations by the results (e.g., 0, 1, 2, 3, etc. fatalities) and compare the fractions that were predicted to be "more dangerous" and "less dangerous". This is a backwards approach to evaluating the predictive model. The observations should be grouped by the predictions (not the results) and the percentages of each group that experienced fatalities should be compared.*

For example, using the golden eagle data, we can assemble a 2 x 2 table of relative risk:

	Predict 0 fatalities	*Predict ≥1 fatalities*
Observed 0 fatalities	*2007*	*2014*
Observed ≥ 1 fatalities	*10*	*43*
Total	*2017*	*2057*
% with fatalities	*0.5%*	*2.1%*

The reviewer's approach would be appropriate if the observations were numerous and the search effort uniformly applied, but our observations were rare relative to the number of turbines searched and the search effort was far from equivalent among turbines. Therefore, there is no value in placing the percentage observations under the predictions; doing so would be misleading. Also, the reviewer's headings representing the predictions are incorrect. Our models did not predict whether 0 birds would be killed or ≥1 would be killed. Our predictions were much cruder than that; they were whether the turbine was less threatening or more threatening to birds. It was a qualitative prediction, not a quantitative one.

So although the turbines predicted to be more dangerous were about 4 times more likely to experience fatalities than the turbines predicted to be less dangerous, 97.9% of those predicted

to be more dangerous experienced zero fatalities. On p.222, par.3, the authors argued that this large rate of false positives is attributable to the short duration of sampling. If so, then the turbines studied for 4 or more years should show a stronger response. Is this effect stronger for the turbines studied for longer periods?

Yes. The reviewer's table is reproduced below (headings are corrected) with data from wind turbines searched 4 years.

	Scored ≤0	Scored >0
Observed 0 fatalities	296	145
Observed ≥ 1 fatalities	2	43
Total	298	157
% with fatalities	0.67%	8.3%

The percentage of turbines associated with actual golden eagle fatalities was 12.4 times larger among turbines predicted to be more dangerous compared to those predicted to be less dangerous. The effect was considerably larger than including all turbines searched over various durations, and this difference supports our conclusion on page 222, paragraph 3.

P30: *p.226, p.229, p.231, p.235, Figures 7-24, 7-26, 7-28, 7-30: In the A part of each of these figures, the authors have again grouped the observations by the results and not the predictions. Since they are attempting to evaluate the quality of the predictions, their approach is inappropriate. Like residual plots for logistic regression, the observations should be grouped by prediction (ranges of the scores) and the fraction of turbines experiencing fatalities should be compared among the prediction groups*.

We disagree with the comment for the reasons given in our response to the preceding comment. Given the statistical rarity of the event and the differential search effort, it would be misleading to implement the reviewer's suggested approach.

P31: *p.242, par.2: The authors claim that elimination of 20% of the turbines might reduce the mortality by 80%. How was this determined*?

First, we made no determination and no "claim" that the elimination of 20% of the turbines would result in any percentage reduction in mortality. We presented our conclusion as an educated guess, not a determination or a claim. Our actual statement was the following, "*We can explain only a fraction of the variation in bird fatalities caused by wind turbines in the APWRA. All birds lumped together (and assuming additive effects from the factors entered into the model), the elimination of 20% of the wind turbines might reduce mortality on the order of 40%.*" Secondly, we guessed the measure might reduce mortality by 40%, and not 80% as claimed by the reviewer.

P31: *p.243, par.2: The authors state that the Bonus, Micon, and KVS-33 turbines are the most dangerous. How was this determined? It is likely the authors intended to include the KCS-56 instead of the KVS-33 based on the total bird fatalities reported in Table D-3. Is it possible that*

there are more fatalities for these turbines because there are more of them, not that they are more dangerous per unit?

In our report we explained how we accounted for differential sampling effort while analyzing the data. It is inappropriate to compare numbers of fatalities among wind turbine models without any regard to the differences in fatality search effort.

P31: *p.245, par.2: The authors state that wind turbines that are at the end of strings or are isolated kill more birds that wind turbines on the inside of strings. It is important to keep in mind that carcasses tossed far enough by a wind turbine that is on the inside of a string can be misattributed to either its left or right neighbor. Wind turbines at the end of a string can only have their kills misattributed to another wind turbine only if it tossed towards the string. Wind turbines that are isolated will not have any chance of getting their carcasses misattributed.*

We were aware of these possibilities of misattribution. On page 45 we wrote, "*...if a bird is killed by an interior turbine, its carcass is likely to fall to either side and to be associated with the neighbor tower; whereas, the end tower only has one neighbor for such a mistaken association to be made*."

P31: *p.246, par.4: Biologists only collected bird behavior data from mid-October through mid-May. What about mid-May through September, especially since summer is when the winds are strong? Perhaps young prey or different types of prey are available more during certain months?*

We were unable to deploy our behavior observers over a full year, so we did the best we could. Indeed, we missed making measurement of behaviors and activity levels during the summer months, which is a shortfall. Our results should be considered as representative of fall, winter and spring during one year, and therefore exploratory in nature.

Also, how were the 61 observation plots selected: randomly or by convenience?

On page 246 we wrote, "*The study plot boundaries encompassed wind turbines easily visible to the observers from a fixed observation point, resulting in a mosaic of irregular shaped, non-overlapping plots (Table 8-1). These 61 plots covered all of the area studied during the behavior research performed under funding from the National Renewable Energy Laboratory (Smallwood and Thelander, in review)*." These plot boundaries were contiguous and included all the Set 1 turbines, so there was no selection, per se.

P31: *p.247, par.2.: The observation plots had a fixed radius of 300 m, so the term variable distance circular point observations is not really appropriate. Variable-radius plots are more commonly used in so-called "distance based sampling" in which the distance to each bird observation is used to estimate probability of detection as a means of calculating bird density (which is not the intent of the authors). The authors did assign birds to one of 3 distance categories (based on distance to turbine), but the furthest category was truncated at 300 m. As Reynolds (1980) states, "With the variable circular plot method no maximum distance restrictions are placed on any observation" (p.310). "Distance-based sampling" is a large sub-*

discipline within wildlife ecology and boasts a sizeable literature (see Volume 119 Issue 1 [2002] of The Auk for several recent papers on this subject), and while Reynolds et al. (1980) is a classic citation and influential in the development of current methods, it is not up-to-date with recognized methods.

We used the wrong description of our sampling method, and we caught our error a year ago during the second peer review of our report. We should have referred to it as 360˚ visual scans. We would correct this error if given the opportunity to revise the report.

P32: *p.247, par.3: The authors state that the 61 observation plots were sampled 4 times each or "once every three to four weeks". How can the sampling cover 210 days and at the same time be once every 21 to 28 days? With one sampling at the start and one at the end, the interval between samplings would need to be about 70 days*.

We tried to maintain a rotation of 3-4 weeks, but it did not always work due to wind and rain. Observers had to leave during strong winds, and there were a number of periods when driving the service roads was impossible or considered too dangerous due to soil saturation.

P32: *p.250, Table 8-2: More explanation is needed to distinguish the types of flight behavior in Table 8-2. Contouring and surfing sound alike*.

If we were to revise the report we would add more detail to define these terms.

P32: *p.251, par.1: The authors assume that, "the number of on-the-minute observations represented the same number of continuous minutes of the same activity." This is a standard assumption with conventional wildlife behavioral sampling, and is likely valid if sample sizes are large enough. This issue has been discussed extensively in the literature (see classic book by Martin and Bateson, 1993), the authors should make use of citations on the subject and defend that the assumption is valid. Also, they should identify their sampling technique within the conventional behavioral sampling lexicon – i.e., there are very standardized differences between focal animal sampling, scan sampling, and instantaneous sampling. The authors likely did the latter, but they should review these terms and identify which best describes their approach*.

We would follow this recommendation if given the opportunity to revise the report. We will do it anyway when we prepare manuscripts for journal submission.

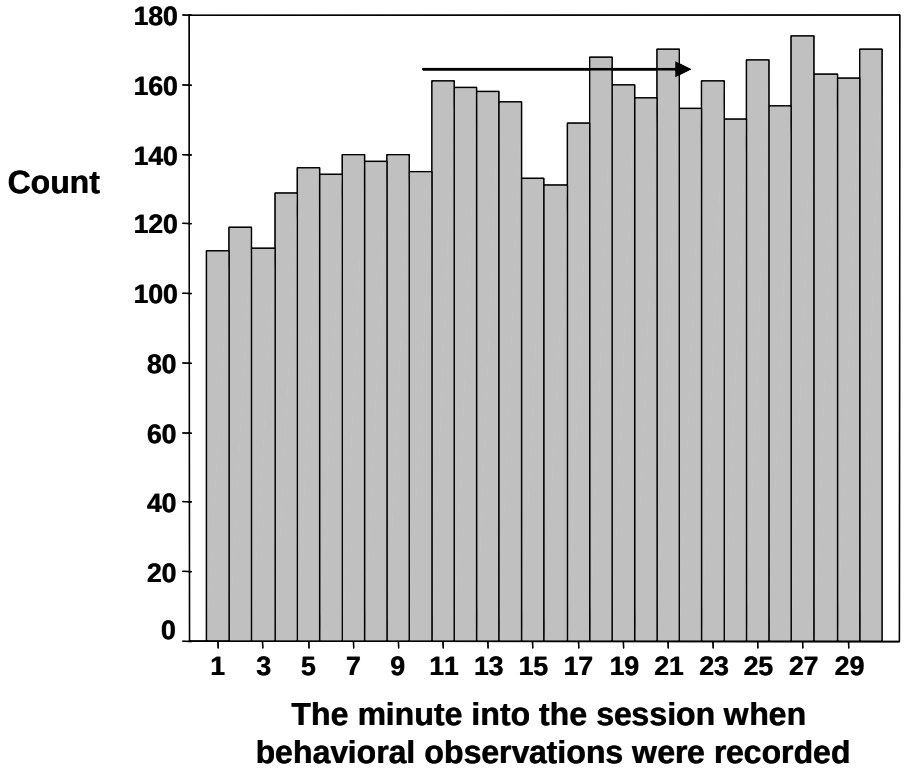
P32: *p.253, par.5: Chi-square tests are performed to test for disproportionate behavior under various conditions. Observations (data points) used in a chi-square test should be independent of one another. Having a single bird provide multiple observations through time removes that independence, thus invalidating the chi-square analysis. If a bird is soaring one minute, it is more likely to be soaring during the next minute. Even if a bird only contributed one observation; it could be recounted as a new bird if it disappeared for only 30 seconds (p.247, par.5)*.

We were aware of the issue of lack of independence, but we were also aware that chi-square analysis has often been used in behavior studies involving sequential sampling and instantaneous

sampling. Acquiring truly independent observations of bird behavior in a study site would be impossible, yet independence of observations is an assumption underlying most statistical tests. Biologists often deal with this issue by acknowledging the violation of the assumption while performing the test anyway. We are open to suggestions how to analyze these data despite the likely non-independence of observations.

P32: *p.260, par.3 and p.260, Figure 8-9: The authors state that an asymptote for some behaviors is reached by about 9 minutes and for others by 20-27 minutes. It is not clear what asymptotes they are referring to. The vertical axis on Figure 8-9A does not include zero, which exaggerates the magnitude of the change. Why did the frequency of behaviors increase with time? Does this suggest birds took some time to habituate to human presence (as suggested by Reynold et al. 1980 and others)? Or does it mean in took 8-30 minutes for observers to begin to fully "notice" (authors' term) behaviors in the observation plots*?

Below is the figure reproduced, this time with 0 included on the Y-axis. What we meant by asymptotes are indicated by the arrow in the case of this graph. We meant that these are minutes into the sessions when counts of birds stopped increasing. We believe the increase in counts reflects the time it took for birds to acclimate to the observers' presence, but it is possible the increase also reflects observer bias.



The term special behaviors is inadequately defined.

On page 256 we wrote, "*Special behaviors included entry into and exit from study plots, as well as landings, diving, mating, flying through the wind turbine string, and a few others.*" We suppose we could have listed out the "few others," and if we were to revise the report we would do so.

P33: *p.256, par.5: The authors absolutely did not observe 855 minutes of flying; they recorded 855 incidences of flight among 3884 observations at minute intervals. There is a difference between these two. This is a problem with equating minutes of an activity with frequency of its observation at 1-minute intervals*.

As earlier pointed out by the reviewer, we explained what we meant by "minutes of flying" on page 251, paragraph 1. We stated an assumption, which made it easier to report our results. However, the observers *did* observe about 855 minutes of flying, even though they were recording observations on the minute.

P33: *p.256, par.6 and p.262, Figure 8-11: The authors state that Figure 8-11A shows the relationship between the number of flights through the rotor zone and the total number of flights observed during a session. What is the slope, r^2 value, or standard error estimate for the relationship? Is this a chance pattern? Regardless, it makes sense that if there are more incidences of flight, there will be more incidences of flight through the rotor zone. And if birds are perching – thus not flying – there will be fewer incidences of flight through the rotor zone*.

We decided in this case to make a qualitative comparison, and not a quantitative one. We agree with the reviewer about sensible conclusions.

P33: *p.264, par.2: Were any bird collisions with turbine blades observed?*

Yes. Smallwood reported to the National Renewable Energy Lab his observation of a rock dove struck by a wind turbine blade. His was the third report of an observed collision submitted to NREL. We are aware of 2 observations since then.

P33: *p.265, Table 8-3: The table totals for the sum of minutes of flying (855) does not match the total of the column (828). Are there other raptor results not tabulated?*

Yes. Had we the opportunity to revise the report, we would add a row to tabulate the observations of unknown species of raptor.

The turkey vulture, red-tailed hawk, and American kestrel account for 87% of the minutes flying and 90% of the flights through the rotor zone, but according to Table 3-1 they only account for 22.9% of the total turbine caused fatalities and for 58.1% of the total raptor fatalities cause by collisions. Why this great disparity?

We do not know.

P33: *p.266, Table 8-4: In this table there are several behaviors or groups of behaviors that have zero recorded minutes of activity for all listed species and yet three other flight behaviors*

listed in Table 8-2 are not included (e.g., high soaring, mating, and land). Why were these omitted?
We ended up lumping high soaring with soaring because we did not feel the observers were distinguishing between the two categories, and in retrospect we decided these categories were vague in their difference. Mating and landing ended up being considered "special behaviors" as explained previously, and are not entirely flight activities, anyway.

P33: *p.267, Table 8-5: There is a discrepancy between the minutes perching for American kestrels between this table (1065) and Table 8-3 (1103).*

In building Table 8-5, we tabulated 0 minutes of perching on transmission towers when we should have tabulated 38. If given the opportunity to revise the report, we would fix this error.

P34: *p.269, par.4: Many of the environmental variables may have coincidentally been correlated with when the birds were sighted. For example, "Golden eagles and American kestrels perched more often than expected by chance during cooler temperatures, which was also more or less when they flew more often." So were Golden eagles and American kestrels mostly observed during the cooler months? Would that also cause an association with certain seasonal types of wind*?

Golden eagles were not seen more often during the cooler months, but American kestrels were. We do not know how to answer the last question because we do not understand it.

And how can they be perching more and flying more at the same time? Would not one increase while the other decreases?

Not necessarily. Flying and perching observations can both increase with increased numbers of the species in the study area.

P34: *pp. 270-275, Tables 8-6 through 8-11: The authors have again conducted 132 univariate hypothesis tests without correcting for multiple comparisons.*

pp. 283-307, Tables 8-12 through 8-16: This time there are 792 simultaneous tests conducted without correction for multiple comparisons.

We did not make multiple comparisons. We performed multiple tests involving multiple variables, of which we screened those appearing in our synthesis appearing in the Discussion section and Table 8-22 and 8-23.

REVIEWER 3

P1: *Smallwood and Thelander do an excellent job of describing background information and previous literature regarding bird mortalities at wind farms. A strong case is made for the relevance of the work they performed. Their objectives were multi-faceted and involved examining bird behaviors, raptor prey availability, turbine characteristics, landscape features, and bird mortalities. Each of these components is a substantial endeavor and the authors are to*

be commended for examining this multitude of factors. Clearly, the authors used a 'heads up' approach that involved making observations, and then attempting to collect data and evaluate hypotheses formulated from these observations. For example, they expanded their research to examine fossorial animal burrows and effects of rodent control on bird mortality.

We appreciate the reviewer's comments.

P1: *Even if the interpretations remain similar, I would have greater confidence if their analyses were improved upon. Furthermore, I would suggest extensive consultation with a statistician to assist in such an endeavor. It is my belief that a great deal of thinking and interaction with a statistician is needed to ensure that the information contained in this impressive data set is properly interpreted.*

We would certainly consult with a statistician and try additional statistical methods if we were given the opportunity to revise the report.

P1: *There was an opportunity to submit questions and have them answered by the authors. However, in my experience, the importance of the nonstatistical aspects of statistical consulting cannot be understated. For example, direct verbal interaction is the best means for arriving at a consensus understanding of material. Use of reflective listening is a very useful technique for making sure everyone is 'on the same page'.*

We agree with the reviewer that a superior means to arriving at an understanding of the project would have been through direct verbal interaction.

P1: *Strength of inference is determined by the study design. Strength of evidence is determined by the data alone. In this case, the study design does not lend itself to a strong inferential setting for two reasons. First, the variables of interest are likely confounded because turbine string placement was not designed with their study objectives in mind. Hence, there are many factors that potentially affect the response variables of interest that are not separately estimable because all combinations of explanatory variables are not represented on the landscape. For example, if one was only interested in the effects of aspect (north, east, south and west) and tower height (say 4 categories), then one would need 16 tower-aspect combinations represented, with replication, for sufficient estimability of main and interactive effects. Thus, this study lacks replication of the set of 'treatment' combinations of explanatory variables being examined (see Johnson 2002 for a discussion on the importance of replication in wildlife research).*

We agree, which is also why we tested one variable at a time, rather than attempting to perform multivariate tests including tests for interaction effects. However, we believe we achieved sufficient replication of most of the "treatments," but the shortfall was in the number of fatalities. The sample size of fatalities was too small to be spread among the many different treatment combinations that would be used to search for interaction effects. A study of this nature would need to be performed for a longer period of time and with less time between fatality searches in order to obtain a sample size of fatalities useful for the types of analysis expected by the reviewers.

P2: *Second, because the sample of sites studied were not randomly sampled any inferences based on statistical tests are not statistically defensible*.
We partly disagree with this comment because we did not have the opportunity to exercise investigator bias, which is the reason random sampling is important to inferential statistics. Of those turbines available to use through September 2001, we simply selected them all, so there was no opportunity for bias on our part. The Set 2 turbines were selected systematically and in ignorance of which turbines would yield more or fewer dead birds, and three quarters of these were selected. Where bias could have affected the study, however, was in the wind turbine owners' incremental granting of access to their turbines. The study would be biased, for example, if the wind turbine owners initially kept us away from the wind turbines they felt were most dangerous to birds. Unfortunately, we suspect this bias might have real, and may have affected our mortality estimates and fatality associations. It is in this respect we agree with the reviewer -- that the underlying reason for random sampling was still an issue for us even though we searched most of the turbines we were granted access to search.

P2: *That the authors have surveyed a majority of turbines provides some support for the notion they have a good representation of the population, however, only approximately 28% of the turbines were measured at least 3 years. If the authors can make the case that their sites are representative of the larger APWRA, then perhaps the scientific inference (not statistical) being made will be acceptable to all*.

We agree our scientific inference was strongest among the wind turbines we searched the longest, and weakest among those we searched only twice. We did not attempt to extrapolate our results beyond the APWRA, and we cautioned against doing so. Within the APWRA, we interpreted each test result specifically to the measured set hence it was derived, and we left to the reader the decision about how far beyond the measured set to draw inference (we provided sample sizes, expected values, observed values, and levels of significance). We expanded our inferences from some of the patterns that appeared stronger or more consistent, but we do not believe we drew strong inference from any single test result.

P2: *In their executive summary, they estimate the number of raptors (and all birds combined) killed annually in the APWRA. I personally, would refrain from making such inferences and would limit estimates to the area specifically surveyed, given the strength-of-inference limitations due to nonrandom sampling and the uncertainty in the 'adjustments' they make in their estimation process*.

Our mortality estimates were extended to the 1300 unsearched turbines, but we also provided the reader the means to restrict mortality estimates to only those turbines we searched, as well as to Set 1 turbines which were searched longer. We were careful to provide the reader the means to decide how far to draw inferences from our mortality data.

P2: *Their projections for all wind-generating facilities in the United Stated (page 86) should also be considered as extrapolations without much credence*.

After performing additional research into fatality comparisons among wind farms (see Attachment B), we agree. If given the opportunity to revise the report, we would replace our Chapter 4 with a more rigorous, critical treatment of the issue.

P3: *They further state that the risk to birds has increased substantially over the past 15 years, indicating a formal trends analysis. This support for this statement is not satisfactory unless more information is given on consistency of detection rates over this time period. Note there are many factors that can cause inconsistencies in observed counts over time, including surveyor differences, environmental differences, and animal behavior differences. The authors did suggest that birds may have altered their behavior in response to the presence of turbines in the area. I suspect that there are many survey methodological differences over this time span as well*.

After performing additional research into comparisons of bird utilization among wind farms (see Attachment B), we agree. If given the opportunity to revise the report, we would replace our Chapter 4 with a more rigorous, critical treatment of the issue.

P3: *Another set of turbines (2548) were surveyed over a 6-month period (see p.47; however, they later allude to the notion they have over a year of data for these turbines on page 76, second paragraph)*.

We made no such allusion. We inadvertently deleted "Seawest" from the phrase "we had not yet completed a full year of fatality searches on these turbines." We were attempting to discuss the Seawest turbines, and not the 2,548 turbines that were added at the end of our study.

P3: *If I understand their methodology, the authors computed estimates of availability that account for survey effort by placing landscape and turbine features on a relative (proportional) basis. However, in most use-versus-availability studies I have seen, 'availability' is assumed known, when it is almost always estimated. What is 'available' from the human perspective of what is on the map, is not necessarily available to the animals of interest, even if they are highly mobile, because they may not have such a map in mind when making decisions. Alldredge et al. (1998) provide a nice overview of statistical approaches to resource selection studies that nicely clarifies the set of assumptions underlying such analyses. If the authors do not modify their analytical methods, which I strongly recommend, then at the least they could more explicitly state the assumptions underlying their approach, the likelihood that the assumptions are valid, and the ramifications if not valid.*

We appreciate the warning that placing landscape and turbine features on a relative or proportional basis expresses our perception of this basis, and not necessarily the animal's perception of it. We would enthusiastically follow the reviewer's recommendation to explore this and other ideas addressed by Alldredge et al. (1998) if given the opportunity to revise the report. We will do so, anyway, as we prepare papers for journal submittal. As for the assumptions we relied upon, on page 186 we cited Smallwood (1993, 2002) for deeper discussions of chi-square tests and use-and-availability analysis. If we were given the opportunity to revise our report, we would explicitly list assumptions underlying our tests.

P3: *Another important design component includes consideration of what the multiple competing hypotheses are and how best to discriminate among them. When possible, readers of this report should be informed of the hypotheses under consideration and how the sampling scheme used can discriminate among these hypotheses. For example, there is an entire section of work in chapter 2 that examines the distances of 'small' versus 'large' birds from the turbine, yet there is no explanation of why the data are being partitioned as such, i.e., what the hypothesis is, and how this partitioning relates to assessing the efficiency of their search radius. Another example of the importance of considering one's hypotheses is demonstrated in chapter 6. The objectives are clearly stated, but the a priori hypotheses regarding the effects of rodent control are not stated. They allude to the notion of ineffectiveness, but I would like to see explicit hypothesis statements. Lacking the benefit of observations, my scientific hypothesis would be that increased intensity of rodent control results in fewer raptors, thus lowering susceptibility to strikes and thus lowering observed fatalities.*

During report preparation we debated listing explicit hypothesis statements. We decided not to list them because the report was already very long. We assumed the reader would recognize the hypotheses being tested based on the statistical test used, and we hoped that our provision of more than the usual information about each test (Appendices C and D) would help the reader understand each hypothesis. Some hypotheses were indeed left vague, as the reviewer pointed out, so if we were given the opportunity to revise the report we would strive to clarify our hypotheses.

P4: *How best to discriminate among hypotheses is an important design consideration. For example, in studying the effects of rodent control, the authors did select sites with a wide range of observed raptor mortality and rodent control intensity. This approach enhanced the ability to distinguish among competing hypotheses. However, they did not random sample turbine strings according to these features (e.g., a stratified design), thus limiting the defensibility of inferences made.*

We agree our inferences from this portion of our study are limited, but we still learned something. For example, raptors continued to visit selected study sites where almost no ground squirrels remained. This is an important finding, and goes to the underlying question of whether raptors forage by gestalt or by enumeration of potential prey items. We think we learned that eliminating prey items from the landscape may not reduce raptor fatalities in the APWRA. We feel more secure in coming to this conclusion, based on evidence, than we would feel speculating raptor mortality can be reduced by eliminating rodents from the APWRA. In other words, the evidence used to conclude rodent abatement should proceed was much weaker than the evidence we relied on to conclude rodent abatement may not achieve its objectives.

P4: *Care must be taken when attempting to demonstrate 'treatment' effects. For instance, the treatments must be effective in their application. My understanding is that rodent control was aimed specifically at eliminating ground squirrels, but not other species (e.g. pocket gophers). Clearly, if one species is targeted, that does not necessarily imply a significant reduction in overall prey availability, in fact, it may increase it. Thus, I question if the rodent control 'treatments' were substantial enough to observe an effect.*

The rodent control effort we described in our report was specifically for ground squirrels, although other rodents were undoubtedly killed, and so were desert cottontails and perhaps other species. Grainger Hunt and Richard Kerlinger have for years advocated specifically for ground squirrel eradication in the APWRA and that is precisely what the wind turbine owners paid Alameda County to accomplish. In some areas there was no question the effort was effective. Where ground squirrels had once thrived, no ground squirrels remained alive – none. At locations where we were familiar with high ground squirrel activity levels, including scurrying animals and warning calls, we found silence and the smell of death following the treatment. And ground squirrel burrows were collapsing for lack of upkeep while we conducted our study. We have every confidence the rodent control program was substantial, and so did Alameda County.

We disagree with the reviewer's suggestion that other species might have increased in response to the ground squirrels' decline, because there were no other species in the APWRA that strongly compete with ground squirrels for forage or space. We also found no supporting evidence for this suggested outcome. What we believe happened, and we discussed this in our report, is the remaining species were more prominent to foraging raptors. Instead of hunting over ground squirrel burrows, which tended to be lower on the slopes than where wind turbines are installed, but which collapsed to disuse after eradication, raptors might have been drawn to visible signs of pocket gophers and desert cottontails, which were more abundant along the ridges and around wind turbines.

P4: *In summary, I suggest that the efficacy of the management actions taken be considered before discarding their usefulness. Similar arguments apply to topics such as benefits of perch guards, etc. Smallwood and Thelander have made decisions and recommendations based on observations from considerable survey effort. Again, I believe there is tremendous value in their data, interpretations of which must be carefully considered for that value to be realized.*

Again we will argue that our conclusions and recommendations regarding mitigation measures were based on evidence, whereas conclusions to implement the measures in the first place were based on speculation, anecdotes, and simplistic deductive reasoning in the absence of evidence. We were more prepared to make our conclusions than were those who concluded implementation of these mitigation measures was warranted. The burden of supporting evidence should be greater for those who decided to kill many thousands of animals per year in an effort to reduce raptor mortality. Our conclusion was merely questioning whether the abatement effort really worked, and whether in the process it was really worth the takings of four species protected by the federal Endangered Species Act.

P4: *Regarding strength of evidence, the authors have completed a notable amount of work. Having surveyed the majority of turbine locations, some of which were surveyed multiple years, I believe there is substantial information to be discerned from the collected data. The key to harvesting that information is proper context and hard thinking about what metrics make sense, appropriate use of statistical tests, placing outcomes of statistical tests in the context of biological relevance, etc. In attempting to determine causal factors of bird mortalities, the authors surveyed locations around turbines in the APWRA that were accessible. They identified all known bird strike mortalities (approximately 1045 if the number due to unknown causes is removed) and examined associations of various turbine/tower, landscape, and environmental*

factors with these mortality counts. Thus, a retrospective observational study has been performed with the intent of determining causal relationships. I believe most statisticians would agree that establishing causation requires a more rigorous approach in study design. Romesburg (1981) stated that causation requires more than correlative evidence; one must eliminate other possible causes and must demonstrate similar associations that are plausible over a wide range of circumstances. Many conclusions stated in the report are plausible, but I am not convinced that causation has been established anywhere in the report.

We agree with this comment. We did not intend to conclude causation through correlation or association, but we did intend to identify patterns that suggest potential causal relationships and that could be productive in follow-up research. The final paragraphs of our report, in Chapter 9, honestly conveyed our level confidence in our conclusions, and they explained that more research will be necessary to answer many of the questions we addressed. In our Chapter 1, we our third project objective was to "*identify possible relationships between bird mortality and bird behaviors, wind tower design and operations, ...*" [underline added for emphasis] We did not attempt to convey the notion that we were identifying causal relationships.

P5: *If I were considering this report for publication, I would reject it in its current form, but encourage the authors to rethink, revise and resubmit as a new manuscript in the future.*

Our report was not a journal submission, and we would never prepare a journal submission this way. Reports to agencies are different than journal submissions, including a different audience and different objectives. In our case, the "editor" accepted the report after a peer review and our minor revisions, with the understanding that from the report papers would be prepared for publication in scientific journals.

P6: *In describing their approach, the authors state they presented mortality estimates as ranges, where the lower end was adjusted for likely outside of their search area, and the upper end was adjusted for fatalities missed due to undetected carcass removal. I would consider both of these to be upper-end adjustments, actually using both simultaneously would provide a higher upper end. The lower end would be represented by unadjusted values. The upper-end estimates must be interpreted with caution.*

We provided the means for the reader to decide for themselves which estimates and adjusted estimates to use as the ranges. Of course, as the report's authors we were expected to present the ranges we felt to be appropriate. The unadjusted estimates are the obvious low end of any range of estimates, but most researchers of bird and bat collisions in wind farms would not accept the estimates adjusted by searcher detection and scavenger removal rates as the high end. Our report did not even address other sources of error and likely biases, most of which would increase the estimates. Crippling bias was not addressed and neither was the human removal of carcasses from our study. Attachment B discusses these biases and error at greater length.

P6: *Justification for their defined metric of mortality as mortalities per megawatt (MW) per year is not properly stated. They give the reason of 'to avoid the false appearance that larger turbines kill more birds'. As with any metric, the variable of interest must be clearly defined. If total number of fatalities at a site is the variable of interest, then neither rate is appropriate. If*

one wants to compare deaths as a function of turbine size, then fatalities per turbine is an appropriate metric and does not give a 'false appearance'. By incorporating the MW produced by each turbine, they have simply factored in the benefit of generated power in this cost representation. There are advantages of using this metric, many of which are stated in appendix A, but the advantage depends on how the metric is to be used.

We disagree our justification for measuring mortality as the number of fatalities per MW per year was *improper*. We do not think there is a "proper" way to express mortality, but we have opinions about which ways are superior to others. Also, we did not factor in MW as a means to factor in cost per fatality. We did not imply this anywhere in our report, and we certainly did not intend for this interpretation. Factoring in MW combined the turbine's rotor-swept area with the turbine's expected operation time. Unfortunately, we did not know the turbines' operation times between fatality searches, so we used MW as a partial surrogate for operation time. In Appendix A we showed quantitatively and graphically why it makes more sense to express mortality in terms of MW instead of per turbine. In short, it makes no sense to compare fatalities per turbine between 40-kW turbines and 2.5-MW turbines. We agree the total number of birds killed by the wind farm is informative, but only when one knows the size of the wind farm involved. And we point out that we provided mortality estimates as total numbers of fatalities, as fatalities per MW per year, and as fatalities per turbine per year. We did this so the reader can use whichever version preferred.

P6: *The authors state that at least 3 years of carcass searches are needed before stabilization of the percentage of non-zero mortality values. Are they saying that if a sample of 100 turbines is surveyed, at least 3 years are needed to estimate the percentage of those 100 that kill at least one bird? I am not convinced this is a useful metric. Rather than focusing on the turbines where zero, or even an occasional mortality occurs, should not the focus be on those characteristics at turbines where numerous mortalities occur (e.g., see figure 3-4). They proceed to interpret this result by stating that one must survey at least 3 years before getting a 'good' estimate of mortality rate. Mortality rate (expressed by fatalities per turbine per year) is not the same metric as percentage of turbines with at least one fatality. While I agree that more data is better for estimation in general, they have not demonstrated 3 years of data are necessary for a 'good' estimate of the mortality rate. The term 'good' in the context of bias would require knowledge of true mortality rate. The term 'good' in the context of precision would require some definition of what precision is needed for the estimates to be useful.*

We did not claim certainty in our estimate that three years should be minimal for making sound mortality estimates. We agree the percentage of turbine strings with at least one fatality is not the same as the number of fatalities/MW/year, but we do believe these two metrics are probably related. We think it is likely that even more years than three will be needed before mortality estimates can be made with reasonable precision. More research will be needed before any of us can conclude how long fatality searches should be performed to reasonably represent mortality, but we'll stick our necks out by predicting at least three years will be needed.

P7: *Important definitions are made regarding their usage of terms like susceptibility, vulnerability, etc. For instance, vulnerability is measured here on a relative, not absolute basis.*

We do not understand the comment.

Several questions came to me as I read the material. For instance, how is habitat use measured? -this is a very important question when interpreting results.

We actually did not measure habitat use. We used methods from habitat use-and-availability analysis, but we did not measure habitat use. The closest we got to measuring habitat use was associating bird observations with topographic features, slope aspects, and other geographic variables.

How close does a bird have to be to the reference point (e.g., rotor) to count as use? The authors use the word 'nearby' wind turbines, but I am uncertain what that implies.

We do not understand the questions. In chapter 8 we defined levels of proximity to wind turbines for this type of analysis, and we defined explicit distances from turbines. Chapter 1 was intended to be conceptual, but not to present our research methods.

Is flying over an area for a few seconds treated the same as when a bird perches or hunts in the same area over several minutes or hours? *Their phrasing suggests they consider the proportion of sampling periods in which use was detected, but this does not indicate duration of use per se. How do they treat observations of multiple birds at the same time? Are pairs treated as one observation?*

The answers to these questions can be found in the methods section of Chapter 8. Pairs were treated as two observations, not one. Chapter 1 was conceptual in its presentation.

P7: *On the bottom of page 9, the authors present a 'model' for vulnerability as the ratio of observed and expected use. I suggest they restate this as a metric, not a model. It is not clear why the Chi-square symbols are in the numerator and denominator of this expression*.

We agree. Had we the opportunity to revise the report, we would replace "model" with "measure of effect." The chi-square symbol was intended to indicate these are the same observed and expected values one would normally use in chi-square analysis.

P7: *Section 1.1.3 is nice section on the difficulty of measuring impact. I would like to know, however, how the number of mortalities per year in the APWRA compares to other hazards, such as collisions with vehicles or airplanes, or deaths due to poaching or contaminants. This would give the reader some perspective on the magnitude of impacts of strike mortalities in the APRWA. I realize that for some species, e.g., the golden eagle, car collisions are unlikely, but what about other human-induced sources of mortality*?

We have two general responses to this comment. First, our objectives did not include comparing wind turbine-caused mortality to other sources of mortality. Our objectives were to seek solutions to the mortality caused by wind turbines. Second, if we are going to compare mortality among the various potential sources, we need to do so carefully. Simply comparing estimates of total numbers, as Erickson et al. (2001) did, is unsatisfying because the comparison does not

account for the pervasiveness of the source. We think it would be more informative to compare species-specific mortality estimates on a human per-capita basis, and the way you get to that per-capita basis is to estimate how many people are served by particular projects or human activities. For example, it would be relatively straightforward to estimate how many people are served by 1 MW of power generated from wind turbines, or from gas-fired power plants, or from PV arrays, or inversely, the proportion of each power generation facility that goes to supporting each person. Then birds killed by the facility can be related to the number of people serviced. If similar transformations in terms can be made for auto traffic, house cats, glass-fronted buildings, and other sources of bird mortality, then mortality estimates can be compared among sources on a common metric. We point out, however, that this approach would be laborious and would be based on many assumptions, but we think this is the sensible approach if we are to compare body counts among fatality sources. A lesser effort can be misleading. For example, the number of birds killed by house cats is irrelevant to concerns over how many golden eagles are killed by wind turbines.

Another consideration is cumulative impacts. Even if we discovered electric distribution poles nationwide kill as many golden eagles as do wind turbines, we might still be concerned about the impacts of wind turbines because wind turbines are the new source of mortality. Just because some other source is equal or greater in its impact on a species does not mean the new impact is somehow insignificant; in fact, it may be all the more significant, especially if the cumulative impacts are considerable.

Section 1.1.3, introduces the notion that by comparing observed and expected frequencies, one is able to identify which environmental factors might have a causal relationship (see p. 12, 4th and 5th sentences of first paragraph). The term 'might have' is important, because this is merely an observational study and thus causation cannot be established.

We agree.

Section 1.1.4 introduces the idea of 'use versus availability' in terms of assessing mortalities and associations with turbine location by considering what percent of mortalities one would expect given random use of the sampled area versus the number actually observed. This reasoning is the basis for much of the statistical testing (Chi-square goodness of fit tests) presented later in the report. I question to what population is the statistical inference being made with these tests.

Technically, the population is the sample used in the chi-square test. For the highly significant tests, we would argue inference can be drawn to the birds using the APWRA at the time of our study. As the APWRA is changed through repowering or land use decisions, e.g., replacing cattle with sheep in Tres Vaqueros, inferences from our chi-square tests will be increasingly restricted to the sample we used in our tests.

I agree that by examining the observed/expected ratios, one can describe places where more or fewer mortalities occurred than expected with random use of the sampled area. However, is it reasonable to assume that birds use landscapes randomly?

No. We agree it is not reasonable to assume random use of the APWRA by birds. The random use was our null condition, not our expected condition. Furthermore, because this was a report to an agency, and not to a scientific journal, we described our expected null condition in a manner we felt the readers would better understand. Rather than the expected null condition being random, it really should be uniform. Uniformity is the mathematical null condition in chi-square tests. (But then the analyst needs to be careful because as animal species organize themselves through home range tenure and territoriality, observers might obtain regular patterns that look like the uniform pattern expected of the null condition.)

On page 20, the authors mention a focused study on bird behavior involving about 1500 wind turbines. Did they randomly sample these turbines from the collection of all turbines they studied? If so, then they could make inferences to the larger collection of turbines they surveyed, but again, I would suggest they resist the temptation to infer to the entire APWRA.

These study plots were not randomly sampled because they covered the entirety of the area in which all the Set 1 turbines were located. We did resist the temptation to infer to the entire APWRA, but we cannot claim 100% success in resisting the temptation.

P9: *My perspective is that the most useful data from this chapter are reported as the percentage of mortalities within their search radius, based on the relative number of birds found outside the search radius.*

We agree.

Their recognition that end towers may require a search radius larger than 50m to find 90% or more of carcasses in the 'world of the turbine' is valuable for future survey efforts.

That was our intended message.

P9: *At the top of page 30, they state a total of 1162 fatalities caused by collisions and by unknown causes were found. Table 2-1 identifies all 1162 fatalities as wind turbine collisions. Why are the unknowns folded into this column of the table*?

For the reason given by the reviewer a few lines previously: "Obviously, the fact that they searched around wind towers suggests they are predisposed to finding mortalities due to collisions than, say, due to natural predation or other factors (e.g., disease)." We assumed birds killed for unknown reasons were killed by wind turbines because they were near wind turbines. Many times it is very difficult to determine cause of death. Unless the animal was chopped in half or dismembered due to blunt-force trauma, it can be difficult to tell the blade actually hit the bird. This is a problem faced by all researchers at wind farms, and most conclude that carcasses found near the turbine were killed by the turbine.

P9: *By assessing the 'efficiency' of their search radius, I assume the authors are referring to what percentage of bird strike mortalities are contained within their search area.*

We were exploring how efficient a 50-m search radius was, and whether we could find nearly as many carcasses with a smaller search area, or by concentrating future searches to one side or the other of a turbine. We were also exploring whether a larger search area would be warranted for future research.

Clearly, the observation that carcasses were found beyond a 50-m radius indicates that their mortality estimates are underestimates of true mortality.

We agree.

I am curious as to why the authors did not expand their search radius to lessen this bias; however, I can appreciate that a larger search area would mean considerably more search effort that logistically may not have been possible.

The reviewer was correct to assume it was a budget issue, but also we did not want to change a standard method in mid-course. If we were to perform the study again, we would search a larger area.

I am curious as to what the detection rate may have been within their defined 50-m radius. The researchers could have directly estimated detection rates using various techniques (double observer, capture-recapture, removal approaches, or distance sampling if actual distances of carcasses were measured for each carcass).

We would add searcher detection trials if we were to repeat the study, but Smallwood has since reviewed all the available reports of such trials, and as a result we believe directed research is needed to improve the reliability of searcher detection trials (See attachment B).

Instead, they related distance to carcass as a function of bird body size, wind turbine attributes, season, etc. They later state (page 49 bottom) that they were unconcerned with underestimating mortality, yet they spend much of chapter 2 examining carcass distances to assess 'efficiency of search radius'.

We knew we were going to under-estimate mortality for a variety of reasons, including biases we were unprepared to handle. We were aware that crippling bias was a factor, and so was human removal of carcasses, and especially scavenger removal of small-bodied birds. We did what we could with the data in Chapter 2 to prepare future research in wind farms, but Attachment B makes the case that focused research on other biases and sources of error will be needed before accurate mortality estimates can be made. We simply lacked the funds necessary to tackle these issues of bias while also performing the study we were funded to perform.

What a priori hypotheses did they have regarding bird body size and distance from turbine? Given a clear association, how is that useful for determining detection rate?

We wondered whether wind turbines threw small- or large-bodied birds any farther from the turbine. We suspected they might, so we tested whether they did. We did not discover any trends in Chapter 2 that were worthy of submission to a scientific journal, but our reporting of

these results in Chapter 2 also reflects a difference between an agency report and a journal publication. In an agency report we can and should report everything we found, including the non-significant patterns and the mundane results because all these results can shape the next study. Whoever performs the next research effort in a California wind farm can read what we did and what we tried, and decide what not to try again.

P10: *The purpose of the arbitrary distinction of small and large body lengths in section 2.2 is unclear to me, as is any age classification. The analyses that followed were size-specific, but I do not understand the reason for such a partitioning*.

We used a natural break in the histogram of body length.

I also do not understand the statement that they lacked sufficient funding to factor in the slope of the hills from each wind turbine. Are they saying they could not afford a clinometer?

We own clinometers. The funding we lacked was to record the data and analyze them. When carcasses were found we did not measure the inclination to the turbine. If we were going to test whether inclination was a factor, we would have had to return to all the sites where we found carcasses and take the measurement. We suspect no researchers are currently measuring inclination to the turbine during ongoing studies.

I am curious as to how Pearson's correlation coefficient (p. 28 bottom) was calculated for assessing the linear association of carcass distances and elevation of tower base. A given tower base may have had multiple carcasses with multiple distances. Did they treat each of the carcasses as independent observations or did they compute an average distance for all carcasses at a given tower base?

We treated each carcass distance as independent. Only a minority of carcasses were found at the same turbine.

In section 2.3, the authors state on page 29 bottom that most carcasses were discovered during summer and winter. Is that because more surveys were performed then, or a greater abundance of birds were present, or a lesser number of birds were present, but they used the area over a much longer duration than passing migrants do in spring and fall?

The answer is a combination of factors, including the length of the season, our sampling effort, increased presence of most raptors in the winter, and some during summer, and stronger, longer-lasting winds in the summer which caused more collisions of some species.

The number of bird carcasses next to KCS-56 and Bonus turbines is drastically higher than all other turbine types. My question is 'is it the turbine type that predisposes it toward more bird strikes, or are there simply more of these turbines or that they were surveyed more often or are these turbines in places were birds are more abundant as a result of some other attribute, e.g., landscape feature?

It was because there were more of these turbines in the APWRA, and they were searched the longest due to the access we were granted. Fatality associations were presented in Chapter 7.

P10: *The ANOVAs reported in this section demonstrate statistical detectability of differences among means of carcass distances by tower height. I am not convinced that any of these analyses are useful given the purpose of this data collection. It was my understanding that the purpose of examining bird carcass distance was to 'assess the efficiency of their search radius'. Testing for differences in mean distances in not an effective approach to determining how large search radii ought to be at different tower locations. One needs to look at the distribution of the distances and/or actually estimate detectability of bird carcasses due to strikes with wind turbines. There are 2 levels of detectability here. The first level of detectability concerns what proportion of strike victims are beyond a prescribed search radius? The authors have collected information for estimating this proportion. The sentences that state 'Our search radius included 84.7% of the carcasses of large-bodied birds (90.5% for small-bodied birds) determined to be killed by wind turbines or unknown causes' are the most informative in this section, although I would eliminate the distinction of large and small bodied birds and eliminate the unknown cause counts. Figure 2-12 is also useful here in demonstrating that the 50-m radius contained approximately 95% of carcasses in most cases (the large variances for KCS is curious, the lone (extreme) observation for the Danwin turbine is also notable).*

Chapter 2 was mostly data exploration for the purpose of improving fatality searches in future studies. The most common result of our ANOVAs was weak significance to non-significance, meaning we did not find much reason to vary the search radius, which is what investigators are continuing to do as we respond to this very comment.

The second level of detectability pertains to within their search radius, what percentage of carcasses is found? Later in chapter 3, page 51, they state they adopted a searcher detection rate of 85% for raptors based on Orloff and Flannery (1992) and 41% for nonraptors based on two other studies. How valid these estimates are for the current study is unclear.

See Attachment B. Our mortality adjustments were probably valid for large raptors, but less valid for small raptors and small non-raptor birds.

P11: *Other concerns are 1) why was ANOVA used for the continuous explanatory variables? Regression modeling seems to be more appropriate (which they also report, but do not emphasize).*

Other reviewers thought our use of regression in this case was inappropriate. We performed both tests because tower height is indeed a continuous variable, but we did not have many heights to work with, so we also used ANOVA.

2) As mentioned previously, there is a potential for confounding effects given only one variable (e.g., tower height) is being examined in each analysis. For example, in the large-bodied bird section, if the 32-m towers were placed more often on sites with greater slopes, then the comparatively large 57m average distance may be a function of slope, not tower height (they recognize this shortfall on page 45 bottom, but fail to see that there are many factors, like blade

speed, that should also be jointly considered in analysis). I would suggest the authors consider regression analyses that include several pertinent explanatory variables, rather than a one-variable-at-a-time analysis approach.

We agree with the recommendation, but we did not have multiple continuous variables to include in the recommended analysis (not in the analysis presented in Chapter 2). Blade speed did not relate significantly to carcass distance, nor was it likely to anyway, because blade speed does not vary as much as many people seem to think. Blade speed remains relatively invariant while RPM decreases with increased rotor diameter, so larger turbines look like they are slower when they really are just as fast as small turbines.

I would also suggest that the authors consider the precision of the model as well as the assumptions underlying its use. If the precision is poor, or if the underlying assumptions are not met, I would not rely on the model and its estimates or predictions. To quote Michael E. Soulé, "Models are tools for thinkers, not crutches for the thoughtless".

We could not agree more. We felt that our single variable at a time approach forced us to examine the data much more carefully than we would have using multivariate methods. As for examining precision of the model, again we agree. We relied a lot on the root mean square error (RMSE) of regression models.

When categorical variables, for instance, rotor direction (upwind or downwind) or wind turbine location (end, gap or interior), are also of interest, then analysis of covariance would provide an improved analysis approach.

This is a useful suggestion, and we would use it if given the opportunity to revise the report.

The differences detected with the LSD tests that are reported are not biologically important in my opinion (e.g., the means ranged from 26m to 33m for large-bodied birds), nor are correlations meaningful when the sample correlation coefficient itself is near zero (see page 44, bottom). If the authors disagree, they need to make a case for why the differences are meaningful, that is they must identify what is a meaningful effect size.

No, we agree, and we said as much right there at the bottom of page 44: "*...although the correlation coefficient was not large*."

P11: *Finally, I do not understand the last sentence in the first paragraph of the discussion (page 45). Clearly, they do have an unknown proportion of actual carcasses given that carcasses were located beyond their search radius. Thus, their observed counts of mortality likely do not represent all strike mortalities over the defined spatial and temporal sampling period. They later clearly state this on page 49, bottom. Perhaps they did not intend to have the word 'not' in this sentence. The authors appropriately recognize alternative reasons why bird distances may be identified farther away for turbines at the ends of strings and on hills*.

We do not understand the comment. The word "not" was accurately used. The truth is we are confident in our conclusion we missed bird and bat carcasses within and beyond our search

radius. There must have been carcasses we did not see beyond our search radius, so our sentence was accurate, in our opinion.

P12: *According to a statement in the executive summary, mortality estimates should not be deemed reliable until 3 years of surveying has been conducted, use of this portion of data would seem to be inconsistent with this statement*.

This statement is not inconsistent. Here we told the readers we have low confidence in the mortality estimates derived from the Set 2 turbines, but we reported them anyway. We felt it is important to be comprehensive, and if we have data from 2,548 turbines we should report them even if the estimates are of low precision. We suspected throughout the study we had been denied access to turbines the owners felt were killing more golden eagles, and when we finally got access to all the turbines we found evidence our suspicions were warranted. We found many more eagle carcasses for the effort at these Set 2 turbines. We released our mortality estimates at these turbines knowing the precision was low, but we did so as an interim step until someone follows up with longer-term fatality searches at those turbines. We still believe 3 years of searches are needed until precise estimates are obtained, though we will give up this belief if additional data or superior interpretation convinces us otherwise.

For those turbines searched one year or more, temporal coverage is stated to be approximately 7 times per year. Two people searched for carcasses within 50m of each turbine and 50m beyond the end turbine. The authors state they did not estimate searcher detection and scavenger removal rates because they were unconcerned with underestimating mortality, yet later, they adopt corrective measures for these processes from other studies in estimating mortality, which is contrary to their prior stated indifference.

We were not indifferent to underestimating mortality. We were unable to obtain adequate on-site data useful for adjusting our estimates. Many of the scavenger removal and searcher detection trials used in other studies are implemented with little regard of the obvious biases affecting the trial results, so who is really indifferent to underestimating mortality. We were honest by not using project funds to perform the same biased-prone trials we see others using. We need directed research in this area.

They conclude by stating their mortality estimates might be conservative. However, I would suggest they may also be overestimates if the set of adjustments they made were not applicable to their surveys.

The reviewer lost us with this comment. The adjustments we used were conservative. Also see Attachment B.

P12: *An abundance of statistical tests were performed, testing for time variation in mortality (Tables 3-3 through 3-8). In conducting several hypothesis tests using a type I error rate of α = 0.05 (comparison-wise error rate), the authors are likely to have some null hypotheses rejected due to type I error. That is to say that some null hypotheses will be rejected even though they are true because the type I error rate for the entire collection of tests (experiment-wise error rate) will be much greater than 0.05*.

Type I errors are always possible. We listed our P-values, so the reader can examine each test result and decide on the likelihood of Type I error. Some of the P-values are much smaller than 0.05, and for these the likelihood of Type I error is much smaller. Among our test results in these tables, there was also the possibility of Type II errors.

P12: *Their metric for reporting bird mortality is clearly the number of fatalities per megawatt of power per year. The authors give previous mortality estimates from other authors, but these were reported in deaths per turbine per year. Hence the numbers from this study cannot be directly compared unless one knows the megawatts per turbine from other studies. So, I am a bit confused by the statement (page 47) that their purpose was to estimate mortality so that comparisons could be made to other sites. (I see later in table 3-12, their use of fatalities/turbine/year for these comparisons.)*

See Chapter 4 and Appendix A of our report. Our results are comparable using our metric. Almost all reports of bird and bat collisions with wind turbines now use our metric.

I would also be interested in knowing if their survey methodology differed from previous work. If so, then they should be cautious in making comparisons of observed mortality rates. For example, if their search methods were more thorough, then observed mortality differences may be due to detection differences, rather than actual mortality differences.

We agree caution is warranted. See Attachment B for additional reasons to be cautious.

P13: *On page 47, they state they were unable to search all turbine strings throughout the study or equally in frequency, so that time spans and seasonal representation varied at turbine strings. Again, I cannot blame them for logistical constraints, but they must take care in analyzing and interpreting patterns in data in light of the fact that they do not have a well-designed study in which all combinations of factors are represented with replication*.

We used large uncertainty ranges in our estimates, and repeatedly added cautionary statements about our estimates. We provided the means for the reader to apply their own assumptions and their own adjustments, and to do so for Set 1 turbines alone, or for Set 2 turbines alone, or for all of them together. What more would the reviewer have us do to be careful in our interpretations of the data?

On the bottom of page 48, in determining time since death, how much do weather conditions affect these estimates?

We do not know.

On page 51, the authors describe adjustments they made to their observed mortality counts. For instance, they state they adopted a searcher detection rate of 85% for raptors based on Orloff and Flannery (1992) and 41% for nonraptors based on two other studies. They further assumed scavenger removal rates (differential by small and large birds) based on Erickson et al. (2003). They added 10% to these rates for the second set of wind turbines. How valid any of these

estimates are for the current study is unclear. Thus, I am skeptical of their estimated number of fatalities for the APWRA given in the executive summary and presented in tables 3-10 through 3-12.

We are skeptical, as well, and our skepticism was reflected in our uncertainty ranges.

I suggest the authors concentrate on the observed mortalities in their study and consider optimal strategies for harnessing the information contained therein.

We think that is what we did, except we did compare our results to other mortality estimates in Chapter 4. Attachment B attempts to harness the information, as the reviewer suggested.

As an example, Figures 3-5 through 3-14 present means and standard errors of mortality estimates (per MW). The use of standard errors implies the authors are inferring to a larger population mean. Therefore, it is important to clarify to what population their interval estimates refer. If they are estimating the mean for the entire APWRA, then the same limitations of inference apply as stated before, given the nonrandom collection of sites that have been surveyed.

The unsampled turbines were a systematic subset of the APWRA. They did not differ from the turbines we searched, and they were interspersed among those we searched. There was no bias in which turbines we searched and which we did not search. We do not think our nonrandom turbine selection was important, and we feel that our target population was the APWRA. We had, however, other reasons to be careful about the inferences we drew.

I suggest the authors consider using descriptive statistics of mean and standard deviation (not standard error), which are informative in terms of describing their sample of sites surveyed.

We agree, and we would revise the report accordingly if given the opportunity.

This data shows that higher mortality occurred for red-tailed hawks and barn owls during the year 1999/2000, although the reasons for this are not stated.

Because we do not know the reason.

There was also more variability in observed mortality rates during this year. In some cases, consideration of the variability is just as interesting as a measure of center, so they might consider why there was more variation in mortalities this year.

We do not know why there was greater variation in mortality one year versus another.

By relying on descriptive statistics for their sample of sites, the hypothesis testing in tables 3-3 through 3-8 is not needed.

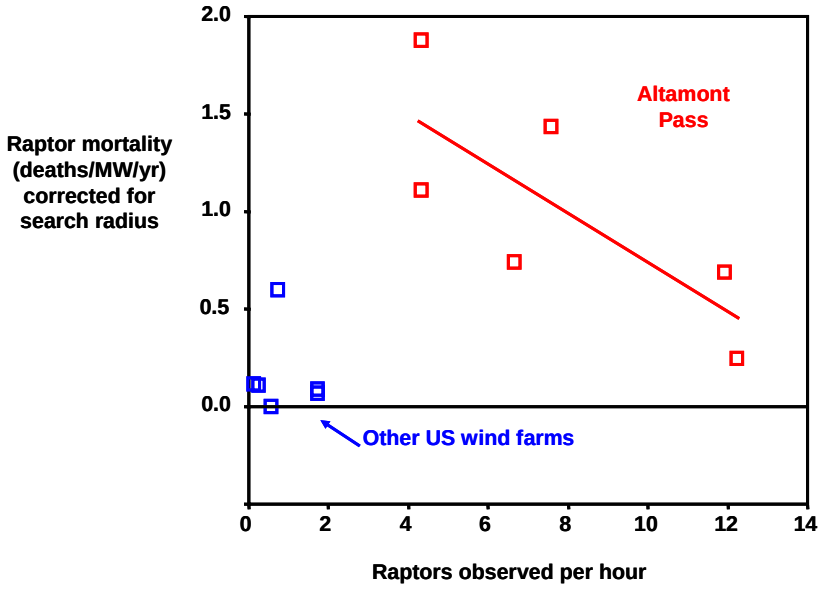
Good point. We would revise according if given the opportunity.

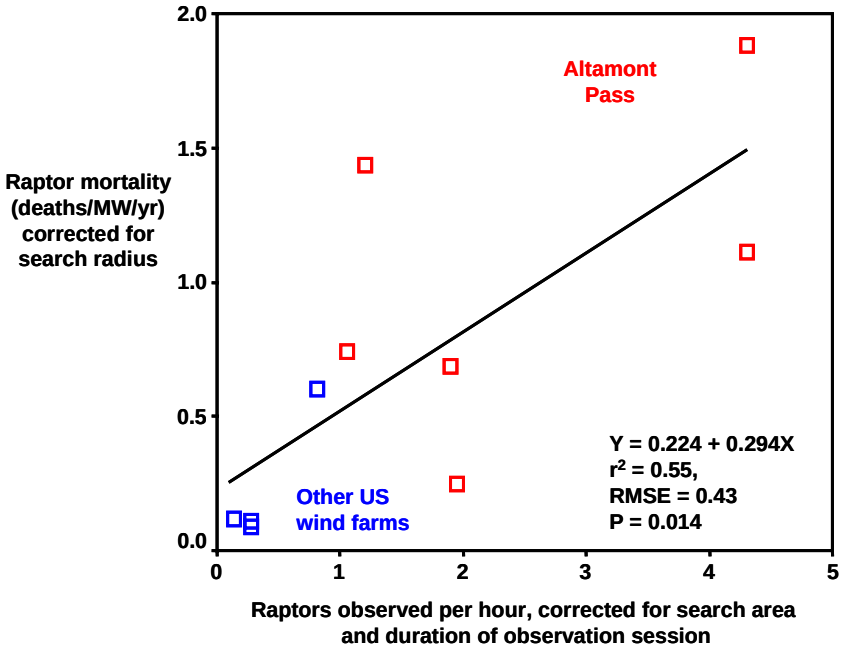
P13: *In their discussion in section 3.4, the authors reference higher mortality rates at Sea-West-owned turbines than other portions of the APWRA. I was not able to find the supporting evidence for this statement. In addition, I fail to find how ownership would affect mortality per se. Perhaps the ownership issue is tied to some other attribute of which I am unaware. I would like to see more clarification from the authors about the point being made. Further clarification regarding the biological significance of any raptor mortality (last sentence of page 76) would be beneficial*.

The Seawest turbines were a group of turbines we added to our search rotation when only 2/3 of a year remained. The ownership was not the important point, but rather the cohort of turbines as they were included in our search rotation. We originally thought mortality was greater at these turbines due to a mathematical artifact of short sampling duration, but we later concluded these turbines really did kill more birds annually. Why they killed more birds was probably not because they had the Seawest label on them, but rather because they were on short towers and in a low-elevation area where more burrowing owls and other birds were vulnerable to collision.

P14: *The authors have relied upon measures of relative abundance, determined from point counts. Usefulness of such indices relies on the assumption that the detection rates of birds are similar across time and space. The assumptions should be explicitly stated. There are a host of reasons why this assumption is likely to fail, including observer differences, animal differences and environmental differences (see Anderson 2001, Ellingson and Lukacs 2003). Logically, I would anticipate some sort of positive association between abundance and mortalities, however the positive associations estimated in figures 4-1 and 4-2 are likely not valid for the above detailed reasons. Figure 4-5 B actually demonstrates a trend counter to this assumed association, so an explanation would be helpful for the reader*.

After researching these patterns further (Attachment B), we agree. After adjusting utilization estimates by the search radius bias described in Attachment B, Figure 4-5B changes dramatically, illustrated in the following two graphs.





Factoring in the effect of the maximum observation distance used (search radius bias), the relationship we reported changed dramatically. We believe the precision of the regression would improve after adjusting the estimates for additional biases we are exploring in research methods. If given the opportunity to revise the report, we of course would revise this chapter. We have not yet decided whether comparisons of utilization and mortality estimates can be compared among wind farms after adjustments are made to estimates of mortality and utilization due to different research methods, or whether making such comparisons are hopelessly meaningless. One reason we are pursuing these comparisons is because other investigators continue to make cruder

comparisons than we did, and are completely ignoring differences in research methods among studies.

P14: *In the results, the authors state that bird mortality did not correlate with radius of search around the wind turbine. I am likely misinterpreting this statement, but if one increases a search area, one can only find more carcasses, not less, so I find this confusing*.

We think we did not find this correlation because there is so much slop in the data caused by other differences in research methods. Sample sizes of fatalities tend to be much smaller in reports of other studies, and these small sample sizes combined with differences among research methods led to unreliable patterns.

The data presented on mortality at APWRA from 1988 to 2000 has not been described in terms of consistency of methodology, surveyor ability, effort, consistency of environmental conditions that might affect detection rates of carcasses, wind turbine numbers, etc. Thus, I find it difficult to accept the reported trends as meaningful. The authors recognize the importance of standardized methods (see page 86), yet they have not made it clear they have met this requirement in their analyses and in fact, they explicitly state some differences among these studies.
We agree. If we had the opportunity we would revise this chapter dramatically. The reason it is important to follow through on this research direction is because other researchers are less careful in comparing utilization and mortality estimates, and at least two meta-analyses have been published out of Europe, based on these estimates. The effects of different research methods on these estimates need to be understood, and methods need to be more consistent among studies.

P15: *For instance, were transects randomly placed (string and grass) within a defined area around the turbine string, or were these haphazardly or judgmentally placed*?

On page 90 we wrote, "*For this effort, we visited 1,526 wind turbines that had been sampled through August 2002*." We visited all the turbines made available to us, so there was no sampling going on. Transects were established along all the Set 1 turbines according to rules given the field workers. They used laser rangefinders to start and maintain each transect as either 20 m or 40 m from the string.

How was average vegetation height measured along a transect, based on every plant encountered or at specific points (i.e., a point transect sampling approach) along the transect?

This variable was described as an index, and we indicated grass height was estimated. We probably confused some readers, however, by describing our index as the "average" height, which implies sampling and measurement. We would revise this description if given the chance.

How might detectability of cattle pats, rabbit pellets, lizards, mammals, etc., differ in different locations?

Not much in annual grassland during the late summer/fall.

As before, to what population is inference being made with the statistical tests?

We give the same answer as before. In our opinion it depends on the strength of the test, or how small the P-value was. We did not extend any of our inferences in this Chapter beyond the APWRA, and much of it was restricted to the set of turbines used to generate the data. None of these test results were used to develop predictive models of avian fatalities.

Use of the word 'significant' needs to be clarified as statistically detectable rather than biologically meaningful. I recommend the authors reserve the word significant only when referring to a biologically meaningful observation. Their one-variable-at-a-time approach may confound the observed associations. For instance, the authors give many examples of how vegetation height differed according to aspect, physical relief, etc. Although I am not convinced these are meaningful differences, they do indicate numerous variables are being considered, and a one-variable-at-a-time analysis procedure has inferential limitations which have been discussed previously.

This is good advice on how we could clarify our meaning by restricting our use of the term "significant." We will adopt this approach in all future research. We need not declare all statistical test results as significant simply because the P-value is <0.05. For many test results, we can simply report the P-value, and for those we deem biologically significant, we will use the term *significant*.

What is the rationale for associating turbine or tower type to lizard counts?

Towers vary in their bases, or whether they have concrete pads. On concrete pads the lizards are common and highly visible and might attract the attention of foraging raptors, which then might be prone to getting killed by the wind turbine while they attack the lizard. Other towers/turbines are anchored down into buried pylons and attached guy wires, and so most of the ground surface at the tower base is matrix substrate on which lizards may be less abundant and less visible.

Counting lizards did not work out too well. We might have been too ambitious with this variable.

P15: *Use of the phrase, 'tended to be significant' on the middle of page 91 is either improper interpretation of P-value as related to effect size or from a decision making perspective of null hypothesis testing a way of circumventing a yes-no answer in the formal test. First, a P-value is not an indication of effect size, after all, its value can be changed simply by changing the sample size; the same estimate can yield differing P-values. A P-value can be interpreted as the probability of observing a result as or more extreme than that observed given the null hypothesis (Ho) is true*.

Point taken. It is fairly common to see P-values treated this way in wildlife biology because the researchers recognize wildlife data are often less useful than, say, laboratory data. Considering P-values between 0.05 and 0.10 has commonly been our way of saying we are exercising professional judgment and are being cautious because the test was less than ideal. Biologists

refer to these types of results as *tending toward significance* or as *trend*. Given the comment, we will re-evaluate our approach to reporting these P-values.

P15: *The biological significance of the observations has not been justified, conclusion have been based on statistical detectability. For instance, the mean difference of vegetation height comparing heavy versus intermittently employed rodenticides was 4.28 cm). Is this a meaningful difference? If I were to look at two such areas, I doubt I would perceive the difference, but the more important question is do prey species and/or raptors perceive the difference? Does habitat use by animals differ substantially because of the lower height? The summary table 5-25 summarizes their findings, but underlying all of these associations are questionable interpretations of biological importance.*

There is ample evidence in the published literature that wildlife do change their use of annual grassland according to vegetation height. However, whereas we explored these relationships to identify potentially productive future research projects (one of which is now ongoing), we did not use any of these to develop predictive models in Chapter 7. We think some of these relationships are biologically important, but we also believe these significant relationships need directed research.

P15: *The mean differences in Table 5-1 are confusing, for instance, when comparing plateau to plateau (these are the same variable) and when comparing plateau to peak and slope (one cannot perform LSD on 2 variables), they must use another multiple comparison procedure which allows combinations of means. Note also that it is improper to follow nonrejection of an ANOVA with LSD multiple comparisons because there is no control for type I error in such a case.*

P15: *The mean differences in Table 5-1 are confusing, for instance, when comparing plateau to plateau (these are the same variable) and when comparing plateau to peak and slope (one cannot perform LSD on 2 variables), they must use another multiple comparison procedure which allows combinations of means.*

If we were to revise the report, we would add the mean values so that the mean differences are more meaningful.

Note also that it is improper to follow nonrejection of an ANOVA with LSD multiple comparisons because there is no control for type I error in such a case.

But we did not do this. On page 91 we explained that our alpha-level was 0.10 in this chapter. We did not follow up non-significant test results with LSD tests.

P15: *Table 5-3 gives several correlation coefficients, most of which indicate a weak linear association at best, yet they highlight the statistical detectability of these cases. For example, the last sentence in the discussion on page 54 restates that vegetation height correlated positively with number of cattle pats, but the sample correlation coefficient was only 0.19, a weak association at best.*

We agree the correlation coefficients were small in Table 5-3. We do not think reporting them was wrong, and in the text we did not characterize them as *significant*. We understand these are not the sorts of results a scientist is going to see in professional journals, but again this is a report where we have the opportunity to report *all* our results so researchers can decide what not to pursue as well as what to pursue in future research on the same topic.

One correlation coefficient given for vegetation height and 'percent in canyon' is moderate (r = 0.46), but I am not certain this is meaningful. The population of turbines being measured is clearly stated as the 1526 wind turbines measured through August 2002. Thus, statistical inferences might be made to this collection of turbines if sampled appropriately.

And this is what we did. Inferences from these tests were not extended beyond the sample, and the sample was *all* the 1,526 turbines that were available to us through August 2002.

P17: *I commend the authors for making observations regarding gopher and squirrel burrows and their proximity to turbines, and developing research regarding the variables. Their objectives are clearly stated as relating ground squirrel and pocket gopher distribution and abundance to rodent control intensity, physiographic and turbine attributes, and comparing raptor mortality to densities and contagion of burrow systems, but a priori hypotheses are not explicitly stated. Burrow densities are implicitly being used as indices to abundance.*

The reviewer is both correct and incorrect about the last statement. Pocket gopher burrow systems usually are occupied by one adult, whereas ground squirrel burrow systems house multiple animals, and the numbers per system vary. Burrow system density can be more closely related to pocket gopher density than to ground squirrel density, but we do not think raptors forage over landscapes according to animal density as much as they do according to environmental indicators of prey availability, including the abundance of burrows. We think it unlikely raptors shift foraging flight directions according to counts they make of potential prey items. Our target variable was therefore animal burrows rather than the number of animals.

I have no knowledge of whether or not this assumption is reasonable. For example, the burrows may represent a population size that existed several years prior to the current observed raptor morality or numbers of burrow per animal may differ depending on landscape or predator abundance features. How ephemeral are these burrow systems?

Animal burrows lose their integrity after several months of vacancy, so we did not map any old burrow systems. Also, we can easily identify fresh burrow activity by the texture and color of the extruded soil. Smallwood has worked with animal burrows for nearly 20 years.

The seasonal effects reported in section 6.3.2 indicate burrow tremendous variability within a year, but are numbers of individuals fluctuating that much?

Absolutely. These fluctuations are well established in the published literature.

Changes in the numbers shown in various figures make me question the relevance of burrows as an index to prey abundance, let alone prey availability.

We do not know why the reviewer came to this conclusion. Seasonal variation in an indicator does not invalidate its use. Understanding this seasonal variation makes it all the more useful because we know not to compare winter values to summer values. We also point out again that the reviewer assumed incorrectly that we were using burrow system densities only as indicators of animal abundance.

Given the observed variation in burrow numbers throughout a year, how did they relate observed bird mortality over a year or several years to burrow contagion?

Seasonal variation in burrow distribution is not the same as inter-annual variation. One cannot conclude high seasonal variation results in high inter-annual variation, because they are different phenomena driven by different factors. Smallwood has been monitoring ground squirrel and pocket gopher burrows in one large annual grassland parcel in California for 7 years, and has found much greater seasonal variation in abundance than inter-annual variation. Burrowing animals tend to establish burrows in the same places, and in most cases immigrants simply take over burrows left vacant by the previous occupants.

P17: *Wind turbine strings studied were selected arbitrarily (not randomly), hence limiting statistical inferences*.

We agree, but given the exploratory nature of this research we believe our arbitrary selection was justified. However, we would have increased the strength of our inferences had we employed a stratified random selection process.

P17: *Figure 6-4 (and later 6-45 and 6-46) presents results comparing burrow system densities between areas with rodent control and areas lacking rodent control, however, for the latter, only 3 observations were available. The strength of evidence here is very weak*.

We agree the relationships based on 3 observations were reaching, and we agree the strength of evidence is weak in this case.

In figure 6-5, they discarded an outlier without explanation. If a poor measurement was made so that this result was unreliable, then state this clearly. If, however, it does not fit their predefined opinion on what should occur, then this is not a viable reason to omit it from the analysis.

We do not remember why this data point was discarded, but Smallwood rarely identifies any data as "outliers" (this may be the only instance of an outlier occurring in the entire report), so there must have been a good reason.

P17: *The authors resort to transformations at various places without explanation or consideration of what is then actually being compared. For example, what does Figure 6-6A tell us? The variable based on a log-log regression in this figure and Figure 6-7 has not been explained*.

Figure 6-6A tells us what we wrote about it in the text, "*Pocket gopher density consistently decreased as larger areas were searched around each string of wind turbines (Figure 6-6A), indicating that pocket gophers were clustered around the wind turbines. Nearly all turbine strings demonstrated a relationship between gopher burrow density and study area size that was similar to the pattern reported by Smallwood and Morrison (1999), which was an inverse power function. Similarly, most of the observed-divided-by-expected number of gopher burrow systems within 15 m of the wind turbines was greater than 1.0 (Figure 6-6B), meaning that gophers were almost always clustered to some degree around the wind turbines*."

The density of pocket gopher burrow systems is consistently an inverse power function of study area size used to estimate density. This pattern resembles the pattern reported on by Smallwood and Morrison (1999) and in Smallwood's other published papers on the pattern. The difference here, however, is our study areas were repeatedly centered on a particular structure, which was the string of wind turbines. Figure 6-6 is telling us pocket gophers repeatedly organize themselves around the wind turbines, probably because conditions are superior at the wind turbines either due to conditions pre-existing the wind turbines, due to conditions created by the wind turbines, or a combination of both.

Why are normal curve shown in all frequency distributions?

Our impression was the frequency of slope values approached a normal distribution, and so we included the normal curve so the reader can decide whether to share this impression. Whether or not the frequency distribution was normal may not make any difference, so we probably could have omitted it.

P17: *I find it curious that ground squirrel avoidance was stated to differ between summer and other seasons, specifically fall (see page 124), given the degree of overlap of these 2 estimated means (Figure 6-27B)*.

That is not what we reported, but we agree our reporting of this result was incomplete. The difference was obviously between summer and winter.

P18: *From a nontechnical perspective, I find the observed relationship between rodent control intensity and pocket gopher burrow density interesting in that its highest level is at a moderate level of rodent control. Cottontail burrows demonstrated an opposite pattern. Why would gopher clustering differ by aspect for control areas, but not for nontreated areas? I could not find explanations for several reported effects*.

We did not have explanations for several reported effects, and we still lack explanations. We reported what we found, and we explained the effects we understood or thought we understood. The effects we could not explain raise more questions that need to be followed up by directed research.

P18: *Comparisons of raptor mortality and small mammal burrow distributions were executed only considering burrow density and thus are prone to many confounding effects as previously stated. At various times, the authors recognize the potential complexity of what they are*

attempting to measure, but they then put this consideration aside and proceed to analyze, interpret and conclude.

What else would the reviewer expect us to do? It was our job to analyze, interpret and conclude, and we did it after warning of potential confounding, bias, and other threats to our interpretation of the results.

P18: *The finding that mallard mortality was related to rodent control intensity was dismissed as a spurious effect. I agree with this conclusion, but it illustrates an important point. When one can develop an ecological explanation for an observed result a posteriori, the result is more heavily weighted as 'truth'. I believe this approach to science is ubiquitous, but not ideal and has lead to many spurious results*.

We agree with the reviewer on this point, but the reviewer has also several times asked for explanations of our results. One cannot have it both ways. If we are going to attempt to explain our results, we are also going to have to deal with readers labeling our results and explanations as truths, and in fact we already have experienced this phenomenon multiple times. As an example, and to our dismay, some readers latched onto certain results in this and other chapters, while ignoring other results, and argued that we proved ground squirrel control was effective at reducing golden eagle mortality.

As the investigators, it was our job to explain our results to the best of our abilities, and despite some lapses we believe we provided vastly greater exploration into the collision issue, with vastly greater explanation of measured patterns, than anyone ever has before us, or since. We believe that our approach to addressing this issue took as a long ways from where we were before, as summarized in Attachment A. Most of our previous notions about why birds were colliding with turbines were based on speculation, deductive logic in the absence of evidence, and anecdotes. Many of our results and explanations were relevant to these previous notions, laying to rest some of them, and directing us toward directed research programs to test others. We hold no illusions that we arrived at truths, but we believe we took significant steps towards truths. An important part of this big step included *a posteriori* explanation of results.

P19: *The authors begin this chapter by stating the importance of identifying causal factors of bird fatalities. They then state that collisions are rarely observed and that inferences must be drawn from carcass locations. Such inferences are merely associative, not causative. I suggest the authors present their observations in the former context as I believe the latter is not attainable within the context of the current study*.

The reviewer is correct. Our opening paragraph of Chapter 7 was poorly worded and gave the false impression we believed we were identifying causal factors. In fact, the presentation of results gives the other impression – that we regarded these patterns as associations only. If given the opportunity revise the report, we would change the first paragraph as well any other text that suggest our study identified causal factors.

P19: *The authors are aware of possible spurious results (see page 218 discussion of mallard fatalities), but they fail to see this potential when explanations can be developed a posteriori*.

We disagree with this statement. We did not fail to see the potential for spurious relationships in any of the *a posteriori* explanations we provided. Just because we did not label each *a posteriori* explanation with "Warning, potentially spurious," does not mean we did not consider the potential.

P19: *More importantly, they fail to see this potential in their overall approach to analysis. For instance, 34 explanatory variables have been measured at each turbine site, 12 bird species examined leading to hundreds of single variable tests of associations (Tables 7-1, 7-2, 7-3). The potential for Type I error (rejection of the null hypothesis of no association when there is no association) is essentially one when considering all of the tests being performed. A total of 408 tests were performed in Tables 7-1 and 7-2. The probability of not making a type I error (using the stated* $\alpha = 0.05$ *level) is* $(0.95)^{408}$. *Thus, the probability of making at least one type I error is* $1-(0.95)^{408} \approx 1$.

The reviewer is assuming all P-values of significant test results were 0.05, but many were much smaller than that. The reviewer over-estimated the likelihood of our committing Type I error. But let's assume we did commit a Type I error in one of the test results. This means one was a Type I error, and the other 407 did not commit Type I errors. Those odds look pretty good to a wildlife biologist. And as we stated previously, we did not select all these results for inclusion in a predictive model. We screened them based on P-value and other factors.

P19: *The main point here is that I do not really understand what their 'model' is. Is the assumption that their predictive models are relatively precise appropriate (page 222)? Testing of their models appears to have been performed using the data to develop the model, which is an inappropriate means of evaluating models (see Olden et al. 2002). The authors should consider using a portion of their data for model development, reserving the other portion for model evaluation*.

Our "models" were simply summations of accountable mortality across the variables selected for the model. Accountable mortality was defined in the report, and should not be confusing. A value was attributed to each category or level of each association variable. So if we had 6 variables in the model, we added the accountable mortality values that corresponded with the categories or levels associated with each turbine and across each of the 6 variables. As an example, let's say the model for a particular species included (1) rotor diameter, (2) tower height, (3) turbine position in the string, (4) whether in a canyon, (5) slope aspect, and (6) turbine density, and let's say a particular turbine had a 17-m rotor diameter, was on a tower 24 m tall, located at the end of turbine string, in a canyon, on a northwest-facing slope, and surrounded by 30 other turbines within 300 m. If for this hypothetical species the accountable mortality values for these conditions were 6%, -2%, 12%, 15%, -3%, and -5% for variables (1) through (6), respectively, then the sum of these values would be 23%. Because this value is greater than 0, we would conclude there is a relatively greater threat of a collision for this species at this turbine.

If we were to revise our report, we would hold aside some data and test the models on them, as the reviewer suggested. Over the past year, WEST, Inc. has been searching for bird carcasses in the APWRA, so if they would share their data with us, then we could test our models further.

P20: *On pages 179 to 182, the authors describe an abundance of previous studies which have presented conflicting conclusions regarding causal factors of collisions. Such disparity suggests to me, as stated earlier, that demonstrating causation is not a simple task and the actual mechanism underlying bird strikes may be very complex, e.g., a combination of many environmental and turbine-based factors. Again, the overall analysis approach has been to look at associations one explanatory variable at a time, thus leading to potential confounding effects and spurious results*.

We agree with the reviewer's assessment that the causes are likely complex. We are less confident than the reviewer that multivariate analysis would be the solution, but with a large enough data set, based on regular search intervals and multiple other methodological caveats, we think multivariate analysis would be preferable to the approach we took.

P20: *On page 182, last full paragraph, the last sentence is unclear. I believe the authors meant to imply that some turbine attributes were collinear or highly correlated, thus similar associations with bird mortality were observed when looking at each variable separately*.

Correct.

P20: *On page 184, the first three paragraphs of section 7.2.2 are wrought with lack of information and misstatements. The authors state that the assumptions of the corresponding hypothesis tests were satisfied, but they do not state what those assumptions are and how they were assessed. For example, Pearson's correlation assumes bivariate normality. Did they assess normality of each variable? How*?

We do not recall assessing normality of each variable, and we probably did not. What we were saying is that we were aware of the assumptions of each test, and so we decided on which test to use based on the nature of data we had available. And then we violated some assumptions with certain tests, which is a common occurrence in biology.

The least squares regression models they used assume the errors are independently and identically distributed as a normal distribution with mean zero and constant variance. How did they assess normality of the residuals? How did they test equal variance?

Smallwood usually examines the pattern of the residuals for heteroscedasticity. We did not test for equal variance, or for normality of residuals.

Analysis of variance (ANOVA; which is really regression with a categorical explanatory variable) has the same assumptions. Were these also assessed for these analyses?

No.

P20: *Misstatements include 'Correlation analyses are summarized by the coefficient of determination, R^2, when prediction is the ultimate objective. I believe they meant to say 'Regression analyses….'*

Correct. The reviewer's assumption of our intended message would be more accurate.

The authors state "We report weak and nonsignificant correlations when doing so meets our objectives." -this sounds dubious and confuses several issues. Statistical detectability is directly affected by sample size, n. Thus, with large n, it is possible to have a sample correlation coefficient of r =0.1, and yet have an associated P-value = 0.001 for the test of Ho: $\rho = 0$ vs. Ha:$\rho \neq 0$. In this case, while the result is statistically detectable, the linear relationship is very weak and is not meaningful or interesting.

We disagree with the reviewer's conclusion. A statistically non-significant test result can be interesting to a biologist when the prevailing view has been that the result should have been statistically significant. Non-significant test results can be both meaningful and interesting, and this is what we were saying in the report.

Alternatively, you can have small sample sizes that result in nonrejection of Ho even though the sample correlation coefficient may be r = 0.75, that indicates a fairly strong positive linear association. I agree that in the latter case, such results may be reported as long as they are not presented as confirmatory evidence.

And this is not what we were saying.

P20: *Given the collection of several explanatory variables, the authors should consider using partial correlation in which one or more variables are controlled when considering the association between two variables of interest.*

We would give this suggestion a try if we were to revise the report.

They also incorrectly state that the coefficient of determination (R^2) is based on the steepness of the regression slope. This is only true within a specific context in which one considers a specific data set and several lines that are being fit simultaneously to the data. In general, the coefficient of determination is defined as the portion of variation in the response variable that is explained by the explanatory variable. For example, if all observations fall on a fitted regression line, then R^2 is one, and this is true regardless of the slope of the line. The exception here is when the slope is zero, in which case there is no variation in the response variable.

Actually, we used the wrong symbol to characterize the coefficient of determination we were discussing. We meant to use r^2, symbolizing the coefficient of determination used in simple linear regression analysis.

P21: *They state that several key assumptions of ANOVA were not met due the absence of a block design. A block design is not necessary for ANOVA, blocks are sometimes useful for reducing variability, but their absence does not preclude assumptions from being met. Equal treatment replication (balanced design) does not preclude successful analysis via ANOVA or other techniques. However, when all treatment combinations are not represented, e.g., fractional factorial designs, then considerable thought must go in to analytical approaches for meaningful*

comparisons that isolate treatment effects to be made. Identification of the proper error term for testing treatment effects is often determined using expected mean squares.

We stand corrected on our statement about block design.

P21: *I do not understand what the numbers in Table 7-8 on page 223 represent. The description is that the numbers represent the largest accountable mortality values calculated from the chi square tests…. Similarly, I do not understand how a specific wind turbine attribute can be reliably associated with X% of mortalities (pages 224-241), given the combination of variables at any given turbine and the inability to control all other factors in their analysis. It follows that I am not confident that their form of model assessment (described as the percentage of correctly predicted dangerous turbines where species-specific fatalities occurred) is a useful metric for assessing model performance*.

Table 7-8 presented our screening of variables for use in developing predictive models. For each of the chi-square tests we performed, we calculated two measures of effect for each category or level of the test. One measure of effect was the observed ÷ value, and the other was accountable mortality: Accountable Mortality = (Observed – Expected) ÷ Total fatalities × 100%. We used the latter in this screening step. The Table shows the largest value obtained in each chi-square test that also met the following conditions:

(1) Significant P-value;
(2) Composed of expected cell values mostly > 5;
(3) Accountable mortality values formed distinct gradients across categories or levels of the association variable.

Essentially, the table identified the association variables we considered candidates for inclusion in our predictive models.

P22: *From an inferential perspective, I would ask the authors to clarify how these specific sites were selected for observations. It looks as though a variety of wind turbine types were selected purposefully (Table 8-1) that cover a previously referenced area (see page 246 bottom), but a stratified sampling approach could have been to ensure random selection and representation of all turbine types as well. In reference to their analysis, I am concerned that their approach is inadequate to identify meaningful relationships for reasons similar to that stated for other chapters*.

All Set 1 turbines were included in the behavior study, so no random selection was necessary.

P22: *First, the approach of examining one variable at a time oversimplifies what is a very complex situation. For example, by only considering minutes of perching by temperature levels, the observed difference of observed and expectations may be the result of another variable, such as wind speed, which often is associated with time of day and temperature. That is to say that such an approach does not identify causal mechanisms. To their credit, the authors do mention another source of complexity, the notion that birds may adapt their behavior in response to the*

presence of wind turbines (page 246). But, I am not at all confident in any of their findings in this entire chapter because of their one-variable-at-a-time goodness of fit approach to analysis.

We are surprised the reviewer made such a sweeping conclusion that no results in this chapter are reliable. Whereas we can agree many of the findings are potentially confounded, we can also find many results with which we are confident and which are useful. For example, how can the reviewer deny seasonal trends we found in flight patterns, or in perching? We found raptors spent disproportionately more time perching in January than in May. One can argue it is impossible to determine whether the difference was between months or between temperature, but it is upon months management actions can be directed.

In another example, we found raptors flew disproportionately closer to wind turbines than they did between 100 and 300 m from wind turbines. What is wrong with this association? Even if it is confounded with topographic features, it is still a useful result. In another example, we found golden eagles perched disproportionately more often in canyons. What potential confounding would invalidate this result? This is where we found them; would the confounding issue lead us to believe we actually did not see them perching disproportionately more often in canyons? In another example, we found red-tailed hawks flew disproportionately less often over flat terrain, and more often over east- and northeast-facing slopes. Even if this last result was confounded, we still observed it, and these results are still useful for predicting where red-tailed hawks are likely to fly more or less often. We reported many results like these, so we do not understand the sweeping nature of the reviewer's comment.

P22: *Second, I question whether many of the stated 'significant differences' are biologically meaningful. For example, in Table 8-6, the chi-square test for time of day effects on perching minutes of all raptors resulted in a P-value less than 0.005, yet the percent deviation from the expected value is less than 3 percent for all categories. In the same table using temperature as the explanatory variable for golden eagles, there is a 20 percent negative measure for temperatures of 60-69 degrees and a 21 percent positive measure for temperatures between 70 and 79 degrees. Do the authors believe that golden eagles perch less due to temperature in the 60s and more in the 70s and then less in the 80s*?

We reported our results – all of them. We agree with the reviewer we should be more restrictive in our use of the term *significant*, but we do not see what is wrong with reporting all our results, regardless of the P-values.

P22: *I recommend that the analysis approach in this chapter (and several others) be changed. I would need to know more to specifically advise on how they should proceed, but I will make the following statements and suggestions. First, birds (or animals in general) do not use habitat randomly. Any assumption of random use is a 'silly null' hypothesis which is certainly false (see Anderson et al. 2001)*.

But it is either this null condition or the null condition of uniformity that we always use because these are the mathematical null conditions of the statistical tests we use. We think it is silly to deny this is the case. Establishing a null condition does not mean we believe animals really do fly around randomly. We do not believe that.

Second, while I agree that understanding how birds use APWRA would be useful in putting bird fatalities in context, I fail to see how associations with variables such as temperature would be useful.

That's why the CEC hired biologists to do the study. We tested whether birds performed certain dangerous behaviors during cold conditions, perhaps because they were less cautious during these conditions. For example, barn owls get killed by autos more often during cold weather. If we found a strong association between fatalities and temperature, then we might pursue an experimental shutdown of wind turbines during cold (or hot) conditions.

P23: *Third, I suggest the authors condense the 30-minute level information to percentages for that survey period and consider relating the percentage of time perching to the set of meaningful explanatory variables collectively. This approach treats each 30 minute period as the measure made on each sampling unit (plot). Such an approach eliminates concerns about the covariance structure between successive minute-by-minute observations. Whether or not one needs to consider the relationship between survey periods on the same site is another issue (perhaps repeated measures structure should be used?). Once a reasonable modeling approach is identified, they must develop appropriate models for consideration and use a well-defined model selection process*.

This is an intriguing idea, and we would try it if we had the opportunity to revise the report. We will look into it when we prepare papers for journal submission.

P23: *I am not sure whether the focal set of wind turbines was always a complete string*.

They were all the wind turbines in the observation plot.

Also, based on figure 8-2, it looks as though a 300-m buffer may include additional sets of wind turbines; thus, they too provide opportunities for perching, 'dangerous flights', etc. How is their presence handled in terms of analysis at the wind turbine level even though they are not the focal set?

If these turbines were within the observation plot, they were focal turbines.

I failed to see the distinction between plot level and string level of analysis mentioned on the middle of page 247.

Plots include all the turbine strings in the plot.

How do they count number of minutes perched if 2 birds are perched in the study area for the first 5 minutes of the study period. Is that 10 bird-perching minutes? If so, then the number of bird-minutes is the metric, rather than minutes alone.

Correct. So the idea above about transforming the observations into percentages of the 30-min session may not work.

They state on page 247 that for each record, they recorded the species, … predominant flight behavior, flight direction, distance to nearest turbine, number of passes by a turbine, and flight height relative to the rotor zone. How does one record flight direction if they traveled a flight was multidirectional, circular, etc?

If the bird is circling, there is no point in recording flight direction, but otherwise the observations were made on the minute, so within our sampling framework we do not care what direction the bird was fling 10 seconds earlier.

Is distance to nearest turbine the smallest distance observed during the entire flight?

No, it is the smallest distance from the bird's recorded on-the-minute position and the nearest wind turbine from that position.

How does one define a pass by a turbine?

If it flies within 50 m of the rotor zone, as defined in the report.

If turbines are in a string a bird flies 100 m above the string, are all of these turbines counted?

No.

How much error is there in measuring flight height relative to rotor zone when birds are considerable distance from the observers?

There was plenty of error, especially the farther away the bird and the higher above ground. We wish there was another way to record a bird's position and height above ground.

Birds may have exited the area for more than 30 seconds, only to return again and be considered as a new bird. Pseudoreplication is a concern here, but clearly the researchers cannot be expected to recognize individual birds per se.

We agree.

Several quantitative or ratio level variables, such as temperature and wind force, were reduced to ordinal categories. Information loss occurs in such a process and is unnecessary. I suggest the authors use these variables as continuous explanatory variables in a model effort other than chi-square goodness of fit tests.

We would try the reviewer's suggestion if were given the opportunity to revise the report.

I do not understand why they compared the correlation of flight frequency in the rotor zone with flight time to that of perching time (page 256). I would assume that when a bird is perching, it is not flying at all, and thus cannot be flying in the rotor zone.

During 30-min sessions we could observe multiple birds, some of which are perching while others are flying, and any given bird can fly during a portion of the session and perch the rest of the time. The comparisons we made are perfectly feasible.

I also do not understand the interpretation of frequency of behavioral observations during a 30-minute session on page 256. If the initial presence of observers is modifying bird behavior, then this is an important observation, which suggests that an initial 'settling' period should occur prior to the actual observation period. However, I am not sure this was the point they were trying to make.

It is the point we were trying to make, but we did not realize this pattern existed until after our data were collected. We needed the data to get the result.

P25: *This section of the report was clearly written and suggests several management alternatives. It is my belief that some form of adaptive management (Walters 1986, Walters and Holling 1990) should take place given the amount of data already collected by this project and others cited therein.*

We agree.

P25: *For example, I assume it would cost very little to paint blades for a sample of turbines.*

At least from the point of view of a couple of wildlife biologists, the license fee for the Hodos painting scheme and for his paints are expensive. We will leave it to others to decide whether the cost would be little or a lot.

P25: *Their first recommendation, at least on the surface, seems reasonable. If sampling in the WRRS program is haphazard and/or voluntary-response based, then from a scientific monitoring perspective, the data collected is of little value. That is not to say that there is nothing to be gained by observations of maintenance workers in the area, because the cost is presumably nothing. I am unsure that the comparisons of observed fatalities are fair (same time period, same locations surveyed, etc.), but if they are, and a consistent relationship between the 2 methods could be established, then such a system would be similar in worth for trend estimation. However, consistency is something not easily obtained in any index-based study (Anderson 2001).*

We agree.

P25: *Their conclusion regarding rodent control is counterintuitive to me, but that does not make it wrong. However, I question whether the conclusion is correct given the potential weaknesses associated with assessing the effects of rodent control treatments (see chapter 6 evaluation). They further state that even if rodent control were effective, displacing raptors would result in a net loss of raptors from the remaining habitat. That is only true if populations elsewhere are at carrying capacity, which I doubt to be the case. If raptors are at carrying capacity, then perhaps there should be less concern about the observed mortalities.*

Why would the reviewer have an opinion about whether neighboring raptor populations are at carrying capacity? Is such an opinion based on evidence? Inference? Does the reviewer have any knowledge of the environmental conditions into which the APWRA's raptors would be displaced? Perhaps the reviewer is unaware the surrounding area hosts the highest density of nesting golden eagles in the world, or that burrowing owls are rapidly declining in northern California. Perhaps the reviewer has not considered that raptors are visiting the APWRA for other reasons besides foraging, or that raptor foraging decisions are not made simply according to prey abundance. Perhaps the reviewer is unaware the Altamont Pass is a low spot in the mountain range, where birds naturally fly through on migration.

The carrying capacity of raptors outside the APWRA is not, and never has been, the determiner of how much concern should be given raptor mortalities in the APWRA. The Migratory Bird Treaty Act prohibits the taking of raptors, and some species are protected by other laws as well. Carrying capacity is an ecological concept that the reviewer would be challenged to characterize in quantitative terms for any species, so its usefulness in the context of assessing biological impact, let alone legal impact, is questionable. And besides, why not strive to reduce raptor fatalities where we know reductions are feasible?

P25: *Their third recommendation (and its subcomponents) is similar in the underlying idea to the second recommendation: reducing prey availability may reduce raptor susceptibility and thus fatalities. Thus, I find it interesting that they advocate ceasing the rodent control program, but encouraging fossorial animals to be farther from wind turbines. However, I agree that one might want to eliminate the rodent control program for other reasons (e.g., adverse impacts on other species of importance)*.

We do not agree our third recommendation is similar to the second. The second goes to a measure applied across the APWRA, whereas the third is applied to small portions of the APWRA. The third measure is more surgical in scope, and simply attempts to shift raptor foraging away from the wind turbines. It acknowledges our belief that it is highly unlikely any measure will manage to shift raptors completely out of the APWRA. We believe the best we could hope to accomplish would be to shift raptor foraging patterns within the APWRA, and even accomplishing this will be a long shot. Measure 3a will be most practically applied to new wind projects in the APWRA, whereas 3b is a simple matter of moving rock piles farther from wind turbines.

P25: *In their test of perch guard effectiveness (recommendation #4), did they control for other factors that may affect mortality? Similar to the results for rodent control, it may not be the method itself that is lacking effect; it may be the implementation, for instance, if the chicken wire readily falls apart*.

We did not control for other factors when testing perch guards, but we also provided additional rational for not investing further in this approach. Birds do not perch on the turbines while the turbines are off, and there are multiple perching opportunities on turbines besides those locations treated with perch guards. In short, the perch guard measure was never going to work, and still will not work.

Their conclusion that wind turbines at the end of strings are edges of clusters kill disproportionately more birds is very plausible, and hence their suggestion of adding pole structures is worthy of consideration for experimentation. However, it may be that by adding pole structures, more birds will collide with the turbine because of the visual impedance mentioned in recommendation number 9.

Visual impedance from a couple of pylons? We do not think so. Also, the reviewer misunderstood what we were saying about mitigation measure number 9, and apparently missed the point we made about disallowing perching opportunities on the pylons.

P26: *Most of the remaining recommendations are yet to be proven as effective management options*.

No, all of the management options are yet to be proven. However, our recommendation to repower the APWRA with new generation turbines is looking promising after the first year of fatality monitoring in the Diablo Winds project.

I suggest that pilot studies be used in which the efficacy of a small set of management actions (a subset of their listing) can be evaluated without the confounding effects of other possible mortality factors.

Such pilot studies have been tried before in the APWRA. Sample sizes tend to be too small for confident conclusions. Also, in the meantime thousands of birds are killed while we mess about with pilot studies.

P26: *Based on my limited knowledge from this report, I am not convinced that enough is known to warrant universal implementation of certain mitigation measures (see page 348 bottom)*.

We do not understand the comment. Which mitigation measures does the reviewer believe are not warranted for universal implementation? The reviewer already agreed with us the WRRS should be replaced with scientific monitoring. Should this replacement not be universally applied? What about the house-cleaning measures we recommended, such as removing broken and non-operating turbines, or retrofitting tower pads to prevent under-burrowing by rodents and rabbits? Should these measures be applied to only some of the turbines? Or does the reviewer regard these measures as unwarranted?

I also question the degree of error one would have in the estimated number of bird mortalities over 10 years as the input to Smallwood's estimator of are for support described on the top of page 348.

We do not understand the comment. Why not incorporate the estimated error into the exercise? Or, why not perform additional research to reduce the error?

P26: *In section 9.3, Smallwood and Thelander state they were unable to extend their model predictions to the turbines not characterized. Given that an appropriate predictive model exists*

(which I have previously questioned), I would suggest that to properly evaluate the model, these sites would provide an independent means of evaluating the model.

We do not have any fatality data from these turbines.

I assume that when they are referring to multivariate statistical methods on page 353, they are referring to multiple response variables, not multiple explanatory variables, but their one-variable-at-a-time approach to analysis indicates there may be confusion regarding this terminology. Multiple regression and analysis of covariance are considered univariate methods because one response variable is being considered. I suggest that these methods could be employed to better examine mortality as a function of several explanatory variables. Multivariate methods generally refer to analysis in which multiple response variables are being considered simultaneously, which requires consideration of the covariance structure among these variables.

Understood.

I am not sure how they arrived at their estimated mortality reductions on page 354. These are likely purely speculative.

Exactly. And since these estimates, Smallwood and Spiegel (2005a-c) developed better models and provided quantitative bases for their estimates of percentage mortality reductions following the implementation of various mitigation measures.

P27: *Their example on page A-2 demonstrates how fatalities rates using number of turbines as a reference can appear to differ at 2 locations even when the same number of deaths occurs at these locations. The reason for this is the sites have differing numbers of turbines. The same differences can be illustrated using their metric of fatalities/MW/year. For example, if farm A generates 40MW/year and farm B generates only 4MW/year and 100 fatalities occur at both places, then the rate is 2.5 fatalities/MW/year for farm A, but is 25 fatalities/MW/year at farm B. This might mislead someone to believe that more fatalities occurred at farm B. Later, they compare regression sums of squares relating MW to bird deaths and turbine numbers to support their proposed metric compared to the turbine per year metric. This approach did not compel me to see the advantages of their metric.*

The reviewer's argument is a red herring. The example we used was a comparison between two wind farms of the same size, whereas the reviewer's comparison is between two wind farms differing 10-fold in size. Other researchers and statisticians familiar with the wind turbine collision issue understood our point and have nearly universally adopted our proposed metric. The reason is really simple – you cannot compare on a per-turbine basis the number of birds killed by a 2.5-MW turbine to the number killed by a 40-kW turbine. By using MW as the basis, we factor in the rotor-swept area, or the size of the turbine.

The main difference between their metric and the one that uses turbines is that theirs incorporates MW produced per turbine, thus the cost (mortalities) is stated within more of a context of the benefit (MW production).

We disagree with this comment. The main difference between the metrics compared is that MW factors in the size of the turbine, or the rotor diameter and rotor-swept area.

Initially, I questioned the purpose of comparing mortality rates among different wind energy generating facilities. If one is only considering management of a particular wind facility, then knowledge of how one place compares to another is not useful. After all, different facilities have many different environmental factors likely to be important in the process that leads to bird strike fatalities. Later, in the discussion, the authors mention replacement of older turbines with larger turbines capable of greater MW generation. Thus, if one wanted to compare fatality rates between time periods at a given site, it may be advantageous to use their metric if considering the cost-benefit aspects of the power generation. They have presented additional arguments for using their metric that make a better case than previously in chapter 3.

Repowering is indeed one reason to compare mortality estimates, but the reality is that mortality has been and will continue to be compared among wind farms, regardless of their differences. So long as researchers and others are intent on making such comparisons, we might as well attempt to do so using a metric that makes more sense than the previous metric.

Also in the results section, the authors refer to the relationship between time span surveyed and number of turbines with non-zero mortality (Figure A6). It seems obvious to me that the more you search, the more likely you are to find at least one mortality at a turbine, so I do not understand why they are emphasizing this point as being important.

Because others claim we should not expect to find additional turbines killing birds in the APWRA (see comments from Review teams 1 and 2). Others claim that whether we found carcasses at the turbines we searched should be the only basis for assessing risk to turbines. One reviewer actually claimed that those turbines where we found no carcasses are obviously not dangerous. One of the points of Figure A6 is that many of the wind turbines we searched only twice will be found to kill birds after additional fatality searches. We really should not be assigning risk to turbines solely on the basis of where carcasses were found during a couple of fatality searches.

Another reason we included Figure A6 is to support our argument that fatality monitoring needs to last longer than one year, which has often been the requirement among new windpower projects. Keep in mind that the mortality adjustments, such as adjustments for searcher detection error and scavenger removal rate, cannot be made on 0-values. Mortality will not be adjusted at all those wind turbine strings where no fatalities were found, because 0 divided by either or both the searcher detection error and scavenger removal rate will equal 0. After 3 years of monitoring, however, about 90% of the turbine strings will have had yielded at least one fatality each, and all these would then be adjusted up by the adjustment terms.

The proportion of turbines where at least one bird has been killed is not a metric that I find particularly useful and does not translate directly to fatalities per turbine or MW per year. I do not agree that this relationship demonstrates that most of their turbines were not sampled long enough to robustly estimate mortality.

We were not claiming the percentage of wind turbine strings with >0 fatalities should be a metric directly related to mortality estimates. We only used the relationship in support of our argument that fatality monitoring should last longer than a year, or part of a year. We found bird carcasses at only 30% of the turbine strings we searched twice, but the pattern in Figure A6 suggests we would find bird carcasses at about 90% of them after 2 or 3 years of searches. We believe our mortality estimates would be more accurate after 2 or 3 years. Whereas we went ahead and estimated mortality among wind turbines searched over half a year, we stand by our conclusion that longer monitoring periods would result in more reliable mortality estimates. Figure A6 shows the reader the percentage of wind turbines for which mortality will not be adjusted up by factoring scavenger removal rate, searcher detection rate, and other sources of bias. After our half a year of searches at most of the turbine strings in our study, Figure A6 shows we were unable to adjust 70% of them up due to the adjustment factors we know were needed, but after another couple of years of searches only about 10% of these turbine strings would not be adjusted upward. We will add that these same adjustments likely applied to the turbine strings where we found no fatalities, but there was no way to adjust them due to their 0-values.

I am not sure what is meant by 'robustly' here, but in section 4.4.2 on page 86, the authors use the word robust to imply reliability based on high precision. Precision, or repeatability, is only one component of accuracy. Bias, or the deviation of an excepted value of an estimator from the true parameter is equally, or perhaps more, important.

We agree.

P28: *In various places, the authors refer to 'less robust' estimates of other researchers and then proceed to compare estimates from various studies. Robustness implies a resiliency to, for example, an assumption violation. It is not clear to me what their use of the term implies.*

We agree the term *robust* was vague (i.e., vague on the level of *carrying capacity*). What we meant by it was increased precision and likely improved repeatability due to our larger sample sizes collected over longer time periods and across a larger area.

In comparing estimates, they report Kerlinger and Curry (2003) underestimated mortality relative to their mortality estimates. Such comparisons are unwise if they constitute different spatial or time periods. Of course, one must know the true mortality to say which estimate is 'better'.

This comment is incorrect. We compared the Kerlinger and Curry (2003) estimates not only to our estimates, but to the estimates made by previous investigators in the APWRA. Each comparison was of estimates made from data collected over the same time periods, and in the APWRA. Our results consistently demonstrated lower estimates made using WRRS data. We agree one can never know with 100% certainty which estimates are accurate, but we believe it is reasonable to conclude WRRS-based estimates are low relative to more rigorous estimates, based on the evidence.

As an example of how much WRRS can differ from our estimates of mortality, consider our follow-up examination of WRRS data maintained by Greenridge Services, LLC during our study. Of the 113 bird fatalities we reported for the EnXco turbines between November 2001 and April 2003, WRRS included reports of 10 of them, or 8.8%.

REFERENCES

Fisher R. A. 1924. The conditions under which χ^2 measures the discrepancy between observation and hypothesis. J. Roy. Stat. Soc., **87**, Ser. A, 442-50.

Fisher R. A. 1950. The significance of deviations from expectation in a poisson series. Biometrics, **6**, 17-24.

Greze B. S. 1939. Eksperimental'nye isseldovaniya nad potrebleniem planktona okunem-segoletkom. Izvest. VNIORCH, **21**.

Howell, J.A. and J. Noone. 1992. Examination of avian use and mortality at a U.S. Windpower wind energy development site, Montezuma Hills, Solano County, California. Final report. Prepared for Solano County Department of Environmental Management, Fairfield, California.

Howell, J.A., J. Noone, and C. Wardner. 1991. Visual experiment to reduce avian mortality related to wind turbine operations, Altamont Pass, Alameda and Contra Costa counties, California, April 1990 through March 1991. Final report. Prepared for U.S. Windpower, Inc., Livermore, California.

Hurlbert S. H. 1984. Pseudoreplication and the design of ecological field experiments. Ecol. Monogr., **54**, 187-211.

Ivlev V. S. 1961. Experimental ecology of the feeding of fishes. Yale University Press, New Haven, Connecticut.

Jacobs J. 1974. Quantitative measurement of food selection: a modification of the forage ratio and Ivlev's electivity index. Oecologia, **14**, 413-17.

Kerlinger, P., R. Curry, L. Culp, A. Jain, C. Wilkerson, B. Fischer, and A. Hasch. 2006. Post-construction avian and bat fatality monitoring study for the High Winds Wind Power Project, Solano County, California: Two year report. Unpubl. report to High Winds, LLC and FPL Energy. 136 pp.

Krebs, C. J. 1989. Ecological methodology. HarperCollins Publishers, New York. 654 pp.

Larsen K. 1936. The distribution of the invertebrates in the Dydso-Fjord, their biology and their importance as fish food. Rep. Danish Biol. Stat., **41**.

Neu C. W., Byers C. R. & Peek J. M. 1974. A technique for analysis of utilization-availability data. J. Wildl. Manage., **38**, 541-5.

Pearson K. 1900. On the criterion that a given system of deviations from the probable in the cause of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. Phil. Mag. Ser. 5, **50**, 157-75.

Shorygin A. A. 1939. Pitanie izbiratel'naya sposobnost' i pishchevye vzamootnosheniya nekotorych Gobiidae Kaspiiskogo morya. Zool. Zhur., **18**, 1.

Smallwood, K.S. 1990. Turbulence and the ecology of invading species. Ph.D. Thesis, University of California, Davis.

Smallwood, K.S. 1993. Understanding ecological pattern and process by association and order. Acta Oecologica 14(3):443-462.

Smallwood, K.S. 2002. Habitat models based on numerical comparisons. Pages 83-95 in Predicting species occurrences: Issues of scale and accuracy, J. M. Scott, P. J. Heglund, M. Morrison, M. Raphael, J. Haufler, and B. Wall, editors. Island Press, Covello, California.

Smallwood, K. S. and L. Spiegel. 2005a. Assessment To Support An Adaptive Management Plan For The APWRA. Unpublished CEC staff report, January 19. 19 pp.

Smallwood, K. S. and L. Spiegel. 2005b. Partial Re-assessment of An Adaptive Management Plan For The APWRA. Unpublished CEC staff report, March 25. 48 pp.

Smallwood, K. S. and L. Spiegel. 2005c. Combining biology-based and policy-based tiers of priority for determining wind turbine relocation/shutdown to reduce bird fatalities in the APWRA. Unpublished CEC staff report, June 1. 9 pp.

Smallwood, K. S. and C. Thelander. 2005. Bird mortality in the Altamont Pass Wind Resource Area, March 1998 – September 2001 Final Report. National Renewable Energy Laboratory, NREL/SR-500-36973. Golden, Colorado. 410 pp.

Smallwood, K. S. and C. Thelander. 2004. Developing methods to reduce bird mortality in the Altamont Pass Wind Resource Area. Final Report to the California Energy Commission, Public Interest Energy Research – Environmental Area, Contract No. 500-01-019. Sacramento, California. 531 pp.

Stauffer, D. F. 2002. Linking populations and habitats: Where have we been? Where are we going? Pages 53-61 in Predicting species occurrences: Issues of scale and accuracy, J. M. Scott, P. J. Heglund, M. Morrison, M. Raphael, J. Haufler, and B. Wall, editors. Island Press, Covello, California.

Tucker, V. A, 1996a. A mathematical model of bird collisions with wind turbine rotors. *J. Solar Energy Engineering* 118: 253–262.

Tucker, V. A. 1996b. Using a collision model to design safer wind turbine rotors for birds. *J. Solar Energy Engineering* 118:263–269.

ATTACHMENTS

A K. Shawn Smallwood and Carl Thelander. 2003. Assessing the knowledge base of avian mortality caused by wind turbines. Unpubl. report. 32 pp.

This attached manuscript was prepared in 2003 and might soon be updated and submitted to a professional journal. We included it with our responses to demonstrate to the reviewers the level of scientific rigor so far applied to the issue of bird and bat collisions with wind turbines. Research on this problem is in its infancy, and whereas shortcomings can be found in our research design, our study was still much more rigorous than most effort previously devoted to the problem. Mitigation measures are needed as soon as possible, so it is important to use the best available data to make decisions about which measures to use and to what extent to use them.

B K. Shawn Smallwood. 2006. Biological effects of repowering a portion of the Altamont Pass Wind Resource Area, California: The Diablo Winds Energy Project. Report to Altamont Working Group. Available from Shawn Smallwood, puma@davis.com . 34 pp.

This attached manuscript demonstrates the continuation of our research effort since the 2004 report. We have made significant advances in our understanding of the methodological biases and shortcomings in our research as well as in the research of others. The next mortality estimates we make will be more reliable, and so will use of our bird utilization data.