Butterfly Conservation Management in Midwestern Open Habitats

Part 2: This science is controversial, isn't it? by Ann B. Swengel

Summary. You may have heard already that there's lots of heat on the topic of how prairie and savanna management affects insects. But I find the results (the actual data and observations) remarkably compatible with each other, once organized. Useful ways of categorizing research include degree of ecosystem degradation, kinds of species studied, how those data are grouped (e.g., species richness, group abundance, individual species), research design (e.g., separate plots within one site or many study sites), and time depth (one or a few years; long term). Since statistics are an integral part of science today, it's very useful to understand how they work, what they do, what they can't do, and what to watch out for. Science is limited to what scientists are able and willing to study. It also can't provide the answer you most want: what will happen in your particular case in the future. Scientists often speak dismissively of anecdotes but your one site is an anecdote that is very important to you. That's why what doesn't look very risky in a large scientific study can start looking awfully risky when it's all or nothing in your individual situation.

You may have heard already that there's lots of heat on the topic of how prairie and savanna management affects insects. In fact, you may have heard there's relatively more heat than light, and the bits of light out there are confusing and incomplete, resulting in uncertainty and controversy. There are lots of reasons for this.

First, and most important, that's normal. Scientists test ideas and debate, probing and arguing alternate points of view. New information may show us something new, but it may also make us re-consider what we already "knew" in a different way. In fact, entirely new scientific breakthroughs may come entirely from re-examining already existing data and "discovering" something new in them to understand. That's the beauty and challenge of science. It's tempting to think that once a scientist has published a paper and "established" some finding, that settles it. However, that's actually just the beginning. As colleagues test and expand on that work, we go back and re-evaluate the old work and may understand it both differently and better. Theoretically, scientists should view data and studies objectively, based on how well the interpretations explain and predict patterns. But in practice, scientists are as human as everyone else. They want the theories they've advocated to "win" and rivals to "lose." So besides the usual amount of debate, there can also be non-scientific reasons for the preponderance of heat over light. If you are a patient seeking advice on a healthy lifestyle, or an amateur hobbyist interested in butterflies, you need to evaluate independently as best you can, respecting experts for their greater knowledge but leaving room for their (and definitely my) limitations and foibles. Remember, a study is not refuted scientifically by shouting or taking a vote but rather by data disproving (precluding the possibility of) that hypothesis. However, shouting and votes can definitely block an independent scientist from getting a study out in the arena for wider consideration, or from getting these findings applied.

Second, butterflies have not been "behaving" as expected around the world. For example, the best intended reserve management for the Large Blue led directly to the extinction of this butterfly in England. An inspiring willingness to embrace this stark outcome and take responsibility for it head on, instead of denying it, coupled with much more research on the species where it still occurred, has resulted in the successful re-establishment of a number of populations of this butterfly back in England. I would caution though that other butterflies have declined and not been so easily restored. Likewise, a detailed analysis of decades of data indicated that British populations of rare butterflies on preserved and unpreserved sites declined and disappeared at equally alarming rates, although for disparate reasons. Did you gasp over that? I know my eyes bugged out when I first read that study. Are butterflies the uncooperative exception of the invertebrate world? I doubt it. Rather, butterflies are well enough known that we can tell better what's missing now. Are the British worse at butterfly conservation than the rest of us? I think it's much more likely that they're actually doing a far better job of documenting, and owning up to, what's happening than the rest of us. In fact, these British results were a major impetus for large improvements in butterfly conservation there, with an even stronger emphasis on long-term butterfly monitoring to achieve more successful results.

Third, why should ecological concepts predominated by vascular plants and vertebrates, which are better known, automatically be valid for insects? If we think we have to study plants and vertebrates in order to understand them and how to conserve them, and we couldn't just imagine it in a vacuum without actual field research, then the same applies to insects.

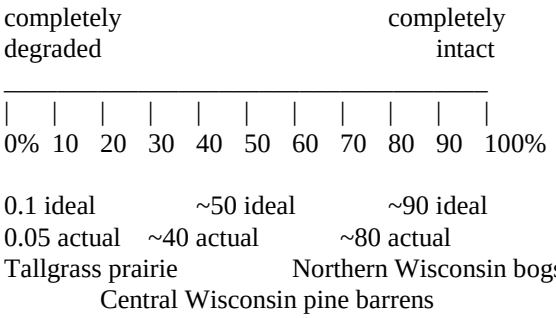
How much disagreement is there? Actually, considering the vast number of insect species and wide array of sites, I find the results (the actual data and observations) are remarkably compatible with each other. That is, once I organize the studies, to make sure I'm comparing (in the idiom) apples to apples and oranges to oranges. Science is about replication, and at first blush, the seemingly endless variety may bring confusion instead of clarity. But consider the parable of the six blind men describing an elephant, each focusing a different part of the same animal. Likewise, each

study describes a part of the picture. With a framework for how to organize each study's part of the whole, a meaningful picture can result. If it's still fuzzy or jumbled, that doesn't mean the data are refuting each other. It may just mean we haven't figured out yet how to imagine the framework to organize ("ordinate") the data compatibly.

CLASSIFICATIONS (THEMES)

The messed-up meter: The most important framework I use is this to characterize ecosystems or vegetative classifications at a landscape scale.

MESSED UP METER: ecosystems

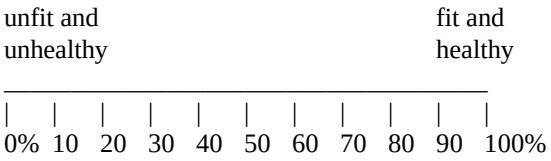


On my "messed-up meter", zero is totally destroyed and 100 completely pristine. Tallgrass prairie is 99.9% destroyed, so that's a maximum, or "start value" (to borrow from athletic scoring), of just 0.1. But what's left is degraded and fragmented (broken into small patches and isolated), so cut that at least by half to 0.05 for actual "execution" (to borrow again from athletics) or conditions of these sites. By contrast, we've also surveyed a lot in sphagnum bogs (peatlands). Limited just to the bogs themselves in northern Wisconsin, that's a start value of about 90, since relatively little outright destruction has occurred there. With some degradation of what's left, cut that down to maybe 80 for execution. While these bogs are relatively intact, they are also naturally fragmented, occurring in less than 1% of the northern Wisconsin landscape. Nonetheless, they contain numbers of specialist butterfly species comparable to other more extensive natural habitats in the same region. Central Wisconsin pine barrens and bogs (both subjects of other articles published online by SWBA) would fall somewhere in the middle, but below 50 for execution.

I place not only sites but also faunas on this messed up meter. How much of the original fauna is still there? Is it even possible to know the answer to this question? Or was the fauna strongly altered before much scientific record of it was obtained? A lot of research on specialists comes from the messed up landscape, but this may not really tell us how they occurred in intact landscapes. In other words, this research may not tell us their optimal niche or preferences, or full range of suitable situations. Conversely, when we look at more pristine landscapes, they may not tell us what needs to be done to enable butterflies to "perform" successfully in the landscape we humans affect and live in.

Here's an analogy to human health.

MESSED UP METER: human health



An obese sedentary diabetic chain-smoking cardiac patient is on the unfit and unhealthy end of the spectrum, while a marathon runner with a healthy diet and low cholesterol is on the fit and healthy end. The way the former performs under a given regimen is very different from the latter. In no way do I diminish the value of people on the unfit end of the meter. I, and you no doubt, have or had loved ones with some combination of various health disadvantages. But a one-size-fits-all approach to diet and exercise can't be appropriate for both. The person on the left end might die from sudden, too strenuous activity while the person on the right might decline in fitness by being placed onto the more ordinary exercise and diet regime of us average mortals. On the other hand, a healthy person can sustain a lot of subpar regimes yet still manage OK, while the unhealthy person offers little margin for error.

Besides the messed-up meter, I classify field research I read in additional ways: by types of animals and how the data on them are organized, and by types of sites or contexts. By categorizing studies and their results along these lines, I obtain a more coherent and predictable view of the bigger picture of how management affects insects.

What kinds of species are being studied? For example, as a group, ants and grasshoppers tend to be more tolerant of fire and grazing while butterflies and leafhoppers are more sensitive, independent of the continent, ecosystem type, and particular species involved. Predatory or herbivorous species? You might think that being a bit higher up the food chain makes a species more vulnerable to management, but as a group (averaging across many species) herbivores actually tend to be more sensitive. This may be because more herbivores are more specialized about their food requirements than predators are, which can be more opportunistic about what they eat or at least more willing to eat a variety of foods.

Piles or species-specific analysis? By the former I mean grouping results at a level higher than species (e.g., family or order), such as the number of individuals or grams of beetles (regardless of species) in a sample. These results tend to be swayed by whatever are the most common species in the sample, without identifying what those particular species are. Just imagine if you did that with plants. Would you analyze only by how many stems or grams were grasses (of any kind) or legumes or composites, without any reference to whether they were native or alien, common or rare,

typical of that habitat type or unusual there? Do you really only care about how many individuals or pounds of birds are in a site, regardless of whether they are starlings or meadowlarks? However, it's a lot easier to weigh piles or count individuals coarsely identified to invertebrate group, which runs up analyzable sample sizes more readily. It's much harder and more time consuming to identify to species.

Species-level presence-absence or abundance? Abundance data can appear more difficult to analyze for effects of management, since how many of a butterfly species you can find depends on a number of other factors as well. How can all these be adequately controlled? Even under a wide variety of conditions, the species may still be findable, even if in quite varying numbers. We butterfliers experience this a lot, when we visit a site early in the morning or in poor weather or early in the flight period compared to visiting again later in the day or in better weather or next week. However, whether you find any individuals at all is just as dependent on the same factors as whether you find many or few. So the problem of "false negatives" (finding zero when the animal is actually present) is just as much a problem in presence-absence analysis. But the full range of possible positive occurrences (from vagrant to abundant) is compressed into a single value (present). As a result, presence-absence is a weak way to detect a pattern. In an abundance analysis, you can detect a decline of a species locally when it goes from abundant to just common, or from that to just reliably present in low numbers. In presence-absence analysis, the change only registers once you can't find any at all and have reliably distinguished this observed absence from a false negative. Abundance studies also are actually easier to design than presence-absence since the latter can only be fairly compared if effort at all the sites is similar. With abundance, as long as a reasonable minimal effort is conducted per site, so that zero reflects low numbers rather than inadequate effort to find what is actually prevalent, observation rates (individuals per time or distance) can account for unequal effort per site.

There's an irony for me regarding presence-absence and detectability. Having cut my teeth, so to speak, on studying forest owls, I have an immense appreciation for the concept of "detectability." On any survey of anything, plant or animal, I am keenly aware that I hope to find what I'm looking for, but I will not find all individuals of any species and may not find any even when they're present. I cannot actually know the total number of individuals present when I survey. Since science is about observable phenomena, I prefer to work from actual observed numbers, converted to observation rates (individuals observed per hour or kilometer surveyed) to make them comparable among sites, since we do not have a set survey length among sites. I then try to account for factors that affect how many were observed, including ones that affect detection (e.g., weather). Others like to work with extrapolations that try to calculate how many probably were there, including ones not seen, based on the distribution of individuals seen relative to how far away they are from the surveyor, or based on how many individuals marked on a previous day that are among the individuals found today. These ratios are used to infer how many individuals are actually present in the survey area compared to how many found. There are lots of assumptions required, and difficulties meeting these assumptions, and that's for discussion elsewhere. Nonetheless, when these extrapolations are performed, the outputs are seemingly astronomical numbers, possibly even 50 or more times the number of butterflies actually seen. So that's why I find it ironic how quickly others may state that a butterfly is "absent" from a site. If for each individual seen, there may be 49 (or more or fewer) others not seen, then it seems to me that when you drop from a couple seen one year to none seen the next year, and the next, shouldn't we place that into the "undefined" state of either subdetectable (working our way down through the 49, then 48, and 47 individuals present but never seen) or absent (since I also sure don't assume an animal is present if no one actually saw any with a reasonable effort during the flight period). Alarms bells should go off (actually, alarm bells should have already gone off, as the animal declined from readily detectable to barely detectable) but it's dangerous for conservation of biodiversity to declare the patient dead and pull the plug prematurely too. I am much more willing to allow for the possibility that the animal is present but undetected, while also very much wanting positive evidence for this too.

Value of presence-absence analyses. Sometimes the goal may be to assess rapidly and efficiently the areas occupied by a species, for example, at an installation that has to reduce negative effects on an endangered species from activities otherwise lawfully occurring at the facility. As soon as presence is found, possibly at the start of searching a site, then surveying is stopped to go elsewhere. This optimizes the number of sites visited over obtaining rigorously comparable abundance data from all sites. A protocol is still needed for determining how much surveying is needed before giving up if the species is not found. This is to obtain a certain level of confidence in a negative result. In this case, it may also be preferable to use informal survey routes rather than fixed survey routes. Experts familiar with a species can be more effective at detecting it when not constrained by a fixed survey route, but instead being allowed to check out locations that appear most appropriate at that particular time (e.g., nectar patches). Fixed routes are usually used for monitoring abundance over time.

Species lists or species-by-species analysis? A lot of analyses are done as "species richness": how many different kinds of species of a group did you find in a site or sample? This is a kind of "pile" too: a single value of how many species were found. This has the effect of reducing the analysis to presence-absence with abundant residents and rare vagrants accorded equal weight. Unless the species are subdivided into ecologically meaningful groups (such as prairie specialists, migratory generalists, and so on), different sites can seem statistically similar in richness but still

have very different faunas relative to conservation value, or the same site can seem steady over time in species richness, yet lose its local specialties while gaining more common species. On the other hand, individual species-by-species analysis provides a lot more information, but this can be overwhelming in the time required and myriad variations of patterns. It then becomes helpful to classify these species into affinity groups, to see whether different groups have different patterns of response. For example, in climate change analyses, moths may be grouped by what life stage they overwinter as, or where in the vegetation they consume food as caterpillars (in the grass layer, or shrubs, or tree-tops), or how many kinds of food plants they are known to consume (one, few, or many). Likewise, European butterflies that overwinter as eggs or caterpillars are faring much more poorly with climate change than butterflies that overwinter as adults.

What kind of species concept are we talking about? The concept of the "morphospecies" has been used to "identify" different species where the faunas are so poorly described that it is not possible to identify to an actual species described by science. Individuals are identified as species #1, #2, #3, and so on, based on appearance. Even though this can fail to distinguish very similar species and can identify as separate species what is one species variable in appearance or differing between males and females, some studies have found that lay workers using the morphospecies concept come up with similar species counts to scientific experts in that invertebrate group. This practice occurs more so, as you can imagine, with more obscure groups of invertebrates and in tropical and poorer countries, which have way more species and/or fewer resources to study them. But relative to a given site, it is difficult to determine whether the morphospecies is a resident or vagrant, whether the site is core or peripheral habitat for it, whether the species is a local endemic or widespread generalist. Even where species are scientifically named, for little studied groups and areas, it may still not be possible to answer these important questions about the species, or to do so any better than roughly.

With better known faunas, well described and studied scientifically, not only is it possible to assign a name to the species but also to describe its range (widespread or endemic), food requirements as immatures and adults, and habitat associations. Butterflies can be organized by gross habitat type (forest, savanna, grassland, wetland). Then within a habitat type, scientists often use a two-way split: specialist or generalist (or similar either-or terminology, such as localized or matrix). I've found that too limiting and prefer a four-way split, which I apply not only to my own work but, as possible, when examining others' studies. For example, in prairie and savanna, I categorize as specialist (restricted or nearly so to native herbaceous vegetation), grassland (widely occurring in both native and degraded grasslands), generalist (occurring in grassland and other habitat types such as forest or wetland), and immigrant (migratory or vagrant, usually not resident year-round). My European bog colleagues have greatly endeared themselves to me with their analogous, if fancy-named, four-way categorization of species in bogs: from tyrphobiontic (bog specialist), then tyrphophilic to tyrphoneutral and finally at the other extreme, tyrphoxenous (non-breeding vagrant). I also find that a four-way split enables more patterns to become apparent along the spectrum from one end to the other. A two-way split may put too many species in the "specialist" category, thus washing out the more sensitive responses of the more specialized species, but even so, the "generalist" category may remain such a melting pot of diverse species as to produce little pattern too. However, only better known faunas can be categorized effectively in more detailed and subtle manners.

What fauna is weighting your statistics? A fundamental question I ask of any management study is what species are most swaying the statistics? Are these species that are locally restricted to your study site(s) or do they occur widely in adjoining areas nearby? Are they primarily restricted to rare examples of specific native vegetation types or are they common species occurring widely in the unconserved landscape?

What is the scale of the study plots? Are management treatments done in experimental plots or actual "life-size" plots? The former approach has wonderful variable control, and can often ensure very similar vegetation among different experimental treatments. But 1x1 or 10x10 or 20x20 meter squares are very different in effect on butterfly populations than quarter quarter sections (40 acres) and especially quarter sections (160 acres) or larger. What management treatments look safe and vegetative results look suitable on the very small scale may not be so on the larger scale of real management units, and does not simulate what happens in the real world. This occurs in medicine too. A little bit of a drug can be safe or even life-saving while an overdose or inappropriately timed dose of something even as innocuous as aspirin can be lethal. So tiny plots can examine whether a vegetative composition and structure are conditions the animal is willing to use or not. But it doesn't indicate whether a population of that animal can survive when only that kind of vegetative condition generated only by that management regime is available.

Separate plots in the same site or entirely distinct sites? The former seems like it's doing what science requires: controlling all other variables such as climate, weather on sampling day, site history, vegetative composition, and so on. But the proximity washes out some differences between treatments, since some butterflies can disperse among different treatment plots. When a study uses separate sites, are these fair comparisons among sites? In other words, are they otherwise about comparable except for the specific variable being studied and are there enough of these different sites in the sample to wash out the random differences among them? That may not be possible. In that case, the range of variation in these other confounding va-

riables needs to be represented in each management. That is, small and large sites, more and less degraded sites, upland and lowland sites, and so on need to be represented in each management type. If all hayed sites are small, degraded, and lowland, and all burned sites are large, undegraded, and upland, or vice versa, then it's difficult to identify what in the butterfly results relates to patch size or vegetative quality or topography or management. Alternatively, if sites or parts of sites start out with different butterfly abundances, butterfly trends on these patches over time can be compared against management.

What is the research design or premise? A comparison of some sort is usually explicit, with one type of site or condition being compared to another. Or the comparison is implicit: a particular situation is described, and this is compared via other scientific papers describing other locations or situations. It's important to identify any controls (the "treatment" group and what is the non-treatment group). Treatments and controls are more suited to an experimental design, with experimental plots and the size-scale issues mentioned above. However, this is highly attractive to scientists because of the ability to control variables. Alternatively, in the "natural experiments" occurring out in the landscape at large, the study may use the approach of an outgroup instead of a deliberately designed control. For example, if the "subject" ("treatment") group is a conservation management approach to prairie management, an outgroup could be sites of native prairie flora in a consistent, unintensive agricultural usage. Or if the treatment group is specialist butterflies, the outgroup could be widespread generalist species occurring in the same sites. The goal is to allow for contrasts to become apparent, if they exist. An outgroup is a kind of control, but I prefer the term "outgroup" because it more accurately describes the opportunistic nature of the comparison being made in natural experiments. When I read studies, I ask myself whether these comparisons are fair. Or fair for answering what? Are there controls or outgroups? Do I agree that they're fair?

What is the context? For example, if a study reports that a bird species prefers taller grass, does that mean you should manage for 3+ foot tall turf? Well, in the specific case I have in mind, the cited study was comparing a turf height of perhaps a foot tall to even shorter, more heavily grazed turf. To us tallgrass prairie folks, that's all short grass. Yet this study was being cited to promote taller grass than studied, either because it wasn't noticed that those heights weren't studied or the results were being projected beyond the heights that had been studied. But in the context of others' studies on the same bird, it actually prefers what I consider shorter grass, in the context of tallgrass prairie preserves—a foot tall or less compared to taller grass.

What kind of setup does the study have? The most common approach to management study is the *"slice of time"* approach. A series of sites is surveyed now, in one or a few years, with each sampling area classified as to management observed now. Then total number of species or individuals observed, or individual species presence and/or abundance, get sorted out by different management characteristics. This method is the one we started with because it's the easiest way to get started with a sample, and you have to start somewhere. However, there's a tremendous number of possible variables that might be relevant to sorting out the data, and it's easy to overlook some, especially ones that are harder to figure out. It's also difficult to get enough independent examples to sort these variables out. For example, what if all small sites are also managed the same way? If management history is known, you can add a *"retrospective"* aspect to the management variables. Take the observed butterfly occurrence and abundance now, and sort them by different management histories. The logic is that the current fauna is the sum of its site history. However, as medical research has found, reconstructing past regimes (what kind of diet and exercise habits the subjects had) can be fraught with incomplete and inaccurate reporting. Furthermore, the variables describing the past that are most relevant to the current condition may not be identified. After just a few years, it's possible to separate data into *"before" and "after"* (comparing the outcome in plots before and after treatment compared to control plots, or comparing subject sites to outgroup sites over time). As with the slice of time, though, the question is whether your "before" really represents a "before" period, since, as with most medical research, the subjects have a lot of unscientifically documented history before the "before" period. The most rigorous approach, both in management and medical research, is *"prospective*." Identify an adequate sample now and follow it into the future. Unfortunately, if you aren't in control of management at all sites, you have to be lucky or clairvoyant in site selection to get adequate samples of each kind of treatment you want to study. Alternatively, you have to have a tremendous amount of resources (to have large enough samples to ensure that even with some sites or people dropping out of the study, you still have adequate representation of each treatment) and patience (to wait for the long-term results finally to happen). As years accumulated into decades in our research, Scott and I have now done all of these approaches.

One approach we haven't particularly used is "*modeling*": creating a computer model to project how something is predicted to occur. It may seem obvious, but I have to state this. The study needs to have a component of validation to verify whether the model is making correct predictions. Some studies have been portrayed as "showing something", but when I actually read them, all they are is a model about how things are thought to occur, with the end result flowing out of that computer construction. When I probe further, I may find no actual validation (independent test of the model with actual field data) or data collection after the model was developed, to test whether the prediction actually happened. In such cases, there may not be a follow-up study evident a few years later to test the validation. So it's left to us readers to watch for ourselves whether it looks like the prediction is happening or not.

THE POWER AND PITFALLS OF STATISTICS

Statistics are an integral part of science today. I do not want to give the impression that only numbers and statistics matter. Actually I think the hardest part of science is the thinking part—finding patterns to test, hypotheses to imagine. However, most research even in field science is presented statistically. To participate in modern science, an understanding of statistics is invaluable. So it behooves me to understand how they work, what they do, what they can't do, and what to watch out for. Many people view statistics as a "black box", something they don't think they can or want to understand. This can be dangerous—leading to a blind faith in statistical outputs, in those who do perform statistical tests, and in scientific studies. Science works best when it gets the most scrutiny and independent evaluation. You need about a college-level vocabulary to read "primary scientific literature" (the original published studies that scientists, including me, write for scientific journals). Some of it is very technical, and you may need experts both in statistics and in the particular field to decipher it all. But if so, you won't be alone—most scientists consult others to help them, either with the statistics and/or with understanding the species involved. That's why we correspond and go to scientific meetings. Especially with a few tips as follow below, it's possible to read this stuff with some independent understanding.

This word "significant" used so frequently and broadly is actually a narrowly technical term statistically speaking. When I write a scientific paper I try to use this word "significant" only in this narrow definition of statistically non-random. Feed numbers into a statistical equation (or "test") and it does calculations. The output is a probability called a "P value", which is used to determine whether what you fed into it looks like a random or non-random distribution of numbers. A P value falls between 0 and 1.0. You get to decide where to put the dividing line in that spectrum by designating the "alpha" value (or "critical" P value): the dividing line between the significant P values and all the non-significant values. The standard is an alpha value of 0.05. This means that P values less than that are significant (unlikely to occur by chance); the rest are not.

False positives. If we use the standard 0.05 as the alpha value, we are using a 95% confidence in that significance. In other words, 95% of instances we call significant with that alpha value are actually significant (non-random) and 5% are expected not to be. If your alpha is 0.01, then 99% are truly significant and 1% not. Those ones that look significant (by the P value) but are actually not are "spurious", a Type I statistical error (analogous to a "false positive" in a medical test). Why not just throw out the Type 1 statistical errors and keep all the validly significant ones? That's the problem. We can't tell which is which. All the statistical test can do is output P values. It can't tell you which ones are "right." One cure for the Type I statistical error is to do lots of tests. For example, if you have a hypothesis that butterfly species are starting to emerge earlier in the year due to climate warming, test each species separately. Then see how many different butterfly species have significantly earlier dates in relation either to more recent years (which have tended to be warmer) or directly to climatically warmer years. If only 5% of your species have a significant pattern, then you haven't found a pattern beyond the background (spurious) 5%. While it's possible that those particular 5% of species are having something biologically meaningful going on, you have not demonstrated this "significantly", that is, statistically, in your study. This is called "going fishing": testing and testing until finally one significant result pops up. Then that one result is taken at face value (even though it is one significant result awash in many non-significant ones) or reported by itself without the larger context to show how hard it was to get that result (fishing and fishing for it), making it therefore likely only a Type I statistical error and not representative of a broader pattern. But if far more than 5% of the species show the same pattern with $P<0.05$, this is very likely a real pattern, even if there's a small uncertainty about one or a few being false positives. Also, some patterns are statistically significant in several different geographical areas, or several different studies, while others aren't. The latter are more likely than the former to be spurious, although you can't know this for any particular test.

False negatives. The Type II statistical error is analogous to a "false negative", where a biologically meaningful (non-random) pattern was found non-significant by the statistical test. Why not just count those particular ones as significant anyway? Same answer—can't tell which is which. One method of counteracting this is to raise the alpha value (e.g., to 0.10) because the dataset is not producing the desired significant result at 0.05. I don't advocate that, because this is usually symptomatic of an inadequate (too small) dataset (discussed further below). It also makes direct comparison of studies more difficult, because they aren't all playing by the same standards. I also would not wish to try to defend this choice in peer review of any of my papers! I don't want to do this anyway, because I want my results to stand the tests of time and comparison to others' work. However, it is possible to have a principled reason for raising alpha. If you are concerned about the earliest possible detection of any negative outcome, you would raise your alpha in order to detect possible adverse reactions sooner, even at the risk of more false positives.

One method used to weed out spuriously significant results is to lower the alpha value by a factor commensurate with how many different variables or tests were performed. The thought is that tests meeting that lower tougher standard of significance are more likely to be significant. (This is true regardless: borderline results at 0.049 are not as strong as ones with $P<0.0001$.) But I advise reading carefully. When one practitioner uses a stringent method and says only these few instances are significant, those results should not be directly compared to everyone else's laxer standard. In some instances where this lowering

occurred, I felt that the dataset had inadequate power and I could not tell whether there was a big enough sample for it to be possible for anything to be significant. In other cases, I felt that lowering alpha changed the standard outside the norm, that we were tipping the balance way too far in the direction of making it too hard to obtain significance, and as a result, missing out on biologically meaningful patterns.

One-tailed P or two-tailed? Are you looking for a statistically significant difference, either way? That is, you are looking at whether values in group A are significantly higher or lower than values in group B. A two-tailed P looks for a significant difference, either way. In most cases you don't have a strong case that it could only go in one direction, so the two-tailed P is appropriate. A one-tailed P would be appropriate when there is already a strong case for narrowing the search to one direction only. A statistical test may provide both kinds of P values, since it can't know which is more appropriate. A two-tailed P is twice the one-tailed P. In other words, if the test kicks out a one-tailed P of 0.04, then two-tailed P is 0.08. Thus, the standard is twice as stringent to get significance with a two-tailed P than with a one-tailed P. I prefer the two-tailed P across the board, as a way to be more objective in my thinking, instead of having to use my judgment for when the case is strong enough to use one-tailed instead. However, a principled reason to use the one-tailed P is to detect possible detrimental effects as soon as possible. Also it takes more time and resources to get the sample size needed to obtain significance as a two-tailed P.

The null hypothesis. This states the hypothesis you are testing. It is usually phrased as if the analysis will produce a non-significant result. For example, if you are testing for a difference in number of butterfly species found in small and large sites, the null hypothesis states that there is no difference in number of butterfly species between small and large sites. If your analysis indicates significantly more species in larger sites than smaller ones, the null hypothesis is rejected. If your analysis indicates significantly more species in smaller than larger sites, the null hypothesis is also rejected, and you will see the reason for the two-tailed P: unexpected outcomes can occur, and scientists need to be open to that possibility! If the analysis does not produce a significant result, then the null hypothesis has not been rejected. This may seem like unnecessarily stilted language. But it actually serves a useful purpose of encouraging clear and objective thinking about the study. Rejecting a null hypothesis is not the same as proving something—the null hypothesis may fail not because it's wrong but because the sample fed into the test was inadequate. The concept is that random is the default; science is about establishing non-random patterns effectively. An alternate way of phrasing the null hypothesis is to base it on existing scientific findings to formulate an expectation at the outset of the analysis. For example, the null hypothesis could be that smaller sites are expected to have fewer species than larger ones. There is the danger of temptation to find the result you're looking for. But that can also be done within the traditional construction of the null hypothesis. And this alternate construction may actually be more honest. At the outset, most scientists have an expectation of what they'll find. After all, they've been reading on their subject and know what others have thought and found. The more interesting part of science for me is when the unexpected occurs, and the process of how this came about can be more accurately reported this way.

Linear and non-linear patterns. When statisticians talk about a linear pattern or relationship, they mean a continuously consistent pattern. For example, if more butterflies get found the warmer it is, that's a linear relationship. If you graph your results, the line connecting the dots doesn't have to be straight, but it does have to go generally in only one direction (except for little wobbles or static in the data), either up or down. Other patterns can also be graphed, and the dots can be connected by a line. But that line may first go up, then down. Line or not, that is a non-linear (non-continuous) pattern. An example of a non-linear pattern is butterfly annual fluctuation. Another is a threshold pattern. For example, an animal species may show inactivity within x amount of time of sunset and beyond, regardless of temperature. But before then, activity may relate to temperature. Yet another example is the categorical test: butterfly abundance broken into categories of vegetative type (e.g. wet, mesic, and dry prairie). Those three categories do not have an inherent numerical relationship to each other. Some variables can be treated both categorically (non-linearly) and linearly. For example, site size can be grouped into categories (small, medium large) or linearly (by each site's acreage). The categorical approach allows for the possibility that medium sites might do better (or worse) than both small and large sites. On the other hand, this approach reduces the size variable to three possibilities, and the variation in size within each category might confound finding differences among categories. The linear approach of using each acreage value might allow an area effect (increasing abundance with increasing site size) to show itself more easily. Why does this matter? Some statistical tests are linear and others not. They each look for a *particular kind* of pattern, not whether there is *any kind* of pattern. A dataset may not produce any significant results in a linear test but may yet have patterns in there, if we can figure out non-linear pattern to test for.

What "significantly different" does and does not mean. A large category of statistical tests looks into whether two or more groups of data significantly (non-randomly) differ from each other. If so, the test doesn't directly say why the two groups are different. It relies on you to determine this. For example, all small sites might also be grazed more intensively. If these sites also have fewer butterflies than the other study sites, is it the smallness or the grazing that's responsible, or both? Likewise, not being significantly different does not mean being the same. Besides the problem of false negatives (Type II errors), which are usually due to inadequate sample size (too small a dataset),

you may also not have identified the factor that is the biologically meaningful difference.

What a "significant correlation" does and does not mean. Correlations test for a significant (non-random) pattern of positive or negative correlation. For example, if a butterfly increases in abundance with hotter summers, that's a positive correlation. If the same butterfly decreases with wetter summers, that's a negative correlation. That is true as long as the correlation test kicks out a significant P value. However, if the correlation is not significant, that's not the same as being uncorrelated. Besides the ever-present concern about false negatives (Type II error) usually due to inadequate sample size, some other confounding factor may influence the result. Suppose in your survey years, all warm years were wet and all cool years dry. Then perhaps these opposing forces were canceling each other out, or in some years one was a more important influence than the other.

Significance bias. Unfortunately, science has a bias toward significance. However, with exhaustive searching, if you still can't find significance, that is certainly noteworthy too. On the other hand, sometimes we may want the result to be not significant (just as with a medical test, we usually have a desired outcome). It's easy to get a non-significant result even for biologically meaningful differences. All you have to do is not get a big sample, or let other uncontrolled variables confound the test. That's why studies need to be read carefully not just for what they're saying but also for what they're not mentioning.

Sample size. Inadequate sample size is a common cause of Type II errors (false negatives). So sample size is the Swengel obsession—how to get enough sample size (a large enough dataset). How much is enough? Well, most statistical tests need at least 5 or 6 examples in each "variate" (group). That is, if you want to look at whether large sites are different in butterfly species richness from small sites, you need a survey from each of 5-6 large sites and each of 5-6 small sites. But in practicality, it takes way more sample than that to get anywhere in deciphering something as complicated as butterflies. They are highly "variable" in statistic-speak. You already know this. Dramatic differences can occur in how many you see due to weather. Dramatic difference in abundance can occur among dates (early, middle, or late in the flight period), years (due to fluctuations in climatic patterns), and due to various habitat characteristics, making some sites really good and others not. Need I go on? But if enough sites are in a confined region and surveyed the same years on similar dates within each year under acceptable weather, then climatic factors are largely the same for all sites, and weather and flight period are being addressed. This allows the variables that interest me (habitat and management characteristics) a chance to start telling their stories. Compared to how most people spend their time, yes, Scott and I are really into butterfly surveying. But in comparison to what it takes to get enough statistical power to learn the things we want to learn—we're often struggling to get there. Thank you again, if you see us in the field, for leaving us be and letting us do what it takes to get our surveys done. There really is never enough time in the right weather and timing to do all the surveys that need to be done.

Anecdotes vs. scientific samples. An anecdote is a single example. A scientific sample is a sufficiently large group of anecdotes that the sample is suitable to feed into statistical tests with enough "power" to put reliance in the outputs. That is, if there's significance, there's enough of them to pass the Type I test, and if there's not significance, you have confidence that it isn't entirely due to inadequate sample size and inadequate power (Type II errors). One of the things I really like about statistical testing is it makes the process more anonymous and blind. No matter how vivid my memory of a specific site or observation may be, once I've gotten everything databased, that just becomes one anecdote among many, all equally weighted in the database.

ISSUES FOR INTERPRETING SCIENCE

Overwhelming other factors affect butterfly data besides management. It's not enough to detect these annual fluctuations, phenological (seasonal timing) patterns, weather effects, and site (vegetative) effects and so on to have enough samples to parse management effects. These other factors also need to be controlled. Annual fluctuations, timing within the growing season, weather on survey day, basic vegetation structure (forest, savanna, grassland, wetland), basic vegetative type (wet, mesic, dry), originalness of vegetation (never tilled vs. formerly plowed), degree of vegetative degradation (brushiness, exotic plants), and if possible, caterpillar food plant characteristics: these all need to be controlled as much as possible. Once you've done that, you're ready to look at management. Let me give a medical analogy. Suppose you want to test whether infant immunization is beneficial. If raw sewage is tainting the water supply that some of these infants are exposed to, then that's going to confound your study, just as annual fluctuations and so on overwhelm management effects. If the raw sewage shows as the overwhelming factor affecting the infants' health outcome, does that mean immunization doesn't matter? No. It means a confounding factor overwhelms the study's ability to answer that question.

An additional complicating factor for variable control is the ever-changing context of biodiversity (lag effects). With us humans, medical research requires long time periods to study long-term medical effects. But the context we study subjects are living in is not constant and controlled. Just imagine trying to tease out medical findings that require decades to develop while having to control variables that keep changing: our food habits, household products, and so on. The same is true for biodiversity: there's a lag time in effects of things that occurred in the landscape in the past, but new things are changing in the landscape now and affecting biodiversity too. So what looks OK in a study now may actually not be so; the lag effect just hasn't come due yet. This is especially demonstrated with patch size and

fragmentation of reserves. This results in something called "extinction debt": the landscape used to be suitable for the long-term persistence of a species' population, but the landscape has now been degraded or fragmented to the extent that it no longer provides an adequate amount of habitat for the species. But the species may not have gone extinct in the area yet, since there is a lag until just the perfect storm of unfavorable factors (often including particularly unfavorable weather that year—weather that has precedent in the area and has occurred before, but not frequently) converge to wipe out the species. Thus, current research can be overly optimistic about what vulnerable populations can survive. They can survive in most years but not all years, but if they can't survive in all years, they no longer will exist in any. On the other hand, sometimes a factor has an immediate large positive or negative effect on butterflies; this finding should be incorporated into responses and actions before long-term research has been obtained. Both sides of this occur in medical research: when strong beneficial results occur for a gravely ill group, or a strong negative result becomes apparent for any group, results are reported and responded to prior to completion of the trial.

How much variation got "explained"? Besides a P value, many statistical tests kick out a measure of how much variability got explained. In a sense, what got explained is that part of the sound coming out of a radio that is the music you want to hear; all that other unexplained variability is static. Usually, only a minority of the variability gets explained. Some tests are set up to determine that if too little variability gets explained, then the overall test is not significant (meaningful)—in other words, you can't put stock in any of the test's findings. The larger the sample, the lower the amount of variability that can get explained and still kick out a significant result. There are a lot of patterns out there that matter, but those are only a small part of what's affecting butterfly abundance and distribution. It's well to keep an eye not just on the significance, but also on how much got explained. Listen to the patterns found, but also leave a lot of room for realizing that the static can have some really important music imbedded in it too, if we can just figure out how to tune our surveys and think up the factors to get the analyses needed to distinguish more of the music.

Unfair exclusions. Consciously or not, scientists can rig their study to get a desired statistical outcome when that is not really representative of the dataset or the biological situation. I watch for whether data are included or excluded from a test by objective descriptors or case by case, which needs to be evaluated for whether that's resulting in rigging. For example, an objective criterion is to exclude all surveys before a certain date and after a certain date (outside the flight period in an area). Individual surveys should not be excluded because the results don't look "right" (expected). If you can identify why that survey came out that way, and then exclude ALL surveys with those unfavorable characteristic(s), then that is objective. I particularly value studies that go out of their way to allow an unexpected or unpopular outcome every chance of being portrayed statistically. If it does, or if it doesn't, either way the outcome is well substantiated. Remember, statistical tests only look at the numbers; they do not determine whether those were the numbers you should have fed into those tests.

Weighting studies by number or by quality. One way to compare studies to each other is to give each one a vote. That looks objective and fair, but in actuality, it is a bias toward small studies. If a scientist spends each year doing a separate study, then that's ten studies in ten years. If, on the other hand, another scientist spends ten years doing one study, that's only one vote, but all other things being equal, that one study has a lot more quality because it is larger and longer term. The fluke effects of one year being hot and dry and another cold and wet can be evened out by having a larger sample over more years. As a result, more understanding comes when available studies are "weighted" by how large they are, how well confounding variables were investigated and controlled, and so on.

Significance vs. predictive power. Another way to test how to understand statistical results is to see how much predictive power they provide. This can weed out those Type I and II statistical errors (false positives and false negatives) and correlated variables (e.g., hotter years tending to be drier), and so on. An action can be based on the findings to see if the expected outcome occurs. Or predictions can be made about what should happen next, to see if they occur. When there is a disparity between what actually happens and what was expected based on previous studies, it's time to re-examine the studies to figure out what confounding variables or alternate explanations exist in the studies.

The discipline of scientific thinking and presentation is challenging, so it is not surprising that scientists fail to be perfect at it. There can be errors of logic and studies can be mis-cited after it is published. It's up to my colleagues, including you, to figure out how well I've followed the scientific method for interpretations here and to help me find and correct my lapses in consistency of standards and objectivity.

To be an expert on a topic, a scientist has to read not just the secondary literature (reviews and syntheses of scientific studies) but also the primary literature (where the original data and tests get presented in the context of the methods), to know what exactly got compared to what over what time frame and with how big a sample size. However, the amount of literature is so vast that it is impossible to read everything. One way I sort out what to read in more detail is to read the Abstract (summary of the study) and perhaps skim the tables, figures, and Discussion (summary of the results in context of others' studies). I call this passive reading. If it's particularly relevant to my field, then I read in more detail, with attention to exactly how the sampling was done where and for how long, what kinds of statistics were done, and what kinds of comparisons were made. That's what I call active (or critical) reading. But it's easy to take the shortcut of just reading the Abstract and

Discussion at face value. There's a trust that the reviewers and editors have vetted this text to maintain standards and accuracy and not overreach. However, following are some of the things I watch for, as an author, reader, and reviewer.

Interpretational creep. Within a particular study, the test may not produce a significant difference between one group averaging somewhat higher than the other. This is correctly stated in the results. But in Discussion, the result is nonetheless treated as an actual finding of difference, and so stated in the Abstract (summary). Or what is floated only as an unsubstantiated hypothesis (one possible explanation that does not exclude other plausible ones) becomes cited as an established finding by others.

Results creep. Unpublished data or reports may get cited without these actually being available for other scientists to examine at the time of publication. A little of this is OK, to alert scientists to watch for that finding to become available in print soon. Unfortunately, sometimes they never do. Scientific papers should primarily be about what is actually in evidence available to all to evaluate.

"Proof" by elimination. Huge fan of Sherlock Holmes that I am, I quibble with a famous concept voiced by this character that if you eliminate all other possibilities, whatever is left (however improbable) must be the truth. One side of this I agree with. It is relatively easier to eliminate possibilities that to validate one. But you can't prove something by elimination only. You still have to have positive evidence that is both necessary to support it and sufficient to eliminate alternate explanations. The fact that something is missing in the landscape today that occurred back then, and back then the animal did better than now, does not prove that this missing something is the reason for the animal's decline now. Have you investigated all the other things that have changed between then and now? I have definitely read studies that use that reasoning to establish a finding, when they have not in fact provided positive evidence that the hypothesis actually works.

What gets said about studies may not pan out when I actually read the study, after digging back from the secondary literature to the primary literature cited as sources. The study is fine so far as it goes. It may show some interesting points not noticed by the authors. But what others say about it may be overly conclusive, given the breadth or length of the study, or inaccurate (for example, it doesn't study rare species but is applied as if it does; or significance is claimed when it's not there). This is another way that science is claimed to indicate something that the data themselves do not support. Most of the time I do not think a deliberate deception is being attempted. Instead, I think that it is easy to see or find what we want to find.

LIMITS OF SCIENCE

Why have I spent so much time on understanding statistics? I want to explain what I understand about management in terms of how these things were originally reported scientifically. This means the concepts arise out of probabilities and likelihoods. I want you to understand the limits and gaps—why science can't answer simple questions with clarity and finality, such as "What is the best way to manage my prairie?" Science looks at particular datasets and what happened in them. It cannot generalize except to the extent that large datasets of great depth and breadth exist. However, we can learn a lot from these past datasets to help us formulate approaches likely to be worth trying for future benefit.

Let me detour to medical research, something which affects and no doubt interests us all. I find it helps to think about an entirely different field to get a suitable context for understanding the field I really really care about. Many factors affect outcomes in human medical research. Each person presents a different set of predispositions and risk factors. There may also be some chance involved—some folks are lucky (and don't die of their cigarette smoking) and others are unlucky (and succumb to minor amounts of second hand smoke), although statistically it's very hard to distinguish between luck and some subtle underlying predisposition. But even very strong medical patterns still have their flukes—the very rare person that survives rabies for example. Fewer things than you might expect are truly 100% safe or lethal. As a result, each person is a combination of positive and negative risk factors, with no guarantee that their individual predispositions will actually happen to them and that the outcomes unlikely for them won't.

The default scientific position that a study has to show a significant difference or we don't think there's a pattern is one of the reasons, I believe, why there are so many reversals and recalls after a medical recommendation or treatment is unleashed on the populace at large. On the one hand, the recommendation or treatment had to show a significant benefit before it reached the general population. But it's only once that broad-scale unleashing has happened that there's enough sample size to get the statistical power needed to detect subtle but meaningful side effects, or verify the lack thereof. It simply costs too much to do that large and long a study beforehand, especially since these negative side effects may affect fewer people than experience the benefit or non-harm. The benefit may be the stronger (more easily detected) significant effect and may apply to more people, but the harm may be more extreme and targeted to more sensitive or vulnerable people, who may not be the primary subjects of the formal research.

Scientific paradigms are good and bad. Paradigms are concepts that explain how some aspect of the world works. They are necessary or data are a jumble with no organization. But paradigms can also be very limiting. They block the consideration of alternatives. One of the big challenges in science is thinking of the questions to ask. We may not think of it, but we may also not want to think of it either. If you were in England in the 1970s and 1980s, would you have wanted to consider the possibility that conservation activity was the proximate cause of the decline and extirpation of the Large Blue? Bear in mind a lot was al-

ready known about the Large Blue: its caterpillar food plant, its symbiosis with ants, when and where to go to find the butterfly. This is not, in and of itself, conservation. This is basic natural history. Conservation is about knowing what are the most important factors limiting and fostering that population there; in other words, what most needs to be prevented from happening, and what most needs to continue happening, to keep the population in existence there. There were lots of other plausible culprits, including our favorite bogeymen, habitat degradation and fragmentation. Did the now prevailing view get accepted by bullying, shouting, scaring everyone else into shutting up about their ideas? No. A breathtaking volume of field work and experiments occurred, to develop hypotheses to test, and when positive evidence of successful population outcomes occurred, that's as close as science comes to proof. This research included traveling to healthy Large Blue populations in other countries, much as Scott and I have traveled the Midwest to study large prairie specialist populations.

As it turned out, habitat management was the key to both the extirpation and re-establishment of the Large Blue. Preserving Large Blue sites in England had involved removing farm grazing from the site. The caterpillar food plants grew lusher, but that vegetative structure was unsuitable for the one ant species that tended the blue immatures. Other ants thrived in the preserves but attacked the blue caterpillars! When some English preserves were restored to suitable vegetative conditions for the critical species of ant, while also maintaining an appropriate density of the caterpillar food plant (enough to generate viable numbers of the butterfly but not so many as to overwhelm the ants), re-introduction of the Large Blue from populations on mainland Europe resulted in successful population re-establishments. And the overall landscape is no doubt more degraded and fragmented now than when the Large Blue disappeared several decades ago.

The hardest part of science is thinking up the possible hypotheses to study in the first place. While a preconceived hypothesis may be necessary in order to set up a valid study design, these can also discourage thinking "outside the box." This is where I think field surveying has an advantage over lab science. You can set up a general survey program, using an established method that is sufficiently efficient and interesting to obtain lots of analyzable data. But you don't have to have all your null hypotheses established ahead of time. You can instead try to maintain an open mind, tabulate and graph the results, then see what possibilities present themselves. To be honest, a lot of our results were things I didn't even conceive of when we first started surveying.

"Fool's experiment." Here's an approximation of a quote I ran across some time ago, attributed to Charles Darwin. "I love a fool's experiment for it is amazing how often obvious truths are proved wrong." Obvious truths are "paradigms" (frameworks for data). Sometimes when you dig back to the primary literature source, it's just an observation or belief, not a scientifically demonstrated finding. It's taken to be an obvious truth. But in science, nothing should be taken to be obvious, outside the rule that it must be demonstrated with evidence.

Science is about finding the most parsimonious explanation that most compatibly explains all the available information. Oversimplification is bad, since it brushes aside loose ends. But excessive complexity is also bad. If you have to contort yourself to make the available data fit your theory, then it's time to look for something more elegant. For example, to maintain the earth as the center of the solar system, epicycles (complex orbits by the sun and other planets) had to be invoked, while Galileo's heliocentric theory explained the observed data with both greater effectiveness and simplicity. Rejecting an explanation because you don't like it is also unscientific. Saying it can't be explained when it actually can, at least in part, is stalling scientific advance. Instead, it's time to move on to the next questions. It's rare that nothing is known. Instead, usually something is known and it's possible to move forward.

Actual errors can occur in all stages of scientific endeavor. At data collection, there can be misidentification, mismeasurement, and errors in recording and writing the data. Mistakes occur at data transcription and databasing, in analysis, in writing it up (both typos and actual errors), in page proofs and printing. I often notice minor errors, typos, and discrepancies in others' publications. None of us are immune to this. I sure wish there were no errors in my work but that is impossible for any human to achieve. Most of the time, I think these errors are of minor consequence, whether they are detectable or not. It's a rare case when an outright error of this sort causes some sort of significant impact.

It's a rare situation where all studies agree on the outcome or finding. If such a "consensus" exists, my first expectation is that not much research has been done on the topic! My second expectation is that not all practitioners are being asked! Simply the way variability in data works, I expect not all of my observations to agree 100% with any pattern or trend. Likewise, as an extension of that, not all studies among different researchers are going to agree with the way most studies go. Sometimes this can be explained by the individual differences in sites or species or methods used. But sometimes it's just the blips that turn up in data. In scientific interpretation, weight is given to what is most replicated. As a result, most scientific understanding is based on a preponderance of available evidence (not all of it) going a certain way. That means there's always room for that preponderance to nudge in different directions once more data become available. Plus each of us has read and seen a different set of evidences, so that we form somewhat different understandings.

In my experience, it's rare indeed that a study is actually invalid; it's what we say about the study (our interpretations) that involves invalidation. Perhaps this happens more in other fields; I can't say. But in my field, usually there isn't something, such as pervasive unreliable

ID, that renders a study invalid and useless. So when there's disagreement or even refutation reported, it's the interpretations (or contexts) of those studies, not the actual data/observations, that are being argued about. It's what we *say* about these results (the conclusions and applications) that has so much conflict.

Science only looks at what scientists are able and willing to collect data on. That is, in management, the option has to exist consistently for me to be able to study it, and I have to be willing to survey the site. As I like to say, I can't tell you the best management. I can only tell you what I've been able to study and read about. Not all management options viable for conservation exist out there for me to study, or to study in enough independent trials to support statistical testing. Examples: summer haying in Wisconsin prairies, goat grazing in bluff prairies. Most important, statistics are interpreted primarily in a relative way (comparing group A to group B), so that context is critical to understanding the finding.

Some groups of species are hard to study unless we deliberately try to, but others have sufficient detectability and popularity that they are more widely known, which can independently test formal science. Knowledge about the obscure and hard to find species mostly arises only when a scientific specialist chooses to study them. But some groups have high visibility, such as butterflies, diurnal birds, and vascular plants. A consequence of this is that non-scientists can contribute to testing what formal science is producing on these species. We amateurs and hobbyists find them whether we're trying or not, when for example we're in urban parks at lunch or family reunions. Witness Jeffrey Glassberg's comment in the acknowledgments in *Butterflies Through Binoculars: The West* about the serendipity of being pulled over for speeding! Butterfliers and birders are species-oriented, not site-oriented or theory-oriented. We just want to find them and understand them. But as we do so, we can also contribute information about species that may not be picked up by formal science, such as continued occurrence in urban parks after no longer being found in preserves, as described in *American Butterflies* (see Part 1).

I can't emphasize enough how hard it is to get scientifically meaningful samples on specialist and rare butterflies. These species tend to have very particular places and times when they can be effectively found. It takes a lot of effort to get enough understanding of a species like Frosted Elfin or Ottoe Skipper to know when to find it with the great fluctuations in seasonal timing and abundance that we experience year to year in the Midwest, and where to find it—what exactly is the habitat it is specialized to, with enough precision to capture the full range of its occurrence but exclude where it doesn't occur. Once it's obvious there's a real problem with the species, it's very hard to get a sample size of enough individuals per populations, and enough separate populations, to sort out what factors are affecting the butterfly population and how.

This means that science, especially statistical science, has the most difficulty telling us about the species that most need conservation help (the rarest and most localized species). I've done a lot of scientific surveys, analyses, and papers. It's wonderful to get really big samples, such as for 'Karner' Melissa Blue and Regal Fritillary. But I'll admit I feel the greatest sense of accomplishment for actually obtaining any statistical power at all on management for the hardest species to get a sample on, such as Mottled Duskywing and Frosted Elfin.

Science can't provide the answer you most want: what will happen in your particular case in the future. That's because science is based on observable (therefore past) phenomena. Science may attempt to predict the future and identify more or less likely scenarios. But science cannot tell you with a certainty what will happen in a particular future anecdote. Remember, scientists often speak dismissively of anecdotes. They—I mean we—want lots of independent examples, a scientific sample, not a single anecdote. However, whether in a medical context or a habitat management context, you are an anecdote. Your one life, or your one site, is one anecdote, but it is also very important. That's why what doesn't look very risky in a large scientific sample can start looking awfully risky when it's all or nothing in your individual situation. Science can give projections and probabilities, and these may in fact be quite well substantiated and very strong. But even for high likelihoods, these are what usually happen, not a guarantee of a specific outcome in a particular situation. As a medical analogy, in a situation where treatment has a low chance of working, a particular patient may be among the lucky few it works for. Conversely, in a situation when most people respond positively, a particular patient may be among the unlucky few who react adversely.

Published by the Southern Wisconsin Butterfly Association (SWBA), a chapter of the North American Butterfly Association. To find out about field trips and meetings, please visit http://www.naba.org/chapters/nabawba/