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 denounced because it leads to an excess of false positive results. Here simulations are used to illustrate and explain how unplanned incremental sampling causes excess false positives. In a parameter regime representative of practice in many research fields, however, simulations show that the inflation of the false positive rate is surprisingly modest. The effect on false positive rate is only half the story. What many researchers really care about is the effect of N-hacking on the likelihood that a positive result is a real effect: 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 would in fact increase reproducibility, counter to current rhetoric. To strictly control the false positive rate on the null hypothesis, the sampling plan (and all other study details) must be prespecified. But if this is not the primary concern, as in exploratory studies, collecting additional samples to resolve a borderline p value can confer previously unappreciated advantages for efficiency the positive predictive value of the generated hypotheses.

Background

There has been much concern in recent years concerning the lack of reproducibility of results in some scientific literatures (*1*). The call for improved education in statistics and greater transparency in reporting is justified and welcome. But if we apply rules by rote, we as a community risk throwing out a lot of babies (good data, promising leads) with the statistical bath water. Experiments in biology often require substantial financial resources, scientific talent, and use of animal 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 brought into question here is a cornerstone of null hypothesis significance testing: test exactly the predetermined sample size N, and then accept the verdict of the hypothesis test, whatever it is. Empirical scientists are accustomed to looking at data, so

simulation is an excellent way to gain intuitions about the implications of statistical methods. Here I simulate the questionable research practice of "N-hacking" – incrementally adding more samples after the fact whenever a preliminary result is "almost significant".

This study began with the intent of demonstrating the known dire consequences of this practice, but obtained an effect an order of magnitude smaller than previously reported (*2-5*). The discrepancy was traced to parameter choices: I had used parameters reflective of real-world practice in experimental biology, whereas published demonstrations had used unrealistic ones. After exploring a broad range of parameters bracketing most biology experiments, it emerged that in the relevant parameter regime for Biology, the elevation in false positive rate is quite modest and lawfully predictable. Moreover, the effects on reproducibility (PPV) – which have not been previously explored – turned out to be beneficial, not harmful. These results were both unexpected and robust. This parameter regime may a "special case" of no interest to the field of theoretical statistics, but it is the only case of interest to experimentalists.

These simulations were meant to describe what researchers in fact do, not to prescribe what they should do. The goal is not to dismiss concerns about sampling procedures, but rather to clarify them in order to better inform choices. Readers will gain working intuitions about why N-hacking is a problem, and how the magnitude and direction of the resulting bias depend on the details of decision heuristics. The results show that in an exploratory study, judicious sample incrementation can be a better option than either starting over from scratch or abandoning a hypothesis after obtaining a nearly-significant outcome. The results also motivate why formal sequential sampling protocols could be a better choice for biology studies that require confirmatory *p* values.

Results

These simulations can be taken to represent a large number of independent studies, each collecting separate samples to test a different hypothesis. All simulations were performed in MATLAB 2018a. Formal definitions of terms and symbols are summarized in a side box.

Part I. 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. I will call the observed false positive rate when the null hypothesis is true FP_0 ("FP null"), also known as the Type I Error Rate. I assume that the significance criterion α has been set in advance. By construction, the *t*-test produces false positives at a rate of exactly α, the significance threshold. The MATLAB code used for simulating the false positive rate on the null hypothesis (FP_0) can be found in (*6*), along with the numeric results of the all simulations described in Figures 2-4.

A cautionary scenario

Suppose 10,000 separate labs each ran a study with sample size N=8, where in every case there was no true effect to be found. If all used a criterion of $\alpha = 0.05$, we expect 500 false positive

results. But suppose all the labs that got "nonsignificant" outcomes reasoned that their studies were underpowered, and responded by adding four more data points to their sample and testing again, repeating this as necessary until either the result was significant or the sample size reached N=1000. The interim *p* values would fluctuate randomly as the sample sizes are grown (Figure 1a). In two of the cases shown (red and blue curves) the *p* value crossed the significance threshold (α=0.05, black line) by chance. Had these studies ended as soon as $p < \alpha$ and reported significant effects, these would represent excess false positives, above and beyond the 5% we intended to accept. Dashed curves show how these "*p* values" would have continued to evolve if sampling had continued.

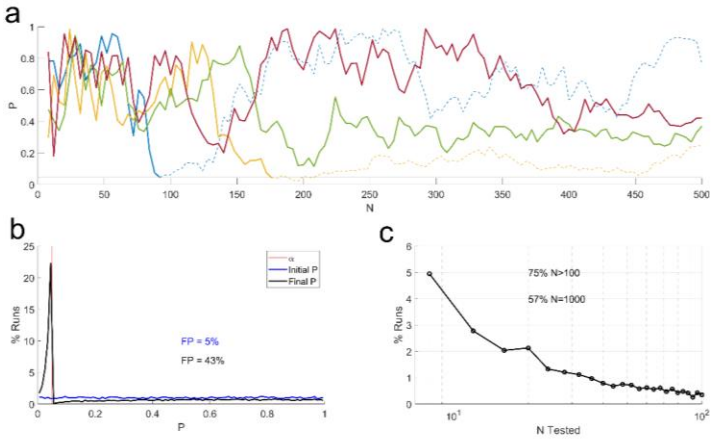

Figure 1. The problem with N-hacking. Simulation of experiments in which there was no true effect, starting with samples of size N=8. If the result was nonsignificant, a study added 4 more and retested, until either the result was significant or N=1000. **a.** Evolution of *p* values of four simulated experiments as N was increased. **b.** Distribution of initial and final *p* values of 10,000 such simulated experiments, in bins of width 0.01. **c.** Distribution of final sample sizes of the simulated experiments, based on counts of each discrete sample size.

In a simulation of 10,000 such experiments, the *p* values for the initial N=8 samples were uniformly distributed between 0 and 1 (Figure 1b, blue), with 495 cases (~5%) falling below 0.05 (red line). After N-hacking, there were 4262 false positives instead of the expected 500 (Figure 1B, black). Therefore, the final "p values" are not really p values – they do not reflect the probability of obtaining the result by chance if the null hypothesis is true. This alarming result has been pointed out by many others (*1-5*), and serves to illustrate that N-hacking can be a serious problem for anyone operating in this regime.

This scenario postulates extremely industrious researchers, however. Suppose the experimental units were mice. For the 5% of labs that obtained a false positive at the outset, the sample size was a reasonable N=8 mice. All other labs had larger final samples. Three quarters of the labs tested over 100 mice, and over half of the labs tested 1000 mice before giving up. This simulation also postulates extremely stubborn researchers: in 75% of the simulated runs, additional data were collected even after observing an interim "*p* value" in excess of 0.90. In my experience in experimental biology research, these choices are implausible.

A plausible scenario in experimental biology

Suppose instead that the sample size would be increased only for those tests that meet a criterion of "p close to α". Furthermore, suppose 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. I simulated such an Asymmetric N-increasing policy as follows: every time a comparison yielded a *p* value that was "almost significant", additional samples were added incrementally, 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. Results of such a policy are shown in Figure 2.

As expected, this Asymmetric N-increasing policy yielded an increase in the rate of false positives FP_0, and this was more severe as the eligibility window w increased (Figure 2a). 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 and 0.10, resulting in a realized false positive rate $FP_0 = 0.0625$ instead of the nominal 0.05. Following this policy resulted in a negligible increase in the sample size on average (Figure 2b). Note that the false positives due to multiple comparisons are included in these reported false positive rates, i.e. these are the *uncorrected* false positive rates.

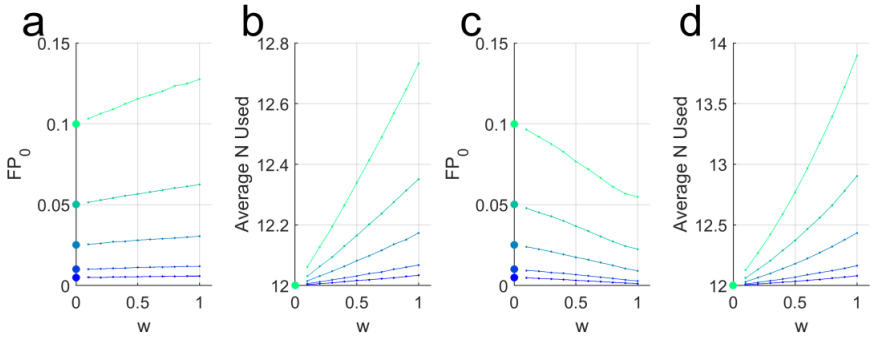

Figure 2. Effect of uncorrected selective sample-increasing on false positive rate FP_0. Results shown are for experiments with initial sample size $N_{init} = 12$, and sample increments $N_{incr} = 6$, and maximum sample size $N_{max} = 24$. Each point or symbol represents results from $M = 10^6$ simulated 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.

To many, it is counterintuitive that adding more observations could do anything but improve statistical rigor – more N is better, right? The main reason false positives are elevated is that experiments were chosen for incrementation 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 is that the rule is asymmetric: it challenges a preliminary result when p is just above threshold, but not when it is just below threshold. To demonstrate this point I also simulated a Symmetric N-increasing policy, in which incremental sample growth occurred whenever a p value was close, whether below or above α: $(1-w)\,\alpha \le p < (1+w)\,\alpha$. Making the policy symmetric more than overcomes the problem – it converts more false positives to true negatives than it converts true negatives to false positives, resulting in a net reduction in false positives (Figure 2c). This is because in addition to the effect noted above, the Symmetric 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 (Figure 2d).

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 some do (*7*), and my observations of real practice in biology concur with this. Therefore, the consequences of both policies will be explored further below.

Dependence on α and the eligibility window 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: $FP_0 = k\alpha$ (Figure 3a). The slopes of these lines are in turn an increasing function of the decision window w (Figure 3b, symbols). On the Symmetric policy, the dependence of FP on α is also linear (Figure 3c) and the slope k declines with w (Figure 3d).

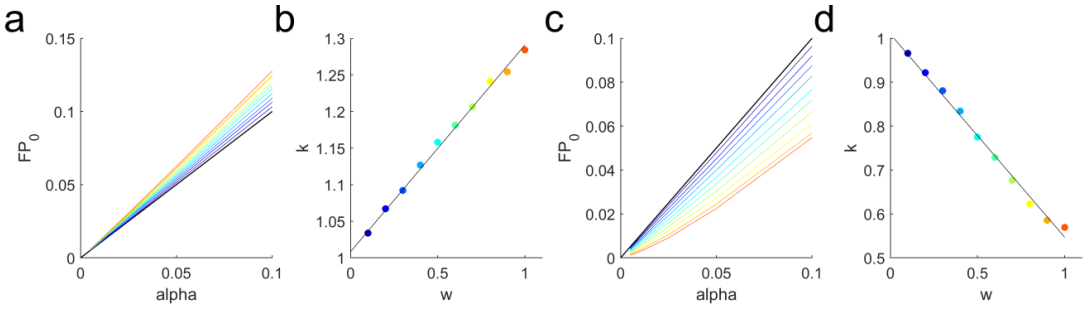

Figure 3. Dependence of false positive rate FP_0 on α and w. Results from simulations using $N_{init} = 12$, $N_{incr} = 6$, $N_{max} = 24$, with $M = 10^6$ simulated experiments per condition. **(a)** The realized false positive rate FP_0 for an Asymmetrical N-increasing policy when the null hypothesis 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 initial sample size N_{init} and increment size N_{incr}

Above I assumed an initial sample size of 12, adding 6 more samples at a time, up to a maximum of 24 samples. To determine if these results were a peculiarity of these assumptions, I repeated the simulations for N_{init} ranging from 2 to 128 initial sample points, adding N_{incr} ranging from 1 to N_{init} samples each time, and capping the maximum total sample size at $N_{max} = 256$. These assumptions more than bracket the range of realistic sample sizes and ad-hoc sample growth that would be commonly used in many experimental biology fields.

Results for the Asymmetric policy with $\alpha = 0.05, w = 0.4$ are shown in Figure 4a. The false positive rate FP_0 is always elevated compared to α (black line), but 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).

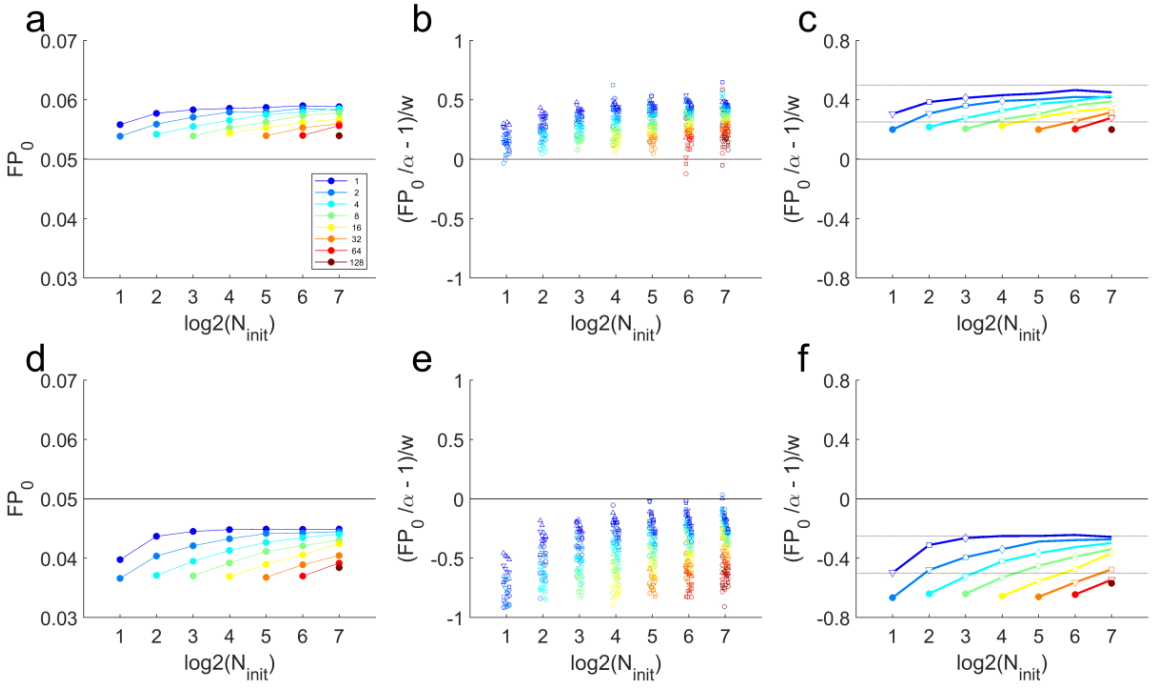

Figure 4. Dependence of false positive rate FP_0 on the initial sample size and sample increment. Each symbol represents the result from 10^6 experiments simulated with a ground truth of no effect. **a.** The realized false positive rate FP_0 of the Asymmetric N-increasing policy, as a function of $\log_2 N_{init}$ (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. For the fixed-N policy by definition $FP_0 = \alpha$, so this equation reduces to 0 (black line). 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. **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 side box). 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).

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 summarize the results for all combinations of of α and w by linearly scaling them (Figure 4b-c). 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 or N_{init} is small (Figure 4d-f). In summary, the effect of uncorrected incremental sampling on the false positive rate is real, but it is modest in size and lawfully related to a handful of parameters.

The "p value" obtained after unplanned incremental sampling is still not a true p value. A number of methods are available for planned incremental sampling or p value correction (*4, 8-13*). If the sampling policy was not set in advance, however, a correction of the p value can only be an estimate, because you can never truly know (or prove) what you would have done if the data had been otherwise. The point here is that in exploratory studies, if one limits unplanned incrementation to cases where the initial p value is rather close to α, the bias introduced by incrementation is not very large. For example, if one's cutoff for ad hoc sample incrementation is $p < 2\alpha$ (corresponding to $w = 1$), the false positive rate will never be elevated by more than a factor of 1.5 (see Appendix 2). Therefore, if one does a Bonferroni correction for the multiple comparisons involved (a factor of 2 or more, depending on how many times one incremented) one will have more than corrected for this deviation from the plan.

Part II. Trade-off between statistical power and positive predictive value

So far these simulations still make another unrealistic assumption: that the null hypothesis is always true. In real research, presumably at least some studies testing for effects that in reality do exist. 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 researchers' 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 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 (*14*).

To explore this in simulations, one must simulate some experiments with no effect (as above) and other experiments with real effects. In simulations, we know the ground truth about which experiments had real effects, so we can directly measure two important quantities: (1) the sensitivity or *power*, which is defined as the fraction of real effects for which the null hypothesis is rejected; and (2) the selectivity, or positive predictive value (*PPV*), which is defined as the fraction of all positive results that are real effects (as opposed to false positives).

		Result of hypothesis test	
		not significant	significant
Truth	no effect	a. True Negative	b. False Positive
	real effect	c. False Negative	d. True Positive
$Power = \frac{d}{c+d}$ $\quad PPV = \frac{d}{b+d}$			

The sensitivity-selectivity trade-off

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. The remaining 99% of experiments had no real effect. The fixed-N policy was compared to either an Asymmetric or Symmetric N-increasing policy.

First it is helpful to recall that in the standard fixed-N policy there is always 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 (e.g., Figure 5a, any red curve slopes downward). By varying α one can define a curve for the sensitivity-selectivity trade-off, which 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 (Figure 5a, compare any two red curves). Drawing on this intuition, 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 α for any desired PPV and achieve higher Power; or choose α for any desired Power and achieve higher PPV, compared to any curve that lies below it.

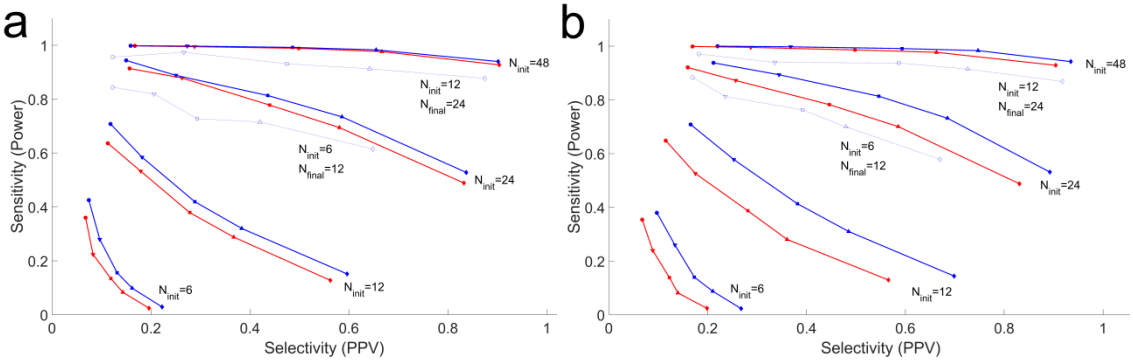

Figure 5. Trade-off between selectivity and sensitivity. (a) Realized selectivity vs. sensitivity in simulations with effect size 1σ, prior effect probability 0.01, and N_{init}=6, 12, 24 or 48 (four curves of each color). Symbols indicate α (○=0.001, ▽=0.005, □=0.01, △=0.025, ◊=0.05). Text labels for pairs of curves indicate initial sample size N_{init}. For every combination of these parameters a total of $M = 10^4/\alpha$ experiments were simulated. 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).

The curves for the standard fixed-N policy (red curves, Figure 5) thus provide the benchmark to which other sampling policies may be compared. An example Asymmetric N-increasing policy is shown (blue curves, Figure 5a). Because samples were added to only a few experiments, 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 (c.f. Figure 2b). 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 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 (the majority, which had a final 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).

The Symmetric N-increasing policy was superior to the fixed-N policy in every way (Figure 5b, compare red to blue), as well as beating the Asymmetric policy (compare blue curves in Figure 5a vs. 5b). Even using the same choice of α the Symmetric policy increased both selectivity and sensitivity relative to fixed-N for all conditions tested.

These simulations demonstrate that for an effect size of $E = 1\sigma$ and prior probability of 0.01, N-hacking is a win-win scenario. Although the absolute numbers depend on the effect size E and fraction of experiments that had real effects $P(H_1)$, the relationships between the curves were the same for effect sizes ranging from $E = 0.5$ to 2 and prior $P(H_1)$ ranging from 0.001 to 0.1 (not shown). 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} using $N_{incr} = N_{init}$ (not shown). In real experiments, E and $P(H_1)$ are not known, but this doesn't prevent us from concluding that regardless of their values, N-hacking in this regime would *improve* reproducibility.

Dependence on the eligibility window w

In Part I, I showed that if one only adds samples when p is rather close to α, the false positive

rate FP_0 is only moderately elevated (Figure 2), but if one used a larger eligibility window w, the false positive rate could be quite high among experiments with no real effect (Figure 1). Does the benefit of N-hacking fall apart when w gets large? To test this, I further simulated results of the Asymmetric policy under this condition, for w ranging from 0.2 to 10, also varying α to define the power-PPV curves. As w increases, these curves move up and to the right (Figure 6, top row). This implies that even if one uses very loose criteria for adding samples, N-hacking has some benefits.

For a fixed choice of α, increasing w always increases sensitivity (warm colors are above cool colors along any curve, Figure 6 bottom row). This makes sense: the more willing one is to add a few more samples, the more false negatives one can rescue to true positives.

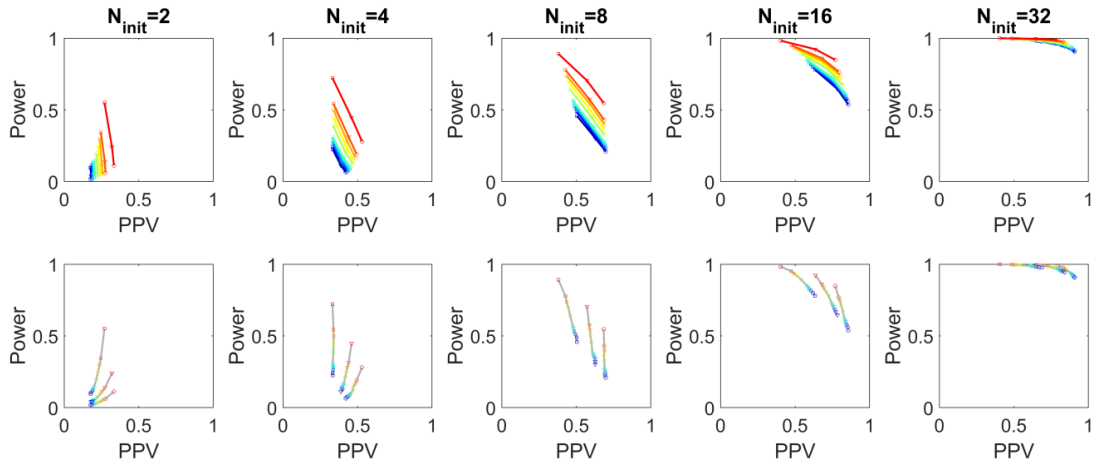

Figure 6. Asymmetric N-increasing policies improve statistical inference even when w is large. Realized selectivity vs. sensitivity in simulations with effect size $E = 1\sigma$ prior effect probability $P(H_1) = 0.10$, with N_{init} as indicated on column title, $N_{incr} = 1$, $N_{max} = 50$. Each 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.

For larger sample sizes, however, uncorrected N-hacking (holding α constant) reduces positive predictive value (e.g. $N_{init} = 16$, gray curves slant to the left) compared to a fixed-N policy (dark blue symbols). Nevertheless, the trade-off between PPV and Power is 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 statistical power of 0.46. Asymmetric N-hacking with a window of $w = 5$ means that more samples would be added for any interim test result of $0.05 < p < 0.25$. Without any correction for incremental sampling or multiple comparisons (as shown), this would erode the 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. 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.

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, if the null hypothesis is true. This has been said many times before, but most writers warn that this practice will lead to extremely high false positive rates (*1-5*). We can replicate those results if we use the same assumptions: that an experimentalist would add more samples after obtaining a non-significant result no matter how far from α the *p* value was, and would continue adding samples until N is quite large (Figure 1). 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 bounded.

The magnitude of 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}. These 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 (Appendix 2). While this cannot be used to formally correct the p value, it could provide useful guidance to the researcher in an exploratory study.

Further simulations demonstrated that under many conditions this type of N-hacking is superior to a fixed-N policy in the sense that it increases the positive predictive value (PPV) achievable for any given statistical power, compared to studies that strictly adhere to the initially planned N. This has not been previously noted, and was unexpected. For 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 as often claimed. 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 is applied for sequential sampling or multiple comparisons.

With the noble motivation of improving reproducibility, researchers are now being told that if they have obtained a non-significant finding with a p value just above α, they must never add more samples to their data set to improve statistical power. They must either run a completely independent larger-N replication, or fail to reject the null hypothesis (which generally means relegation to the file drawer, in the current publishing climate). To dissuade researchers from unplanned sample incrementation, multiple didactic articles have shown that the resulting false positive rate would be wildly inflated (c.f. Figure 1). These demonstrations were unrealistic and misleading. To make informed choices, researchers need more relevant and nuanced information about the trade-offs they must negotiate.

So, is N-hacking ever OK?

Adding samples after completing the planned experiment violates a basic premise of null hypothesis significance testing (NHST), and forfeits *control of* the Type I Error rate. But if the

goal is to generate hypotheses that are likely to be reproducible, many researchers might validly be willing to abandon having an exact p value in exchange for reducing the risk of false negatives, improving the positive predictive value, and conserving time, animal lives, and other resources. In an explicitly exploratory study, some statisticians might concede that unplanned sample incrementation is not even N-hacking.

For researchers conducting transparently exploratory studies, then, these simulations could inform better informal decision heuristics about sample growth. An exploratory study should be labeled as such, disclose that sample incrementation occurred, report the interim N and p values, and describe their decision heuristics as honestly as possible. The simulations presented here would help a reader interpret the implications of those choices.

But if an exact p value is required, as in a confirmatory study, no deviation from the prospective experimental design is OK, including N-hacking. That doesn't rule out incremental sampling, however. It would not be N-hacking if the incrementation policy were committed to advance, because pre-specification makes it possible to determine the results expected when the null hypothesis is true, at least by simulation.

If one is going to pre-specify an incremental sampling plan, however, one could probably do better than the ad-hoc heuristics simulated here, which were meant merely to describe what I believe to be common lab practices. It is beyond the present scope to explain and compare sequential sampling methods, and others have ably done so (*15, 16*). Here I will just provide a brief indication of some options.

One option is a phased study. For example, one could prespecify a 2-phase protocol with an initial phase of $N = 16$ and $\alpha = 0.10$, followed (if a "significant" effect is found in Phase I) by a second phase with $N = 33$ and $\alpha = 0.01$. Compare this to a Symmetric N-increasing policy with $N_{init} = 16, N_{incr} = 1,\ \alpha = 0.05, w = 1, N_{max} = 128$. In both cases, additional data will be collected whenever the initial sample yields $p < 0.10$. If prespecified (and no other deviations from the research plan occurred) both would have strictly interpretable p values. In a simulation with an effect of size 0.5 SD and a prior probability of 0.1 (10^6 runs) these both had an average sample size of about 20, a statistical power of about 26%, a PPV of about 92%.

Another option is Wald's Sequential Probability Ratio Test (*17*), which has been proven to be optimal in some respects. In Wald's method, one sets in advance a threshold a to accept and another threshold b to reject the null hypothesis. Then one computes a test statistic S after each new sample point is added (the cumulative log likelihood ratio of the alternatives). If $S \leq a$ the null hypothesis is accepted, if $S \geq b$ the alternative is accepted, and if $a < S < b$, one continues sampling. The thresholds a and b can be set analytically to obtain the desired statistical power β and false positive rate α. Superficially, the N-incrementing policies simulated here resemble Wald's method in that there are two thresholds and an indeterminate range between them when sampling continues, but Wald selects these two thresholds in an optimal way. A downside of Wald is that one must commit to sampling until one or the other threshold is crossed, which puts one at risk of having to test a very large N.

A third option is Bayesian Sequential Sampling. This 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 (*18*). This is also closely related to Wald's method and the drift diffusion model (DDM) of decision-making (*19*), and does not require knowledge of the prior probability.

Broader Implications

In the effort to promote rigor in science, we need to question "questionable" research practices more deeply. Some may be inevitably and severely misleading (*20*). Others may have small effects, or only in specific circumstances. The potential for abuse does not establish actual abuse; sometimes the same practice (e.g. "unplanned sample incrementation") could either reduce or increase reliability of research, depending on exactly how it is deployed. A more realistic and nuanced exploration is far more instructive for researchers, and can lead to more useful suggestions for improved practice of science.

Many experimental studies in Biology are exploratory, involving not only unplanned incremental sampling but also iterative revisions of the experimental methods, analysis methods, and hypotheses. In such studies one cannot obtain a confirmatory p value, even if the sampling plan is prespecified. However, this flexibility may be essential to the success of the research in terms of making valid, novel discoveries efficiently. Therefore, science reforms that seek to turn all research projects into confirmatory research could backfire. Instead, we in Biology need to be more open about labeling exploratory studies as such (including refraining from reporting p values or telling null-hypothesis-testing stories), and work harder to articulate the methods and heuristics we routinely employ to ensure scientific rigor in the context of exploratory studies.

Acknowledgements

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

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	In a fixed-N policy: the number of samples in each group In an incrementing policy: N_{init} Initial sample size N_{incr} Number of samples added each time N_{max} Maximum sample size before stopping
p Value returned by statistical test	The fraction of experiments in which one would observe a difference at least as great as the observed difference, if in fact H_0 were true.
α Significance criterion	A criterion to reject H_0 only if $p < \alpha$
w Eligibility window	In these incrementing policies, defines how close to α a p value must be to add samples
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 $FP_0 = p$ 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: $E \equiv \frac{\lvert\mu_A - \mu_B\rvert}{\sigma}$
$P(H_1)$ Prior probability of an effect	The probability H_0 is false, before considering the data. In simulations, this fraction of 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 prior $P(H_1)$ and effect size 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 prior $P(H_1)$ and effect size E Related to "False Positive Risk"(*21, 22*): $FPR = 1 - PPV$

Appendix 1: Extensions and limitations of these results

These simulations used a normal distribution for the source distributions and an independent sample *t*-test as the 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.

I have simulated the practice of unplanned sample incrementation after computing a *p* value on the initially planned sample. But even if no interim statistical tests are performed, the same issues arise. For example, deciding whether to collect more data depending on the effect size seen in the initial data, or based on visual inspection of scatter plots, is also N-hacking.

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. Testing a wide range of values, it was possible to draw general conclusions about the *direction* of the effect of N-hacking, entirely on frequentist grounds. In "underpowered" conditions (low N, stringent α, small effect size, low prior probability), whatever the PPV would have been 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, however.

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. That situation is not considered here.

Numerical simulations and graphs are easy for experimentalists to understand, because they present the expectations of the hypothetical scenario in terms directly comparable to data. But I have not attempted an analytic treatment that would allow for a proof or specification of the conditions under which these results obtain.

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. When interpreting any single experiment, however, one should take into account the specific p values that were obtained (a "p-equals" rather than "p-less-than" approch)(*21*).

As others have noted, "chasing significance", such as by N-hacking, may be incentivized by 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 (*23-25, 26*). But since N-hacking biases the p value itself, the issues explored here would arise even if no decision threshold were used.

Many statisticians advocate supplementing reported p values with some other statistical measure such as the Odds Ratio (*21, 27*), Bayes Factor(*18, 28*), the False Positive Risk (1-PPV) (*22*), or a non-Bayesian bound on the Bayes Factor (*29*). Some of these measures do not make any assumptions about how data were collected, and in this respect are immune to concerns about N-hacking.

Appendix 2: A conservative bound?

If the simulated decision rules were implemented as strict policies, the simulated data show the following inequalities (dotted lines, Figure 4 c,f):

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

These are loose bounds (in many conditions the false positive rate falls well below this value), but have the virtue of being easy to calculate. For example: an Asymmetric N-increasing policy with $w = 0.4$, $N_{init} = 10$, $N_{incr} = 10$, $N_{max} = 50$, would have an estimated $FP_0 < 0.0550$ 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 $w = 19$ for $N_{init} = 2$ to 128 with $N_{max} = 256$ and still did not exceed this empirical bound (not shown). The MATLAB code provided in (*6*) can be used to simulate the false positive rate for other parameter combinations.

In principle, these inequalities could be used to estimate a bound on the false positive rate or estimate a corrected *p* value after unplanned sample incrementation if the heuristic decision rule can be articulated. This estimate will be conservative if one assumes an asymmetric policy, a larger window *w* than one thinks one would ever increment, and a maximum sample size N_{max} larger than one thinks one would ever collect. But this will still be only an estimate, unless the decision policy is fixed in advance.

References

1. 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).
2. C. Albers, The problem with unadjusted multiple and sequential statistical testing. *Nat Commun* **10**, 1921 (2019).
3. D. Szucs, A Tutorial on Hunting Statistical Significance by Chasing N. *Front Psychol* **7**, 1444 (2016).
4. E. Schott, M. Rhemtulla, K. Byers-Heinlein, Should I test more babies? Solutions for transparent data peeking. *Infant Behav Dev* **54**, 166-176 (2019).
5. H. J. Motulsky, Common misconceptions about data analysis and statistics. *Naunyn Schmiedebergs Arch Pharmacol* **387**, 1017-1023 (2014).
6. P. Reinagel, N-hacking simulation: In silico experiments exploring the effect of a questionable research practice on the reliability of research results. . *CodeOcean* **[Source Code]**, (2020).
7. 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).
8. D. L. DeMets, G. Lan, The alpha spending function approach to interim data analyses. *Cancer Treat Res* **75**, 1-27 (1995).
9. D. Lakens, Performing high-powered studies efficiently with sequential analyses. *Eur J Soc Psychol* **44**, 701-710 (2014).
10. B. J. Sagarin, J. K. Ambler, E. M. Lee, An Ethical Approach to Peeking at Data. *Perspect Psychol Sci* **9**, 293-304 (2014).
11. 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).
12. R. W. Frick, A better stopping rule for conventional statistical tests. *Behav Res Meth Ins C* **30**, 690-697 (1998).
13. P. Grünwald, R. de Heide, W. Koolen, Safe Testing. *arXiv*, 1906.07801 (2019).
14. D. Colquhoun, An investigation of the false discovery rate and the misinterpretation of p-values. *R Soc Open Sci* **1**, 140216 (2014).
15. 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.
16. D. Siegmund, *Sequential analysis : tests and confidence intervals*. Springer series in statistics (Springer-Verlag, New York, 1985), pp. xi, 272 p.
17. A. Wald, *Sequential analysis*. Wiley mathematical statistics series (J. Wiley & sons, New York, 1947), pp. xii, 212 p.
18. S. N. Goodman, Toward evidence-based medical statistics. 2: The Bayes factor. *Ann Intern Med* **130**, 1005-1013 (1999).
19. J. I. Gold, M. N. Shadlen, Neural computations that underlie decisions about sensory stimuli. *Trends Cogn Sci* **5**, 10-16 (2001).
20. E. Vul, C. Harris, P. Winkielman, H. Pashler, Puzzlingly High Correlations in fMRI Studies of Emotion, Personality, and Social Cognition. *Perspect Psychol Sci* **4**, 274-290 (2009).
21. D. Colquhoun, The reproducibility of research and the misinterpretation of p-values. *R Soc Open Sci* **4**, 171085 (2017).
22. D. Colquhoun, The False Positive Risk: A Proposal Concerning What to Do About p-Values. *Am Stat* **73**, 192-201 (2019).

23. R. L. Wasserstein, A. L. Schirm, N. A. Lazar, Moving to a World Beyond "p < 0.05". *Am Stat* **73**, 1-19 (2019).
24. R. S. Nickerson, Null hypothesis significance testing: a review of an old and continuing controversy. *Psychol Methods* **5**, 241-301 (2000).
25. S. N. Goodman, Toward evidence-based medical statistics. 1: The P value fallacy. *Ann Intern Med* **130**, 995-1004 (1999).
26. S. N. Goodman, Of P-values and Bayes: a modest proposal. *Epidemiology* **12**, 295-297 (2001).
27. 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). *J Math Psychol* **89**, 98-98 (2019).
28. L. Held, M. Ott, On p-Values and Bayes Factors. *Annu Rev Stat Appl* **5**, 393-419 (2018).
29. D. J. Benjamin, J. O. Berger, Three Recommendations for Improving the Use of p-Values. *Am Stat* **73**, 186-191 (2019).