1133 A

August 25, 2004

Craig J. Wilson
TMDL Listing Unit
Division of Water Quality
State Water Resources Control Board
P.O. Box 100
Sacramento, CA 95812-0100

Subject: Comments on statistical issues in the draft Water Quality Control Policy

Dear Mr. Wilson:

I am writing to follow up on my participation in the September 8th workshop and to add my comments on the September draft of Appendix A. As I stated to the board, my overall view is that the direction in which these draft policies are moving is a very sensible one from my perspective as a statistician. There are critical features of the July draft that I regard as excellent steps to improve the process of making Section 303(d) listing decisions:

-- The exact binomial method is a good choice of a statistical framework, offering a clear, easily understood basis for routine use in making decisions.

-- The use of an arbitrary "rule-of-thumb", such as "list if at least 10% of the samples exceed the standard" is ill-advised. Such rules amount to "head-in-the-sand statistics", since their performance characteristics—i.e. Type I and Type II error probabilities, are highly variable and uncontrolled.

-- By contrast, the method you and your staff have employed to achieve "balanced error probabilities" at specified "pegs" (true percentages of exceedance) is clear and effective as a means of controlling the effects of random variability in testing in an even-handed way. In my teaching of statistics to science and engineering students at Caltech over the last thirty-six years, I have regularly recommended this kind of specification of error probabilities as more useful and transparent than the usual approach of arguing that one or the other hypothesis should have the *a priori* "benefit of the doubt".

-- There is a wonderful balance between listing and delisting decisions, always using currently relevant test data in the same, consistent way. I will say more about this below.

Some of the participants in the September 8th workshop criticized the use of the same critical thresholds for listing and delisting—e.g. 3 exceedances out of 25 puts a toxicant on the list and 2 exceedances out of 25 takes a toxicant off the list, suggesting that this

approach will lead to frequent cycling between listing and delisting. I disagree. It is easy to show by probability calculations or simulations that the phenomenon of "cycling back and forth" will rarely occur. Moreover, as I argued at the workshop, what any sort of "borderline results" naturally point to is the need for further testing. This is what I understand has influenced some states to use a "planning list".

A more fundamental consideration that I heard discussed at the workshop is the idea that "delisting should be harder than listing". I fail to see why this should be so. In fact, this idea would seem to contradict a fundamental statistical principle called "sufficiency", which says in essence that data used to make decisions should be used in a way that considers only *relevant* information. Rather than introduce technicalities, let me illustrate my point with an example:

Suppose that water bodies A and B are measured once a month for a particular toxicant and that over a 4-year period they each have 5 exceedances. Then in the July draft's tables, 3.1 and 4.1, we can see that both water bodies would barely escape listing based on evaluation of the 4-year period. Suppose further that body A had 3 of its 5 exceedences in the first half of the 4-year period (24 monthly measurements) while B only had 2, and that these data were examined after the first 2 years. Then under the July draft's approach (which I endorse), A would be listed at the 2-year mark (3 out of 24 suffices), while B would not be listed. At the 4-year mark, which of the two bodies should we regard as more worrisome? Both show the same number of exceedances, 5, but in the most recent 2-year period, A has performed better than B. Using the July draft's balanced treatment of listing and delisting, there is no paradox, since A will be removed from the list and B will not be placed on it. But if we "make delisting harder" in some way, then presumably A's performance could be judged insufficiently favorable to justify delisting. Thus, paradoxically, the same overall performance with better *recent* performance would be penalized: A would stay on the list while B would continue to be unlisted.

Let me now address some of the changes made from the July draft to produce the September draft. I understand that the board may feel that for toxicants it is desirable not to require a minimum of three exceedances for listing. And indeed I remarked at the September 8th workshop that if the "pegs" of 5% and 20% were shifted downward, as they have been in the latest draft—to 3% and 18%, then the minimum number could change; so there is a distinction between the statistical logic and the specific implementation of that logic. But I am concerned that the change to a "peg" at 3% will have negative consequences. In general statistical practice, wherever physical measurements are subject to sampling variability and "lab measurement" variability, it is commonplace to expect a non-trivial frequency of so-called "outliers"—i.e. data values that are not representative of the true state of the system being measured. An "outlier rate" of 2 or 3% for physical data would not be exceptional in my experience. I therefore am concerned that using 3% as a peg may be an inappropriate choice. There are other statistical approaches that could be used to modify Tables 3.1 and 4.1 of the July draft in a rational way without changing the pegs, but to me those tables already look very reasonable.

I have an even greater concern about the September version of Table 4.1. Not only have the pegs been moved downward to match the 3% and 5% used in Table 3.1, but also the wonderfully consistent treatment of listing vs. delisting (remarked upon previously) has been destroyed by the sentence immediately below Table 4.1. The critical change in that sentence is the one replacing "alpha and beta at most 0.2 " by "alpha and beta at most .10". If the consistency of inferences for listing and delisting were maintained, Table 3.1 shows that Table 4.1 would require a minimal sample size of 16 for delisting and, for example, sample sizes of 16-24 with at most one exceedance would call for delisting. I recommend deleting this change and restoring the original balance and statistical logic.

I hope the strenuous efforts you and your staff have made at the direction of the board—to develop a solid and consistent system for routine decision-making regarding listing, will bear fruit and will lead to the adoption of such a system, making intelligent use of the science of statistics to guide (but not replace) the careful evaluation of scientific evidence about the possible impairment of water bodies.

Sincerely,

Gary Lorden

Professor of Mathematics
Caltech
Pasadena, CA 91125

12260