Critical review of Knapp et al. (2008), MN DNR Publication 166

By Daniel McEwen and Brian Wisenden

General Comments

Per request of David Majkrzak, as a representative of the Pelican Lake Property Owners Association (PLPOA), we have reviewed the Minnesota Department of Natural Resources Special Publication 166, March 2008 entitled "Fish community response to the introduction of Muskellunge in Minnesota Lakes" by Michael L. Knapp, Steven W. Mero, David J. Bohlander, and David F. Staples. This study, and the more condensed version published in the peer-reviewed journal North American Journal of Fisheries Management (NAJFM) in 2012, rely on the same data and the same analyses. Both versions of this study reach the conclusion that there is no consistent evidence that stocking muskellunge (*Esox masquinongy*) has a negative impact on other fish species in the lake. Because this study is often used to guide decision-making for management of muskellunge stocking programs, it is important to evaluate limitations in the analysis and interpretation of these data. In our review we assessed two central tenets of the scientific method: (1) that the statistical methods test the question, and (2) that the conclusions follow logically from the results. While the study certainly took much effort and contains an impressive amount of data and analysis, the basic conclusion of Knapp et al. (2008) that "*The lack of consistent negative changes in mean CPUE after stocking suggests these fish species have generally coexisted well with muskellunge in the lakes at the densities that have resulted from stocking*" cannot follow from the analysis.

We find two overarching problems with the study: (1) the methodology in data collection and analysis violate assumptions inherent to the statistical tests used, and (2) even if the methodology was appropriate, conclusions are based on negative results (i.e., they were *not* able to detect an effect of stocking muskellunge on other species of fish). In our review of Knapp et al. (2008) we found ten significant flaws in data analysis and interpretation related to these two problems. Consequently, in our professional opinion we find that this study is inconclusive as to the question Knapp et al. (2008) set out to answer (i.e., Does muskellunge stocking negatively impact other members of the fish community they coexist with?).

We summarize these flaws, in no particular order of importance, in the list below, followed by detailed explanations of each.

1. **Improper interpretation of a nonsignificant result.**
2. **Lack of correction for experimentwise error rates.**
3. **Lack of reliable estimates of abundances for populations.**
4. **Lack of any concurrent data or analysis of muskellunge population trends.**
5. **Lack of any concurrent data on stocking of other game fish with or before muskellunge.**
6. **Lack of data on size/age distribution of fishes.**
7. **Lack of accounting for environmental variables.**
8. **Lack of independent experimental units.**
9. **Lack of a rigorous experimental control.**
10. **Insufficient information to evaluate merits of the study.**

Specific Comments

1. **Improper interpretation of a nonsignificant result.** The conclusion made by Knapp et al. (2008) follows from their inability to demonstrate statistically significant differences in fish catch-per-unit-effort (CPUE) in comparisons of pre- and post-stocking of muskellunge for most fish species in most lakes in most communities in their sample. The methods (i.e., Wilcoxon Tests) used to make the comparisons in this paper are part of a group of commonly used statistics that are generally known as null hypothesis tests. Null hypothesis tests test the probability of collecting a sample of data if there is in fact no difference (i.e., no difference in CPUE in other fish species after muskellunge stocking compared to before stocking). These statistical tests provide a "p-value", which is the probability of collecting the sample data set if the null hypothesis is true. If the probability that the sample data came from statistical populations such that there were no difference between the pre- and post-median CPUE (i.e., the null hypothesis) is less than 5%, then we would conclude that it would be rare to collect the sample data for the statistic. In this case, we would reject the "null hypothesis" (that there is no difference between the pre- and post-stocking median CPUE for other fish species) and accept the alternate hypothesis (that there is a difference between the pre- and post-stocking median CPUE for other fish species). The philosophical underpinning of this method is that a test can either reject a null hypothesis, or fail to reject the null hypothesis, but it cannot ever "accept" the null hypothesis, which is precisely what Knapp et al. (2008) do.

   Mathematically, for any null hypothesis test, the p-value of a test statistic is a function of the sample size (e.g., how many data points exist in the pre and post surveys), the effect size (e.g., the difference between the median value of CPUE pre vs. post) and the variance (e.g., how consistent the CPUEs are within the pre or post sampling groups). A nonsignificant p-value (i.e., $p > 0.05$) does not mean that there is no effect of muskellunge on the fish community; it simply means that the sample size is not large enough, the variance is not small enough, and/or the effect size is too small to allow the test to detect the effect. A statistical test cannot decipher which of these is/are true but if authors provide information on all three, readers can evaluate which might be particularly important in leading to a nonsignificant result. Knapp et al. (2008) provide effect sizes and sample sizes, but do not provide estimates for variances, so there is no way to judge whether variances might be too large to allow an effect to be detected.

   While effect sizes and sample sizes were provided, the sample size for many lakes, either pre- or post-stocking is very small (provided samples were not temporally autocorrelated and thus pseudoreplicated, which is almost certainly not true, see comment #8 below). For example, for the test of an effect of muskellunge stocking on walleye (Table 3), only 4 of the 40 lakes had a sample size larger than 5 for pre-stocking walleye density. For 5 of the 40 lakes in Table 3 Knapp et al. (2008) reached conclusions on the impact of muskellunge stocking when there were no pre-stocking data at all. Drawing conclusions in the absence of data was not unique to Table 3 (walleye). Every fish species considered in Knapp et al. (2008) includes some lakes that were granted "within-quartile" status (no impact of muskellunge stocking) when the complete set of data needed to make that assessment was missing.

Knapp et al. (2008) conclude that their inability to reject the null hypothesis means that muskellunge stocking has no effect on other fish species. This is a very common error in interpretation. There has been much information published about misunderstandings of the very common error of concluding that a failure to reject a null hypothesis provides evidence that the null hypothesis is true. For a cogent overview of this point see "Interpreting Failure to Reject a Null Hypothesis" by D.F. Parkhurst 1985 published in the Bulletin of the Ecological Society of America 66:301-302.

2. **Lack of correction for experimentwise error rates.** Another property of null hypothesis tests is that setting the cut-off at 5% for determining when an event is so rare that it likely did not occur by random chance, still leaves a 5% chance that a comparison of data sets for which there is truly no difference will reject the null hypothesis by mistake. These are, in effect, rates of false positives. Mathematically it can be shown that the probably of getting a false positive in a comprehensive study reporting multiple p-values is not the single p-value reported for a test but the product of the single p-value multiplied by the total number of p-values generated. Statisticians use the concept of an experimentwise error rate to take into account the risk of inflating false positives (e.g., Klockars and Hancock 1994, Shaffer 1995).There are two general ways to address the problem of experimentwise error. First, more sophisticated multilevel or multivariate statistical models can be used to reduce the number of p-values generated. Second, there many methods that can be used to control for this experimentwise inflation (e.g., Holm 1979, Hommel 1988, Wright 1992, Benjamini and Yekutieli 2001), none of which were used by Knapp et al. (2008). It is important to recognize that if the proper corrections had been performed, there would have been even fewer significant differences detected, adding apparent credence to their general conclusion that there is no effect of muskellunge on other fish species even though this reduction in significant p-values would have come from statistical artefacts rather than an actual ecological process.

3. **Lack of reliable estimates of abundances for populations.** Knapp et al. (2008) use catch-per-unit-effort (CPUE) from gill net surveys to estimate population size. CPUE may not correlate well with abundance or be biased in favor of certain size classes that compromise accuracy of true population estimates (e.g., Harley et al. 2001, Grant et al. 2004, Maunder et al. 2006). For CPUE to be proportional to abundance the catchability for all individuals in a population of a given species must be constant.

   This is almost certainly not true for a number of reasons. The relative proportion of different age classes changes from year to year. The efficiencies of gear and localities where fish might reside changes. Environmental conditions can influence catchability. For example, patterns in precipitation and temperature can alter the environment and therefore where fish live either because of their own environmental preferences and tolerances or because their prey relocate in response to shifts in environmental variables. Catchability is also likely a function of abundance itself, the very thing that CPUE is used to estimate. Other factors that can influence the catchability of a fish include net saturation, gear interference, and a disparity in the amount of time a net sits. Knapp et al. (2008) used the

best data available to them to compare pre- and post- muskellunge state but unfortunately, these data were all CPUE, but if CPUE does not actually measure the true population sizes of fish species in a lake, then any conclusions based CPUE are flawed regardless of the quality of the analysis.

4. **Lack of any concurrent data or analysis of muskellunge population trends.** There is no information about concurrent muskellunge populations or their size structure. Instead, there is an assumption that stocking muskellunge is successful and that the effect of stocking produces a similar density and size distribution of muskellunge in each lake. If muskellunge stocking failed, meaning that there was no establishment of a muskellunge population, then using data from that lake to test for the effect of muskellunge on other fish species is not valid. Knapp et al. (2008) concede this very point that stocking of muskellunge does not mean that they were able to establish a population (Hanson et al. 1986, cited in Knapp et al. 2008). Moreover, while the first year that muskellunge stocking began is provided, there is no information about how long muskellunge were stocked in a particular lake, or the frequency, intensity or method of stocking. If there was not parity in stocking regimes, and/or parity in the effectiveness of recruitment for muskellunge following stocking regimes, using these data to evaluate the effect of muskellunge becomes a weak test because in some lakes muskellunge density may have been quite high, while in other lakes the density may have been quite low or even absent. Further, the authors defined "prestocked" lakes (i.e., the control against which the post-stocked effect was measured) to include lakes known to be stocked with muskellunge fry and lakes known to contain native populations of muskellunge, assuming that muskellunge populations in those lakes were negligible but providing no evidence to support the claim.

5. **Lack of any concurrent data on stocking of other game fish with or before muskellunge.** No information is provided about stocking other fish species concurrent with muskellunge stocking. If, for example, walleye appear to increase after stocking muskellunge, but there was simultaneously a new (or increase in) walleye stocking at the same time, there would be no way to relate the population dynamics of walleye to muskellunge stocking. A robust statistical model may have taken this covariate into account.

6. **Lack of data on size/age distribution of fishes.** The only measure of the potential impact of muskellunge provided by Knapp et al. (2008) is an estimate of abundance of other species without regard to size structure. In general, there is a negative relationship in wild populations between abundance and individual size/age of animals (Carbone and Gitteman 2002, Brown and Gilloly 2003). In other words, comparing pre-stocking CPUE (assuming this is a good proxy for abundance, see comment #3 above) to post-stocking CPUE as a test of the effect of muskellunge stocking could be misleading. For example, a CPUE of 15 young-of-the-year walleye fingerlings before stocking vs. a CPUE of 3 mature 5^{+} walleye after stocking would be counted as a decline by the method used by Knapp et al. (2008) even though harvestable fish are of greater value to a fishery. A more appropriate metric in a study like this would be total biomass or some other analysis based on cohorts or size classes.

7. **Lack of accounting for environmental variables**. Undoubtedly environmental variables, both temporal and nontemporal, confound the effect of stocking muskellunge on other fish species. Muskellunge are not the single driver for population dynamics of other species in the lake, but each species, based on its life history, would be expected to relate differently to environmental variability. This environmental variability produces noise that masks potential impact of muskellunge and subsequently contributes to the high number of nonsignificant p-values generated in the analysis. Types of environmental variables that change with time include ice out, summer temperatures, and etc.). Knapp et al. (2008) do not provide actual sampling dates or make an effort to control for temporal factors. Nontemporal environmental characteristics are provided (Table 2), but there is no evidence they controlled for these in their statistical models. It is an unavoidability that after accounting for these environmental variables there would have been more power to detect an effect of muskellunge stocking on other fish populations, given that it would have greatly reduced unexplained lake-to-lake variability attributable to environmental factors unrelated to muskellunge**.**

8. **Lack of independent experimental units.** The authors analyze data at three separate levels from lowest to highest, which are individual lake, lakes pooled by lake class, and all lakes pooled together. For the results of a statistical analysis to be valid, an underlying assumption of all statistical tests is that each sample is independent (i.e., the information from one sample provides no information about another data point collected from the same sample). Multilevel statistical models (e.g., multiple linear regression) can be used to control for some factors that may lead to a lack of independence, but these authors did not do that. For example, samples from multiple lakes taken in a given year are not independent from one another because they are all subjected to the same weather events unique to that particular year. Other examples of a clear violation of the independence assumption includes samples collected within a single lake where CPUE in years closer to each other likely are more related than CPUE collected in years spaced further apart (i.e., temporal autocorrelation) or when data are collected within the same lake class assuming within lake samples are exchangeable with between lake samples. Treating these samples as if they were independent replicates is pseudoreplication (Hurlbert 1984). Samples collected in this way are subreplicates which must either be pooled or treated with a multilevel hierarchical model. The lack of temporal independence is especially important because there are many auxiliary variables that are temporally correlated such as lake shore development, angling pressure, ecological succession, improvement in sampling efforts, and general climatic effects. Because the higher level analyses (i.e., pooled by lake class and then all lakes pooled) are derived from this lower lake to lake analysis, the flaws described here apply to those levels as well and, for some analyses, become compounded (e.g., if all lakes are pooled into a single analyses, not only is there temporal autocorrelation but also spatial autocorrelation and lake class autocorrelation).

9. **Lack of rigorous experimental control.** The currently accepted methodology to assess environmental impacts is the Before-After Control-Impact (BACI) design made popular by Stewart-Oaten et al. (1986) in which samples are collected before and after an intervention at

an impacted site (e.g., where muskellunge are stocked) and simultaneously data for before and after are collected at a non-impacted or control site (e.g., where muskellunge are not stocked). The BACI design aims to solve the temporal autocorrelation problem (see #8) by using paired differences in time of abundances between the impact and control site before and after the impact. While data collection done in sequence for the control and impact sites individually are expected to be temporally autocorrelated, paired differences are less likely to be. In the Knapp et al. (2012) version of this paper, the authors comment on how difficult it is to find "control" lakes to use in a formal BACI assessment and instead opt for using quartiles for "normal" abundances by lake class as a "control". There is not enough information in the Knapp et al. studies (2008 or 2012) to decipher how these quartiles were constructed but if we assume they rely on CPUE data and use a similar procedure to make population estimates as Knapp et al. (2008), these quartiles suffer from the same limitations as outlined for Knapp et al. (2008) here (e.g., CPUE as a proxy for abundance, pseudoreplication in time, etc.). Further, what is important is not how a particular lake relates to a nonrandom selection of other similar lakes, but how the lake responds after the introduction of muskellunge relative to its pre-stocking populations after taking into account background changes to populations not related to stocking as determined in the control lakes. For example, if walleye population estimates declined after the introduction of muskellunge but remained within the interquartile range estimated for the "control" lakes, would the authors argue that this was a non-effect? If walleye population estimates declined in control lakes that did not have a muskellunge stocking program at the same time and magnitude as they did in the stocked lakes, it would suggest that the walleye populations are responding to some environmental pressure external to the muskellunge stocking program. Yet, without a rigorous control, the authors would conclude the same (i.e., muskellunge stocking has no negative effect on the populations of other fish species). We would argue that the use of similar lakes and the comparison to interquartile ranges is not an adequate substitute for a rigorous BACI design.

10. **Lack of enough information to evaluate merits of the study.** Evaluating the merits of the study was difficult because not enough detail was provided in the methods section. For example, the authors say that they used the same basic netting on each lake, but these analyses are over many lakes, many years, and many surveys. While many or even most of the methods may be consistent, it is important to know the circumstances where inconsistent methods were used. Data and methodology used to generate the interquartile ranges by lake class are lacking. While the total number of sampling occasions before and after stocking was reported, the timing of the sampling events was not reported, and this missing information is important. For example, if there are three sampling occasions before stocking, did those three sampling occasions occur in consecutive years? Were they just prior to stocking or 20 years prior? Were they separated by gaps in years? In general, we should expect enough detail either given or referenced such that we could repeat the surveys in relevant details.

Summary

We acknowledge that addressing the questions that Knapp et al. (2008) do in this report is a difficult task because factors that influence population dynamics of fish are many and they interact in complex ways. The data available to Knapp et al. (2008) with which to test these complex effects were drawn post-hoc from multiple sources. These data and analysis do not serve as a robust assessment of the effect of stocking muskellunge on fish community structure. It may be the case that in some lakes, muskellunge stocking has little impact on other fishes in the community, or it may be that they have a big impact. Our critique is that the data and statistical methods presented in Knapp et al. (2008) do not represent a convincing test of the question of "Do muskellunge stocking programs impact other game species of fish?" and therefore their interpretation (there is no impact) does not follow from the data presented.

From a practical stand point, the angling experience is important to the valuation of lake-front properties and the tourism economy generally, and fish abundance is impacted by fisheries management decisions. Therefore, it is our recommendation that the limitations of the data and analyses used by Knapp et al. (2008) should be acknowledged by decision makers, and that better data, analysis and information should be obtained. The burden of demonstrating "no harm" in fisheries management has traditionally fallen on those supporting a manipulation (Walters and Martell 2004), which in this case is stocking muskellunge. In the absence of good data, managers may have to resort to models that will require knowledge about the ecosystem as a whole to generate predictions on how an individual lake may respond. It is difficult to imagine, even if these data and analyses were strong, that using average conditions of a typical "lake" can be used for making accurate predictions about a specific lake. The idea that these lakes, even within a lake class, are interchangeable is likely not defendable. Each lake is unique in shape, chemistry, biota, use, history and more. These historical contingencies cannot be ignored in making management decisions.

References

Benjamini, Y., and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics* 29, 1165–1188.

Brown, J.H. & Gillooly, J.F. (2003) Ecological food webs: high-quality data facilitate theoretical unification. *Proceedings of the National Academy of Sciences of the USA*, 100, 1467–1468.

Carbone, C., & Gittleman, J. L. (2002). A common rule for the scaling of carnivore density. *Science*, 295, 2273-2276.

Grant, G.C., Radomski, P., & Anderson, C.S. (2004). Using underwater video to directly estimate gear selectivity: the retention probability for walleye (*Sander vitreus*) in gill nets. *Canadian Journal of Fisheries and Aquatic Sciences* 61, 168-174.

Harley, S.J., Myers, R.A., & Dunn, A. (2001). Is catch-per-unit-effort proportional to abundance? *Canadian Journal of Fisheries and Aquatic Sciences* 58, 1760-1772.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics* 6, 65–70.

Hommel, G. (1988). A stagewise rejective multiple test procedure based on a modified Bonferroni test. *Biometrika* 75, 383–386.

Hurlbert, S. H. (1984). Pseudoreplication and the design of ecological field experiments. *Ecological monographs*, 54, 187-211.

Klockars, A.J., & Hancock, G.R. (1994). Per-experiment error rates: The hidden costs of several multiple comparison procedures. *Educational and Psychological Measurement* 54, 292-298.

Knapp, M.L., Mero, S.W., Bohlander, D.J., Staples, D.F., & Younk, J.A. (2012). Fish Community Responses to the Introduction of Muskellunge into Minnesota Lakes. *North American Journal of Fisheries Management* 32, 191-201.

Maunder, M.N., Sibert, J.R., Fonteneau, A., Hampton, J., Kleiber, P., & Harley, S.J. (2006). Interpreting catch per unit effort data to assess the status of individual stocks and communities. *ICES Journal of Marine Science: Journal du Conseil* 63, 1373-1385.

Parkhurst, D.F. (1985). Interpreting failure to reject a null hypothesis. *Bulletin of the Ecological Society of America* 66, 301-302.

Shaffer, J.P. (1995). Multiple hypothesis testing. *Annual Review of Psychology* 46, 561–576.

Stewart-Oaten, A., Murdoch, W. W., & Parker, K. R. (1986). Environmental impact assessment: "Pseudoreplication" in time? *Ecology* 67, 929-940.

Walters, C.J., & Martell, S.J. (2004). *Fisheries Ecology and Management*. Princeton University Press.

Wright, S.P. (1992). Adjusted P-values for simultaneous inference. *Biometrics* 48, 1005–1013.

Reviewer Biographies

Dr. Brian Wisenden has a Bachelors and Master's Degree in Fisheries Management and a Ph.D. in the behavioral ecology of fishes. He has also served as the editor of the peer-reviewed journal Behaviour for 15 years. During his academic career of 23 years he has published over 75 peer-reviewed articles and book chapters. His primary teaching responsibilities include Organismal Biology, Aquatic Biology, and Animal Behavior. Dr. Daniel McEwen has a Ph.D. in impacts of water-level regulation on the community ecology of large lakes in Minnesota. During his academic career of six years, he has published 12 manuscripts in the field of aquatic ecology. His primary teaching responsibilities include General Ecology, Aquatic Biology, Research Design, and Quantitative Biology.

Disclaimer

While every attempt was made to provide an unbiased review of the scientific merits of this study, the authors acknowledge that they were compensated by the Pelican Lake Property Owners Association (PLPOA) for the time they spent in preparing this review. While both reviewers are currently employed by Minnesota State University, Moorhead (MSUM), the comments and content here do not necessarily reflect the position, thoughts, or opinions of the University (MSUM).