Is N-Hacking Ever OK? A simulation-based study

Pamela Reinagel

Section of Neurobiology, Division of Biological Science, University of California, San Diego.

Abstract

After an experiment has been completed and analyzed, a trend may be observed that is "not quite significant". Sometimes in this situation, researchers incrementally grow their sample size N in an effort to achieve statistical significance. This is especially tempting in situations when samples are very costly or time-consuming to collect, such that collecting an entirely new sample larger than N (the statistically sanctioned alternative) would be prohibitive. Such post-hoc sampling or "N-hacking" is condemned, however, because it leads to an excess of false positive results. Here Monte-Carlo simulations are used to show why and how incremental sampling causes false positives, but also to challenge the claim that it necessarily produces alarmingly high false positive rates. In a parameter regime that would be representative of practice in many research fields, simulations show that the inflation of the false positive rate is modest and easily bounded. But the effect on false positive rate is only half the story. What many researchers really want to know is the effect N-hacking would have on the likelihood that a positive result is a real effect that will be replicable: the positive predictive value (PPV). This question has not been considered in the reproducibility literature. The answer depends on the effect size and the prior probability of an effect. Although in practice these values are not known, simulations show that for a wide range of values, the PPV of results obtained by N-hacking is in fact *higher* than that of non-incremented experiments of the same sample size and statistical power. This is because the increase in false positives is more than offset by the increase in true positives. Therefore in many situations, adding a few samples to shore up a nearly-significant result is in fact *statistically beneficial*. In conclusion, if samples are added after an initial hypothesis test this should be disclosed, and if a p value is reported it should be corrected. But, contrary to widespread belief, collecting additional samples to resolve a borderline p value is not invalid, and can confer previously unappreciated advantages for efficiency and positive predictive value.

Background

There has been much concern in recent years concerning the lack of reproducibility of results in some scientific literatures. The call for improved education in statistics and greater transparency in reporting is justified and welcome. But if we apply overly-conservative rules dogmatically, we as a community risk throwing out a lot of babies (good data, promising leads) with the statistical bath water. Experiments in biology and psychology often require substantial financial resources, scientific talent, and use of animal and/or human subjects. There is an ethical imperative to use these resources efficiently. To ensure both reproducibility and efficiency of research, experimentalists need to understand statistical issues rather than blindly apply rules.

The rule in question is a cornerstone of null hypothesis significance testing: sample exactly the predetermined sample size N, and then accept the verdict the hypothesis test, whatever it is. Adding a few samples and retesting after a negative result can produce misleading outcomes. But this depends on the parameter regime in which one is operating; what researchers need to know is what can occur in their operating regime.

Empirical scientists are accustomed to looking at data, so simulation is an excellent way to gain intuitions about the implications of statistical methods. Simulation is also a frequentist approach: no Bayesian assumptions are required. Here I simulate the denounced practice of "N-hacking" – incrementally adding more samples after the fact whenever a preliminary result is "almost significant". The simulations demonstrate the known effect that post-hoc sample growth of this kind elevates the false positive rate, and show why this is the case. After exploring a broad range of assumptions bracketing common practice in many fields of research, however, it emerges that the elevation in false positive rate is quite modest, and it becomes apparent that it could readily be corrected for. Moreover, additional simulations show that there is a truth underlying researchers' intuition that growing the sample size is a good idea. The purpose of this article is not to dismiss concerns about sampling procedures, but rather to demonstrate that there are better options than either starting over from scratch or abandoning a hypothesis after obtaining a nearly-significant outcome.

Results

These simulations can be taken to represent a large number of independent studies, each collecting separate samples to test a different hypothesis. I assume that a significance criterion α has been set in advance, and the sample size would be increased *only* for those tests that meet a criterion of "p close to α". I further assume that the maximum number of samples the study could or would add is no more than a few times greater than the original sample size, or a few hundred total samples. All simulations were performed in Matlab 2018a.

Effect of incrementally growing sample size on the false positive rate

Experiments were simulated by drawing two independent samples of size N from the same normal distribution. An independent sample Student's *t*-test was then used to accept or reject the null hypothesis that the samples came from distributions with the same mean. Because the samples always came from the same distribution, any positive result is a false positive. By definition, the *t*-test produces false positives at a rate of exactly α, the significance threshold, regardless of the mean or standard deviation of the source distribution or the sample size N. I will call the false positive rate when the null hypothesis is true FP_0 ("FP null").

I defined a plausible Asymmetric N-increasing policy as follows: every time a comparison yielded a *p* value that was "almost significant", additional sample points were added incrementally to the sample, and the *t*-test repeated. This was iterated until the *p* value was either significant, or no longer close, or the maximum number of samples was reached. The definition of "almost significant" was: $\alpha \leq p < (1+w)\,\alpha$, where $0 < w \leq 1$. For example if $\alpha = 0.05$ and $w = 0.2$, one would accept a hypothesis if $p < 0.05$, reject if $p > 0.06$, and add

samples for p values in between. This would be representative of conditions under which I have seen researchers increment sample size in the fields of biology in which I have worked.

Results of such a policy are shown in Figure 1, assuming an initial sample size of $N_{init} = 12$, incremental sample growth of $N_{incr} = 6$, and maximum sample size of $N_{max} = 24$. For every choice of w and α, $M = 10^6$ independent experiments were simulated. This is meant to represent 10^6 separate studies, each using this policy to test only one hypothesis.

As expected, this Asymmetric N-increasing policy yielded an increase in false positives FP_0, which was more severe as w increased (Fig. 1a). Nevertheless the overall elevation in false positives was rather modest. For example with a policy of $\alpha = 0.05$ and $w = 1$, sample size was grown whenever p was between $0.05 - 0.10$, resulting in a realized false positive rate $FP = 0.0625$ instead of the nominal 0.05. Following this policy resulted in a negligible increase in the sample size on average (Fig. 1b). Note that no multiple comparison correction was done within study for the interim retesting on the policy; instead the false positives due to multiple comparisons are included in the reported false positive rates, i.e. these are the *uncorrected* false positive rates.

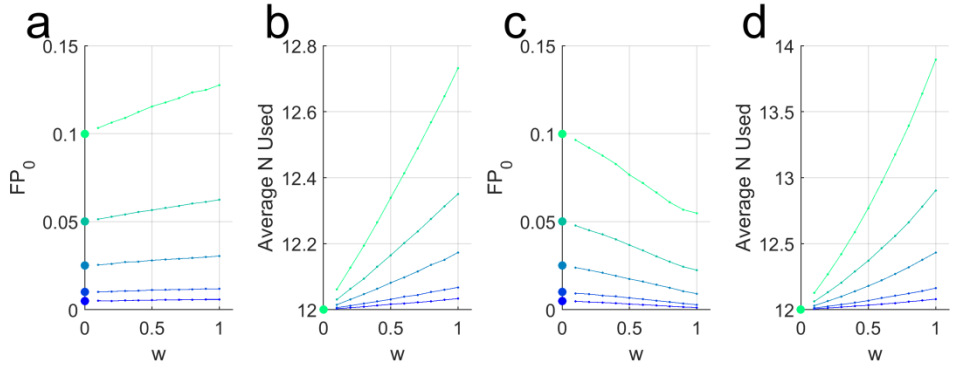

Figure 1. Effect of selective sample-increasing on false positive rate FP_0. Results shown are for experiments with initial sample size $N_{init} = 12$, and sample increments $N_{incr} = 6$, and maximum sample size $N_{max} = 24$. Each point or symbol represents results from $M = 10^6$ simulated experiments. **(a)** For any choice of α (0.005, 0.01, 0.025, 0.0500, 0.1; colors), the Asymmetric N-increasing policy yields an increase in false positives FP_0 which grows with the decision window w. The case of ($w = 0$) is identical to the standard fixed-N policy, and yields false positives at a rate of $FP_0 = \alpha$ (solid symbols). **(b)** Average N in the final sample using the Asymmetric policy, as a function of α and w. **(c)** The Symmetrical N-increasing policy yields a net decrease in false positives, which grows with w. **(d)** Average N in the final sample using the Symmetric policy.

The main reason false positives are elevated by this policy is that the experiments that were incremented and retested were chosen in a biased way. By selectively incrementing only the subset of true negatives in which the difference between experimental and control groups was rather large, and thus nearly significant, even a small difference between groups in the added samples would be sufficient to push the overall group difference over the threshold for significance, purely by chance.

The problem with this policy is that it is asymmetric: N is incremented when p was just above threshold, but not when it was just below threshold. To demonstrate this point, consider a Symmetric N-increasing policy, in which incremental sample growth occurred whenever a p

value was either just below or just above α: $(1-w)\,\alpha \le p < (1+w)\,\alpha$. It turns out that making the policy symmetric more than overcomes the problem – it would convert more false positives to true negatives than it converts true negatives to false positives, resulting in a net reduction in false positives (Fig. 1c). This is because in addition to the effect noted above, this policy also incremented the sample size in a biased subset of the false positives: ones in which the difference between experimental and control groups was rather small and thus barely significant. The Symmetric policy resulted in a slightly larger final sample size on average (Fig. 1d). In discussions of statistical malpractice, it is often asserted that an experimentalist would never add more samples after obtaining a significant p value, but interestingly there is evidence that they do (*1*). Therefore the consequences of both policies will be explored further below.

Dependence on α and w

For the Asymmetric N-increasing policy, analysis of the simulated data reveals that for any given choice of w, the false positive rate depends linearly on α: $FP_0 = k\alpha$ (Fig 2a). The slopes of these lines are in turn an increasing function of the decision window w (Fig 2b, symbols). On the Symmetric policy, the dependence of FP on α is also linear (Fig 2c) and the slope k declines with w (Fig 2d).

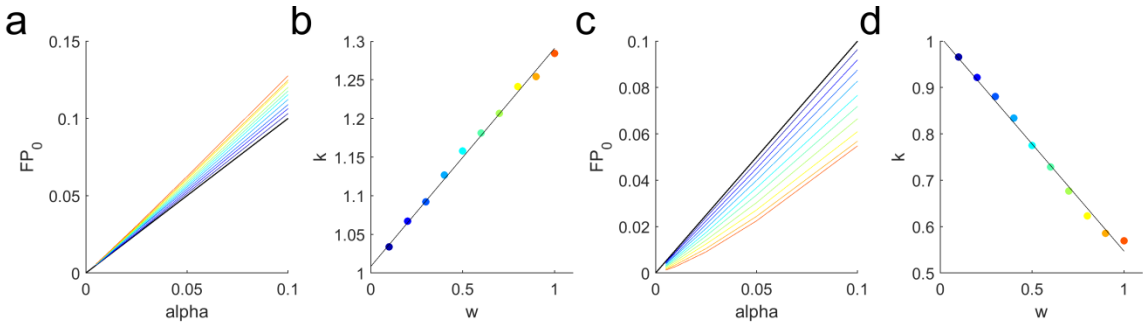

Figure 2. Dependence of false positive rate FP_0 on α and w. Results from simulations using $N_{init} = 12$, $N_{incr} = 6$, $N_{max} = 24$, with $M = 10^6$simulated experiments per condition. **(a)** The realized false positive rate FP_0 for an Asymmetrical N-increasing policy when the null hypothesis is true. Color indicates w (cf. panel b). For each value of w, $FP_0$is plotted for each value of α and the data points connected (not a fit). The identity line (black) is the false positive rate of the standard Fixed-N policy, $FP_0 = \alpha$. **(b)** The slopes k obtained from linear fits to the data shown in (a), plotted as a function of w (colored symbols). The dependence of the slope k on w is not linear in general, but was approximately linear in this parameter range (linear fit, black). **(c)** Like (a) but for the Symmetrical N-increasing policy. Note that for $w > 0.5$, FP_0 is biased upward at larger values of α, due to the imposed cap of $2N$ additional samples. Therefore to determine slopes, lines for each w were fit using only values of α for which this cap did not impact any simulation. **(d)** The slopes k from the linear fits to the data shown in (c), for all values of w for which a slope could be estimated (*i.e.* at least two values of α were not impacted by sampling cap). The dependence of the slope k on w is not linear in general, but was approximately linear in this parameter regime (linear fit, black).

Dependence on N_{init} and N_{incr}

These results above depend quantitatively on the assumptions made for N_{init}, N_{incr}, and N_{max}. Additional simulations below explore values of N_{init} ranging from 2 to 128 initial sample points, incremental sampling N_{incr} ranging from 1 to N_{init} per increment, capping the

maximum total sample size at either$N_{max} = 256$ or $N_{max} = 5N_{init}$. Simulations were run with values of α between 0.005 and 0.05 and values of w from0.1 to 0.4. These assumptions more than bracket the range of realistic sample sizes and ad-hoc sample growth that would be commonly used in many experimental research fields. For each combination of N_{init}, N_{incr}, α and w, $M = 10^6$ experiments with no true effect were simulated to estimate the fraction yielding false positive results expected on the null hypothesis.

The simulations show that initial sample size and incremental sample size are important. Results for the Asymmetric policy with $\alpha = 0.05, w = 0.4$ are shown in Figure 3a. The false positive rate FP_0 is always elevated compared to α (black line), and this is more severe when the intial sample size is larger (curves slope upward) or the incremental sample growth is smaller (cooler colors are higher). Nevertheless the false positive rate didn't exceed 0.06 for any condition. In this range of parameters, the dependence of k on w was approximately linear, so one can visualize the results for all combinations of of α and w on the same scale by plotting them as $(\frac{FP_0}{\alpha} - 1)/w$ (Fig 3b). Recall that for the fixed-N policy by definition $FP_0 = \alpha$, so this equation evaluates to 0 for any parameter combination. Positive values on this scale indicate an increase in the false positive rate compared to the fixed-N policy, and negative values reflect a lower false positive rate. Combining results for all choices of α and w and fitting curves as a function of N_{init} (Fig 3c) allows one to summarize trends independent of choice of α and w.

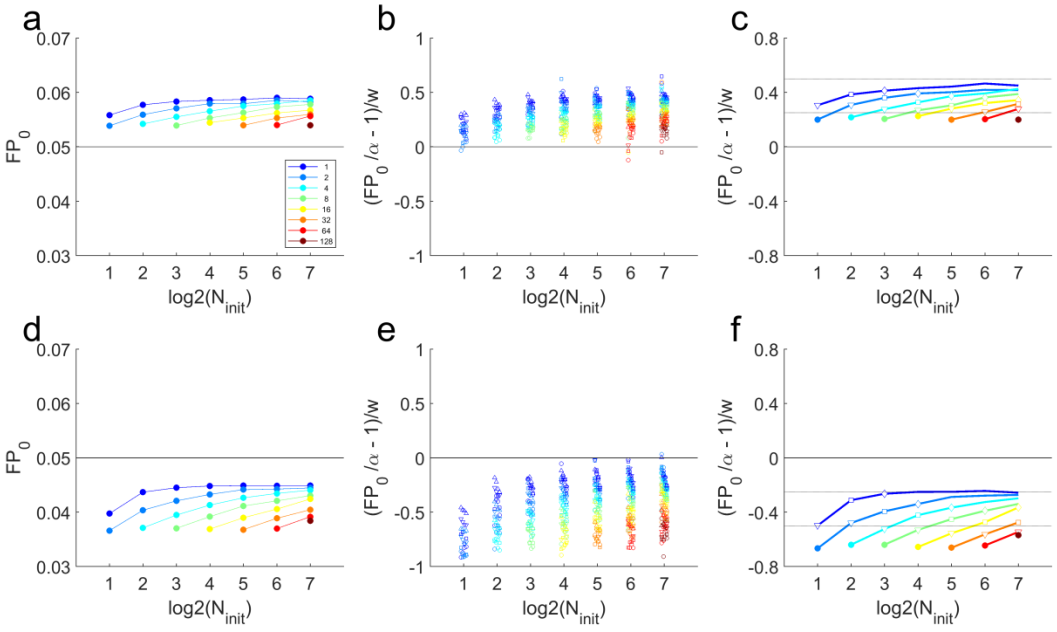

Figure 3. Dependence of false positive rate FP_0 on the initial sample size and sample increment. Each symbol represents the result from 10^6 experiments simulated with a ground truth of no effect. **a.** The realized false positive rate FP_0 of the Asymmetric N-increasing policy, as a function of $\log_2 N_{init}$ (horizontal axis) and N_{incr}(colors), for the case $\alpha = 0.05$, $w = 0.4$, $N_{max} = 256$. **b.** Results for all choices of α (0.005, 0.010, 0.025, or 0.050; symbol shapes) and w (0.1, 0.2, 0.3 or 0.4, small horizontal shifts), plotted as $(\frac{FP}{\alpha} - 1)/w$ (vertical axis) to reveal regularities. **c.** Summary of simulations in (b) obtained by fitting the equation $FP = (cw + 1)\alpha$. Symbols indicate simulations in which $N_{incr} = N_{init}$ (closed circles), $N_{incr} = N_{init}/2$ (open triangles), $N_{incr} = N_{init}/4$ (open squares) and $N_{incr} = N_{init}/8$ (open diamonds). Dotted black lines relate to the rules of thumb (see text). Panels **d-f**: as in a-c, for the Symmetrical N-increasing policy. Similar results were found for both policies using $N_{max} = 5\ N_{init}$ (not shown).

In the case of the Symmetric policy, the false positive rate FP_0 is always lower than α; this beneficial effect is strongest when N_{incr} is large (warm colors in Fig 3d-f) or N_{init} is small (positive slopes in Fig 3d-f).

In summary, these simulations show that the effect of incremental sampling on the false positive rate is real, but it is modest in size and lawfully related to a handful of parameters. From the simulations one can take away some rules of thumb (dotted lines, Fig 3 c,f):

Assymetric Policy	***Symmetric Policy***
$FP_0 < \alpha(1+\frac{w}{2})$ for $N_{incr} \le N_{init}$	$FP_0 < \alpha\left(1-\frac{w}{4}\right)$ for $N_{incr} \le N_{init}$
$FP_0 < \alpha(1+\frac{w}{4})$ for $N_{incr} = N_{init}$	$FP_0 < \alpha\left(1-\frac{w}{2}\right)$ for $N_{incr} = N_{init}$

For example: a study that committed to an Asymmetric policy with $w = 0.4$, $N_{init} = 10$, $N_{incr} = 10$, $N_{max} = 50$ could conservatively estimate that the false positive rate expected on the null hypothesis is $FP_0 < 0.0550$ by rule of thumb, compared to the simulation result of $FP_0 = 0.0541 \pm 0.0001$. Additional simulations for $\alpha = 0.05\ or\ 0.10$, $N_{incr} = 1$ (*i.e.* the worst case conditions) were extended to $w = 19.0$ for $N_{init} = 2 - 128$ with $N_{max} = 256$ and still did not exceed this empirical bound (not shown). These rules of thumb are meant to be helpful guides, but have not been formally proven, which limits generalization to other conditions not tested. The Matlab code provided in Supplementary Materials can be used to estimate by simulation the false positive rate for other parameter combinations.

For any single experiment, the p value obtained after incremental sampling should be corrected, and a number of methods are already available for this (*2-8*).

Trade-off between statistical power and positive predictive value

N-hacking increases the false positive rate expected on the null hypothesis because some true negative results will by chance be converted to false positives when a few samples are added. But the motivation for adding samples is the hope of increasing sensitivity: some "almost-significant" effects are false negatives, which might be converted to true positives with added samples. How these harmful and beneficial effects balance depends on what fraction of the tested hypotheses are in fact true (prior probability of effect) $P(H_1)$, and how large the effects are when present (effect size E). The reason for this is nicely explained in (*9*). To explore this trade-off, one must simulate experiments in which some of the hypotheses tested are true, i.e. there are some real effects. In that context one can discuss the sensitivity, or statistical power, which is the fraction of real effects for which the null hypothesis is rejected – the chance that a real effect, if present, will be discovered. The selectivity, or positive predictive value (PPV), is the fraction of the experiments rejecting the null hypothesis that reflect real effects – the chance that a positive result will turn out to be reproducible. More formal statements of these definitions are given in Appendix 1.

Dependence on N_{init} and α

To show how N_{init} and α affect the statistical power and positive predictive value of an experiment, simulations were done exactly as described above, but now 1% of all experiments

were simulated with a real effect of 1σ difference between the population means, such that rejecting the null is the correct conclusion. This was simulated for several choices of N_{init} and α, comparing $w = 0$ (i.e. fixed-N policy) to $w = 0.4$, on either the Asymmetric or Symmetric N-increasing policy. Note that in real experiments, the prior probability of a real effect and the true size of the effect are not known. But in simulations these facts are imposed, and thus known precisely.

First, it is helpful to remember that even in the standard fixed-N policy there is a trade-off between sensitivity and selectivity, which is controlled by the choice of α. For a given sample size N, increasing the arbitrary cutoff for significance α increases sensitivity, at the expense of reduced PPV. By varying α one can define a curve for the sensitivity-selectivity trade-off (e.g., Fig 4a, any red curve). This curve summarizes the options available for interpreting data sets acquired in this way. The choice of α is up to the investigator, depending on the relative priority one sets on avoiding missing real effects vs. avoiding believing false ones.

Simulating this for different choices of N further illustrates that in a fixed-N policy, a larger sample size N is always better: it increases both sensitivity and selectivity, moving the entire curve up and to the right (Fig 4a, compare any two red curves). This observation suggests a generalization that the statistical quality of any two experimental policies can be compared by relating these curves. A higher curve is better – it means one could choose some α to achieve greater sensitivity for any desired selectivity, or to achieve greater selectivity for any desired sensitivity, relative to any curve that lies below it.

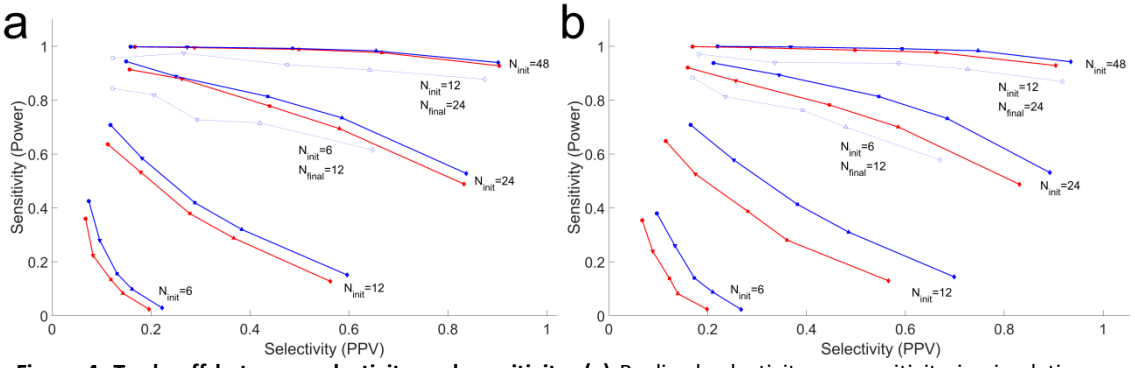

Figure 4. Trade-off between selectivity and sensitivity. (a) Realized selectivity vs. sensitivity in simulations with effect size 1σ, prior effect probability 0.01, and N_{init}=6, 12, 24 or 48 (four curves of each color). Symbols indicate α (○=0.001, ▽=0.005, □=0.01, △=0.025, ◊=0.05). Text labels indicate N_{init} for pairs of curves. For every combination of these parameters a total of $M = 10^4/\alpha$ experiments were simulated to estimate the probabilities of true positives, true negatives, false positives and false negatives caused by the policy. Results for the standard fixed-N policy (red) and for an Asymmetric N-increasing policy using $N_{incr} = N_{init}$, $N_{max} = 2N_{init}$, and $w = 0.4$ (blue) are shown. The small subset of experiments that added samples (to $N_{final} = 2N_{init}$) are shown in dotted blue curves and open symbols for $N_{init} = 6$ and $N_{init} = 12$. **(b)** Results for standard fixed-N vs. Symmetric N-increasing policy, details otherwise as in (a). In other simulations, the Power-PPV curves for both the Asymmetric and Symmetric N-increasing policies were above those of the fixed-N policy for effect size ranging from $E = 0.5$ to 2 and prior probability $P(H_1)$ ranging from 0.001 to 0.1 (not shown).

The curves for the standard fixed-N policy (red curves, Fig 4a) provide the benchmark to which other sampling policies may be compared. An example Asymmetric N-increasing policy is shown (blue curves, Fig 4a). Because few experiments experience incrementing, the average final sample size was negligibly greater than the fixed-N policy: $\langle N_{\text{final}} \rangle \leq 1.02\ N_{init}$ for all parameter combinations tested (not shown; cf. Fig.2a). Therefore, the overall sensitivity and selectivity of the policy can be reasonably compared to the fixed-N policy with $N = N_{init}$ (paired curves). For all choices of N_{init} simulated, the curve for the Asymmetric N-increasing policy (blue) fell entirely above and to the right of the corresponding curve for the fixed-N policy (red). Thus on average the Asymmetric N-increasing policy resides entirely on a better frontier than the fixed-N policy: for any point on the fixed-N curve there exists some choice of α for which the Asymmetric policy curve has equal selectivity with higher sensitivity, and another choice of α for which the Asymmetric policy has equal sensitivity with higher selectivity.

Comparing the two policies with the same choice of α is also informative (symbols of same shape on the red vs. paired blue curves). For the parameter combinations with lower power ($N_{init} = 6\ or\ 12$ with any α, or $N_{init} = 24$ with $\alpha < 0.01$), using the same choice of α in an Asymmetric N-increasing policy – even without any correction for the false positive rate or multiple comparisons – yielded improvements in *both* statistical power *and* PPV relative to fixed-N. This was the case up to at least $w = 1$ (not shown). For the parameter combinations with higher power ($N_{init} = 48$ with any α, or N_{init}=24 with $\alpha \geq 0.01$), using the same α for the Asymmetric N-increasing policy led to a loss in selectivity relative to the fixed-N policy (the matched symbols are to the left of their fixed-N benchmarks). Still, this loss in selectivity was accompanied by a far greater improvement in statistical power than could be achieved by moving along the red curve (changing α) to obtain the same selectivity. In this sense, the Asymmetric policy represented a superior trade-off even in these cases.

The small subset of experiments for which sample size was increased had $2N_{init}$ final samples. Is the whole effect due to the fact that un-incremented experiments lie on the fixed-N curve for $N = N_{init}$ and the incremented subset lie on the curve for $N = 2N_{init}$? The answer is no. Considering the incremented subset of experiments separately (dotted blue curves) reveals that they live on a frontier *above* the curve for fixed-N experiments with a sample size of $N = 2N_{init}$. The subset of experiments that were not incremented (which had a sample size of exactly N_{init}) lay on a curve that was slightly above or indistinguishable from the fixed-N benchmark in all cases examined (not shown).

These simulations demonstrate that for an effect size of 1σ and a rather pessimistic prior probability of 0.01, the Asymmetric N-increasing is a win-win scenario: for any initial sample size, whatever selectivity (PPV) one can achieve on the fixed-N policy, that same selectivity can be achieved with higher statistical power on the Asymmetric N-increasing policy with some choice of α. Additional simulations showed that this remained the case as either the prior probability or effect size approached 0 (although PPV approaches 0 in both cases), for a range of N_{init} and $N_{incr} = N_{init}$ (not shown).

The Symmetric N-increasing policy was superior to the fixed-N policy (Fig 4b, compare red to blue as described for Fig 4a), as well as beating the Asymmetric policy (compare blue curves in 4a to 4b). Even using the same choice of α the Symmetric policy increased both

selectivity and sensitivity relative to fixed-N for all conditions tested. The subset of experiments on the Symmetric policy that had added samples to a final $2N$ fell on a curve well above the fixed-N experiments $2N$ samples, and the subset of experiments that reached a verdict with N samples fell on a curve either above or indistinguishable from the fixed-N curve with N samples.

Dependence on the decision window w

To determine the impact of the decision window w on these conclusions, I further simulated results of the Asymmetric policy for a range of N_{init} for w ranging from 0.2 to 10 and $N_{incr} = 1$. In my experience people rarely N-hack when $p > 2\alpha$ ($w = 1$), and it was shown above that higher values of w would produce quite high false positive rates on the null hypothesis (FP_0). But how would this affect statistical inference when some of the experiments have real effects?

As w increases, the curves in the PPV vs. Power plot always move up and/or to the right (Fig 5, top row). Comparing simulation outcomes for the same choice of α, one sees that increasing w always increases sensitivity (warm colors are above cool colors along any curve, Fig 5 bottom row). For any choice of α, as N_{init} increases uncorrected N-hacking switches from increasing the PPV to eroding it, compared to the fixed-N policy (gray curves slant left).

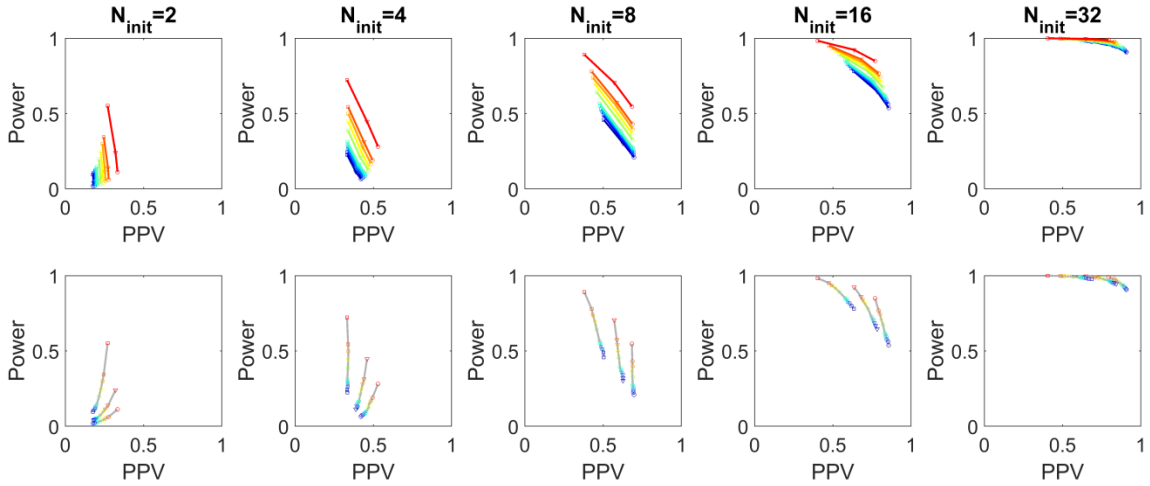

Figure 5. Asymmetric N-increasing policies improve statistical inference even when w is large. Realized selectivity vs. sensitivity in simulations with effect size $E = 1\sigma$ prior effect probability $P(H_1) = 0.10$, with N_{init} as indicated on column title, $N_{incr} = 1$, $N_{max} = 50$. Each symbol represents the uncorrected result from $M = 10^6$ simulated experiments. Symbols indicate α (○=0.01, ▽=0.02, □=0.05). Colors indicate w (blue→red= 0, 0.2, 0.4, 0.6, 0.8, 1, 2, 3, 4, 5, 10). Note that dark blue ($w = 0$) is the standard fixed-N policy. **Top row:** simulations with the same w and different α are connected with curves, color indicates w. **Bottom row:** the same data, but simulations with the same α and different w are connected with gray curves.

Nevertheless, the trade-off of PPV vs. Power remains advantageous. For example, consider $N_{init} = 8,\ \alpha = 0.05$. In this case the fixed-N experiment has a PPV of 0.50 (not 0.95, as the experimenters might falsely believe), and a power of 0.46. Asymmetric N-hacking with a window of $w = 5$ implies that more samples would be added for any interim test result of $0.05 < p < 0.25$. Without any correction for incremental sampling (as shown), this would erode

the ultimate PPV from 0.50 down to 0.43 (in other words the False Positive Risk would be 57% instead of 50%). But in exchange for this, the statistical power would be increased from 0.46 to 0.78. So the investigators would be slightly more likely to believe a result that is a fluke, but far more likely to find a real effect if it is there. Different ways of correcting the p value will place one somewhere to the right on this overall superior curve.

It is noteworthy that in the most extreme case of Asymmetric N-increasing ($N_{init} = 2$, retesting after every $N_{incr} = 1$ sample) the practice is strictly beneficial, always increasing both power and PPV without even adjusting p or α. The significance of this will be discussed below.

Discussion

Main conclusions

These simulations demonstrate that increasing the sample size incrementally whenever a result is "almost" significant will lead to a higher rate of false positives expected by chance (i.e. if the null hypothesis is true). The problem arises from the fact that equally close "just" significant results are not similarly challenged. Most writers warn that this practice will lead to extremely high or even 100% false positive rates (*5, 10-13*). But those projections are based on assumptions that are not representative of typical practice in some scientific fields, such as that an experimentalist would add more samples after obtaining a non-significant result no matter how far from α the *p* value was, or would continue adding samples indefinitely until achieving a significant outcome. If instead one considers circumstances in which the *p* value would have to be rather close to α for one to add samples (e.g., no more than twice α), and a limited number of total samples could be added before giving up (e.g., no more than five times the initial sample), the effects on the false positive rate are modest and clearly bounded.

The increase in the false positive rate depends on the initial sample size $N_{init,}$ significance criterion α, closeness criterion w, increment size N_{incr}, and total sample cap N_{max}. Simulations demonstrate in which direction and how steeply the false positive rate depends on these factors. Some rules of thumb emerge for how bad the effect could possibly be, given those parameters. This is not meant to be a proposed method for correcting the p value; there already several correction methods available to account for incremental sampling – even if the decision to increment the sample size was made after the fact (*2-8*).

Further simulations demonstrate that under many conditions this type of N-hacking is superior to a fixed-N policy in the sense that it increases the statistical power achievable for any given positive predictive value (PPV), compared to studies that strictly adhere to the initially planned N. This has not been previously noted. In particular, for the great many experimental studies that use a small initial sample size ($N \leq 12$) and $\alpha = 0.05$, if one would only add samples post-hoc when $P < 0.1$ and would always quit before exceeding $N = 50$ samples, it is simply not true that N-hacking leads to an elevated risk of unreproducible results. A verdict of "statistical significance" reached in this manner, far from being dubious, is more likely to be reproducible than results reached by fixed-N experiments with the same initial or final sample size – even if no correction were applied for sequential sampling or multiple comparisons.

Scientists in exactly this situation are currently being told (by teachers, advisors, reviewers, editors, and even staff biostatisticians) that if they have obtained a non-significant finding with a p value just above α, they cannot validly add more samples to their data set to improve statistical power; they must either run a completely independent replication, or accept the null hypothesis. The results shown here imply that this is bad advice. It is true that adding samples after the test violates the basic premise of null hypothesis significance testing (NHST). But that is not the same as being *invalid*. Adding more samples *with disclosure* is never invalid, and there are methods for rigorous correction of the P value within the NHST framework. Moreover, these simulations show that there are statistical benefits of incremental sampling that are often overlooked.

Extensions and limitations

These simulations used a normal distribution for the ground truth source distribution and an independent sample *t*-test as our basic hypothesis test. But the analysis of the false positive rate FP_0 only depends on the assumption that the statistical test used generates *p* values that are uniformly distributed between 0-1 on the null hypothesis. In other words, as long as the statistical test being used is valid for the distribution being sampled and the structure of the experiment, the dependence of false positive rate on parameters in these simulations should generalize to any source distribution and statistical test. Power analysis may be affected by the shape of the source distribution, however, so generality of those results to other distributions should not be assumed.

These simulations only considered experiments in which a single hypothesis is tested on each sample. Multiple tests on a single sample (such as a gene chip array experiment) is a very different situation, because in that case incrementing N and retesting would lead to re-testing of all the hypotheses, regardless of their original *p* values. This case has been discussed by others.

Numerical simulations and graphs are easy for experimental scientists to understand, because they present the expectations of the null hypothesis in terms directly comparable to data. These simulations covered a broad range of parameters to provide concrete intuitions about the directions and orders of magnitudes of effects in relation to experimental parameters. But this is no substitute for analytic treatments, which can determine exactly under what conditions these results will apply, provide rigorous proofs and precise bounds or corrections.

Nothing in this argument assumes that one can know the PPV of any single experiment. In the real world the prior probability of a true effect $P(H_1)$ and the effect size E are unknown to the investigator. But in simulations the effect size and prior probability are known, which allows one to demonstrate that certain inequalities hold regardless of those values. Under certain conditions (*e.g.,* when power is low due to small N or stringent α), whatever the PPV would be using fixed-N, the PPV after Asymmetric N-increasing would be greater. Under other conditions, the PPV after Asymmetric N-increasing is lower than that of a fixed-N experiment, but there still exists some choice of α that would provide the same PPV as fixed-N with higher power, and some other choice that would provide the same power with higher PPV. How to find these values of α is not addressed.

These simulations asked: if a population of scientists followed a certain sampling policy, what fraction of their experiments would yield a "significant" difference when the null hypothesis was in fact true (FP_0), and what fraction of their "significant" findings would be real effects (PPV)? These are population-level questions. One could separately analyze the outcomes in subsets of the simulated experiments whose interim or final p values fell in narrow sub-ranges, though I have not. When interpreting any single experiment one should take into account the specific p values that were obtained at each decision point (a "p-equals" rather than "p-less-than" approch)(*14*).

The real reason you should not N-hack

Some may be suspicious of the claim that N-hacking, in the sense simulated here, provides better statistical inference than fixed-N experiments, in terms of the power achieved for any given PPV, as well as the number of samples required to achieve this power. But this result is not at odds with established statistical theory. The N-incrementing policies described here are closely related to other well-described sequential sampling methods – particularly in the limit of $N_{init} = 2, N_{incr} = 1$ (Fig 5, left panels). For example, in Wald's Sequential Probability Ratio Test (*15*), one sets a threshold α to accept and another threshold β to reject a hypothesis. Then one computes a test statistic S after each new sample is added. If $S < \alpha$ the hypothesis is accepted, if $S > \beta$ the hypothesis is rejected, and if $\alpha < S < \beta$ one continues sampling. Similarly, a Bayesian sequential sampling method sets a criterion c, and then sequentially computes the Bayes Factor for the hypothesis vs. null hypothesis. If $BF > c$ the hypothesis is accepted, $if\ BF < \frac{1}{c}$ the hypothesis is rejected, and otherwise one keeps sampling (*16*). The drift diffusion model (DDM), which is widely used to model decision-making, is closely related (*17*). Fully sequential sampling methods are known to be statistically powerful and efficient. The kind of N-hacking commonly practiced is merely a weak version, conferring minor benefits compared to fixed-N methods. So the real reason not to advocate N-hacking as an intentional method is that fully sequential sampling methods are even better (*18, 19*).

Finally it is worth noting that this entire problem arises because of the currently standard practice of setting arbitrary cutoffs for "statistical significance" and reducing analog P values to binary hypothesis tests. It is not at all clear that experimental science is well served by this overall approach (*20-22*). Converting a *p* value into a significance verdict necessarily discards information. Maintaining an analog estimate of the evidence for a hypothesis as data are accumulated would be better (*23*), and would eliminate the need for most of this discussion.

Even if a p value has been corrected for incremental sampling, it has been argued that p values are so widely misunderstood as to be misleading. This is a separate question from the one addressed here. Although it remains controversial, many advocate supplementing reported p values with some other measure that is closer to what the experimentalists want to know, such as the PPV, Odds Ratio (*14, 24*), Bayes Factor(*16, 25*), the False Positive Risk (1-PPV) (*26*), or a non-Bayesian bound on the Bayes Factor (*27*). Those arguments and suggestions are all still applicable to corrected p values after incremental sampling.

Supplementary Materials

Matlab code for simulating the false positive rate on the null hypothesis (FP_0) can be found at http://www.ratrix.org/codeshare/NhackingAndFPrate.zip

Acknowledgements

PR acknowledges Hal Pashler, Casper Albers and Daniel Lakens for valuable discussions and helpful comments on earlier drafts of the manuscript.

Appendix 1. Definitions

H_0 Null hypothesis (no effect)	For example, in an independent sample *t*-test comparing samples from populations A and B, the null hypothesis is that the means of the groups are the same: $H_0: \mu_A = \mu_B$
H_1 Alternative hypothesis (effect)	For the *t*-test example, the alternative is that means of the populations are not the same: $H_1: \mu_A \neq \mu_B$
N Sample size	number of samples assessed from each population
p Value returned by statistical test	The fraction of experiments in which one would observe a difference between groups at least as great as the difference observed, if in fact H_0 were true.
α Significance criterion	A criterion to reject H_0 only if $p < \alpha$
FP_0 False Positive Rate on the Null	For any decision policy, probability of rejecting the null if the null is true: $FP_0 \equiv P(reject\ H_0 \mid H_0)$ For fixed-N case, by definition, $FP_0 = p$ In simulations, obtained when all simulated experiments draw both samples from the same distribution.
E Effect size	The true difference in the means of the two populations being compared, expressed as a ratio of the (shared or pooled) standard deviation: $E \equiv \frac{\lvert\mu_A - \mu_B\rvert}{\sigma}$ In real experiments the effect size is unknown *a priori*, and estimates from data are often upwardly biased. In simulations E is imposed and thus known.
$P(H_1)$ Prior probability of an effect	The probability H_0 is false, before considering the data. In real experiments this is very problematic to estimate. In simulations this is imposed: $P(H_1)$ of the simulated experiments draw samples from distributions whose means are in fact different.
Power (Sensitivity)	The probability that a real difference will be found to be significant: $Power \equiv P(reject\ H_0 \mid H_1)$ Depends on $P(H_1)$ and E
PPV Positive Predictive Value (Selectivity)	The probability that an effect that was deemed significant is in fact real: $PPV \equiv \mathrm{P}(H_1 \mid reject\ H_0)$ Depends on $P(H_1)$ and E Related to "False Positive Risk"(*14, 26*) as: $FPR = 1 - PPV$

References

1. E. C. Yu, A. M. Sprenger, R. P. Thomas, M. R. Dougherty, When decision heuristics and science collide. *Psychon Bull Rev* **21**, 268-282 (2014).
2. D. L. DeMets, G. Lan, The alpha spending function approach to interim data analyses. *Cancer Treat Res* **75**, 1-27 (1995).
3. D. Lakens, Performing high-powered studies efficiently with sequential analyses. *Eur J Soc Psychol* **44**, 701-710 (2014).
4. B. J. Sagarin, J. K. Ambler, E. M. Lee, An Ethical Approach to Peeking at Data. *Perspect Psychol Sci* **9**, 293-304 (2014).
5. E. Schott, M. Rhemtulla, K. Byers-Heinlein, Should I test more babies? Solutions for transparent data peeking. *Infant Behav Dev* **54**, 166-176 (2019).
6. D. Lakens, E. R. Evers, Sailing From the Seas of Chaos Into the Corridor of Stability: Practical Recommendations to Increase the Informational Value of Studies. *Perspect Psychol Sci* **9**, 278-292 (2014).
7. R. W. Frick, A better stopping rule for conventional statistical tests. *Behav Res Meth Ins C* **30**, 690-697 (1998).
8. P. Grünwald, R. de Heide, W. Koolen, Safe Testing. *arXiv*, 1906.07801 (2019).
9. D. Colquhoun, An investigation of the false discovery rate and the misinterpretation of p-values. *R Soc Open Sci* **1**, 140216 (2014).
10. C. Albers, The problem with unadjusted multiple and sequential statistical testing. *Nat Commun* **10**, 1921 (2019).
11. D. Szucs, A Tutorial on Hunting Statistical Significance by Chasing N. *Front Psychol* **7**, 1444 (2016).
12. J. P. Simmons, L. D. Nelson, U. Simonsohn, False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychol Sci* **22**, 1359-1366 (2011).
13. H. J. Motulsky, Common misconceptions about data analysis and statistics. *Naunyn Schmiedebergs Arch Pharmacol* **387**, 1017-1023 (2014).
14. D. Colquhoun, The reproducibility of research and the misinterpretation of p-values. *R Soc Open Sci* **4**, 171085 (2017).
15. A. Wald, *Sequential analysis*. Wiley mathematical statistics series (J. Wiley & sons, New York, 1947), pp. xii, 212 p.
16. S. N. Goodman, Toward evidence-based medical statistics. 2: The Bayes factor. *Ann Intern Med* **130**, 1005-1013 (1999).
17. J. I. Gold, M. N. Shadlen, Neural computations that underlie decisions about sensory stimuli. *Trends Cogn Sci* **5**, 10-16 (2001).
18. J. Bartroff, T. L. Lai, M.-C. Shih, *Sequential experimentation in clinical trials : design and analysis*. Springer series in statistics, (Springer, New York, 2013), pp. xv, 237 pages.
19. D. Siegmund, *Sequential analysis : tests and confidence intervals*. Springer series in statistics (Springer-Verlag, New York, 1985), pp. xi, 272 p.
20. R. L. Wasserstein, A. L. Schirm, N. A. Lazar, Moving to a World Beyond "p < 0.05". *Am Stat* **73**, 1-19 (2019).
21. R. S. Nickerson, Null hypothesis significance testing: a review of an old and continuing controversy. *Psychol Methods* **5**, 241-301 (2000).
22. S. N. Goodman, Toward evidence-based medical statistics. 1: The P value fallacy. *Ann Intern Med* **130**, 995-1004 (1999).

23. S. N. Goodman, Of P-values and Bayes: a modest proposal. *Epidemiology* **12**, 295-297 (2001).
24. M. J. Bayarri, D. J. Benjamin, J. O. Berger, T. M. Sellke, Rejection odds and rejection ratios: A proposal for statistical practice in testing hypotheses (vol 72, pg 90, 2016). *Journal of Mathematical Psychology* **89**, 98-98 (2019).
25. L. Held, M. Ott, On p-Values and Bayes Factors. *Annu Rev Stat Appl* **5**, 393-419 (2018).
26. D. Colquhoun, The False Positive Risk: A Proposal Concerning What to Do About p-Values. *Am Stat* **73**, 192-201 (2019).
27. D. J. Benjamin, J. O. Berger, Three Recommendations for Improving the Use of p-Values. *Am Stat* **73**, 186-191 (2019).