Perspective

QUALITY STANDARDS FOR SCIENTIFIC EVALUATION

Naomi Louchouarn[1], Tara K Meyer[2], Kelly J Stoner[3]

[1] *Carnivore Coexistence Lab, Nelson Institute of Environmental Studies, University of Wisconsin-Madison, Madison, WI, 53706, USA. Contact: louchouarn@wisc.edu*

[2] *Wildlife Biology Program, Department of Ecosystem and Conservation Sciences, W.A. Franke College of Forestry and Conservation, University of Montana, Missoula MT, 59812, USA. Contact: tara.meyer@umontana.edu*

[3] *Wildlife Conservation Society, 1050 East Main Street Suite #2, Bozeman, Montana, 59715, USA. Contact: kellyjstoner@gmail.com*

1. Introduction

Managing large carnivores is a priority for wildlife agencies and conservation organizations around the globe. Reducing livestock damages caused by carnivores and fostering coexistence are key objectives for successful management and conservation (Treves, 2009; USFWS, 2017). In 2016–2017, four independently published scientific reviews examined the efficacy of intervention methods used to prevent carnivore attacks on livestock (Miller et al., 2016; Treves et al., 2016; Eklund et al., 2017; van Eeden et al., 2017). Synthesizing these efforts, van Eeden et al. (2018) found that only 114 out of the 27,000 studies examined across these four reviews used rigorous, objective and quantitative experimentation standards. These were articles that initially came up using the search terms used by each team of authors. The initial number was then narrowed down using specific pre-determined qualities (i.e. the methods were quantitative and/or the species were large carnivores). This filtering process narrowed down the results significantly. Even commonly used intervention methods lacked rigorous scientific evidence of their effectiveness (van Eeden et al., 2018).

All four reviews reported feasibility and perception of intervention efficacy as key management decision factors for livestock producers and wildlife managers, but most evaluations of depredation intervention methods were opportunistic (van Eeden et al., 2018). The review by van Eeden et al. (2018) revealed that many currently employed methods are not effective or may even be counter-productive, meaning they either increased the number of depredations or reduced tolerance for carnivores when an intervention was ineffective. These results expose a significant need for greater rigor in experimentation. Wildlife managers and producers should use quantitative evidence of effectiveness whenever possible when making decisions about carnivore management and preventing livestock damages (van Eeden et al., 2018).

Treves et al. (2016) described standards of evidence for examining the effectiveness of depredation intervention methods and initially outlined two levels of rigor: gold and silver. Here, we summarize a new report (Treves, 2019) which clearly defines these standards of evidence and an additional bronze standard. Additionally, we will share examples of recently

A lioness feeds on and defends her prey in Botswana. *(Photo: Kelly Stoner)*

published studies employing these three standards and how they relate to the findings of van Eeden et al. (2018), identify common challenges to implementing these standards in the field and make recommendations for future research.

2. Establishing standards for experimental evaluation

The strength of a scientific experiment depends on whether the study successfully reduces biases in selection (how the test groups are chosen), treatment (how the interventions are applied), measurement (how data are collected) and reporting (including statistical analyses) (Treves et al. 2016; Treves, 2019; Treves et al., 2019). The three standards of evidence are therefore categorized according to their ability to reduce these four biases (Treves, 2019; Table 1). While the best scientific practice requires interventions to be assessed using a 'gold standard', designing and carrying out gold standard experiments may not be feasible in all real-world situations. We review the aspects of each standard described in van Eeden et al. (2018) and Treves (2019) and describe examples that illustrate the feasibility of each in practice below. The standards explored here should be applied when managers and researchers define method success or effectiveness as reducing livestock depredations by carnivores (Rigg et al., 2019).

2.1 Gold standard

The strongest standard of evidence, the gold standard, aims to eliminate biases by comparing randomly assigned intervention methods (treatments) with randomly assigned controls (i.e. no treatment) and employing a statistically appropriate number of replicates (Treves, 2019; Treves et al., 2019). For example, a number of independent livestock herds (replicates) can be randomly assigned to receive either an intervention or a control. Random assignment for each herd reduces selection bias (Treves et al., 2019), which is common in conflict-prevention studies since livestock owners may volunteer for treatments, researchers or wildlife managers may choose areas where they believe treatments would be most effective (e.g. Santiago-Avila et al., 2018) or effectiveness of methods may be self-reported rather than measured (e.g. Boast et al., 2016).

Treatment bias must also be eliminated or reduced by standardizing intervention implementation on the ground (Treves, 2019; Treves et al., 2019). This improves comparability between replicates, increasing the probability that results are based on carnivore responses to the intervention and not on differences in implementation (Treves, 2019). Ideally, gold standard experiments should also aim to reduce measurement bias by ensuring the measurements on replicates are made without knowledge of whether they are controls or treatments (Treves et al., 2019). In other words, if a herd is receiving an intervention, it is best for data to be collected by a researcher who is unaware whether the herd is or is not receiving an intervention. This is especially challenging because many depredation intervention methods are too conspicuous to be invisible to the researcher taking measurements. One way to attempt to reduce measurement bias is to have a researcher other than the field researcher take measurements (Treves, 2019).

Further biases can also be eliminated through the design of the experiment itself (Table 1). For example, a cross-over design is a method that allows replicates to be compared to themselves by having a randomly selected portion of replicates begin as controls and then switch to treatments and *vice versa* for the remaining replicates (Treves et al., 2019). This method allows researchers to account for potentially confounding variables that may make herds incomparable, such as the location of pastures. Confounding variables can make it difficult to design a field study with independent herds, such that researchers can correctly identify changes in predation risk as being due to treatments and not to other factors (Treves et al., 2016; Ohrens et al., 2019; Treves, 2019; Treves et al., 2019). The cross-over design also ensures that all herds receive a treatment at one point in time, which may make the experiment more palatable to participating livestock producers (Ohrens et al., 2019; Treves, 2019).

An exemplary peer-reviewed, gold standard study comes from Ohrens et al. (2019). This study in Chile used an experimental test on 11 herds of domestic alpacas *(Vicugna pacos)* and llamas *(Lama glama)* randomly assigned to control or treatment conditions, with a cross-over design to test a light deterrent against pumas (*Puma concolor*) and Andean foxes *(Lycalopex culpaeus)*. In this study, the researchers were able to isolate the effects of light devices in deterring pumas and Andean foxes by comparing each replicate to itself, thereby avoiding the difficulty of comparing herds that may have differences (e.g. predisposition to predation, individual differences in animals, etc.). Therefore, researchers in this study could make a strong inference that light deterrent devices could successfully deter pumas, but not Andean foxes. This result is not surprising given van Eeden et al. (2018)'s finding that deterrent devices were effective in 67–75% of 11 experimental or quasi-experimental studies. Interestingly, van Eeden et al. (2018) found that deterrent devices were effective in 95–100% of correlative studies examined (n = 29). The differences in results clearly illustrate the importance of standards of evidence that lead to strong inference when determining effectiveness of predator deterrence methods.

Gold standard experimentation can be challenging to implement in practice, particularly when a control is necessary for comparison. In order to achieve the highest level of scientific rigor, the experimental control ought to be the absence of any treatment. However, creating a true control may not be practical in these experiments because absence of any treatment would require leaving a herd (and therefore a producer's livelihood) entirely unprotected. For example, if the treatment is predator-proof fencing, then one might assume that the absence of the treatment (control) would be no predator-proof fencing, and thus in order to assess the real effectiveness of this method, no other type of prevention intervention should be allowed to be implemented by the producer. A more ethical solution would instead be to maintain the same base conditions between treatment and control groups (Treves, 2019; Treves et al., 2019). For example, if a producer habitually checks on his or her herd every few days, then the producer may continue to do so for both treatment and control while the erected fence acts as the treatment. Herds receive *more* protection under a treatment scenario than when they are a control group, instead of receiving *no* protection. This method is likely to be more acceptable to producers if scientists are testing an *added* prevention method while producers maintain 'business-as-usual' practices. Treves (2019) suggests that this is a particularly important distinction as it shows that gold standards of experimentation are more challenging, but not impossible, to implement.

Gold standard experiments, while resulting in the most consistent and rigorous scientific inference, require studies to be developed with very specific conditions (Treves, 2019). Unfortunately, this means that gold standard studies will rarely use previously col-

Table 1 Three common biases, how to avoid them and the strength of inference that can be achieved when using gold, silver and bronze standards of experimentation. Adapted and expanded from Treves (2019).

Standard of evidence	**Gold**
Definition	Randomly selected control and treatment groups which are statistically comparable.
Types of biases[1]	
Selection bias	**None.**
Treatment bias	**None.**
Measurement bias	**Sometimes.** Avoided if the researcher collecting data is unaware of whether the replicate is a treatment or control.
Potential conclusions	Can isolate treatment effects from potential impacts of confounding factors such as time, spatial characteristics and other differences between replicates.

Standard of evidence	**Silver**
Definition	Depredations are measured multiple times over the study period before and after a treatment is implemented (in which case controls come before treatments), and/or treatments are compared to controls but one or both are not randomly selected.
Types of biases[1]	
Selection bias	**Yes.** Treatments and/or controls are not randomly selected.
Treatment bias	**Sometimes.** Avoided by using a cross-over design and standardized implementation between treatment replicates.
Measurement bias	**Often.** Same as with gold standard but more likely to occur when no controls are used. May be avoided if the researcher collecting data does not know what the intervention is, but this is rare.
Potential conclusions	Can isolate treatment effects from many confounding factors such as treatment implementation, but not necessarily from spatial or time variables.

Standard of evidence	**Bronze**
Definition	Depredations are measured on replicates where treatments are already being used or have just been implemented in response to a depredation. Rarely a control. Correlative studies.
Types of biases[1]	
Selection bias	**Yes.** Non-random selection of treatment replicates and treatments are often implemented as a result of depredations.
Treatment bias	**Yes.** Treatments are harder to standardize, usually because they have been implemented before the study begins.
Measurement bias	**Often.** As for silver.
Potential conclusions	Can identify potential patterns and correlations between treatments and outcomes but cannot isolate effect from time, spatial patterns, implementation (unless this is controlled for) or other potential confounding factors.

[1] Note that there is always potential for reporting bias, but we have not included it here since this is a bias that should be eliminated based on ethical scientific reporting standards. For more on this bias, see table in Treves (2019).

Livestock held in an effective *kraal* in Botswana. This is an example of very effective fencing using purchased or found materials.

An example of an ineffective fence that is permeable to predators in Botswana.

(Photos: Kelly Stoner).

lected data. Due to this, we observed that silver or bronze standards of experimentation are more commonly found in the recent literature for evaluating depredation prevention tools and methods.

2.2 Silver standard

Silver standard experimental designs lack the random assignment of treatments and/or controls, and are often longitudinal over time, i.e. the effectiveness of the treatment is measured at multiple points along a timeline (van Eeden et al., 2018; Treves, 2019). In most longitudinal studies, either controls are not used at all or there is no specific record of control conditions occurring prior to implementation of the treatment (Smokorowski and Randall, 2017). This means that changes observed during a study could be the result of treatments or other factors such as time or seasonal conditions (Treves, 2019). Furthermore, the lack of random assignments may inadvertently introduce selection bias. Researchers could unintentionally select replicates predisposed to depredations (or *vice versa*) for replicates receiving treatments. However, silver standard studies still allow researchers to reduce other biases such as treatment and measurement biases as they allow a great deal of control over intervention implementation and measurement of predator responses (Treves, 2019).

A recent study by Weise et al. (2018) examined the efficacy of fortified *kraals* (predator-proof night enclosures) in reducing carnivore attacks on livestock in the Kavango Zambezi Transfrontier Conservation Area, in Botswana. This study randomly assigned some herds as control groups (e.g. controls used unfortified *kraals*, therefore they were fenced but not predator-proof), but it did not randomly assign treatments; instead, researchers found and included producers who already used fortified kraals. The authors examined the number of livestock attacks in both treatment and control herds over 18 months. Because treatment herds were not randomly assigned, control groups were spatially separate from treatment groups and the environmental conditions for these control groups (e.g. geographic features, dominant landcover types, predator density, wild prey density, etc.) were not recorded for treatment groups. Thus it is difficult to conclude whether attack occurrences or absences were due to the fortified *kraals* or another external variable. However, comparing randomly assigned controls and treatment *kraals* over time enabled the researchers to minimize some treatment biases (e.g. differences in *kraal* type, style and maintenance) and allowed them to isolate the effect of *kraal* implementation.

The experiment found fortified *kraals* to be more effective at reducing predator attacks but that *kraals* required a great deal of maintenance to stay effective. This result is consistent with the findings of Eklund et al. (2017) and Treves et al. (2016) (as referenced in van Eeden et al., 2018): 66% of high inference studies on enclosures found them to be effective. However, about 22% of the studies showed enclosures to be ineffective, perhaps because their effectiveness was highly reliant on frequent maintenance (Weise et al., 2018). Despite having a weaker strength of inference than gold standard, silver standard experiments are easier to implement and accommodate situations where researchers and managers have less control. As with the Weise et al. (2018) study, silver standard experimentation allows for greater use of existing intervention efforts.

Another example of the silver standard of inference was published by Santiago-Ávila et al. (2018). In this study, the authors used pre-existing data collected by the Government of Michigan to examine the effectiveness of lethal methods versus non-lethal methods for wolf-livestock conflict prevention. These authors retroactively compared the data from lethal efforts to a variety of non-lethal methods employed by state wildlife managers. The authors considered the herds protected by non-lethal intervention methods to be pseudo-controls, because wildlife managers would sometimes choose to forgo lethal control and instead provide livestock producers with non-lethal deterrents (Santiago-Ávila et al., 2018). Because the field agents made non-random decisions about where to implement lethal control, the method in which herds were assigned either lethal or non-lethal control introduced selection bias. The authors accounted for spatial variation and the potential for treatment bias by comparing an intervention site to itself over time (cross-over design). However, they could not account for the selection bias imposed by field agents (Santiago-Ávila et al., 2018). In this study the researchers were able to eliminate sufficient confounding variables (e.g. spatial variation) in order to isolate the effect of certain depredation prevention methods. Therefore, while not all biases are removed, statistical analyses from silver standard studies may be used to draw conclusions about the relationship between variables and outcomes (Treves, 2019).

2.3 Bronze standard

The third standard of evidence is the bronze standard, which relates primarily to correlation studies (van Eeden et al., 2018; Treves, 2019). Correlative studies have a lower power of inference because they examine the effects of interventions non-systematically (resulting in treatment bias), they usually do not use control replicates and they are frequently implemented in response to livestock losses (thereby they do not reduce selection bias; Treves et al., 2016; van Eeden et al., 2018; Treves 2019). A recent example of a bronze standard study comes from Boast et al. (2016). This paper examined the effects of cheetah (*Acinonyx jubatus*) translocations on livestock losses. Data for this study were collected after a livestock kill occurred and no controls were used (e.g. no comparisons were made for depredation events in areas where translocations had and had not occurred) (Boast et al., 2016). Therefore, it is possible that other factors may have confounded the results. Van Eeden et al. (2018) described only five peer-reviewed studies on translocations as a predator deterrence method, all of which were correlative and one of which found translocations to be counterproductive in preventing conflicts.

Bronze standard experiments are quite common in scientific literature about depredation prevention, likely because they usually cost far less than gold or silver standard experiments and they can be conducted opportunistically. For example, it is simpler and less expensive to do a bronze level analysis of cheetah translocations that are already occurring in response to livestock losses than it is to design and implement a new cross-over gold standard experiment. While correlation studies cannot isolate causal links, they can identify potential patterns of depredation as a result of intervention methods. Van Eeden et al. (2018) suggest that, due to the lower strength of statistical inference in correlation studies, it would be best to use these as preliminary studies that identify methods for more rigorous testing.

3. Recommendations and future research

When implementing intervention methods to prevent livestock depredations by carnivores, either for experimental or functional purposes (or both), it is important not only to select the appropriate method(s) but also to implement them consistently and effectively. Intervention methods are applied across a diversity of ecosystems and species, and their effectiveness in various contexts should be carefully and rigorously examined (Rigg et al., 2019).

We encourage further research to be focused on:

1. designing high quality experiments to rigorously test the functional effectiveness of intervention methods, as suggested by van Eeden et al. (2018);
2. examining the relationship between functional effectiveness of intervention methods and likelihood of method use by producers (i.e. whether quantitative evidence of intervention effectiveness influences which method(s) a producer chooses to implement); and

A foxlight placed on a woodpole next to a llama or alpaca sleeping site in the altiplano of Chile. *(Photo: Omar Ohrens)*

3. using rigorous social science methodologies to qualitatively evaluate the links between livestock depredation reductions and any resulting cultural shifts in how carnivores are perceived or accepted on the landscape.

Conservation practitioners, whether they be wildlife managers, non-profit organizations or researchers, will be invaluable in achieving these research goals, as they are likely to best identify which methods are used locally, how to implement an experiment cost-effectively and how to communicate with participating producers in order to examine its effectiveness.

When choosing to study or implement conflict mitigation methods we recognize that managers, researchers and conservationists have varying definitions of effectiveness. In general, intervention effectiveness is commonly understood as either reducing the frequency of depredations, improving producer tolerance for depredation events, reducing the killing of carnivores in retaliation to depredations, or a combination of these. Thus it will be important for researchers to have clear goals and a clear definition of the desired method effectiveness from the outset of each study.

Finally, we recognize that a key goal of evaluating depredation prevention methods is to understand their relative efficacy, enabling wildlife managers, conservationists and producers to select the most effective method(s) for their situation. However, we should note that the effectiveness of a method that is tested using high standards is not absolute, because the effectiveness will vary given there are infinitely diverse environmental and human factors and conditions (species dynamics, weather patterns, geography, socio-political dynamics, etc.; Treves, 2019). In order to assume that a method will match the effectiveness in multiple contexts, these dynamic factors would all have to be exactly the same. Therefore, these evaluations should be used as guides to help producers, managers and conservationists understand which tool or suite of tools are *more likely* to be effective in a particular scenario with regard to the specific carnivore species, habitats and livestock involved. Decisions about which techniques to use are likely to be influenced by a number of other factors besides their efficacy, including cost and availability. Understanding the relative effectiveness of key conflict intervention methods will save wasted resources spent on ineffective methods and lend credibility to decisions as managers and researchers update management plans and respond to livestock damages caused by carnivores.

Acknowledgements

We would like to thank Lily van Eeden, Ann Eklund, Jennifer Miller, Miha Krofel, Adrian Treves, José Vicente López-Bao, Guillaume Chapron, Mikael Cejtin, Mathew Crowther, Christopher Dickman, Jens Frank, David Macdonald, Jeannine McManus, Arthur Middleton, Thomas Newsome, William Ripple, Euan Ritchie, Oswald Schmitz and Mahdieh Tourani for providing the inspiration for this work and contributing their expertise from the corners of the globe. We thank the editors at *Carnivore Damage Prevention News,* whose comments and suggestions helped improve and clarify this manuscript.

References

Boast LK, Good K, Klein R (2016) Translocation of problem predators: is it an effective way to mitigate conflict between farmers and cheetahs (*Acinonyx jubatus)* in Botswana? Oryx 50(3), 1–8.

Eklund A, Lopez-Bao JV, Tourani M, Chapron G, Frank J (2017) Limited evidence on the effectiveness of interventions to reduce livestock predation by large carnivores. Sci. Rep-UK 7, 2097.

Miller JRB, Stoner KJ, Cejtin MR, Meyer TK, Middleton AD et al. (2016) Effectiveness of contemporary techniques for reducing livestock depredations by large carnivores. Wildlife Soc. B. 40(4), 806–815.

Ohrens O, Bonacic C, Treves A (2019) Non-lethal defense of livestock against predators: Flashing lights deter puma attacks in Chile. Front. Ecol. Environ. 17(1), 32–38.

Rigg R, Ribeiro S, Colombo M, Lüthi R, Mettler D, Ricci S, Vielmi L, Zingaro M, Salvatori V (2019) Evaluation of prevention measures: can assessment of damage prevention be standardised? Carniv. Damage Prev. News 18, 24 –30.

Santiago-Avila FJ, Cornman AM, Treves A (2018) Killing wolves to prevent predation on livestock may protect one farm but harm neighbors. PLoS One 13(1), e0189729.

Smokorowski KE, Randall RG (2017) Cautions on using the before-after-control-impact design in environmental effects monitoring programs. Facets 2, 212–232.

Treves A (2009) Hunting for large carnivore conservation. J. Appl. Ecol. 46, 1350–1356.

Treves A (2019) Standards of evidence in wild animal research. Report for the Brooks Institute for Animal Rights Policy & Law.

Treves A, Krofel M, McManus J (2016) Predator control should not be a shot in the dark. Front. Ecol. Environ. 14(7), 380–388.

Treves A, Krofel M, Ohrens O, Van Eeden LM (2019). Predator control needs a standard of unbiased randomized experiments with cross-over design. Front. Ecol. Evol. 7:462. doi: 10.3389/fevo.2019.00462.

USFWS (2017). Endangered and threatened wildlife and plants; removing the Greater Yellowstone Ecosystem population of grizzly bears from the Federal List of Endangered and Threatened Wildlife. Federal Register 82(125), 29699–30008.

Van Eeden LM, Crowther MS, Dickman CR, Macdonald DW, Ripple WJ et al. (2017) Managing conflict between large carnivores and livestock. Conserv. Biol. 32(1), 26–34.

Van Eeden LM, Eklund A, Miller JRB, Lopez-Bao JV, Chapron G et al. (2018) Carnivore conservation needs evidence-based livestock protection. PLoS Biol. 16(9), e2005577.

Weise FJ, Hayward MW, Aguirre RC, Tometetso M, Gadimang P et al. (2018) Size, shape and maintenance matter: a critical appraisal of a global carnivore conflict mitigation strategy – livestock protection *kraals* in northern Botswana. Biol. Conserv. 225, 88–97.