Independent Scientific Advisory Board
for the Northwest Power and Conservation Council,
Columbia River Basin Indian Tribes,
and National Marine Fisheries Service
851 SW 6th Avenue, Suite 1100
Portland, Oregon 97204

Review of NOAA Fisheries document:

A Power Analysis of Two Alternative Experimental Designs to Evaluate a Test of Increased Spill at Snake and Columbia River Dams, Using Smolt-to-Adult Returns of Anadromous Salmonids

Members (in alphabetical order)

Kurt Fausch
Stan Gregory
William Jaeger
Cynthia Jones
Alec Maule
Peter Moyle
Katherine Myers
Laurel Saito
Steve Schroder
Carl Schwarz
Tom Turner

ISAB 2018-2
March 19, 2018

ISAB Review of NOAA Fisheries document:

A Power Analysis of Two Alternative Experimental Designs to Evaluate a Test of Increased Spill at Snake and Columbia River Dams, Using Smolt-to-Adult Returns of Anadromous Salmonids

CONTENTS

Executive Summary

In response to a January 2018 request from NOAA Fisheries' Northwest Fisheries Science Center, the ISAB reviewed the document *A Power Analysis of Two Alternative Experimental Designs to Evaluate a Test of Increased Spill at Snake and Columbia River Dams, Using Smolt-to-Adult Returns of Anadromous Salmonids (*January 2018 draft*)*. NOAA's analysis considers two general experimental designs: (1) a before/after design for which there is no variation of spill level within the prospective years and (2) a block design that includes variations between two spill levels during prospective years. NOAA Fisheries requested the ISAB's review to help inform their analyses and recommendations regarding potential future operations at the Columbia River System mainstem hydroelectric projects.

The ISAB found that NOAA's analysis is a standard assessment of the power to detect effects and appears to be structured appropriately. The methodology and conclusions are conceptually sound, and the block design provides advantages to a before/after study. The key advantage to the block design is that high year-to-year variation is controlled for by conducting both spill regimes in the same year. However, the advantages are somewhat tempered because of several sampling and estimation issues. In addition, the success of the proposed experiment may depend on a number of factors including the availability of sufficient water and tagged fish to actually implement the experiment, the assumption that past fish behavior (in the retrospective years) is indicative of what will happen under the new spill regime, and other operational issues that would need to be resolved before an actual experiment is implemented.

These operational issues include:

- the expected improvement in smolt-to-adult returns (SARs) associated with particular spill levels
- the impacts of low and high flow years on the study's implementation
- whether the proposed spill regime is equally beneficial for fish of all sizes/ages when they start their migration
- impacts of the proposed spill regime on fish travel speed
- effects of new spillway detectors on the study design
- identification of adaptive management triggers, and
- the experiment's effects on migrant survival when only half of the year is at higher spill levels compared to a full year at higher spill levels.

So, while a theoretical implementation (i.e., NOAA's block spill paper) may show high statistical power, an actual implementation may have less power because of these problems. These issues and several topics that should be explored to strengthen NOAA's analysis are described in the full report.

INTRODUCTION

In response to a January 2018 request from NOAA Fisheries' Northwest Fisheries Science Center, the ISAB reviewed the document *A Power Analysis of Two Alternative Experimental Designs to Evaluate a Test of Increased Spill at Snake and Columbia River Dams, Using Smolt-to-Adult Returns of Anadromous Salmonids (January 2018 draft)*, referred to as Smith (2018) in this review. NOAA Fisheries (also referred to as NMFS) requested the ISAB's review to help inform their analyses and recommendations regarding potential future operations at the Columbia River System mainstem hydroelectric projects. The NOAA Fisheries' request memo includes a useful background summary:

> "For the past decade, spill has been a prominent strategy for increasing salmon survival through the Columbia River federal hydropower system. Increasing spill in the future has been proposed as a method to further increase productivity by increasing smolt to adult returns (reducing latent mortality) of migrating salmon and steelhead. Understanding whether and how much salmon benefit from higher spill is therefore an important question for the region. Two broad categories of experimental design are available to evaluate the effects of increased spill compared to the baseline spill level that has occurred over the past decade: 1) a before/after design that compares survival in future years with increased spill throughout the season to survival in past years with baseline spill, or 2) a block design that varies future spill levels within years and compares future survival both to the baseline period and among blocks. Staff at the Northwest Fisheries Science Center have proposed an example of a block design and have conducted a preliminary power analysis that indicates such a design will generally be better able to detect benefits from increased spill compared to a before/after approach."

NOAA Fisheries asked four questions about the power analysis of Smith (2018) that the ISAB answers below:

1. Would a block study design, (either as described or in a modified form), enhance the region's ability to assess the potential for increased spill to reduce latent mortality (increase SARs) relative to a before/after comparison test (as proposed by the CSS Oversight Group - May 2017 submittal to ISAB)?
2. Can the ISAB identify any logical or study design flaws relating to a block design study that NMFS has not adequately considered? If so, what are they?
3. Could the proposed block study design be improved (timing, duration, etc.) to increase its power? If so, what are the ISAB's recommendations?
4. Aside from SARs, juvenile reach survival, and travel times, are there other metrics that the ISAB would recommend considering as a means of assessing this question?

The ISAB's review was informed by presentations from Steve Smith at our October 27, 2017 meeting and from Jay Hesse and Ryan Kinzer (Nez Perce Tribe), Ed Bowles (Oregon Department of Fish and Wildlife, ODFW), and the Comparative Survival Studies (CSS) Oversight Group—Charlie Petrosky (Idaho Department of Fish and Game, IDFG), Steve Haeseker (US Fish and

Wildlife Service, USFWS), and Adam Storch (ODFW) at our February 23, 2018 meeting (link to talks and review materials). These presentations raised important questions about how the actual block spill would be implemented. In summary, issues were raised about:

1. What are the proposed operations for the spill block study? What effect size is expected given the models from the CSS and others that describe the relationship between existing flows and survival and SARs?
2. Will it be possible to maintain the proposed flow levels for the majority of the study under different flow patterns that may occur in the future? If not, what should be done with the data outside of the planned operations?
3. Is there benefit to having one of the spill blocks be the same spill operation as the previous eight years (i.e., 2008 BiOp operations)? Is there benefit in having a wider contrast in the proposed spill levels relative to current operations?
4. Does the block spill design affect the collection of transport and in-river migration (T:I) ratios? Is there another way to get these data to ensure they are collected in support of the long-term trend analysis for the efficacy of transportation?

While these issues are beyond the scope of the power analysis, they will need to be considered when implementing a block spill experiment.

The ISAB also appreciated NOAA Fisheries' participation in the February 23 meeting to answer questions raised by the ISAB and the presenters.

ISAB SUMMARY COMMENTS

The analysis of the block spill design must perform several tasks to evaluate the ability of a proposed experimental design to detect the impact of increased spill using smolt to adult returns (SARs):

- estimate the effect of period (the two-week intervals during which spill could be modified), the year-specific effects and the period-year specific effects on SARs from the retrospective years (already collected) in order to simulate such effects in future (prospective) years.
- estimate the extent of overlap among the four migrating cohort groups as they migrate through the hydrosystem to estimate the net effect of any proposed nominal spill effect on the individual cohorts. This overlap is termed "spreading" in Smith (2018).
- simulate future populations of fish to generate realistic outcomes from various spill scenarios. These are then analyzed with the retrospective data, and the power to detect spill effects is estimated.

Appendix A and B of Smith (2018) provided details on estimating period effects and the variance of the year-specific and period-year specific effects. Appendix A begins with a tabulation of the SARs by period and year and different combinations of stocks which is then analyzed using the mixed-effects logistic model:

$$logit(SAR_{ij}) = (\alpha + \rho_i) + (\gamma_j + \varepsilon_{ij})$$

where $a + r_i$ represents the baseline and period (early vs. late) effects, $\gamma_j + \varepsilon_{ij}$ represents the year and year-period effects (observational variation) respectively, and results are reported in Appendix B of Smith (2018). The mean number of smolts (on the logarithmic scale) released in the tail race of Lower Granite Dam (LGR) in each period is also estimated. This is described in Steps 1 and 2 on page 7 of Smith (2018).

Appendix C of Smith (2018) provides details on how cohorts released at LGR "spread" out as they move through the hydrosystem. This spreading will make it more difficult to detect the impact of the new spill regime because not all fish will experience the new spill regime at all dams. This spreading was estimated using a product-binomial method where the overall probability of staying in one of the periods during movement through the hydrosystem is found as the product of the sequential probabilities of staying in the same period as a fish moves from dam *i* to dam *i+1*. This probability is then used to estimate the impact on the assumed effect of the new spill regime if a fish did not experience the new spill regime at all dams by using a simple power model for the effect of the new spill regime (i.e., each dam contributed equally to the overall spill effect). A simple approximation is used to estimate the spread effect on any expected impact of a new spill regime. This corresponds to Step 4 on page 8 of Smith (2018).

The estimates of variability from the retrospective years are assumed to be valid for all future years. These are used to generate estimates of the SAR values that might be expected in the absence of any new spill regime effects. The adjusted spread factors are then used to "implement" the new spill regime effects. Note that the treatment effects tested in this power analysis are simple proportional increases in SARs (e.g., a 25% increase) and are not based on any biological relationship between actual spills and expected improvements in SARs. The period-year random effects are applied; sample sizes are generated based on historical patterns; and the number of "returning" adults is then generated. These simulated data for prospective years plus the actual retrospective data are then analyzed using a mixed-logistic model:

$$logit(SAR_{ij}) = (\alpha + \rho_i + \tau G_{ij}) + (\gamma_j + \varepsilon_{ij})$$

where t is the NET effect of spill (more on this later) and G_{ij} is an indicator variable (0/1) if period *j* in year *i* was targeted for the new spill level. The proportion of simulated data sets where a spill effect is detected is used as an estimate of power. Note that because of the spreading of cohorts as they move down the hydrosystem, the parameter t is NOT an unbiased estimate of the impact of spill on a particular cohort. For example, the new spill level could double the SAR for this cohort, but the value of t may only be 1.5 because of the spreading of the cohorts across both spill levels as the fish move downstream. The spreading of cohorts (typically) reduces the response of fish cohorts to the new spill overall, but a statistical test for the null hypothesis that $\tau = 0$ is still a valid test for no spill impact. Conversely, because t may underestimate the impact of spill, the experiment may fail to detect the NET impact of spill because of an inadequate sample size (too few fish and/or too few years).

Smith (2018) is a standard assessment of the power to detect effects and appears to be structured appropriately. As summarized in the answers below, the methodology is appropriate and conclusions sound.

There are several topics that should be explored to strengthen Smith (2018), and a number of operational issues need to be resolved before the spill test is implemented.

AREAS IN WHICH THE ANALYSIS COULD BE STRENGTHENED

There are a number of areas where additional information could strengthen Smith (2018).

a) What are other sources of variability that could impact the results?

The simulation study must make a number of simplifying assumptions about variation when simulating data for prospective years. The experiment assumed that all sources of variation in SARs have been accounted for in the historical record and that the past reflects the future. The paper needs a discussion of potential impacts of changes in the variation in the future. For example, is the variation seen in the retrospective years still valid in the face of climate change? Will changing land use patterns have an impact on flow patterns? Have there been changes in the hydrosystem that make variation seen in retrospective years no longer applicable? Will ocean conditions over the next 10 years be represented by the historical record?

b) Is there an advantage to redefining the cohorts to avoid effects of spread?

The impact of the new spill regime is attenuated because not all fish experience the targeted spill regime at all dams due to fish “spreading out” as they move through the hydrosystem. Appendix C of Smith (2018) estimates the impact of this spreading. Is there any advantage to redefining the cohorts to reduce the amount of spreading across the different spill regimes, such as only using fish released from the first week of the block? The number of fish assigned to each cohort will be smaller (but as noted elsewhere in the paper, the effect of the number of fish on the power analysis is small), but the treatment effect will be cleaner. The modification may result in an increase in statistical power. There is a brief mention of this in the discussion of Smith (2018), but no formal analysis appears to have been done.

c) Methods to actually estimate the effect of the new spill regime need to be developed.

The current model is unable to produce unbiased estimates of the effect of the new spill regime on SARs for two reasons:

- the spill effect is modeled as multiplicative on the actual SARs, but the effect of spill (the t parameter) is estimated on the odds of returning. This is a technical issued that can be resolved by modeling spill effects directly on the odds scale rather than the SARs scale.

- the effect of spill regimes is contaminated by spread of the cohorts.

As noted in the discussion of Smith (2018), an approximate way to resolve the second issue is to use the approximate spread factor directly in the model. This is an important issue to resolve because once the spill experiment is enacted, a key parameter to estimate is the actual impact of spill. It is unclear if the spread can be actually measured if the spill experiment is implemented because not every dam has spillway detectors to provide estimates of when fish pass over a particular spillway.

d) Do interim analyses provide early indicators of success?

Under the current block spill design, the entire returns from a brood year are required for analysis. Does an interim analysis (e.g., after the first major year of returns) provide any early indicators of success (or failure)? Presumably, the retrospective data currently available can be disaggregated even today and the feasibility and usefulness of an interim analysis investigated in the simulation study.

e) Can multiple spill levels be tested to accelerate learning?

The current block design allows for only one spill level to be tested in each year. It is unlikely that multiple treatment spill levels could be implemented in a single year due to logistical constraints. Can the current analysis be modified to account for two different spill levels over prospective years (e.g., lower and higher additional spill) to accelerate learning while gaining experience with modification to spill levels during operations? By doing so, the answer to the question "Is more spill better?" could be determined quicker, even though estimates of the individual spill effects may not be well estimated. For example, the term $\alpha + \rho_i + \tau G_{ij}$ in the model would be modified to $\alpha + \rho_i + \tau G_{ij} + \varphi H_{ij}$ where the final term is the second (higher spill level) and the test for no spill effect is the joint test that both τ and φ are zero. Similarly is it possible to model spill as a continuous variable so that even if the higher spill varies among years beyond their control; the model can still be fit?

f) Current block spill design is a special case of carry over designs.

The current experiment is a special case of a carryover design (refer to https://onlinecourses.science.psu.edu/stat509/node/123). The design and analysis of carryover designs was never referenced in the paper, and the experience with carryover designs can provide useful insights into the potential biases and efficiency gains in the spill block design. For example, carryover effects in carryover designs are analogous to the impact of cohorts spreading as they move down the hydrosystem, and experience with carryover designs in accounting for carryover effects may be useful in obtaining an unbiased estimate of the spill effect in the block spill design.

CONSIDERATIONS IF THE SPILL EXPERIMENT IS IMPLEMENTED

These items do not require an immediate response but will need to be considered if the actual experiment is proposed.

a) What effect size is to be expected?

The current paper is very “general” in that there is no assumption of what level of spill corresponds to particular spill effects in their model. Several presenters attempted to map projected impacts of spill scenarios to the effect sizes in this paper. This should be done prior to implementation so that the most appropriate power analysis is selected.

b) What are the impacts of low flow and high flow years on implementing the spill experiment?

The simulation study assumes that the new spill regime can actually be implemented as scheduled. However, are there operational (e.g., power generation) or physical conditions (e.g., lack of water or too much water) that could prohibit/shorten the new spill regime or affect the control spill in a particular year? The historical power/flow records should be examined to estimate the risk that the actual spill experiment could not be implemented in a particular year.

c) Is there any information on whether the impact of the proposed spill regime is equally beneficial for fish of all sizes/ages when they start their migration?

Appendix C of Smith (2018) makes a simplifying assumption that the impact of the new gas-cap spill regime is the same for all fish regardless of size or age when they start of migrate. Are there any empirical data about this issue? If size/age is important, then this may impact the formation of the cohorts in a particular year.

d) Is there any information on the impact of the proposed spill regime on travel speed?

For example, perhaps under the new gas-cap spill regime, the spread of cohorts may decrease (which presumably should increase power). What do the CSS models on the effect of spill indicate about the possible impact of the new gas-cap spill regime on the spread of the cohorts over time? This may influence how cohorts are defined if and when the spill experiment is implemented.

e) Do spillway detectors affect the design of the experiment?

Spillway detectors are being installed on many dams. If the exact travel path were known for every fish, there would be no need to define cohorts as is currently done. Will the spillway detectors have sufficient detection efficiencies to change how the experiment is run?

f) Need to place the experiment into an adaptive management framework.

An adaptive management framework is needed when the actual spill experiment is proposed. For example, how often will data be examined for early indications of success/failure? How will the experiment be modified if early indications (e.g., an interim analysis before all returns from brood years are available) are that higher spill levels are indeed beneficial or if the observed effect size is much smaller than predicted? Will the proposed spill experiment collect the right type of information to make interim analyses possible?

g) Need to acknowledge the consequences of the experiment on migrant survival compared to full spill for entire seasons at 115/120% or 125% spill.

If survival is lower at lower spill rates, the survival that results from the experiment will be lower for ten years than it would have been with the higher spill throughout the entire spill season each year. The potential biological consequences of the experiment relative to other management options should be identified explicitly as part of the adaptive management framework. Estimates of the potential magnitude of the differences between options would inform decision makers about both experimental and biological strengths and limitations.

ISAB RESPONSES TO NOAA QUESTIONS

Finally, we provide summary responses to NOAA's questions submitted with Smith (2018).

1. *Does a blocked design provide advantages to the before/after study?*

Yes, the key advantage to the blocked design is that the high variation seen in year-specific effects (i.e., year-to-year variation) is "controlled for" by conducting both spill regimes in the same year. The advantages are somewhat tempered because of the "spreading" of the cohorts, but the amount of spread appears to be small enough that it does not compromise the ability to detect differences in SARs between the two spill regimes. The current analysis model does not provide a direct estimate of the effect of the new spill regime, but the analysis can be modified as briefly discussed in Smith (2018).

Similarly, the benefits of additional spill may be tempered because releases must be made downstream of Lower Granite Dam and so these fish have already experienced one power house contact. It was pointed out during the presentations that spillway detectors are being installed on the Lower Granite Dam spillway, so it may not be necessary to define cohorts using fish that have had the initial dam contact.

2. Are there any logical or study design flaws that are not adequately considered?

The study design is conceptually sound; the success of the proposed experiment may depend on the availability of tagged fish and water to actually implement the experiment, and the assumption that past behavior (in the retrospective years) is indicative of what will happen under the new spill regime. For example, spilling more water may reduce travel times and so the "spread" of the fish may differ from that forecasted spread based on the retrospective analysis. High water levels may make it impossible to have a "control" BiOp spill level in some years. Fluctuation in flow may not make it possible to maintain the assigned spill levels in a particular treatment period. These could impact implementation of a particular spill experiment. So, while a theoretical implementation (i.e., NOAA's block spill paper) may show high statistical power, an actual implementation may have less power because of these problems. These implementation considerations will need to be considered when actually designing a proposed experiment.

It is important that the analysis methods be developed to provide an unbiased estimate of the impact of spill (see below).

While the proposed design does not prevent studying the impact of transportation on survival compared to in-river fish, the overall survival rate of in-river fish under the block spill will consist of the survival rate of fish that "experience" the control flow and fish that "experience" the new spill levels vs. the survival rate of all fish that experience the same spill level. It is not clear if the estimate of the transportation effect will then be comparable to previous years where only a single spill level was in effect.

3. Can the proposed block study design be improved to increase power?

Estimated statistical power is already high (i.e., only a few years of implementation are needed to detect an effect). Modifications to the block study design to improve power will not lead to substantial increase in power due to any additional increases in effect size being swamped by uncontrollable events during the experiment. However, the issue of fish cohorts "spreading" as they travel downstream may be amenable to improvements by changes to the cohort definition (e.g., truncating cohort 2 earlier) to reduce the incidence of these fish experiencing the other spill regime.

Learning may be accelerated by implementing two levels of treatment spill (high and higher) in future years at the price of reduced precision of the impacts of a particular spill level.

4. Are there other metrics?

Juvenile reach survival and SARs integrate the effects of spill on fish over the remaining lifecycle. As such, it is difficult to separate out where the potential benefits of the new spill regime may actually occur, i.e. improved estuary, plume, or ocean survival. Unfortunately, the current monitoring systems cannot provide separation of mortality into components once the fish pass Bonneville unless acoustic telemetry is used. The proposed spill experiment needs to be well coordinated with ongoing estuary, plume, and ocean projects influencing factors investigating survival of juvenile and adult salmon and other focal species so that the experiment does not disrupt these other studies.

There was no direct assessment of fish health under the new spill regime (e.g., is there evidence of gas trauma?). Some monitoring of fish directly downstream of the dams could be done (e.g., netting) to capture juvenile fish after they pass through the new spill regime to gather such data.

While the new spill regime may provide improvements in the MEAN SARs, there is no monitoring on the variability in the SARs over time. Much of this is driven by uncontrollable factors (e.g., the Pacific Decadal Oscillation), but if the new spill regime exacerbates the variation in SARs (even while providing an increase in the mean), it may not be a preferable option because of the increased risk of quasi-extinction. Unfortunately, estimating the variability over time in the SARs under the new spill regime would require an order of magnitude more effort (e.g., duration of the spill experiment in the 10-20 year range) and therefore is not likely a feasible metric for a short term experiment.

The current design must wait until all returns are received from a brood year. However, an interim analysis may provide early indications of success or failure of the experiment. It is unclear if the SARs values from previous years can be disaggregated without considerable effort in reanalyzing all the past tagging data.

EDITORIAL COMMENTS

p.i - One response variable is smolt travel time. Some background should be added so that readers understand why this is biologically important (i.e., is there evidence that travel time is related to fish survival?)

p.ii - There are numerous instances where the wording can be improved. For example,

> “The Block Design exceeded Before/After more for Chinook data than for steelhead data.”

should be reworded as:

“The increase in the power of the Block Design relative to that of the Before/After design was more when the spill experiment operated on Chinook rather than for steelhead.”

These types of "shorthand" used in the report need to be expanded to be understandable for an audience not completely familiar with the concepts of power or simulation studies.

p. ii (and elsewhere) - "The effect of this spreading is 'noise' in the signal: the realized contrast between cohorts is less than the ideal targeted contrast." While "noise" is a common statistical term, some readers may not understand its usage here. Perhaps reword along the lines of "The effect of this spreading obscures the signal…."

p.iii (and elsewhere) - SAR can only be determined after an entire brood year has returned as adults. So sentences such as "2 or 3 years of data" really mean "2 or 3 years of implemented spill." This needs to be clarified so that readers will not assume that after 3 years you actually know the results.

p.1 - Again, shorthand sentences such as

> "Statistical tests then test for differences between prospective ("After") years and retrospective ("Before") years.

need to be reworded to:

> "Statistical tests then test for differences in mean logit(SAR)s between prospective ("After") years and retrospective ("Before") years.

p. 3 - Is "season" synonymous with "period"? It is not clear how the statement preceding Table 2 results in 16 patterns unless season is the same as period. Also, is "block" the same as "period"?

p. 12 - It is surprising that the Before/After design has higher power than the blocked design. Perhaps this is simulation "error" (i.e. just an artifact of the simulation randomness?).

p. 13 - "A second notable result is that the estimated period effects (ρ_2, ρ_3, ρ_4) tended to be more significant for Chinook than for steelhead." It is not clear what is meant by "more significant."

p. 18 - It is not clear why the black dashed line has power < alpha = 0.10. What happened here? See for example, page 19 where it is exactly at alpha (as expected). Same problem is in Figures 3, 4, and 6.

Figures 1-8 - In the legends, is the "increase" referring to increase in spill, or increase in fish numbers or survival? Similarly, in the y-axis title, is the "increase" that would be detected the increase in fish numbers or survival?

Figure C.1 - This is a nice figure, but adding some explanation in the caption of abbreviations in the figure would be helpful (e.g., does "pd. 1" refer to Period 1, and "cap" refer to "Gas-cap spill"?). Also, what is the scale of the y-axis—perhaps proportion of fish passing over time?