

January 30, 2026

Response to Reviewer

Manuscript: "Forking-Peeks: When Optional Stopping Meets Analytic Multiplicity"

Dear Editor and Reviewer,

We thank the reviewer for the supportive assessment and helpful suggestions. We have revised the manuscript to clarify interpretation and strengthen the practical recommendations. Below we provide a point-by-point response. For ease of reading, reviewer comments are reproduced in italics, followed by our responses.

1) "p-zone" strategy and interpretation of expected sample size $E[n]$.

Response: We agree. Our p-zone stopping rule explicitly includes a futility component (stop when $p \geq \tau$), which can reduce $E[n]$ by terminating unpromising studies early. To avoid the impression that p-zone stopping is generally 'efficient' or that it represents typical p-hacking behavior, we added text in the Discussion clarifying that many real-world questionable research practices may (a) continue sampling despite $p \geq \tau$, (b) expand the analytic search space (additional outcomes, covariates, transformations), or (c) pursue additional studies, which can increase sample usage while still inflating false positives. We emphasize that our modeled p-zone is a descriptive heuristic and that other, less 'efficient' (but common) practices could yield larger $E[n]$ alongside elevated Type I error.

2) Conservatism of the proposed correction $\alpha/(m \cdot L)$.

Response: We agree that $\alpha/(m \cdot L)$ is conservative because it ignores the strong dependence across interim looks and the correlation structure among analytic forks. We revised the Practical recommendations section to frame $\alpha/(m \cdot L)$ as a simple, hand-calculable worst-case bound (maximally safe), not a universally required adjustment. We also strengthened the discussion of sequential-design alternatives that spend α over looks (e.g., Pocock-type or O'Brien–Fleming-type boundaries and general α -spending

functions) as well as structured multiplicity procedures (e.g., Holm and FDR control) and simulation-based calibration when the dependency structure is known or can be approximated.

3) Realism and educational interpretation of the "Lock-in" strategy.

Response: We appreciate this suggestion. We expanded the Discussion to explicitly stress that Lock-in can appear 'consistent' because the same analysis is used at later looks, yet the initial choice of the analysis fork is data-dependent. As a result, Lock-in still entails multiplicity and selection bias. We highlight this point as an educational takeaway: consistency over time does not eliminate multiplicity if the initial selection was driven by the observed data.

Minor: Figure 1 visibility of small bars (Fixed-n vs p-zone).

Response: We added a clarifying note to the Figure 1 caption explaining that the y-axis range is set to accommodate large inflations under Forking-Peeks, which visually compresses differences near 0.05. We direct readers to Table 2 for precise values.

Minor: Terminology ("Forking-Peeks" as a coined term).

Response: We now explicitly state at first use (Abstract and Introduction) that 'Forking-Peeks' is a term we coin in this paper to describe the joint use of repeated looks and analytic multiplicity.

Minor: Table 2 $E[n]$ under Fixed-n equals 100.0 by design.

Response: We added a note to the Table 2 caption clarifying that Fixed-n involves no early stopping and therefore always reaches $n_{\max} = 100$, yielding $E[n] = 100$ by construction.

We hope these revisions address the reviewer's suggestions and further improve the paper's clarity and practical value.

Sincerely,

[Corresponding author name] on behalf of all authors