2026 ■ Vol. 22 ■ no. 2

Forking-Peeks: When Optional Stopping Meets Analytic Multiplicity

Yuusuke Harada [a]

[a]Hiroshima University

Abstract ■ Researchers sometimes make two flexible choices during a study: they look at results before data collection is finished, and they try several defensible analyses before deciding what to report. We call the joint use of these practices Forking-Peeks. Using computer simulations of simple two-group studies, we show that either practice alone can increase false-positive findings, but their combination can do so dramatically. In one illustrative setting with 16 correlated analysis options and four interim looks at the accumulating data, a nominal 5% false-positive rate rose to about 70% when the analyst repeatedly selected the best-looking analysis. Even a seemingly more consistent strategy, choosing the best-looking analysis once and then keeping it fixed, still produced severe inflation because the initial choice was data-dependent. The same selection process also made statistically significant effects look larger than they really were. The manuscript provides reproducible code and visual risk maps to help readers understand how small, plausible research decisions can compound when data collection and analysis remain flexible. We conclude with practical recommendations: plan stopping rules in advance, use valid sequential designs when interim looks are needed, treat multiple analyses as multiplicity, and report effect sizes and intervals rather than relying on a single threshold.

Keywords ■ optional stopping; multiple comparisons; p-hacking; Type I error; Monte Carlo simulation.

✉ yobou.hokkaido.harada@gmail.com

10.20982/tqmp.22.2.p046

Acting Editor ■ Denis Cousineau (Université d'Ottawa)

Reviewers
■ Four anonymous reviewers

Introduction

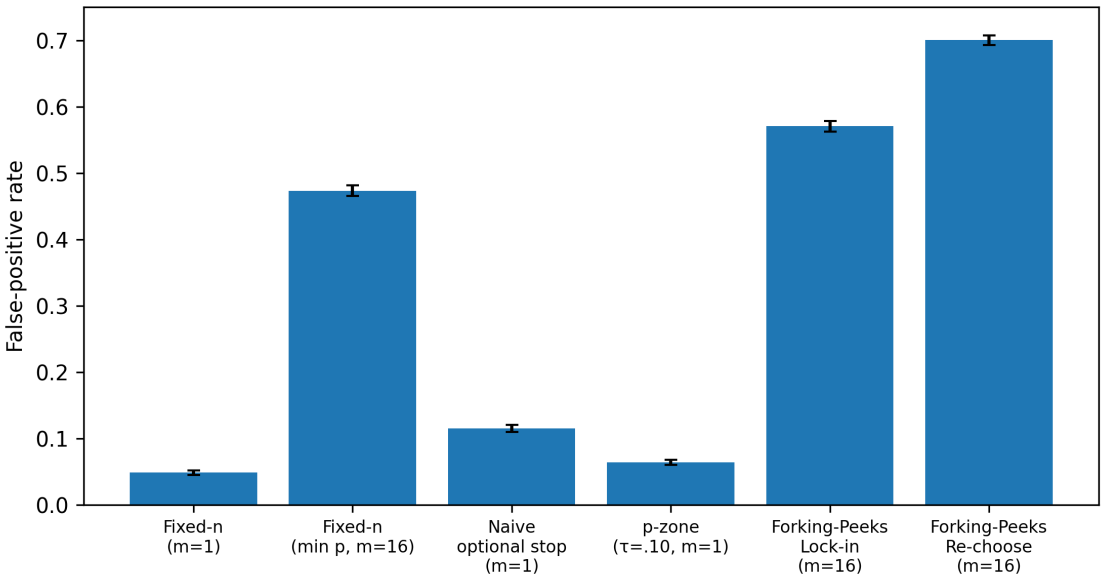
Null-hypothesis significance testing and the p-value remain common tools for making evidential claims in psychological and behavioral science. These tools can work as intended when the sample size, outcome, exclusion rules, and analysis plan are fixed before the data are examined. Problems arise when those choices are adjusted after seeing partial or complete results. Such adjustments are often called researchers' degrees of freedom because they give analysts room to make reasonable-looking decisions that nevertheless change the probability of finding a statistically significant result.

One common flexibility is optional stopping, also called data peeking. In this practice, researchers examine results during data collection and stop when the result becomes significant or looks close enough to significance to continue. Repeatedly applying an ordinary .05 test at several interim looks is not the same as conducting one planned test, and it can increase the long-run rate of false-positive findings (Simmons et al., 2011; Stefan & Schönbrodt, 2023).

Importantly, not every interim analysis is problematic. Planned sequential designs can be valid. In such designs, researchers decide in advance how often they will look at the data and what evidence threshold will be used at each look. Classical group-sequential methods and alpha-spending approaches protect the false-positive rate by budgeting the total error rate across the planned looks rather than simply reusing the .05 threshold each time (Pocock, 1977; O'Brien & Fleming, 1979; Lan & DeMets, 1983; Lakens, 2014).

A second flexibility is analytic multiplicity. Researchers

Figure 1 ■ False-positive rate (Type I error) under key workflows ($\delta = 0$, $\rho = 0.3$). Error bars show Wilson 95% confidence intervals. Because the y-axis spans severe inflation to display Forking-Peeks conditions, differences among near-nominal workflows (e.g., fixed-n vs. p-zone with $m = 1$) appear visually small; see Table 2 for exact estimates and confidence intervals.



may have several plausible outcomes, transformations, exclusion rules, or model specifications. If these options are explored and the most favorable result is reported, the reported analysis is no longer equivalent to a single prespecified test. Gelman and Loken (2013) describe this problem as the garden of forking paths: the analysis path can be shaped by the data even when researchers do not set out to fish for significance. Multiverse and specification-curve approaches make such choices more transparent by displaying how conclusions vary across reasonable specifications (Steegen et al., 2016; Silberzahn et al., 2018; Simonsohn et al., 2020).

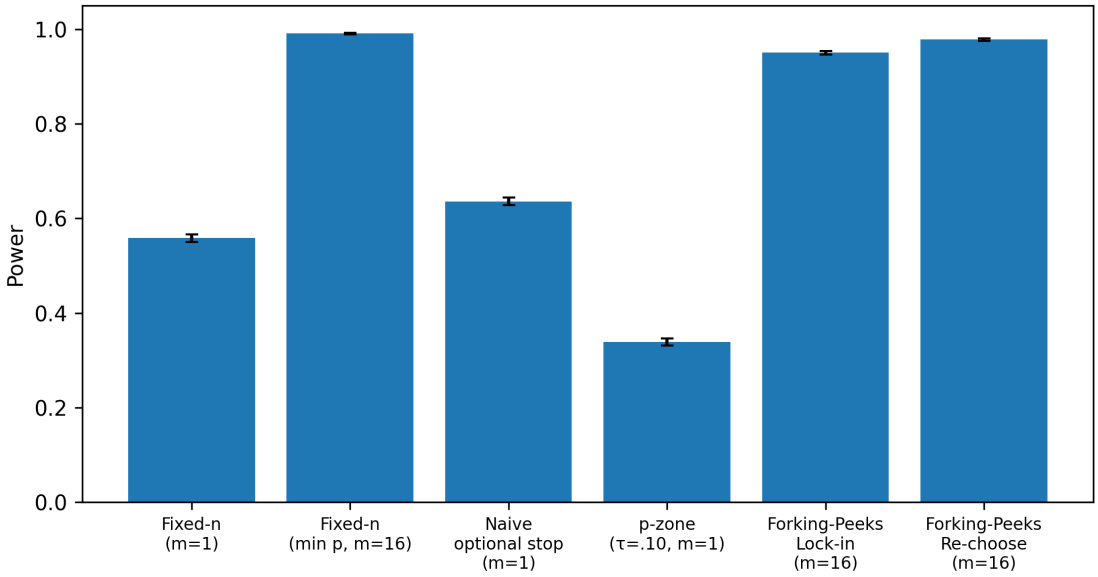
Most demonstrations treat optional stopping and analytic multiplicity separately. That separation is useful pedagogically, but real research workflows can combine the two. An analyst may look at interim results while also exploring several analysis choices and then decide whether to continue, stop, or report the most favorable result. We coin the term Forking-Peeks for this combined workflow: repeated peeking at accumulating data while selecting among multiple analytic forks.

The goal of this paper is not to accuse individual researchers of intentional misconduct. Rather, we use simulation as a teaching tool to show how apparently small and individually defensible decisions can compound. We model a simple two-group design, vary the number and similarity of analysis forks, and compare two selection strategies. In the Re-choose strategy, the analyst selects the best-looking fork at each interim look. In the Lock-in strategy, the analyst selects the best-looking fork at the first look and then keeps that fork fixed. The Lock-in strategy is especially educational because it can look consistent after the first choice, even though the first choice itself was data-dependent.

Our contributions are:

- A factorial simulation study that maps false-positive risk when optional stopping and analytic multiplicity are combined.
- A comparison of two realistic fork-selection strategies, Lock-in and Re-choose.
- A plain-language explanation of expected sample size and winner's curse, so readers can see not only whether a result becomes significant but also how large selected effects appear.
- Supplementary checks showing how a simple Bonferroni-style safeguard behaves, while emphasizing that planned sequential methods are usually better when interim looks are intended.
- Open, parameterized Python code that reproduces the figures and tables and can be adapted to other research designs.

Figure 2 ■ Power under key workflows ($\delta = 0.3$, $\rho = 0.3$). Error bars show Wilson 95% confidence intervals.



Method

Design overview

We organized the simulations into three parts. Study 1 provides benchmarks: a fixed-sample analysis and a naive optional-stopping analysis with a single prespecified test. Study 2 introduces p-zone stopping, a heuristic in which data collection continues only when the interim p-value is close to the significance threshold. Study 3 combines p-zone stopping with multiple analysis forks, producing the Forking-Peeks workflow.

Table 1 groups the design choices into fixed settings and systematically varied factors. The notation is introduced here and used only as needed. The symbol n denotes the number of observations per group at a look. The symbol L denotes the number of planned looks at the data. The symbol m denotes the number of analysis forks. The symbol ρ denotes how strongly the forks resemble each other. The symbol δ denotes the true standardized mean difference in the population, whereas $\hat{d}$ denotes the estimated standardized difference from the selected analysis. The symbol R denotes the number of Monte Carlo replications.

Data-generating process

Each simulated study compared two independent groups, labeled A and B. At the population level, Group A had mean 0 and Group B had mean δ. When $\delta = 0$, there was no true group difference; any significant finding was a false positive. When $\delta = 0.3$, Group B had a small true advantage.

To represent multiple plausible analysis forks, each simulated participant could have several correlated outcome values. A high fork correlation ($\rho = 0.6$) means that the forks tend to give similar information, as would happen when several analyses are minor variations of the same measurement. A lower correlation ($\rho = 0.3$) means that the forks are more distinct. Every fork was analyzed with the same two-sided independent-samples t-test comparing Group A and Group B. Thus, the simulations isolate the effect of having multiple analysis options while holding the test family constant.

Researcher degrees of freedom

Stopping strategies. In the fixed-n condition, data were collected to nmax and analyzed once. In the naive optional-stopping condition, the same test was conducted after each planned look; data collection stopped as soon as $p < .05$ or continued to the maximum sample size. In the p-zone condition, the study stopped for success when $p < .05$, continued only when the p-value was close to the threshold ($.05 \leq p < \tau$), and stopped for futility when $p \geq \tau$. This p-zone rule is not a valid sequential design. It is a stylized model of a practical heuristic: “continue only if the result looks almost significant.”

Fork-selection strategies. With a single fork ($m = 1$), the analysis is prespecified. With multiple forks, the analyst has several correlated analysis options. In the Re-choose strategy, the analyst evaluates all forks at each look

Table 1 ■ Simulation design: fixed settings and systematically varied factors.

Factor	Levels / values	Plain-language role
Fixed settings		
Study design	Two independent groups with equal sample sizes	A simple setting familiar to most applied researchers.
Test used in every fork	Two-sided independent-samples t-test	No other test families were simulated. Forks represent alternative outcomes/ specifications, not different statistical tests.
Initial sample per group (n_0)	40	First interim look.
Increment per look (Δn)	20	Additional observations per group before each later look.
Maximum sample per group (n_{max})	100	Final planned sample size if the study has not stopped earlier.
Number of looks (L)	4	Looks occur at $n = 40, 60, 80$, and 100 per group; fixed-n analyses have $L = 1$.
Nominal alpha	0.05, two-sided	The ordinary threshold used before any adjustment.
Monte Carlo replications (R)	15,000 per condition	Number of simulated studies used to estimate each rate.
Varied factors		
True population effect (δ)	0, 0.3	$\delta = 0$ estimates false positives; $\delta = 0.3$ estimates power and selected effect size.
Stopping rule	Fixed-n; naive optional stopping; p-zone stopping	Different ways of deciding when data collection ends.
p-zone upper limit (τ)	0.06, 0.08, 0.10, 0.12, 0.15, 0.20; key examples use 0.10 and 0.15	Data collection continues only when the p-value falls between .05 and τ.
Analysis forks (m)	1, 2, 4, 8, 16	Number of plausible analysis options available to the analyst.
Fork correlation (ρ)	0.3, 0.6	Higher values mean that the analysis forks are more similar to one another.
Fork-selection strategy	Single; Lock-in; Re-choose	How the analyst chooses among forks.

and reports the one with the smallest p-value. In the Lock-in strategy, the analyst chooses the best-looking fork at the first look and then applies that fork at later looks. Lock-in is therefore more consistent than Re-choose after the first choice, but the first choice is still data-dependent.

Outcome measures

The primary outcome is Type I error: the proportion of simulated studies that reported a significant result when the true effect was zero. Secondary outcomes are power when $\delta = 0.3$, the expected final sample size per group ($E[n]$), and the size of the selected effect estimate. We also report winner's curse, which in this manuscript means selection-induced exaggeration: among results that pass the significance filter, the estimated effect tends to be larger than the true effect because favorable random errors are preferentially selected.

Simulation procedure

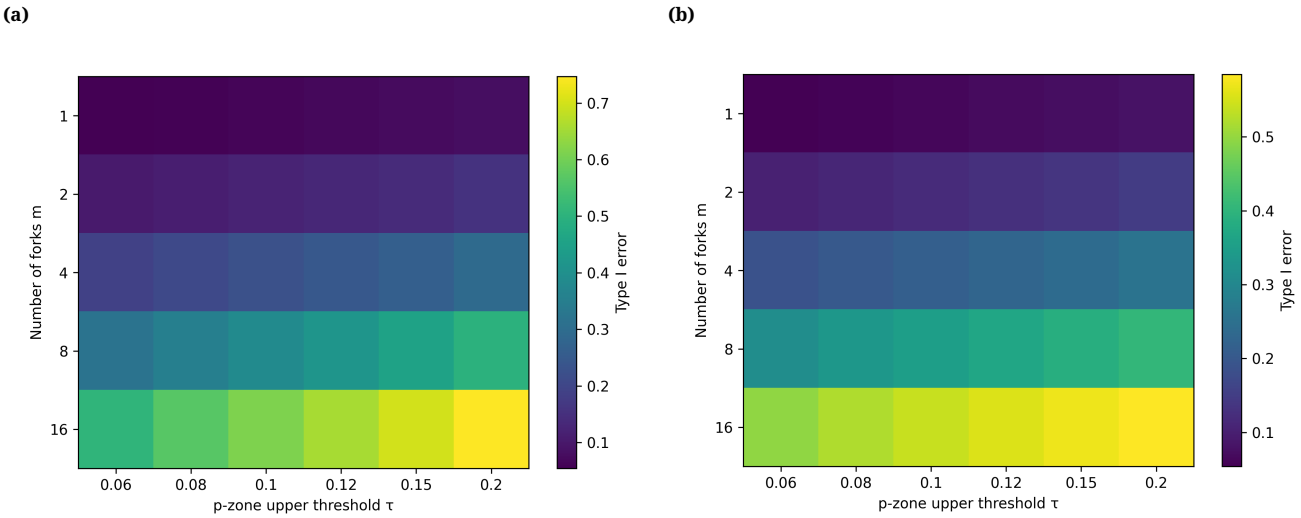
For each condition, we ran $R = 15,000$ simulated studies. In each simulated study, we generated data up to nmax per group, computed interim t-tests at the planned looks, applied the stopping and fork-selection rules, and recorded whether the selected result was significant. Proportions are reported with Wilson 95% confidence intervals (Wilson, 1927).

Reproducibility

All code to reproduce the simulations, figures, and tables is provided in an accompanying repository archived with a persistent identifier. The analysis was run in Python 3.13.5 (Python Software Foundation, 2025), using NumPy 2.3.5, SciPy 1.17.0, pandas 2.2.3, and matplotlib 3.10.8. After cloning the repository and installing dependencies, a single command reproduces the full pipeline:

```
python run_all.py
```

Figure 3 ■ Type I error risk map for Forking-Peeks under (a) Re-choose selection ($\delta = 0$, $\rho = 0.3$) and (b) Lock-in selection ($\delta = 0$, $\rho = 0.3$)



The script writes all intermediate results to CSV files and regenerates all figures. Random seeds are fixed in the code so that the reported results can be reproduced exactly, up to platform-level numerical differences.

Results

All square-bracketed intervals in this section are Wilson 95% confidence intervals.

Benchmark conditions. Under the null ($\delta = 0$), fixed-n single-analysis testing maintained Type I error near the nominal .05 level (0.049 $[0.045, 0.052]$; Table 2). By contrast, selecting the smallest p-value across $m = 16$ forks at fixed n inflated false positives to 0.473 $[0.465, 0.481]$. Thus, analytic multiplicity alone was enough to produce a large increase in false positives when the selected fork was treated as if it had been prespecified.

Optional stopping. Naive optional stopping with four looks ($n = 40, 60, 80$, and 100 per group) approximately doubled Type I error (0.115 $[0.110, 0.121]$) while reducing expected sample size only slightly ($E[n] = 95.4$ per group). p-zone stopping with $\tau = 0.10$ produced a smaller false-positive rate (0.064 $[0.060, 0.068]$) but also stopped very early on average ($E[n] = 41.3$ per group), because many simulated studies stopped for futility when the interim result did not look promising.

Forking-Peeks. Combining repeated looks with multiple forks produced the largest inflation. For $m = 16$, $\tau = 0.15$, and $\rho = 0.3$, Type I error reached 0.571 $[0.563, 0.578]$ under Lock-in and 0.700 $[0.693, 0.708]$ under Re-choose (Table 2). Risk maps in Figure 3 shows the same pattern across the broader grid: inflation increased as the number of forks grew and was generally larger under Re-choose than under Lock-in.

Results for the higher fork-correlation condition ($\rho = 0.6$) are now shown in Appendix Table A.1. The higher correlation reduced the amount of inflation because the forks contained more overlapping information, but the qualitative conclusion did not change. For example, with $m = 16$ and $\tau = 0.15$, Type I error was 0.404 $[0.396, 0.412]$ under Lock-in and 0.468 $[0.460, 0.476]$ under Re-choose, both far above .05.

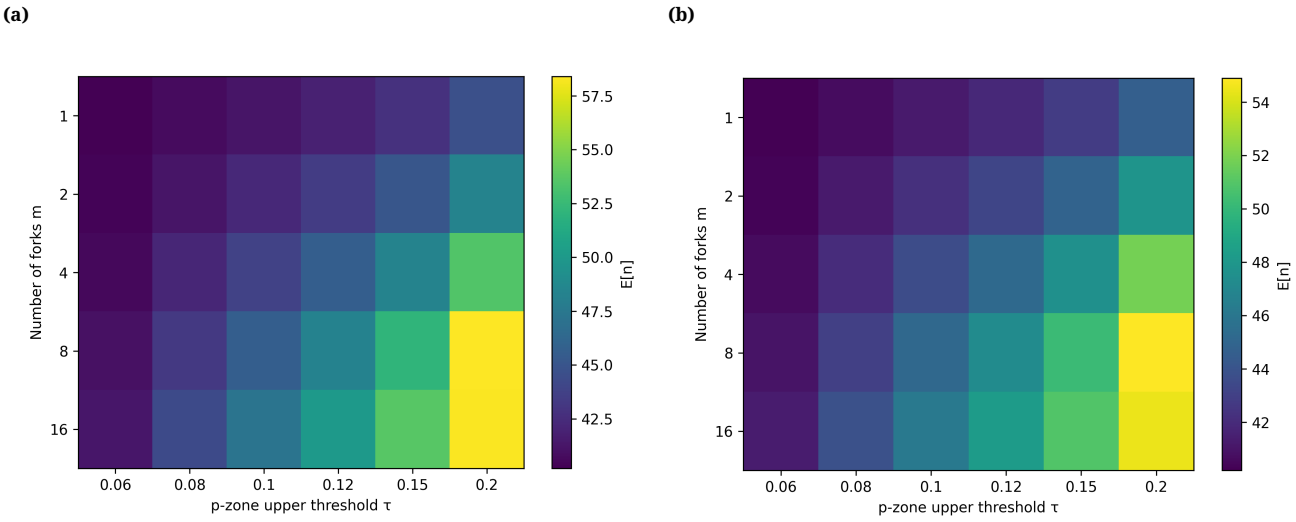
Correction check. We also evaluated a simple Bonferroni-style safeguard as a diagnostic check (Appendix Table A.2). For fixed-n multiplicity, the threshold was divided by the number of forks. For Forking-Peeks, the threshold was divided by the number of forks times the number of looks. This check controlled Type I error but was conservative: for $\rho = 0.3$, the Forking-Peeks false-positive rate fell to 0.025 $[0.023, 0.028]$ under Lock-in and 0.034 $[0.031, 0.037]$ under Re-choose. These results are not offered as the best sequential design; they illustrate why unadjusted look-by-fork searches are unsafe and why planned sequential methods are preferable when interim monitoring is needed.

Discussion

Key findings

The central lesson is simple: flexible data collection and flexible analysis choices are each risky, and the risks compound when they are combined. A single fixed-n analysis behaved as expected, with a false-positive rate close to .05.

Figure 4 ■ Expected sample size $E[n]$ per group under (a) Re-choose selection ($\delta = 0$, $\rho = 0.3$) and (b) Lock-in selection ($\delta = 0$, $\rho = 0.3$)



Naive optional stopping roughly doubled the false-positive rate. Analytic multiplicity alone was even more consequential when the best of 16 forks was selected. Forking-Peeks combined both mechanisms and produced the largest inflation.

The p-zone rule deserves careful interpretation. In our simulations, p-zone stopping often reduced the expected sample size because studies stopped early when interim results looked unpromising. In that narrow sense, the rule can look efficient. That efficiency is ironic: it comes from abandoning unpromising simulated studies, not from valid error control. Real questionable research practices may be less efficient. Researchers might continue collecting data despite p-values outside the p-zone, search for new outcomes, change exclusions, or add new forks. Such behavior could combine large sample sizes with high false-positive risk. Our p-zone simulations should therefore be read as one stylized heuristic, not as a complete model of all questionable research behavior.

The Lock-in strategy is also instructive. At first glance, Lock-in can look more honest than Re-choose because the analyst applies one analysis pipeline after the initial choice. The problem is that the initial choice was made after looking at the data. Consistency after a data-dependent choice does not remove the bias created by the choice. This is a useful teaching point: an analysis can be stable after selection and still require multiplicity adjustment because the selection process itself used the data.

The winner's curse results show why false-positive rates are not the only concern. When selection favors significant results, the reported significant effects tend to be too large. In the $\delta = 0.3$ simulations, Forking-Peeks produced high apparent power, but the selected significant effects were substantially larger than the true effect. This means the same workflow can create a double illusion: more discoveries and larger-looking effects.

Relation to prior work and novelty

These results align with earlier work showing that undisclosed flexibility can produce false-positive findings (Simmons et al., 2011) and with simulation catalogues of p-hacking strategies (Stefan & Schönbrodt, 2023). The extension here is the composite workflow. Rather than asking only what happens when researchers peek, or only what happens when they choose among analyses, we ask what happens when these behaviors occur together while data accumulate.

The risk-map format is meant to be educational. It lets readers see how false-positive risk changes as forks become more numerous, as forks become more or less correlated, and as the analyst chooses forks differently. The results for $\rho = 0.6$ show that correlated forks are less damaging than nearly independent forks, but they are not harmless. This matters because real analysis choices are usually correlated: alternative exclusions, transformations, or related outcomes often share much of the same information.

Practical recommendations

For confirmatory research, the safest message is to separate exploration from confirmation. Exploration is useful for learning and hypothesis generation, but confirmatory p-values require prespecified outcomes, analyses, and stop-

Table 2 ■ Key results under the null (Type I error) and under a small true effect (power; $\delta = 0.3$). Results shown for $\rho = 0.3$. Wilson 95% confidence intervals are in brackets; $E[n]$ is expected final sample size per group. For fixed-n conditions, $E[n] = 100$ by design because data collection proceeds to n_{max} without early stopping.

Condition	Type I error ($\delta = 0$)	$E[n]$ per group ($\delta = 0$)	Power ($\delta = 0.3$)	$E[n]$ per group ($\delta = 0.3$)	Mean $\hat{d}$ | significant ($\delta = 0.3$)
Fixed-n (single analysis)	0.049 [0.045, 0.052]	100.0	0.558 [0.551, 0.566]	100.0	0.402
Fixed-n (min p over $m = 16$)	0.473 [0.465, 0.481]	100.0	0.991 [0.990, 0.993]	100.0	0.517
Naive optional stopping ($m = 1, \tau = 1.0$)	0.115 [0.110, 0.121]	95.4	0.636 [0.628, 0.644]	75.7	0.468
p-zone stopping ($m = 1, \tau = 0.10$)	0.064 [0.060, 0.068]	41.3	0.339 [0.331, 0.346]	42.8	0.553
Forking-Peeks Lock-in ($m = 16, \tau = 0.15$)	0.571 [0.563, 0.578]	50.9	0.950 [0.947, 0.954]	42.6	0.658
Forking-Peeks Re-choose ($m = 16, \tau = 0.15$)	0.700 [0.693, 0.708]	53.7	0.978 [0.975, 0.980]	42.3	0.653

ping rules.

If interim looks are planned, researchers should use a valid sequential design rather than repeatedly applying the same .05 threshold. In plain terms, valid sequential designs decide in advance how much evidence is needed at each look so that the total false-positive risk remains controlled across the study. Examples include group-sequential boundaries and alpha-spending approaches (Pocock, 1977; O'Brien & Fleming, 1979; Lan & DeMets, 1983; Lakens, 2014).

If multiple outcomes, exclusion rules, transformations, or model specifications are examined, researchers should treat them as multiplicity. A simple Bonferroni-style rule divides the target false-positive rate by the number of tested opportunities. For example, dividing by m controls the worst-case risk across m forks at one fixed sample size, and dividing by m times L is an even more conservative safeguard when each fork is evaluated at L looks. This is derived from the Bonferroni/union-bound logic: the probability of at least one false positive is bounded by the sum of the individual test probabilities. However, this rule can be much too strict when tests are highly correlated, as they usually are across interim looks. It is best viewed as a hand-check or worst-case guardrail, not as the preferred design for planned interim monitoring.

When interim monitoring is part of the research design, sequential methods are usually more appropriate than simply dividing alpha by all look-by-fork combinations. When many analytic specifications are defensible, multiverse or specification-curve approaches can show robustness instead of hiding selection. When analyses are exploratory, authors should label them as exploratory and avoid presenting the selected p-value as if it came from a single prespecified test.

Finally, researchers should report effect sizes and uncertainty intervals and avoid treating p < .05 as a pass/fail label. The winner's curse results show that selected significant findings may exaggerate the effect even when the underlying effect is real.

Limitations

The simulations are intentionally simple. Every fork used the same two-sided t-test, and the forks were represented as correlated versions of a two-group outcome. Real research workflows may include different models, covariates, transformations, missing-data rules, and outcome definitions. The simple design is useful for teaching because it isolates the effect of peeking and fork selection, but it does not cover every possible analysis workflow.

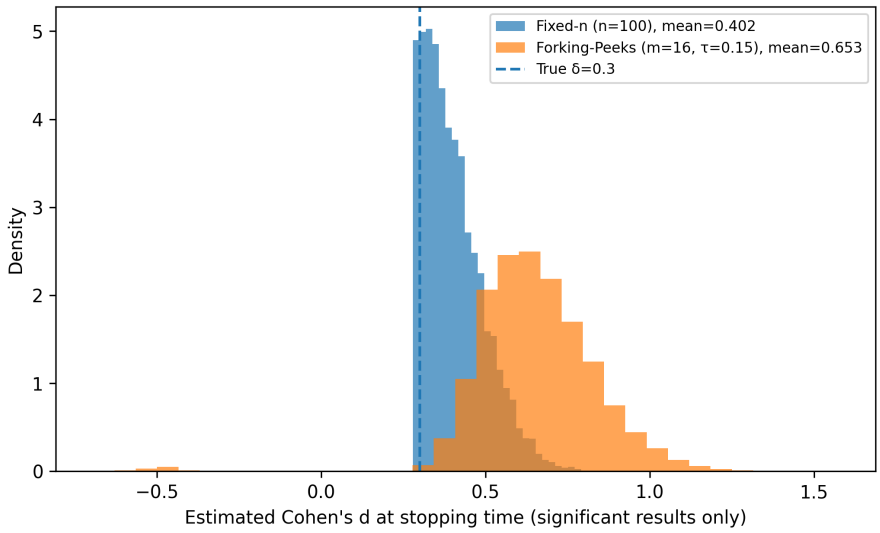
The p-zone rule is also only one heuristic. Some researchers may stop early for futility, others may continue regardless, and others may add new variables or analyses when the original result looks weak. Future work could model such adaptive expansion of the analysis space.

Finally, we focused on frequentist null-hypothesis testing. Bayesian methods, confidence sequences, and decision-theoretic sequential designs provide different ways to reason about accumulating evidence (Rouder, 2014; Howard et al., 2021; Johari et al., 2022). Comparing these approaches in the same Forking-Peeks framework would be a useful extension.

Conclusion

Forking-Peeks describes a combined workflow in which researchers look at accumulating data and select among multiple analysis forks. In simulations, this combination produced far larger false-positive rates and stronger winner's-curse effects than either flexibility alone. The

Figure 5 ■ Winner's curse: distribution of estimated Cohen's d among significant results under fixed-n vs. Forking-Peeks ($\delta = 0.3$, $\rho = 0.3$) in the Re-choose condition ($m = 16$, $\tau = 0.15$).



practical message is not that researchers should never explore. Rather, exploration should be labeled as exploration, and confirmatory claims should use prespecified analyses, valid stopping rules, and appropriate multiplicity control. Small flexibilities can become large distortions when they are allowed to accumulate.

References

Gelman, A., & Loken, E. (2013). *The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time*. https://sites.stat.columbia.edu/gelman/research/unpublished/p_hacking.pdf

Howard, S. R., Ramdas, A., McAuliffe, J., & Sekhon, J. S. (2021). Time-uniform, nonparametric, nonasymptotic confidence sequences. *The Annals of Statistics*, *49*(2), 1055–1080. doi: 10.1214/20-AOS1991.

Johari, R., Koomen, P., Pekelis, L., & Walsh, D. (2022). Always valid inference: Continuous monitoring of A/B tests. *Operations Research*, *70*(3), 1806–1821. doi: 10.1287/opre.2021.2135.

Lakens, D. (2014). Performing high-powered studies efficiently with sequential analyses. *European Journal of Social Psychology*, *44*(7), 701–710. doi: 10.1002/ejsp.2023.

Lan, K. K. G., & DeMets, D. L. (1983). Discrete sequential boundaries for clinical trials. *Biometrika*, *70*(3), 659–663. doi: 10.1093/biomet/70.3.659.

O'Brien, P. C., & Fleming, T. R. (1979). A multiple testing procedure for clinical trials. *Biometrics*, *35*(3), 549–556. doi: 10.2307/2530245.

Pocock, S. J. (1977). Group sequential methods in the design and analysis of clinical trials. *Biometrika*, *64*(2), 191–199. doi: 10.1093/biomet/64.2.191.

Python Software Foundation. (2025). *Python documentation, version 3.13.5*. https://docs.python.org/3.13/

Rouder, J. N. (2014). Optional stopping: No problem for Bayesians. *Psychonomic Bulletin & Review*, *21*(2), 301–308. doi: 10.3758/s13423-014-0595-4.

Silberzahn, R., Uhlmann, E. L., Martin, D. P., Anselmi, P., Aust, F., Awtrey, E., Bahník, Š., Bai, F., Bannard, C., Bonnier, E., Carlsson, R., Cheung, F., Christensen, G., Clay, R., Craig, M. A., Dalla Rosa, A., Dam, L., Evans, M. H., Flores Cervantes, I., Nosek, B. A., et al. (2018). Many analysts, one data set: Making transparent how variations in analytic choices affect results. *Advances in Methods and Practices in Psychological Science*, *1*(3), 337–356. doi: 10.1177/2515245917747646.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*(11), 1359–1366. doi: 10.1177/0956797611417632.

Simonsohn, U., Simmons, J. P., & Nelson, L. D. (2020). Specification curve analysis. *Nature Human Behaviour*, *4*(11), 1208–1214. doi: 10.1038/s41562-020-0912-z.

Steegen, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science*, *11*(5), 702–712. doi: 10.1177/1745691616658637.

Stefan, A. M., & Schönbrodt, F. D. (2023). Big little lies: A compendium and simulation of p-hacking strategies. *Royal Society Open Science*, *10*(2), 220346. doi: 10.1098/rsos.220346.

Wilson, E. B. (1927). Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, *22*(158), 209–212. doi: 10.1080/01621459.1927.10502953.

(Appendix A follows)

Table A.1 ■ Key results for the higher fork-correlation condition ($\rho = 0.6$). Wilson 95% confidence intervals are in brackets.

Condition	Type I error ($\delta = 0$)	$E[n]$ per group ($\delta = 0$)	Power ($\delta = 0.3$)	$E[n]$ per group ($\delta = 0.3$)	Mean $\hat{d} \mid$ significant ($\delta = 0.3$)
Fixed-n (single analysis)	0.050 [0.047, 0.054]	100.0	0.562 [0.554, 0.570]	100.0	0.401
Fixed-n (min p over $m = 16$)	0.318 [0.310, 0.325]	100.0	0.935 [0.931, 0.939]	100.0	0.479
Naive optional stopping ($m = 1, \tau = 1.0$)	0.115 [0.110, 0.120]	95.5	0.638 [0.630, 0.645]	75.6	0.467
p-zone stopping ($m = 1, \tau = 0.10$)	0.062 [0.058, 0.065]	41.2	0.344 [0.337, 0.352]	42.9	0.549
Forking-Peeks Lock-in ($m = 16, \tau = 0.15$)	0.404 [0.396, 0.412]	50.0	0.840 [0.834, 0.846]	45.1	0.615
Forking-Peeks Re-choose ($m = 16, \tau = 0.15$)	0.468 [0.460, 0.476]	51.8	0.877 [0.871, 0.882]	45.0	0.609

Table A.2 ■ Bonferroni-style correction check for $m = 16$ and $\tau = 0.15$. For fixed-n multiplicity, alpha was divided by m; for Forking-Peeks, alpha was divided by $m \times L$. Wilson 95% confidence intervals are in brackets.

Condition	Correction	Type I error ($\delta = 0$)	$E[n]$ per group ($\delta = 0$)	Power ($\delta = 0.3$)	$E[n]$ per group ($\delta = 0.3$)
$\rho = 0.3$					
Fixed-n (min p over $m = 16$)	alpha / m	0.045 [0.042, 0.049]	100.0	0.811 [0.805, 0.817]	100.0
Forking-Peeks Lock-in	alpha/$(m \times L)$	0.025 [0.023, 0.028]	75.3	0.486 [0.478, 0.494]	77.9
Forking-Peeks Re-choose	alpha/$(m \times L)$	0.034 [0.031, 0.037]	86.3	0.686 [0.679, 0.694]	75.8
$\rho = 0.6$					
Fixed-n (min p over $m = 16$)	alpha/m	0.032 [0.029, 0.035]	100.0	0.622 [0.614, 0.629]	100.0
Forking-Peeks Lock-in	alpha/$(m \times L)$	0.018 [0.016, 0.020]	67.0	0.367 [0.359, 0.374]	78.0
Forking-Peeks Re-choose	alpha/$(m \times L)$	0.023 [0.021, 0.025]	72.4	0.496 [0.488, 0.504]	77.5

Appendix A: Additional results and correction check

This appendix reports the higher fork-correlation results ($\rho = 0.6$) and the Bonferroni-style correction check requested for clarity. The substantive conclusions are unchanged: stronger fork correlation reduces, but does not eliminate, inflation; the simple correction check controls Type I error but is conservative.

Open practices

The *Open Data* badge was earned because the data of the experiment(s) are available on the journal's web site.
The *Open Material* badge was earned because supplementary material(s) are available on the journal's web site.

Citation

Harada, Y. (2026). Forking-peeks: When optional stopping meets analytic multiplicity. *The Quantitative Methods for Psychology*, *22*(2), 46–55. doi: 10.20982/tqmp.22.2.p046.

Received: 31/01/2026 ~ Accepted: 22/06/2026