An overview of the scientific basis for the US position related to the bigeye stock assessment structural uncertainty grid

Not all tagged fish that are caught get reported. The assessment has to assume how many, and it offers two ways of handling that assumption. The 2026 grid uses both and treats the choice as uncertainty. This document explains why one of the two is the sounder choice, and why carrying both does not represent uncertainty about the stock. The United States position is stated first; the evidence follows.

96 of 112 named checks held · 100 drawn · 88 ran · **80 converged (MGC ≤ 1e-4)** · 168 fixed-parameter MFCL evaluations

The United States position

As it relates to the 2026 bigeye tuna stock assessment, the United States would be prepared to accept the current assessment if the grid were constrained to only include the tag reporting adjustment models.

Working from a shared understanding of the Multifan-CL code, including the tag reporting rate adjustment within the mixing period is unbiased under the conditions required (e.g., the true reporting rate equals the estimated reporting rate, tag release groups are large enough to be informative, and the survival floor doesn't trigger). Excluding the tag reporting rate adjustment assumes complete reporting of recaptured tags, and mathematically will always remove fewer tags than including it, which in turn will generally result in a lower estimated fishing mortality. The assumption of complete reporting is implausible given tag seeding experiments and inspection of the assessment input files. We would therefore be averaging an estimate that is capable of being unbiased with one that is biased by construction, with the net result potentially inducing a bias into the overall uncertainty grid. This is the basis for our preference to only retain runs that include the adjustment as they are the only ones theoretically capable of providing an unbiased estimate.

This relates to a larger fundamental uncertainty relative to the influence of tagging data which ties back to our initial recommendation that a more appropriate grid axis would integrate across the inclusion/exclusion of tagging data in the model. However, this work is not possible at this time so what we are currently proposing would be a compromise until future work informing this issue can be completed.

The remainder of this document sets out the analysis supporting that position. Each element of the position above is taken up in turn: the conditions for an unbiased corrected estimate in Part 6, the ordering of the two options and what it does and does not establish in Part 4, the comparison of fitted rates against the tag-seeding experiments in Part 6, the effect on derived management quantities in Part 5, and the wider tagging-data axis in Part 8.

Executive summary

Tuna are tagged, released, and some are later recaptured and reported. The assessment must assume what fraction of recaptures actually get reported: the **reporting rate**. Right after release there is a settling-in period, the **mixing period**, during which tagged fish have not yet spread through the population, so the model is not supposed to fit their recaptures.

But those fish have still been caught, and they still have to be taken out of the tagged group. MFCL offers two ways to do that, and the 2026 ensemble uses both, splitting the models roughly evenly and treating the choice as a source of uncertainty. **This analysis asks whether that treatment is supported by how the two options actually behave, and finds that it is not.**

Neither option is unbiased, and the two are not symmetric

The option that *does not* correct for reporting, **exclude**, removes recaptures at face value, which amounts to assuming every recaptured tag was reported. The model's own fitted rates say that does not hold for most of the tagging data. It also inserts terms into the fitting criterion whose best possible answer is always “reporting is 100%”. Those terms are not information about the fishery; they follow from the setting itself, and they push the estimated reporting rates upward, away from what the independent tag-seeding studies indicate.

The option that *does* correct, **include**, divides each reported recapture by the estimated reporting rate before removing it. That can give an unbiased answer, but only if the reporting rate is about right, the release groups are big enough to be informative, and the scaled-up removal does not try to take out more tags than exist. When the last condition fails, MFCL caps the removal, and standard model output does not record that it happened.

Why this matters for advice

Reporting rates and fishing mortality trade off against each other. Tags that come back are evidence of fishing. If the model assumes a larger share of recaptured tags was reported, then the tags it sees account for more of the fishing, and it needs less fishing to explain them. Assume a smaller share and the opposite follows. So an assumption about reporting is also, unavoidably, an assumption about how heavily the stock is fished.

That is why the choice cannot be treated as a free modeling convention. It moves the quantities the advice rests on. The case made in this document rests on which setting is capable of being unbiased, and would stand regardless of which direction the numbers happened to move. Part 5 reports the size of the effect for completeness.

The corrected setting is also the fragile one

Of the 50 models drawn for each setting, **16 of the 50 include models were lost**, either failing to build or failing to converge, against **4 of 50** for **exclude**. That is 32% against 8%, a factor of four. Convergence failure alone runs 17% against 2%. Everything reported below is stated on the **34 + 46 = 80 models that met the assessment's own convergence criterion** (maximum gradient ≤ 1e-4), which turns out to be exactly the set carrying a retained parameter file.

This is a real cost of the corrected setting and it should be stated plainly: dividing observed recaptures by an estimated rate is numerically demanding, and MFCL does not always get there. It is also the reason the two arms are not evenly represented, since the surviving **include** models are the ones where the correction happened to be well-behaved.

The two options bracket each other, not the truth

The two settings do span a range. Correcting for reporting can never remove fewer tags than not correcting, so F under **exclude** sits at or below F under **include**, and the grid does therefore report an interval. That much is correct. The question is whether that interval is likely to contain the right answer, and there are two reasons to think it is not.

An interval is only as good as its endpoints. If **include** is unbiased under its stated conditions, then **exclude** sits below it by construction. Carrying both then averages an unbiased estimate with a biased one. That does not widen a valid interval; it moves the central estimate.

Both endpoints can sit on the same side of the truth. The raising each option applies depends on the *estimated* reporting rate, not the true one, and the true rate is not known. In this assessment the fitted rates land above the independent tag-seeding evidence in all five priored groups. If the seeding trials are nearer the truth, then both settings raise the removal by too little, both remove too few tags, and both understate F. The interval between them would then lie below the answer rather than around it.

The defensible claim is not that the true reporting rate is known, nor that the two options cannot be ordered. It is that the true rate is almost certainly not 1.00, which is what **exclude** assumes. A setting with a very low probability of being correct is a poor candidate for a grid axis, whichever side of the other option it falls on.

What the analysis supports

If models fitted to tagging data must inform advice, use include only. It is capable of being right under stated conditions; **exclude** is biased low unless reporting is genuinely complete, which it is not.

The MFCL manual recommends the opposite, on the grounds that the correction can be numerically unstable. That numerical concern is real, and the manual is written as general guidance across applications. What this analysis adds is a measurement of both sides of the trade in this particular assessment: at the reporting rates the models actually estimate, the numerical guard costs at most **1.04 objective units** and engages on **2 of 98** release groups, while the alternative affects **96 of 98**, about 48 times as many, in every model. **Here, the problem the manual's preference avoids is the smaller of the two.**

The deeper point, from the United States review, is that *both* options fix the tagged population deterministically from the observed returns, with no uncertainty and no test against what the model predicts. A large share of the tagging data steers the assessment without ever being fitted. The durable fix is to change MFCL so mixing-period recaptures are predicted from the tagged cohort and scored, with reporting rate and mixing-period availability estimated rather than assumed, or to base advice on models that do not fit tagging data at all.

The rest of this document sets out the evidence. Every directional claim names a check that either held or did not, and failures are reported in the same form as passes.

A note on scope. This is an analysis of one configuration choice in one assessment, not an evaluation of MFCL or of the 2026 assessment as a whole. The behavior described below follows from a deliberate and reasonable design decision, namely not to fit recaptures from fish that have not yet mixed, and the difficulty it runs into is genuine. Every source claim is transcribed from a named file at a named commit, every number is produced by scripts in this repository, and the links in the footer resolve to the public commits the analysis was run against, so any statement here can be checked or contradicted directly.

*A note on attribution. The companion tag analysis cited throughout is the work of **United States stock assessment scientists** and is referred to here as the US review. Its repository is named for the meeting it was prepared for, but it is not an output of SC22 or of the Scientific Committee, and no conclusion in this document should be attributed to either. The findings and recommendations below are our own. The one genuine SC22 citation is the tag-seeding information paper SC22-SA-IP-05, which supplies the reporting-rate priors.*

PART 1 · THE ORIGINAL QUESTION

Why the reporting-rate penalty moves with the axis

The starting observation: of all the objective-function components, the **tagged-fish reporting-rate penalty** is the one the `RR` axis really moves, with a median near 60 under **include** against 100 under **exclude**. The regional recruitment penalty differs by a couple of units; the unclassified residual barely differs. The effect is localized in the reporting-rate prior, not smeared across the fit.

Restricted to the 80 converged members. The `RR` axis moves the tagged-fish reporting-rate penalty by roughly ten times as much as it moves either of the other two components. This is the distributional picture behind the medians quoted above.

That penalty is a prior. Tag-seeding studies (SC22-SA-IP-05) give expected reporting rates for the purse-seine fisheries in regions 2, 3-west and 3-east; the penalty charges the model for departing from them. Transcribed from `multifan-cl/src/callpen.cpp`:

$$\text{penalty} = \Sigma \text{ over unique reporting-rate groups: } w_g \cdot (\hat{X}_g - \text{target}_g/100)^2$$

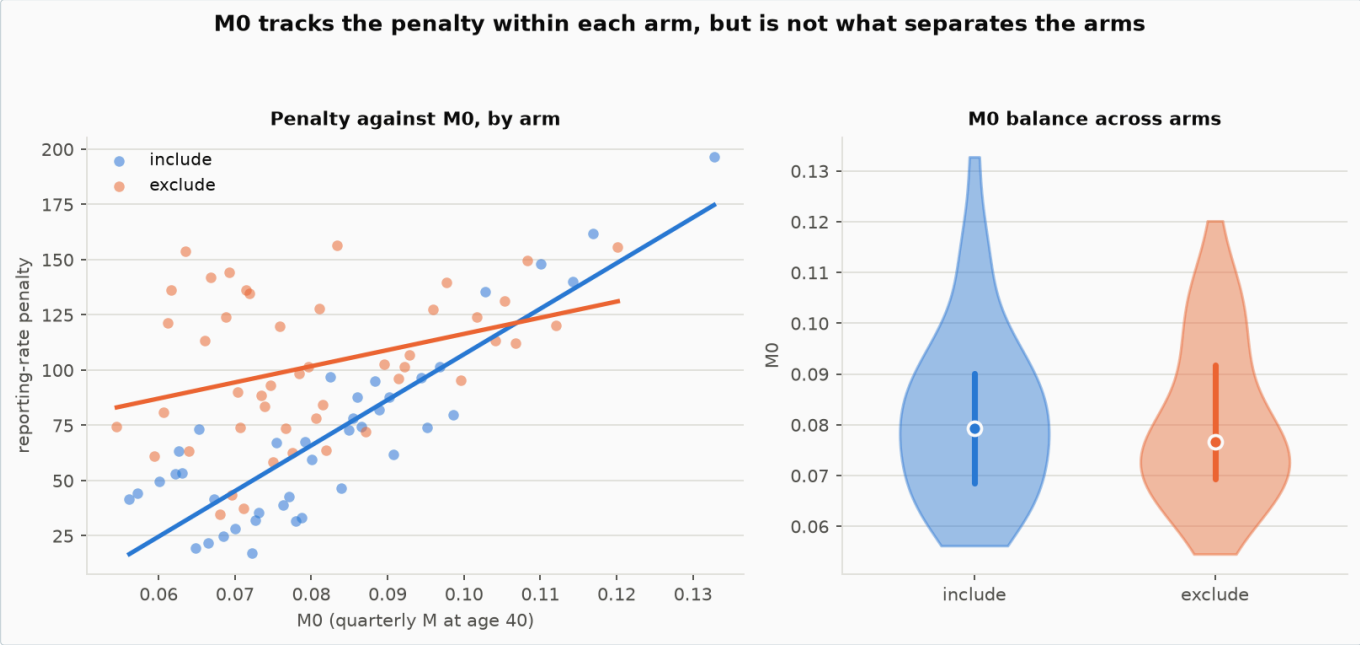
Reconstructing it from the PAR files reproduced the reported component to $\mathbf{4.9 \times 10^{-10}}$ across all 80 members. The arithmetic is settled and everything downstream rests on it.

> **HELD** `task3.reconstruction-matches-reported`
>
> max|resid| = 4.86e-10 (tol 1e-06); median resid = 2.32e-12

Three structural corrections to the original framing. **Five** groups carry an informative prior, not three: there are three distinct priors, each shared by an RTTP and a PTTP/pooled group. Group 18 is `PS.EAST.3`, not region 4. And the selector in the source is the penalty matrix, not the active-flag matrix.

Not natural mortality

M0 has a tighter relationship with this penalty than the flag axis does, and the 88 models are a selected subset, so M0 leaking in was the obvious worry. It is not: M0 is balanced across arms, and adjusting for it *increases* the axis effect from 33.98 to 34.3 while R^2 rises from 0.45 to 0.85.



All 88 retained models, before the convergence screen. M0 predicts the penalty within each arm (left), which is where the R^2 jump comes from, but the two M0 distributions overlap (right), so M0 is not the thing separating **include** from **exclude**.

HELD `task4.M0-balanced-across-arms`

standardised mean difference (exclude - include) = -0.076

HELD `task4.adjusted-arm-effect-positive`

adjusted arm effect (exclude - include) = 35.54 [28.72, 42.36]

One balance check did not hold: K is imbalanced at K = 0.20 (include 7/41, exclude 16/47). K is in every adjusted model and the adjusted estimate is the larger one, so the imbalance does not manufacture the effect. Reported because it is real.

DID NOT HOLD `task4.K-balanced-across-arms`

largest level-share difference = 0.170

Which groups, and the mixing window

Group 17 dominates the penalty's *level* (68% of the include-arm median) but supplies only 19% of the arm *gap*; that comes from group 18 (45%) and group 7 (34%). Level and gap are answered by different parameters.

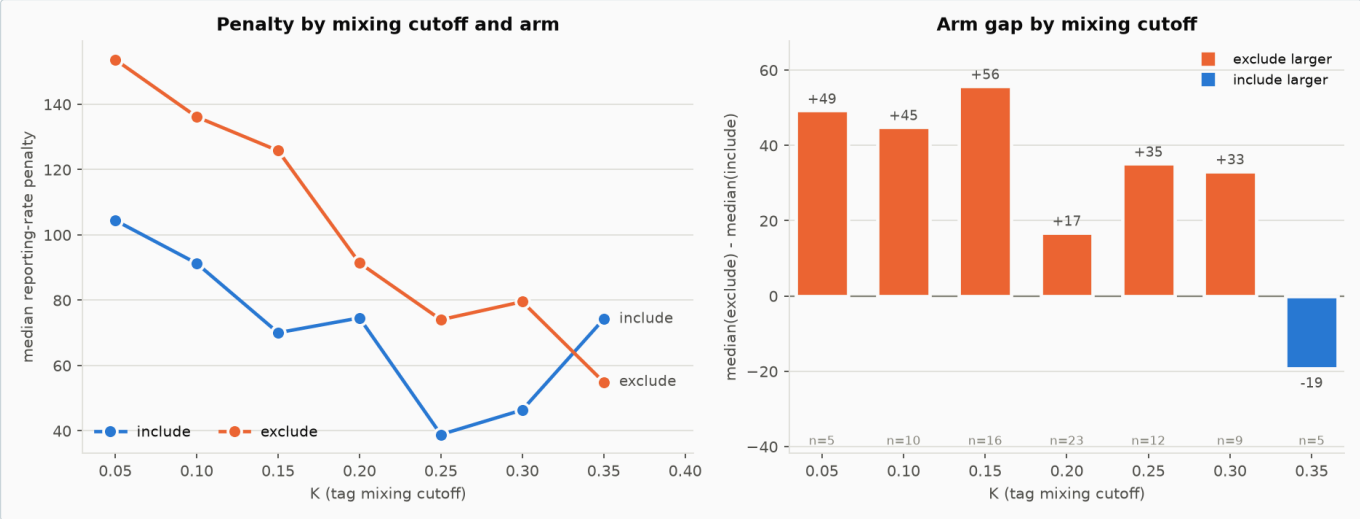
DID NOT HOLD `task5.arm-gap-dominated-by-group-17`

group 17 supplies 19.3% of the median total-penalty arm gap; group 18 supplies 45.0% and group 7 34.1%

The gap scales with the mixing window in both directions tested: arm × K = −218 [−298, −138], and arm × mean mixing period = +21.1 [13.7, 28.6].

HELD `task6.arm-gap-shrinks-with-K`

arm x K interaction on the raw penalty = -218 [-298.3, -137.7], p=7.74e-07



The gap is largest where mixing windows are longest and vanishes where they collapse. This is the first sign that the mechanism lives inside the mixing period.

PART 2 · THE MECHANISM

Channel 2: why exclude's in-window terms favor complete reporting

The first investigation explained the gap by a *removal* argument: the reporting rate has an extra lever under **include**, so a smaller excursion suffices. That is correct but direction-agnostic. A second mechanism predicts the sign, and the observed sign is upward.

Mixing-period recaptures *are* scored under both settings: the likelihood loop starts at the release period and the multiplier never consults the flag. What differs is what the prediction is built from. MFCL solves for a fishing mortality that removes a target number of tags, and the flag sets that target. Writing `R` for observed returns and $\hat{X}$ for the estimated reporting rate:

SETTING	REMOVAL TARGET	PREDICTED REPORTED RETURNS	CONSEQUENCE
include	$R / \hat{X}$	$\hat{X} \cdot (R/\hat{X}) = R$	$\hat{X}$ cancels, so the term carries no information about it
exclude	R	$\hat{X} \cdot R$	**minimized at $\hat{X} = 1$**, best satisfied by complete reporting

Under **exclude** every mixing-period recapture becomes a small vote for "reporting is complete", and the tag-seeding prior is the only thing pushing back.

Confirmed at frozen parameters

Take a converged model, freeze every parameter, and sweep one group's reporting rate under each setting. Nothing re-estimates, so nothing can compensate.

HELD `task1.exclude-mixing-block-monotone-toward-one`

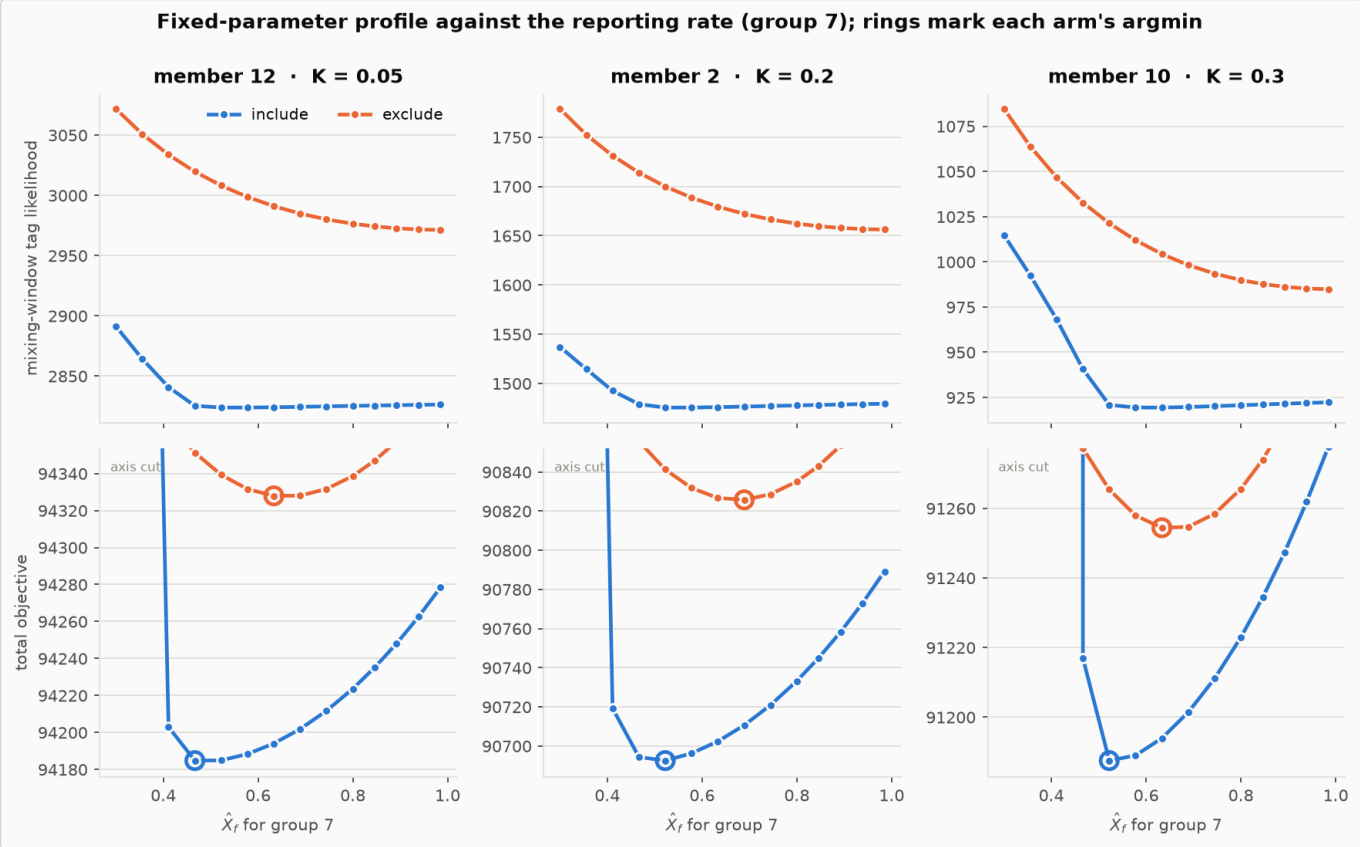
first differences of the exclude mixing block are negative in 100.0% of grid steps (worst profile 100.0%)

HELD `task1.include-mixing-block-invariant-to-X-above-0.8`

worst relative range of the include mixing block for X >= 0.8 = 0.0086; median 0.0009 -- this is the sub-range where the raised removal is small enough that the survival floor should not rescale the Newton-Raphson target

HELD `task1.penalty-is-identical-under-both-arms`

max |rr_penalty(exclude) - rr_penalty(include)| = 0 across the grid



Top: under **include** the mixing-window likelihood is flat in the reporting rate (0.09% median variation above $\hat{X} = 0.8$); under **exclude** it slopes toward $\hat{X} = 1$ in 100% of grid steps. Bottom: the best-fitting reporting rate sits **higher** under **exclude** in every member (+0.111 to +0.167 in $\hat{X}$ for group 7).

The pre-registered stop condition was that the **include** curve be flat everywhere. Over the whole sweep it is not, reaching up to 22.5%. The reason is the survival guard, and isolating it turns out to answer a question the US review left open (Part 3).

DID NOT HOLD `task1.include-mixing-block-invariant-to-X`

worst relative range of the include mixing block over the full grid = 0.2254 (tolerance 0.02); median 0.0701

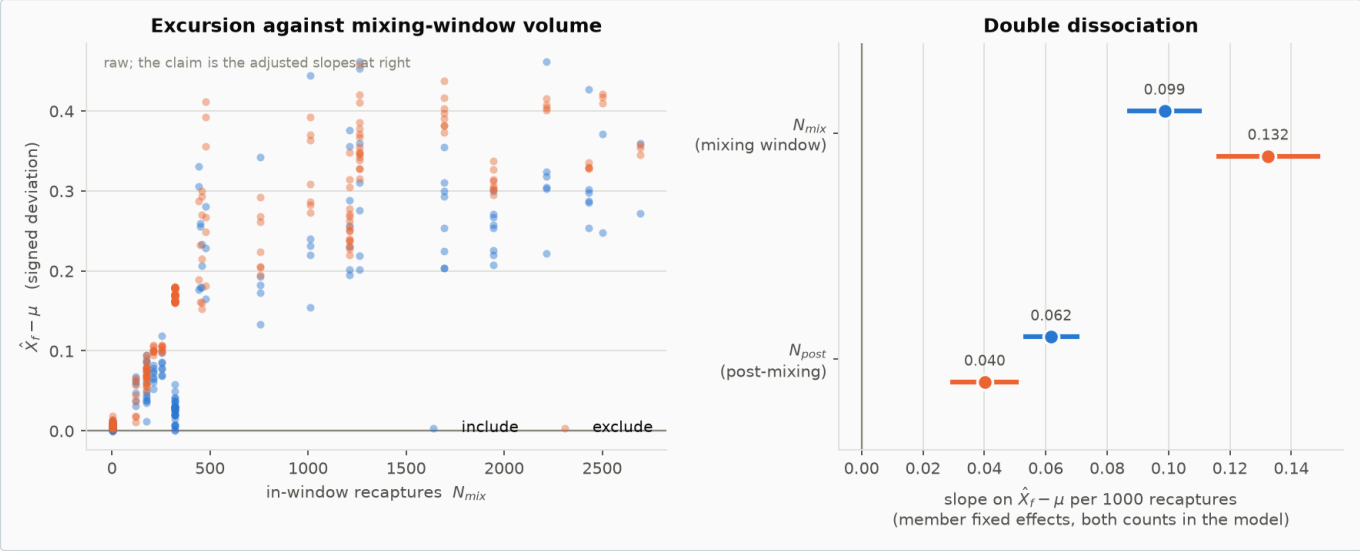
The argmin shift splits by prior weight, not by member: group 7 (weight 354.5) moves in all three members; group 17 (weight 739.2) is pinned by its heavier prior. That reproduces the group ordering seen in the ensemble (+0.142 vs +0.014) from an independent, fixed-parameter experiment.

HELD `task1.argmin-shift-larger-for-the-lower-weight-group`

median argmin shift: group 7 (w=354.5) +0.1667 vs group 17 (w=739.2) +0.0000 -- the same ordering as the observed ensemble arm differences (+0.142 vs +0.014)

What orders the pull across groups

It scales with the *count* of mixing-period recaptures resisted by prior weight, not the in-window *share*, which is why an earlier $\hat{\psi}$-based test found nothing. Raw `N_mix` predicts nearly as well in both arms, which alone approaches the falsification threshold; the discriminating test controls for post-mixing returns, which carry the same identification benefit and none of the pseudo-data pull.



Under **exclude** mixing-window returns pull *harder* (+34%) while post-mixing returns pull *less* (−35%), both intervals excluding zero. That opposite movement is the signature of re-weighting toward pseudo-data.

HELD `task3.Nmix-and-Npost-arm-effects-have-opposite-signs`

N_mix slope 9.868e-05 -> 1.325e-04 (+34%) while N_post slope 6.177e-05 -> 4.014e-05 (-35%); both interaction CIs exclude zero

PART 3 · **RESOLVING AN OPEN QUESTION**

Which MFCL routine the assessment actually used

The US review flagged an ambiguity it could not resolve from source:

> *"the source defines more than one version of the routine that performs this removal, and they differ both in whether they honour* `tag_flags(i,2)` *and in whether the cap is applied at all … we have not been able to confirm whether it is also the version carrying the cap."*

That is exactly right, and reading the current source does not settle it. If anything it makes the ambiguity sharper. At `multifan-cl` HEAD the routine actually called is `do_newton_raphson_for_tags2` , which divides by the reporting rate **unconditionally** and applies **no cap**. The variant that honors the flag and carries the cap sits at `tag3.cpp:2769` and is never called; a third copy is commented out. The assessment binary is version **2.2.7.9** and predates that state of the source, so the source cannot answer the question. The binary's own behavior can.

Decompose the objective along a fixed-parameter profile. Everything except the tag terms is frozen, so

```
residual = objective − ( mixing-window tag block + post-mixing tag block + reporting-rate penalty )
```

is constant unless some *other* reporting-rate-dependent term exists. It is:

ARM	MAX UNEXPLAINED TERM	AT THE FITTED RATE	PEAKS AT
exclude	7.7e-07	0	n/a (flat everywhere)
include	190,673	1.04	$\hat{X}$ = 0.30, the largest raising

HELD `task10.no-unexplained-objective-term-under-exclude`

exclude: the objective is fully accounted for by the tag blocks and the reporting-rate penalty -- max unexplained range 7.65e-07 objective units across all profiles

HELD `task10.binary-carries-the-cap-and-honours-the-flag`

resolves the US review's open question empirically: the 2.2.7.9 binary both honours tag_flags(i,2) (the arms differ) and applies the posfun cap (an unexplained penalty term appears under `include` at low X only)

The binary both honors the flag and applies the cap. The arms differ, so the flag is live; and an unexplained penalty appears under **include** only, growing as the raised removal grows, which no other term in the objective can produce. The residual above $\hat{X}$ = 0.8 is not quite zero below, which is why the cancellation is near-exact rather than exact in that sub-range.

> **DID NOT HOLD** `task10.unexplained-term-vanishes-above-X-0.8`
>
> include: unexplained term range above X = 0.8 is 2.7 -- the cap does not engage there, which is why the cancellation is exact in that sub-range

This also fixes the cost of the guard. The cap is what the manual's recommendation is designed to avoid, and at the reporting rates these models actually estimate it costs at most 1.04 objective units.

PART 4 · THE DEEPER PROBLEM

Both options bypass the likelihood

The US review identifies a problem that both settings share, and it subsumes the mechanism above. Mixing-period recaptures fix the tagged population **deterministically**, from the observed returns, with no uncertainty attached and no test against what the model predicts.

Our source trace shows exactly how. Likelihood terms for these recaptures *are* formed, but their predictions are constructed from the observations rather than from the model. Under **include** the prediction reproduces the observation, so the term is inert. Under **exclude** the prediction is a reporting-rate fraction of the observation, and the only way to close that gap is to conclude reporting is complete.

So a substantial share of the tagging data steers the assessment without ever being fitted. Because the removal is deterministic, it carries *more* authority than fitted data would: a fitted observation can be outweighed by other information, a deterministic constraint cannot.

The influence does not stay inside the mixing window

The effect is not contained to the mixing window, and this follows from the source rather than from inference. `tag_fish_rep`, the estimated reporting rate $\hat{X}$, is a single free parameter per (release group, fishery) cell, not one value per period. The likelihood-scoring loop in `src/tagfit.cpp` runs every period from release onward with no branch on it, `for (ip = initial_tag_period; ip <= ub; ip++)`, and applies the identical multiplier, `tag_rep_rate(it,ir,ip,fi) * tagcatch(it,ir,ip,fi)`, at each one, mixing and post-mixing alike; `tag_rep_rate` resolves to `tag_fish_rep(it, pp)`, a function of release group and fishery only, never of the period `ip`. So whatever pull the mixing-period pseudo-data puts on $\hat{X}$ under **exclude** is carried, completely unchanged, into the predictions the model compares against the genuinely-fitted post-mixing recaptures for that same cell. The parameter shaped by in-window pseudo-data is the same parameter used to score genuine data outside the window.

This is exactly what Part 2's `N_mix` / `N_post` double dissociation shows empirically: the same $\hat{X}$ responds to both mixing-period and post-mixing recapture counts at once, with the mixing-period component pulling harder under **exclude** and the post-mixing component pulling less. It is one parameter informed by both in-window pseudo-data and genuine post-mixing returns, with the in-window component taking the larger share under **exclude**. If the effect were confined to a few free parameters it would already matter; those parameters then being used, unmodified, to score genuine data extends its reach.

This was a reasonable response to a real difficulty. Tagged fish are not mixed, so predicting their recaptures from population-average availability is knowingly wrong, and declining to fit them is defensible. But unlike a data series that can simply be dropped, these fish have left the tagged population and still have to be removed from it. Desensitizing the likelihood did not desensitize the model.

The one-sidedness argument, stated carefully

This argument needs no part of the channel-2 account, but it is worth separating what it does and does not establish, because the two are easy to conflate.

What is not in dispute. Assuming complete reporting cannot remove more tags than correcting for reporting does. If estimated F rises with the ratio of recaptured to total tags, as it should, then **exclude** yields an F at or below **include** for the same cell. The two options are therefore ordered, and running both does report an interval in F. Describing that interval as a bound on the *adjustment*, between a raising factor of 1.00 and one of $1/\hat{X}$, is a fair description of what the grid does.

What that interval does not establish. Ordered is not the same as containing the answer, and two things stand between them.

The first is the status of the endpoints. Take the conditions under which **include** is unbiased as given: the estimated rate is about right, release groups are large enough to be informative, and the survival floor does not engage. Under those conditions **include** is centered on the answer and **exclude** sits below it by construction. Averaging across the axis then combines an unbiased estimate with a biased one, and the product is a displaced central estimate rather than a wider valid interval. This, rather than any claim about bracketing, is the direct basis for preferring to retain **include** alone.

The second is that both endpoints are computed from the *estimated* reporting rate, and the true rate is unknown. It may be above or below $\hat{X}$. What it is almost certainly not is 1.00, which is the value **exclude** assumes, and that alone makes **exclude** a low-probability candidate for a grid axis. But the more consequential case is the other one. **If the true rate is *lower* than the fitted rate, the correct raising factor $1/X_{true}$ exceeds the applied factor $1/\hat{X}$, so both settings remove too few tags and both understate F.** The interval they span then sits below the answer entirely, and no weighting across the axis recovers it.

That is not a hypothetical here. The fitted rates in this assessment exceed the independent tag-seeding priors in **5 of 5** priored groups, in both arms (Part 6). To the extent the seeding trials are the better estimate of the true rate, this assessment sits in exactly the case where the axis is not merely one-sided but displaced as a whole, and the direction of the displacement is the optimistic one.

The broader point is that the reporting-rate flag is a narrow proxy for a wider question. The uncertainty that matters is how much the tagging data should influence the assessment at all, and an axis expressing that directly would be more informative than one choosing between two removal conventions. Building it is not feasible now (Part 8), which is a reason to state the present axis's limits plainly, not a reason to treat it as if it spanned the uncertainty it is standing in for.

What it does to the advice

Stated on the 80 models meeting the assessment's convergence criterion (MGC ≤ 1e-4). That screen selects exactly the 80 members carrying a retained parameter file, so no conclusion changes coverage by switching between the two criteria.

HELD `task11.convergence-screen-selects-exactly-the-retained-par-set`

MGC <= 0.0001 selects 80 members and the retained-par set is 80; the two sets are identical, so no conclusion changes coverage by switching criterion

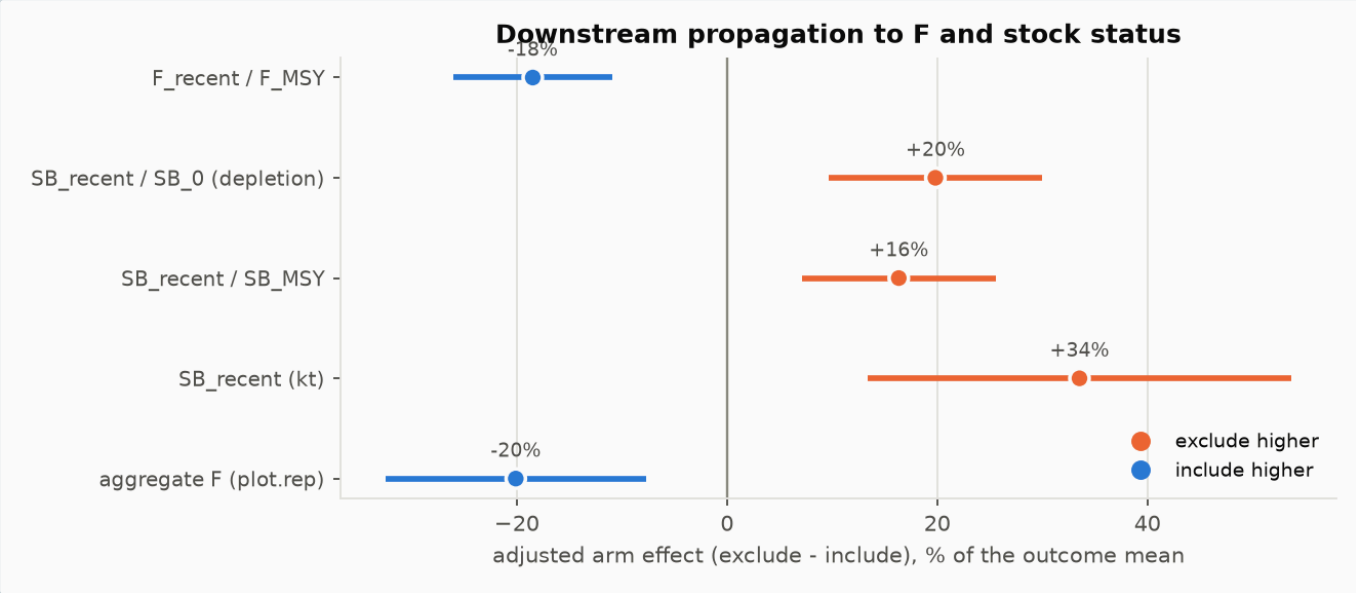
Adjusting for every other grid axis, **exclude** carries **-0.132** [-0.198, -0.066] on F/F_{MSY}, or -15% of the ensemble mean, and **+0.0466** [0.0180, 0.0752] on depletion, +17%. Both are smaller than the same fits on all 88 models that ran (−0.166 and +0.053), because the convergence screen removes **include** models disproportionately.

HELD `task5.arm-effect-on-F-is-negative[f_recent_fmsy]`

F/F_MSY: -0.1664 [-0.2347, -0.09819], p=5.83e-06 (mean 0.9007)

HELD `task5.arm-effect-on-depletion-is-positive`

SB/SB_0: +0.05338 [0.02608, 0.08067], p=0.000204 (mean depletion 0.2692)



Every management quantity moves as the mechanism predicts and no interval crosses zero. This is the size of the effect, not evidence for which channel caused it, since both channels push F the same way and compound.

An important negative result

We tried to decompose that effect cheaply, by measuring how much F moves per unit reporting rate with parameters frozen. It moves **not at all**: across 168 evaluations the objective took 56 distinct values per model while F/F_{MSY} took **1**, a range of exactly 0.

DID NOT HOLD `task9.fixed-parameter-profile-can-decompose-the-downstream-effect`

it cannot: dF/dX and the removal channel are both structurally zero at frozen parameters, so the observed -18% arm effect on F/F_MSY is an estimation effect and only re-estimation can decompose it

The reason is structural: the reporting rate enters only the tag likelihood, its own penalty and the tagged-cohort solver, and the tagged cohort is fitted *to* the population without feeding back into it. So the downstream effect is **optimizer behavior**: the reporting rate reshapes the likelihood surface and the F-related parameters relocate during estimation. That rules out a class of explanation, and it removes the cheap instrument, since only re-estimation can measure the pathway.

Fragility of the tagging module, and what it does to the comparison

The losses are heavily arm-specific, and they compound across two stages:

STAGE	INCLUDE	EXCLUDE
Drawn	50	50
Failed to build / run	9	3
Ran but failed MGC ≤ 1e-4	7 (17%)	1 (2%)
Surviving	**34**	**46**
90th-percentile max gradient	7.7e-04	9.9e-05

HELD `task11.tagging-module-is-arm-specifically-fragile`

cumulative loss from the designed 50 draws: include 16/50 (32%) vs exclude 4/50 (8%) -- a factor of 4.0

HELD `task11.convergence-failure-alone-is-arm-specific`

of models that ran: include 7/41 (17.1%) failed MGC vs exclude 1/47 (2.1%)

The convergence screen therefore *degrades* the randomization the design bought. The **include** share falls to 42%, and the standardized arm differences on the other axes widen: M0 to -0.211, effort creep to +0.184, and K to +0.111.

HELD `task11.convergence-screen-worsens-arm-balance`

include share falls from 46.6% to 42.5%; |SMD| on K goes 0.044 -> 0.111

Does the imbalance understate the difference?

Yes for the raw comparison, and no for the adjusted one. The distinction matters.

The **raw** arm difference in F/F_{MSY} falls from -0.1415 across all models that ran to **-0.0858** on the converged set, a 39% shrinkage. Imputing the lost members at their own design points and completing both arms to the designed 50/50 restores it to -0.1264. The lost **include** models sit where F is high, so removing them pulled the **include** mean down and closed the gap.

HELD `task11.raw-gap-is-materially-truncated-by-the-convergence-screen`

raw F/F_MSY gap falls from -0.1415 (all that ran) to -0.0858 (converged only), a 39% shrinkage; completing the arms restores it to -0.1264

The **adjusted** estimate barely moves. Three routes to a balanced representation (reweighting to the designed arm × K marginal, completing both arms by imputation, and a genuinely balanced **subset of models that actually ran**, below) give F/F_{MSY} arm effects spanning -0.142 to -0.135, against -0.132 unbalanced.

HELD `task11.adjusted-gap-is-robust-where-the-raw-gap-is-not`

adjusted estimate moves +2.4% on completion while the raw difference moves +47% -- the adjustment already conditions on the axes that drive the differential loss, so the adjusted effect is the one to quote

That is the expected pattern rather than a coincidence: the adjustment already conditions on the axes along which the losses are concentrated, so the composition shift that moves the raw difference is precisely what the adjusted model absorbs. **The practical consequences are that raw arm-versus-arm comparisons on the retained ensemble understate the difference by roughly 40%, and that the adjusted estimate is the one to quote.** A more balanced ensemble would change the headline little; it would change any unadjusted presentation a great deal.

A balanced subset of real runs, not interpolation

The routes above lean on a fitted model to stand in for the composition the design intended. It is worth also asking the question with no model at all: is there a subset of the 80 converged runs, *as fitted*, that is balanced on its own? Optimal 1:1 matching, pairing each **include** model to the **exclude** model closest to it on the five continuous axes (K, τ, M0, effort creep, h) by minimum total distance, with a caliper to drop poor matches, finds **27 pairs** (27×2 of the 80 converged models) with every axis balanced to |SMD| < 0.15.

HELD `task11.matched-pairs-achieve-good-covariate-balance`

worst |SMD| across (K, tau, M0, creep, h) after optimal 1:1 matching is 0.150 on 27 pairs (caliper 1.627 std units, 7 candidate pairs dropped as poorly matched)

On those 27 pairs, the **raw, unadjusted, unmodelled** paired difference is F/F_{MSY} -0.162 [-0.234, -0.090] and depletion 0.0560 [0.0298, 0.0821], matching or exceeding the adjusted whole-sample estimate, using no regression at all. This is the cleanest evidence that the imbalance was suppressing the raw comparison rather than the adjustment manufacturing an effect that is not really there.

HELD `task11.raw-paired-difference-corroborates-the-adjusted-estimate`

on 27 real, covariate-matched pairs the RAW (unadjusted, no model) paired difference is -0.1621 [-0.2342, -0.0899], at least as large as the adjusted whole-sample estimate (-0.1315) despite using no regression at all

The reweighting and imputation routes rest on an assumption worth stating: that within a design cell the retained members stand in for the lost ones. Differential convergence is exactly what makes that doubtful, so those two are illustrative bounds, not corrections. The imputation in particular extrapolates to the design points where estimation failed, where a smooth model of the outcome is least trustworthy. The matched-pairs result does not carry that assumption: it uses only models that were actually fitted.

What these numbers do not include

Every risk and arm-effect figure in this document (the adjusted estimates, the matched pairs, the $P(F>F_{MSY})$ and $P(SB<LRP)$ figures above) is built from each model's single best-fitting parameter vector, one point per grid cell. None of it propagates **estimation uncertainty**, the spread of parameter values consistent with the data within a single fit. What is reported here is the **structural** component only: how the point estimate moves across the grid. That is a real, distinct restriction, not a formality. The repository's own combined structural-and-estimation audit (`results/tables/estimation-uncertainty-audit.csv` , pooling 100 joint-Hessian draws per model with equal weight across all 80 retained models) puts the 10th to 90th percentile of SB_{recent}/SB_{MSY} at roughly 0.82 to 2.57 around a median of 1.53, a band wide enough that estimation uncertainty is not a minor addition on top of the structural spread reported here. That table is pooled across **both arms together**, so it confirms the scale of what is missing without being able to say how the two arms' risk figures would move relative to each other.

Answering that would need the per-model draws split and recombined by arm, and the repository only partially supports doing so. Per-model joint-Hessian draws exist for **62 of the 80 converged models** (25 of 34 **include**, 37 of 46 **exclude**; the 18 gaps are exactly the models without a positive-definite Hessian, split evenly across arms, so not themselves an added source of imbalance). Even with full coverage, combining a hundred joint draws per model into a single arm-conditional risk estimate is a real undertaking, not a quick addition, and the existing pipeline for it (`report/augment-uncertainty-projection.R`) requires an R toolchain with several packages that this analysis environment does not have. It was not attempted here. **The structural-only figures above should be labeled as such wherever they are used, and should not be read as the full assessment uncertainty on the arm comparison.**

PART 6 · THE CHOICE

Both options have limitations. The limitations are not symmetric.

EXCLUDE · no reporting-rate correction

- **Assumes complete reporting.** The model's own fitted rates contradict this for most of the tagging data.
- **Pseudo-data on 96 of 98 release groups**, in every model. The in-window terms are minimized at $\hat{X} = 1$ regardless of the data.
- **Known directional bias.** It can never remove more tags than the corrected option, so wherever reporting is incomplete it leaves too many tags in the population, giving lower F and larger population scale.
- **Estimated rates pushed up**, systematically closer to the 0.99 bound.
- **Diagnostics disagree with the fit** by ~21.6% in the mixing window.
- **Numerically well-behaved**, with 6% of draws failed and no cap engagement.
- **What the MFCL manual recommends.**

INCLUDE · correct returns by $\hat{X}$

- **Capable of being unbiased**, under three conditions: the reporting rate is about right, release groups are large enough, and the raised removal leaves enough tags.
- **No pseudo-data in the general case**, since $\hat{X}$ cancels out of the in-window terms.
- **The third condition is the fragile one.** The cap engages on 2 of 98 release groups, and where it does the same $\hat{X}\cdot R$ form partially returns. Nothing in standard output reports it.
- **Small release groups are a real risk.** The US simulation study shows a heavy upper tail in F rather than an obvious loss of precision.
- **The cap penalty is discarded outright** for release groups 21, 72 and 112 (hardcoded), and 21 is one that engages.
- **Numerically fragile**, with 32% of the 50 draws lost against 8% for **exclude**, counting build failures and convergence failures together. Informative non-response, not random loss.

Both settings estimate reporting rates above the external evidence

The two options are reasonably described as bracketing the reporting-rate adjustment: not correcting is equivalent to assuming RR = 1, which applies the smallest possible raising factor of 1.00, while correcting applies $1/\hat{X} > 1$. As a statement about the two options relative to each other, that is correct. It becomes a statement about the *truth* only if $\hat{X}$ is exogenous, and it is not. $\hat{X}$ is estimated inside the same likelihood, and in this assessment it lands well above the tag-seeding evidence in **both** arms, which places the true adjustment outside the range the two options span rather than within it.

PRIORED GROUP	SEEDING-TRIAL RR	FITTED, INCLUDE	FITTED, EXCLUDE	RAISING APPLIED, WIDEST OPTION	RAISING IMPLIED BY SEEDING
7 · RTTP PS.2	0.496	0.524	0.666	1.91	**2.02**
10 · RTTP PS.WEST.3	0.512	0.579	0.588	1.73	**1.95**
14 · PTTP PS.2	0.496	0.502	0.502	1.99	**2.02**
17 · PTTP PS.WEST.3	0.512	0.768	0.783	1.30	**1.95**
18 · PTTP PS.EAST.3	0.528	0.815	0.900	1.23	**1.89**

In **5 of 5** priored groups the raising implied by the seeding trials exceeds what either option applies. For groups 17 and 18, the two that dominate the penalty, the widest option raises by 1.30 and 1.23 against seeding-implied factors of 1.95 and 1.89.

The pattern is diagnostic rather than incidental. Group 14 is the one case where the applied raising nearly reaches the seeding value (1.99 against 2.02), and group 14 has **six** in-window recaptures, too few to move its rate off the prior. Where the tag data is abundant enough to move $\hat{X}$, $\hat{X}$ rises and the gap to the external evidence opens. So the span between the two options is a span in the *fitted* adjustment, and the fitted adjustment is itself displaced upward relative to the independent seeding estimates.

This is the concrete form of the general point in Part 4. A reporting rate estimated *above* the true rate means a raising factor applied *below* the correct one, so too few tags are removed and F is understated. That applies to both settings, since both use $\hat{X}$. If the seeding trials are the better guide to the true rate, the interval this axis reports does not straddle the answer in this assessment; it sits below it, and the more conservative of the two options is the one further away.

Removing the feedback into the tagged population does not remove the influence on the fit

A natural defense of the uncorrected setting is that it avoids circularity: with no correction, none of the parameters estimated from the tag likelihood influence the tagged population abundance at the end of the mixing period, so that abundance, which drives all post-mixing predictions, is not affected by the reporting rate. That is true as far as it goes, and it is the design intent.

But it tracks influence in one direction only. The mixing-period *likelihood terms still exist* under both settings, and under the uncorrected setting they contain $\hat{X}$ while their predictions are built from the observed returns. So the mixing-period data continues to influence $\hat{\mathbf{X}}$ even though $\hat{X}$ no longer influences the tagged population. And $\hat{X}$ is not confined to the mixing period: it scales every predicted recapture *outside* the window as well. The path runs mixing-period returns → $\hat{X}$ → post-mixing predictions → the tag likelihood, and it is not closed by breaking the feedback into the cohort.

The circularity is therefore cut in one direction and left open in the other, and the direction left open is the one that carries the bias, because those in-window terms are minimized at $\hat{X} = 1$ regardless of what the data say.

The recommendation

If advice must come from models fitted to tagging data, it should come from include models only, and the reporting-rate axis should be removed from the uncertainty grid rather than rebalanced.

Three independent reasons, none of which requires accepting the whole channel-2 account:

- **Averaging over the axis displaces the estimate.** The two options are ordered rather than straddling: **exclude** yields an F at or below **include** by construction. Where **include**'s conditions hold, carrying both combines an unbiased estimate with a biased one and shifts the central estimate rather than widening it validly. Separately, both endpoints are computed from the estimated rate, so if the true rate is below it both understate F and the interval sits below the answer, which is the case the seeding evidence here points to.
- **Asymmetric capability.** **Include** can be unbiased under stated conditions. **Exclude** requires mixing-period reporting to be essentially complete, a rate of 1.00, which is almost certainly not the case, so its probability of being the correct setting is low.
- **Asymmetric prevalence.** Established by source transcription and a fixed-parameter experiment: pervasive under **exclude** (96 of 98 groups), localized under **include** (2 of 98).

Weighing the manual's recommendation

The manual prefers **exclude** because the correction can be numerically unstable. That concern is real. Both sides of the trade can now be measured on this assessment:

	RELEASE GROUPS AFFECTED	COST AT FITTED RATES
The cap (**include**'s problem)	2 of 98	≤ 1.04 objective units
Pseudo-data (**exclude**'s problem)	96 of 98	shifts F/F_{MSY} by -15%

About **48 times** as many release groups are affected by the issue the manual's preference introduces as by the one it avoids. The guard fires only where the raised removal is extreme, at reporting rates far below any the models estimate, and where it does fire it costs a fraction of an objective unit. **The manual's guidance is general, and the balance may fall differently in other applications; in this assessment the numerical concern is the smaller of the two costs.**

Two qualifications

Include's higher failure rate is *informative* non-response: the models that failed are those where the correction was least feasible, so the surviving set is conditioned on the correction being benign. And neither arm's rates should be treated as truth. The external anchor is the tag-seeding priors, not either fit.

> An **include**-only ensemble already exists and needs no new runs: the 34 retained members cover every level of every other axis. Acting on this needs a decision, not compute.

PART 7 · SUPPORTING FINDINGS

The rest of the evidence

The diagnostic inconsistency

The tag-fit panels circulated with assessments cannot detect misfit inside the mixing window. The report writers apply a different multiplier from the likelihood: in-window they use 1.0 under **exclude** while the objective uses $\hat{X}$. Since the solver drives the in-window predicted catch to the observed returns, the plotted fit is an **identity in both arms**. This is verified against the circulated `plot.rep`, where time-at-liberty bin 1 gives pred/obs of 1.0000 to 1.0207 across members of both arms, against 0.80 to 0.83 in the first fully post-mixing bin.

> **HELD** `task2.report-likelihood-discrepancy-still-holds`
>
> exclude arm: the objective's in-window predicted returns sit 21.6% below the values written to plot.rep (median 802 of 3605 fish)

Under **exclude** the objective's in-window predictions sit **21.6%** below the plotted ones, or 802 of 3605 fish per model. This is an inconsistency in MFCL's reporting worth raising upstream regardless of which setting is adopted.

The reporting-rate bound

Bound contact (at 0.99) occurs in two **include** members and none under **exclude**, which looks backwards. Re-testing on the estimation scale (an arcsine-of-log transform, so equal proportion gaps are not equal distances) does not reverse the ordering, but the distribution resolves it: **exclude** sits systematically closer to the bound in median (transformed gap 308.6 vs 359.6), and the two **include** contacts exceed the entire **exclude** distribution. Contact counts were the wrong statistic.

> **HELD** `task4.exclude-closer-to-bound-across-all-priored-groups`
>
> median transformed gap over all priored groups: include 359.6 vs exclude 308.6

> **DID NOT HOLD** `task4.bound-contact-test-in-transformed-space`
>
> proportion scale: include 5.9% vs exclude 0.0%; estimation scale: include 5.9% vs exclude 0.0% -- ordering does not reverse

The survival floor across the ensemble

A `bet.tag`-based screen (deliberately conservative) flags floor engagement in 100% of **include** members and none under **exclude**, consistent with the objective decomposition in Part 3. Engagement concentrates in the pooled release group (98) and group 21, whose penalty MFCL discards.

> **HELD** `task8.floor-engagement-higher-under-include`
>
> share of members with >=1 engaging release group: include 100.0% vs exclude 0.0%

> **HELD** `task8.lowest-penalty-include-members-are-more-floor-engaging`
>
> lowest-penalty tercile 2.00 vs highest 1.36 engaging release groups; Spearman(penalty, n_engaging) = -0.588

$\hat{\psi}$ under three anchors

$\hat{\psi}$ inherits the rates it is raised by, so the published diagnostic, computed with **exclude**-arm rates, is itself subject to channel 2. Recomputing under the externally anchored tag-seeding priors raises release-weighted $\hat{\psi}$ from 0.409 to 0.414 (+1.1%), consistently in 80 of 80 members. **But the correction is second-order:** the mixing configuration moves $\hat{\psi}$ by 0.563 across the K axis, about 120 times more than the raising scheme does.

> **HELD** `task6.psi-higher-under-external-anchor-than-as-published`
>
> release-weighted psi-hat: as published (exclude rates) 0.4091 -> external anchor (seeding priors) 0.4137 (+1.1%)

> **HELD** `task6.mixing-configuration-dominates-the-raising-scheme`
>
> psi-hat moves 0.5635 across the K axis (from 0.7860 at K=0.05 to 0.2225 at K=0.35) versus 0.0047 between raising schemes -- a factor of 120

The published baseline does not reconcile: the closest configuration (K = 0.30) gives $\hat{\psi} = 0.226$ against the published 0.231, but 35 groups at $\hat{\psi} \geq 0.3$ against 27. These tables are therefore **not** drop-in replacements; the scheme contrast is internally consistent, the level is unreconciled, and that gap should close before anything circulates.

DID NOT HOLD `task6.published-baseline-reconciles`

closest configuration is K=0.3: psi-hat 0.2259 vs published 0.231 (close), but 35 groups at psi>=0.3 vs published 27 (does not reconcile)

PART 8 · STATUS AND NEXT STEPS

What is established, and what is not

Established by transcription or controlled experiment

- The penalty's exact functional form, reproduced to 4.9×10^{-10}.
- That **exclude**'s mixing-period terms are minimized at $\hat{X} = 1$, and **include**'s are inert in $\hat{X}$ except where the cap engages.
- That flipping the flag alone, at frozen parameters, moves the best-fitting reporting rate up.
- That the assessment binary both honors `tag_flags(i,2)` and applies the cap, resolving an ambiguity the source at HEAD cannot.
- That the reporting rate cannot move F at frozen parameters; the downstream effect is optimizer behavior.
- That the in-window tag-fit diagnostic is an identity in both arms.

Established observationally, on a selected non-crossed ensemble

- Adjusted axis effects on the 80 converged models: +34.3 [27.0, 41.7] on the penalty, -0.132 on F/F_{MSY}, +0.0466 on depletion.
- The `N_mix` / `N_post` double dissociation.
- That **exclude** sits systematically closer to the reporting-rate bound.
- That the **include** setting is arm-specifically fragile, with 32% of drawn models lost to build or convergence failure against 8% for **exclude**, and that on 27 covariate-matched pairs of real, converged runs the raw F/F_{MSY} difference (-0.162) matches the adjusted whole-sample estimate without any model behind it.

Not established

- **That the downstream effect is caused by the reporting rate.** Both channels push F the same way and compound; conditioning on $\hat{X}$ over-mediates because $\hat{X}$ and F are jointly estimated. Only paired refits can settle it.
- **That either arm's rates are correct.**
- **A reconciled ψ correction.**
- **The magnitude of the small-release-group risk** in this assessment specifically. The US simulation study indicates a heavy upper tail in F, but we have not quantified it here.
- **How the risk figures change once estimation uncertainty is included.** Every $P(F>F_{MSY})$ and P(SB<LRP) figure here reflects structural uncertainty only, meaning the spread of single best-fitting parameter vectors across the grid, with none of the within-model parameter uncertainty folded in. The ensemble's own audit shows that uncertainty is large in scale; it is pooled across both arms in the repository and was not decomposed by arm here (Part 5).

Recommended sequence

1. **Report the diagnostic inconsistency upstream.** No compute, and independent of the include/exclude decision.
2. **Take the axis out of the grid.** Present **include**-only, or present the arms separately with their ordering stated and with the caveat that both are computed from the fitted reporting rate. A 20%-versus-29% split in limit-exceedance probability follows from a modeling choice rather than from uncertainty about the stock, and is better presented as such.
3. **Reconcile the ψ baseline** before any corrected figure circulates.
4. **Then paired refits**, the only instrument that can measure the downstream pathway now that the fixed-parameter route is closed. Time one full-path run before committing to a count, and consider targeting the 9 failed **include** members, since recovering them removes the selection concern on an **include**-only ensemble.

The durable fixes

Both come from the US review and neither is a grid adjustment.

1. **Modify MFCL** so the mixing-period likelihood terms that already exist score predictions generated from the projected tagged cohort, with mixing-period availability and the reporting rate estimated rather than assumed. This follows the original design reasoning instead of discarding it: the obstacle was that recaptures cannot be predicted from an unmixed cohort, so estimate the departure from mixing. **This adds parameters to a model that already has an identifiability problem, not a clean one.** The assessment's own diagnostics (jittering, simulation self-test, parameter correlations, bounds checks) already identified, as the authors noted and the US review restates, that several parameters of the 2026 assessment's spatial structure remain “poorly determined and weakly identifiable” before any availability parameters are added. Layering one more per-release-group parameter onto that model, with tag release groups that are mostly small and noisy, could compound the existing estimability problem rather than trade it for a cleaner one. The US review says as much of its own proposal: “it remains to be seen if the additional parameters would be identifiable.” There is also a resolution requirement independent of that: an availability multiplier too coarse to span the true departure from mixing inflates mixing-period mortality and biases F upward instead of removing the bias it targets. **This is a research direction pending its own simulation testing, not a fix ready to adopt, and that testing should establish identifiability before implementation is considered, not after.**
2. **Include spatially implicit, fleets-as-areas models** in the grid, which sidesteps the tagging treatment entirely and integrates over spatial-structure uncertainty as well.

Given both of those carry open questions, one on identifiability and the other a simplification of the population dynamics, a third and more pragmatic option follows directly from everything above: **stop fitting tag recoveries inside MFCL's joint likelihood and treat them as an external analysis instead**, using an independent tag-based estimate of F or natural mortality as a cross-check or to inform a prior, rather than a term optimized jointly with the rest of the model. This sidesteps the channel-1/channel-2 pathology in Part 2 and the added-complexity risk above without waiting on either fix, at the cost of losing the tagging data's direct influence on the fit. It is not a new recommendation from that review. It generalizes an aside in its second path, that a model not fitting tagging data directly can still use it "to inform prior distributions for mortality parameters", extending that to the current spatially-explicit structure rather than tying it to a change in spatial resolution.

The axis the grid would ideally carry

Underlying all of this is a question the reporting-rate flag is a narrow proxy for. The substantive uncertainty is not which of two removal conventions MFCL applies during mixing; it is **how much the tagging data should influence the assessment at all**, given that a large share of it enters deterministically rather than through the likelihood. An axis contrasting models that fit tagging data with models that do not would express that directly, and would integrate over a real and consequential uncertainty rather than over a pair of conventions that are ordered by construction.

Constructing that axis is not feasible for this assessment cycle: it requires the non-tagged configuration to exist, be tuned, and converge across the rest of the grid, which is a body of work rather than a re-weighting. That is a reason to state plainly what the present axis does and does not span, and to prefer the option capable of being unbiased in the meantime. It is not a reason to treat the current axis as though it covered the uncertainty it stands in for.

Reproducing this analysis. Analysis code lives in `analysis/rr-penalty/`, `analysis/channel2/` and `analysis/summary/` of N-DucharmeBarth/ofp-sam-bet-2026-ensemble, reproduced by the `run-all.sh` in each. Tidy CSVs, figures and the full check ledgers are written to `out/` and `out/channel2/`. This document is generated from those CSVs by `analysis/summary/build-summary.py`, so every number in it is substituted from the analysis output rather than typed in.

Sources. Source excerpts are pinned to `multifan-cl` commit `624dee4f`; the assessment binary is version 2.2.7.9. The two differ, which is what Part 3 addresses. Qualitative framing, the one-sidedness argument, the three conditions for the corrected option and the two paths forward are drawn from N-DucharmeBarth-NOAA/sc22-tag-analysis (commit `907274f`), referred to throughout as the US review. **Note on attribution:** despite the repository name, which refers to the meeting the work was prepared for, that analysis is the work of United States stock assessment scientists and is not an output of SC22 or of the Scientific Committee. Nothing in this document should be read as an SC22 finding or position. The only genuine SC22 citation here is the tag-seeding information paper SC22-SA-IP-05, which supplies the reporting-rate priors. The analysis, checks and recommendations in this document are our own.

Coverage. Objective components and the design table cover all 88 retained models; 80 have a retained `final.par` and meet the MGC ≤ 1e-4 convergence criterion, and all downstream conclusions are stated on those 80. No estimation run was launched for either investigation: every MFCL invocation set `parest_flags(1) = 0`.