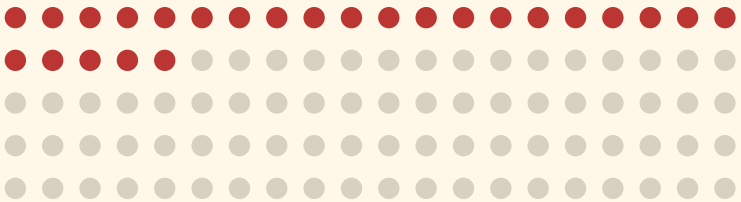
Twenty chances to call a false win

When a test has zero lift, more interim reviews can change the error rate.

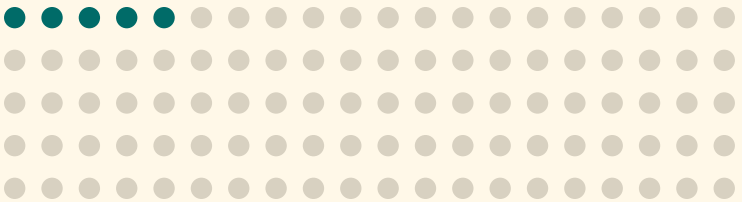
24.78%

Ordinary threshold



5.03%

Pocock rule

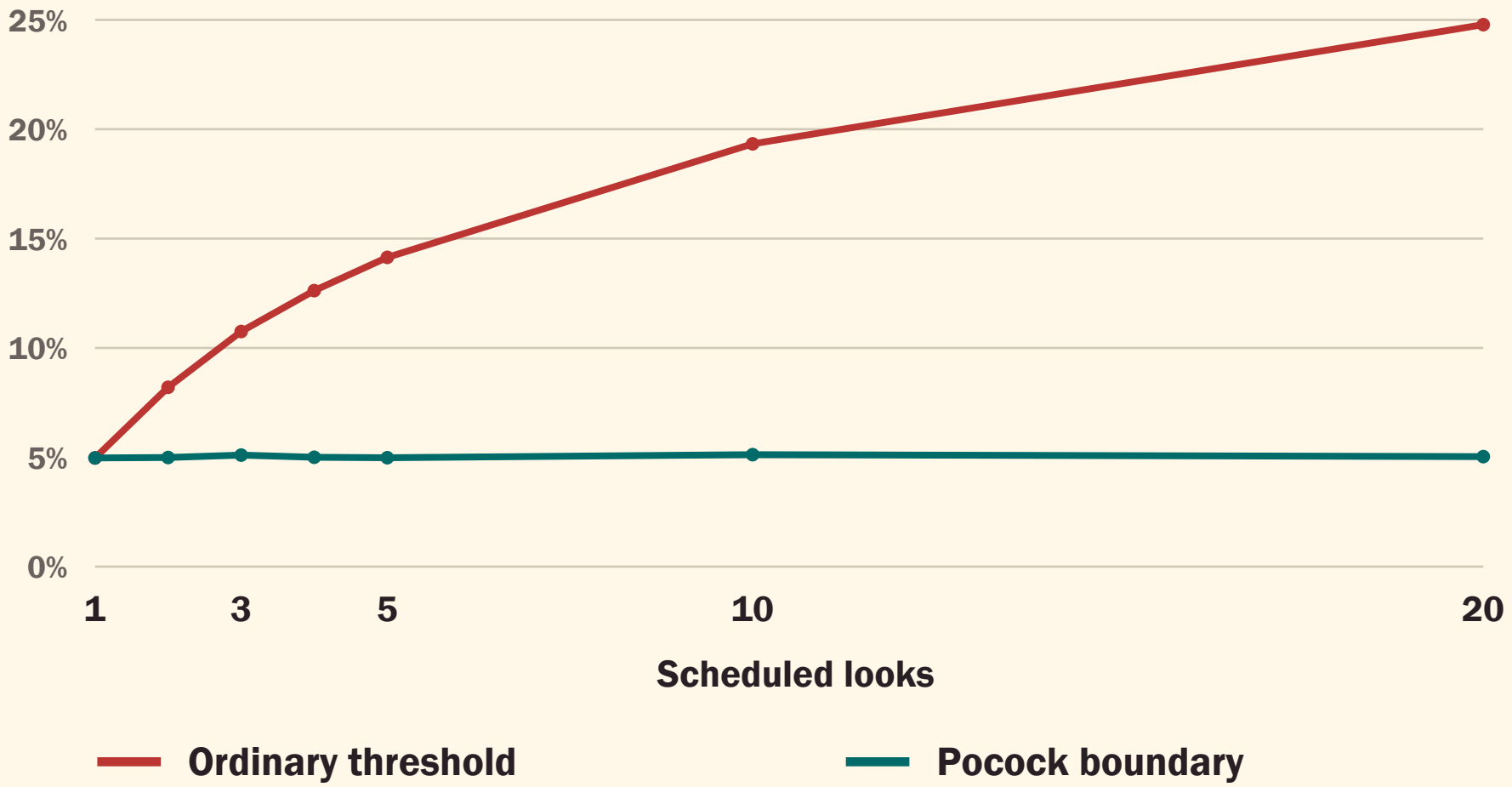


In the 20-look schedule, roughly one in four ordinary tests declared a winner. A boundary set for that schedule stayed near the 5% target.

Source: controlled A/B benchmark • 200,000 no-lift trials per review schedule • two-sided design.

How does error grow with more looks?

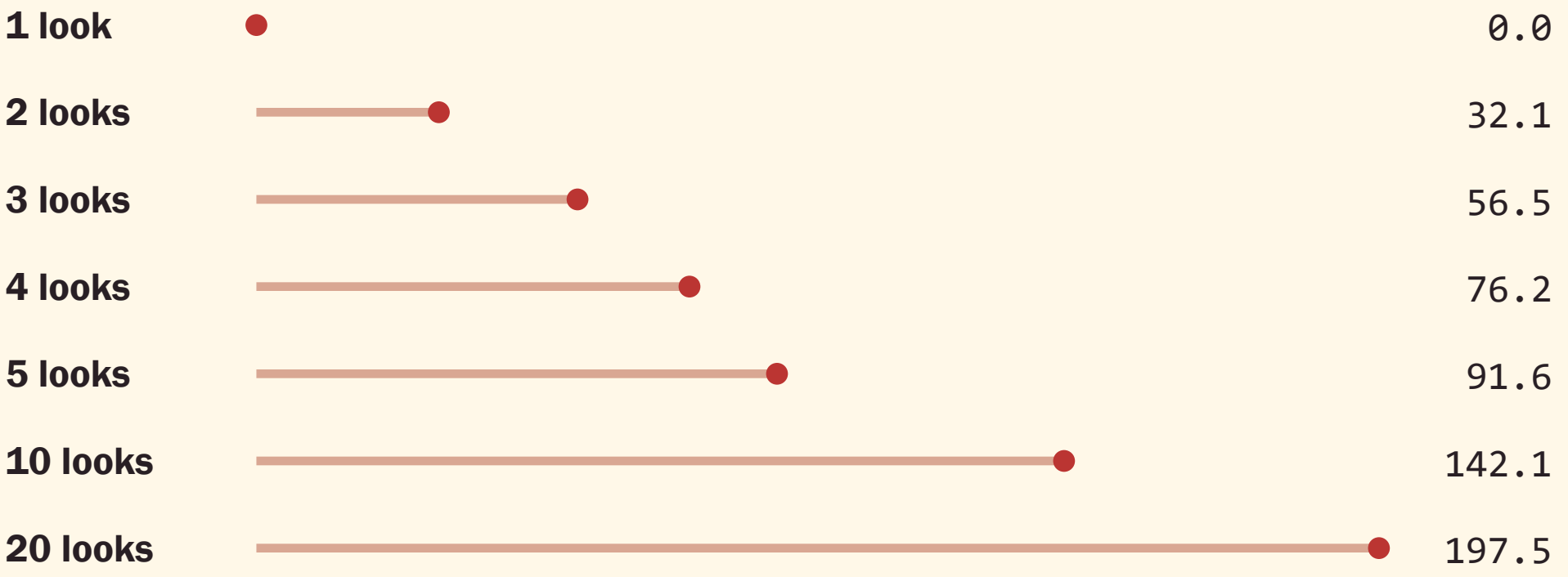
Observed false-win rate in each prescheduled design, using a shared percentage scale.



Source: controlled A/B benchmark • 200,000 no-lift trials per review schedule • two-sided design.

How many excess calls does this imply?

Expected extra false winners per 1,000 no-lift tests using the ordinary threshold.

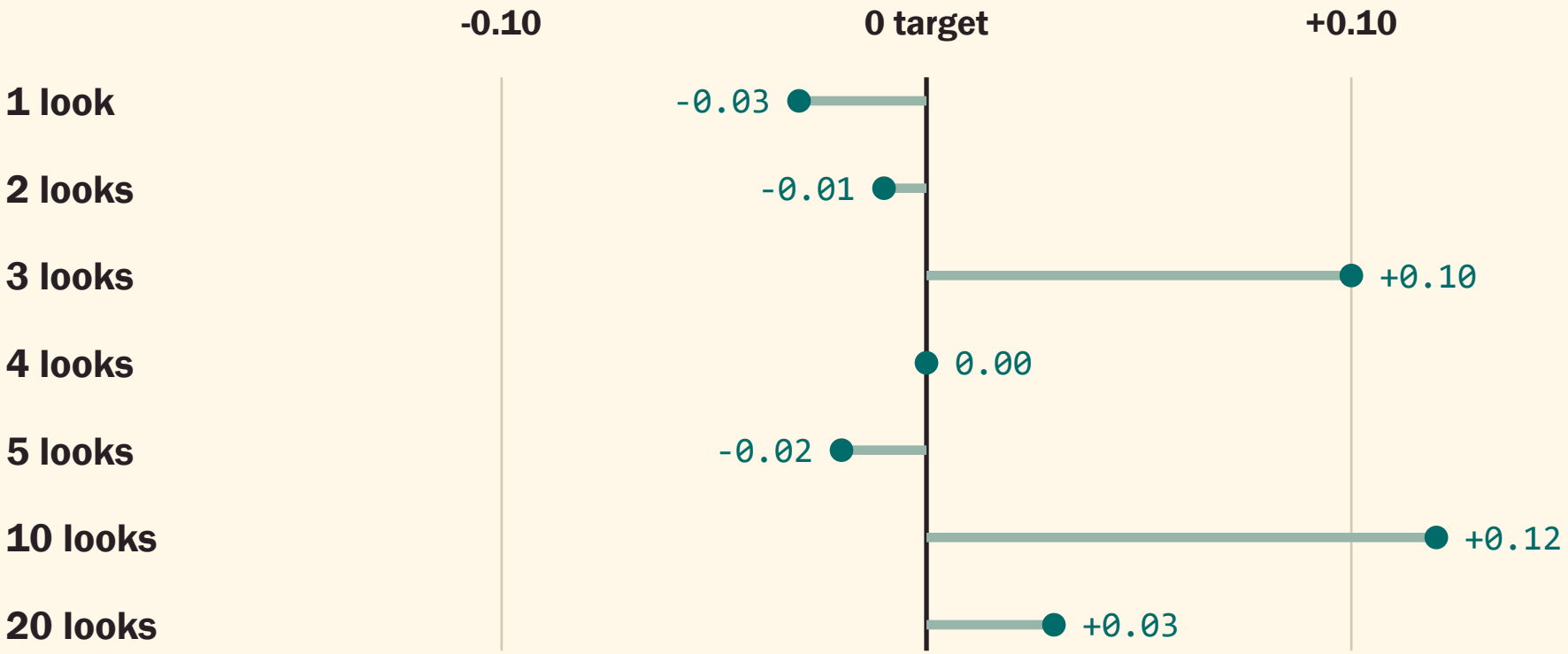


At 20 looks the gap is 19.75 percentage points, equivalent to 197.5 additional calls per 1,000 tests. Fractional counts are expectations.

Calculation: 1,000 × (ordinary false-win rate − calibrated false-win rate), expressed as proportions.

How close did calibration land?

Pocock false-win rate minus the 5% target, in percentage points.



Calibrated results range from 4.97% to 5.12% across schedules. These decimals are benchmark estimates from a finite number of trials.

Source: controlled A/B benchmark • 200,000 no-lift trials per review schedule • two-sided design.

Write the stopping rule before launch

A repeatable review protocol for a growth team checking an A/B test.

Fix the schedule	Choose the maximum number of reviews and when they occur. Define the conversion outcome and the decision direction.
Calibrate to the design	Use a boundary fitted to the chosen review schedule. Changing the schedule or threshold changes the error behavior.
Test power separately	The benchmark here uses zero lift. Evaluate detection probability, expected duration and sample size under plausible effects before deployment.

Source: project experiment-design and buyer notes. Scheduling guidance is an application of the benchmark.