Stopping Rule Sampling Procedures do not Jeopardize Generalizability

Timo von Oertzen[1], Aaron Peikert[2], Hannes Diemerling[1,2], Tina Braun[3], and Timothy R. Brick[4]

[1]Thomas Bayes Institute, Berlin

[2]Humboldt University Berlin

[3]Charlotte-Fresenius University

[4]Pennsylvania State University, University Park, PA, USA

Author Note

Correspondence concerning this article should be addressed to Timo von Oertzen, Thomas Bayes Institute, Berlin. Email: timo.vonoertzen@thomasbayesinstitute.org. Contact information of the other authors: aaron.peikert@mpib-berlin.mpg.de, hannes.diemerling@gmail.com, tina.braun@charlotte-fresenius-uni.de, tbrick@psu.edu

Abstract

In this paper, we challenge the long-held belief that inferences based on early stopping designs lead to poor generalizability. We show by mathematical analysis and simulation that two identical results in terms of sample size and test statistics generalize equally well no matter whether or not a stopping rule has been applied.

The problem of early stopping has long been a challenge in the behavioral and health sciences. In short, adaptive stopping criteria allow increased efficiency and reduces potential harm. For example, if a study shows evidence that an experimental treatment is effective after only half of the planned participants have been run, failing to stop the study early relegates more participants to a less-effective placebo group. However, the literature on null hypothesis significance testing has long held that this type of adaptive stopping design leads to decisions with poor generalizability. We argue that in fact this error is due to an overly precise focus on local decisionmaking. We present a mathematical definition of generalizability that permits direct comparisons of generalizability between models, and demonstrate that the generalizability of even frequentist results is independent of early-stopping-based sampling.

These results throw into question a large swath of literature on the appropriate handling of multiple testing and adaptive sampling in the frequentist framework, and highlight the inherent challenges of inference based on local evidence such as p values.

Keywords: Generalization, stopping Rules, Bayesian Statistics

Stopping Rule Sampling Procedures do not Jeopardize Generalizability

The question of early stopping has been a matter of significant discussion over many decades (see, for example, (**OBrien_Multiple_1979**)). In an early stopping paradigm, a series of decision points is defined over the course of the design. A sequential testing series is then enacted, whereby at each decision point some criterion condition is computed and compared to a cutoff. A simple example of a sequential test for early stopping in the frequentist framework is a simple null hypothesis significance test, performed after every participant's data is collected. In the simplest form of this process, a p value is computed after each participant. The p value is compared to a pre-decided α value, in this very simple example a static value of .05. If the p value is larger than the α, another participant is recruited and their results recorded until either a the full pre-planned sample size is collected or one of the tests results in a p value less than .05. In essence, the conclusion of the study then rests on whether such a low p value was collected before the preplanned sample size (in which case the null hypothesis would be rejected), or not (in which case no conclusion would be drawn).

Note that not all early stopping rules in the frequentist framework base their stopping rules on the p value. Others focus on process of adaptation (Chow & Chang, 2008), or on the selection of stopping criteria (**KenKelly**) that do not bias Type-I error rates.

The benefits of early stopping are not generally in dispute. In brief, it is more efficient to stop a study as soon as enough data is available to reach a conclusion. This efficiency is not only a matter of participant costs and laboratory fees; it also requires fewer participants, and therefore carries potential ethical burdens. In a randomized clinical trial where the experimental treatment might be life-saving, withholding that treatment from additional placebo-group participants than necessary may result in preventable deaths; the same is true if the treatment is ineffective and experimental participants are denied standard-of-care treatment (**.**)

The challenge to early stopping was identified at least as early as **Wald_method_1945** by **Wald_method_1945**, although that paper cites discussions from an even earlier paper (**Dodge_method_1929**). More recent work has tackled this problem as recently as (Grieve, 2024), suggesting that this problem is ages old. Specifically, the problem lies in the increase in Type-I error rates from early stopping if a statistically significant p value was obtained before the completion of the study. To summarize briefly, repeated comparisons to the computed p value lead to an accumulation of Type-I error rates. Over the decades in between, a large number of frequentist approaches have attempted to circumvent these effects, most commonly by sacrificing Type-II error rates to improve Type-I error (see, e.g. O'Brien & Fleming, 1979; Pocock, 1977).

In the end, this has led to the general consensus that applying a stopping rule without appropriate correction constitutes sloppiness at best and scientific malpractice at worst (Fanelli, 2009). Researchers including renowned methodologists, and the authors of this article as well, are therefore tempted to assume that a low p value obtained by a stopping rule sampling procedure is less valuable than the same collected from complete data. Most often this is expressed by asserting that the stopping rule results are 'less generalizable' (used without formal definition) than a result obtained in a classical procedure, even if the sample sizes ended up the same. The intuition behind the term 'less generalizable' is that the same p value reported implies a lower chance to replicate; that is, that it is less likely the same results will be found in the future in the same population when using new data samples.

This intuition is so strong that it has been presented as a death knell for frequentist statistics; for example **Venderkerchove_2014<empty citation>** suggested that the seeming inability to stop early was the most important characteristic of any statistical reasoning engine. It is so broad that it is often directed at similar sampling approaches, such as "sneak-previewing", wherein the study is stopped if the first few participants do not seem favorable (see, e.g. Armitage et al., 1969), and so pervasive that a number of scientists

have developed approaches (sometimes put forward as 'laws of good practice' (Simmons et al., 2011)) that aim to avoid even the suspicion that an early stopping rule has been applied. For example, *a priori* registered reports with a pre-registration of the targeted sample size has been suggested as a means to avoid the issue (Nosek & Lakens, 2014).[1]

Decades of work have focused on appropriate correction methods to retain appropriate Type-I error rates (see Shaffer, 1995, for an overview). The most popular of these corrections (in the frequentist framework) focus on adjustments to the *alpha* value or to the bounds of the confidence interval to account for the challenge of repeated testing(see, e.g. O'Brien & Fleming, 1979; Pocock, 1977). These approaches sacrifice statistical power–that is, allow an increased Type-II error rate–in exchange for this control. This reduced power results in yet another type of inefficiency, because it means that either pre-planned sample sizes must be increased to enable the greater adaptivity, or that studies with lower power may not find a statistically significant result because of the corrections required for a given stopping rule.

The logic of this intuition is as follows: Assume we have some sample of size N on which we have run a test. A lower p value derived from that sample is related to a more extreme estimate for some other statistic relative to the test's expectations (e.g. the standard error). For example, a lower p value might result from a larger regression estimate or a lower -2 Log Likelihood value. If all else is held equal and the stopping rule reduces the p value artificially, it must do so by either providing an incorrect estimate of the estimate or biasing the test (e.g. by reducing the standard error). In either case, the stopping rule sample is misrepresenting the results of the test, which should lead to erroneous expectations about replication.

We argue that the notion that stopping rules or other sampling procedures threaten

---

[1] Oddly, although it is a noted concern about modern science, the "file-drawer effect"–that is, the stopping rule wherein one discards a study unpublished if a pre-selected α value (e.g. $p < .05$) is not reached by the time the pre-planned N is collected, is rarely expressed as a biasing sampling technique (Rosenthal, 1979).

generalizability in the sense that reports made under those rules do not allow the same (or even any) conclusion on a population is incorrect. This seems surprising considering the existent alpha inflation, or bias of estimates, obtained and frequently shown in the literature (Ioannidis, 2005). This logic makes an important error, however. In particular, the assumption that all else is held equal is incorrect, because in one case the data set has met the early stopping criterion. For any fixed sample size, the fact that a favorable result is apparent earlier in the study implies a more favorable prior in the next study. These two factors cancel each other out precisely. While this perfect cancellation may not seem obvious in frequentist language, it is obvious in Bayesian language: if at some sample size the results are reported based on some condition and the likelihood of those results are combined with the original prior distribution, then the reporting condition is already accounted for precisely in the posterior, and hence induces no changes.

In this article, we first introduce a formal definition for *equal generalizability* of two methods in the sense that the results from these methods allow the same expectation for a future sample, and thus for the population. We also formally define sampling procedures and show that if sufficient statistics (e.g., the data set itself) is reported for a parameter, then all sampling procedures under this definition generalize equally well. We then demonstrate this using a simulation on a realistic illustrative example, where clinicians want to establish whether a therapy is effective. We examine both generalizability and frequentist Type-I and Type-II error rates (sometimes referred to as *operating characteristics*). Given the somewhat counterintuitive nature of the results for many researchers, we will mathematically discuss the fallacies that lead to our wrong expectations and the seeming contradictions in the results, and close with a final discussion.

Mathematical Background

Why Most Sampling Procedures Generalize

To investigate the generalizability of data sampling procedures, we use the following definition for a sampling procedure:

Definition 1. Let X be a random variable over a set Ω. A *sampling procedure* is a map $p : \Omega^{\infty} \rightarrow \Omega^{*}$, where Ω^{*} is the set of all finite samples. The sampling procedure has a condition on X^{*} that can be true or false. For the smallest n that satisfies the condition, the procedure returns $(x_1, ..., x_n)$; we say the procedure reports this result. If there is no such n, the procedure returns an empty data set. In this case, we say the procedure *ignores* the sample. We write $p(X)$ to denote the report of a sampling procedure p on an infinite iid draws of X.

This definition includes most common strategies to sample data. The standard procedure of sampling 100 participants would be a single condition that tests whether $n = 100$ and then reports the data. A procedure collects 100 participants but only reports the results when they are favorable (e.g., significant for a specific test) would have a condition that tests whether $n = 100$ and the result is favorable; if the condition is true, it reports this sample, otherwise it ignores the sample (because higher n cannot satisfy the condition). A stopping rule procedure that continues testing until it finds a statistically significant results would have a condition that regardless of n reports as soon as the test is significant. A stopping rule procedure with some correction or set of corrections on the term 'statistically significant' (e.g., by requiring a more strict *alpha*) would include these corrections in the condition. Both may ignore the sample if the stopping rule is never satisfied. A procedure that samples until it finds a significant result, but at most 100 participants, would combine the $n = 100$ condition with the stopping rule. A procedure that 'sneak-previews' data by first collecting a few participant, and if the results are promising on these few participants, continues collecting up to a usual number of participants, would be modeled by a condition that on $n = 100$ participants checks the first few (e.g., the first 10) participants and only reports when the results were favorable on those (of course, in practice one would stop sampling at 10 participants if the condition later will not be satisfied).

In this definition, all data collected for a study are reported or discarded together.

That is, a procedure which selectively discards records that show results inconsistent with the hypothesis is not a valid sampling procedure in this definition. This requirement is discussed formally in Corollary 5 and in the Sections below.

Our primary question of interest is whether two sampling procedures differ in the degree to which conclusions can be made from the reported data set to the general population, typically termed generalizability. We define the property that two sampling procedures generalize equally well as follows:

Definition 2. Let X be as before and p_1 and p_2 be two sampling procedures. We say that p_1 and p_2 *generalize equally* if for all $n \in \mathbb{N}$, $n \geq 1$, and $x = (x_1, ..., x_n) \in \Omega^n$,

$$X|(p_1(X) = x) = X|(p_2(X) = x) \tag{1}$$

So two procedures generalize equally well if, given that they report the same data (and in particular report at all), allow the same conclusions about the population random variable X. For simplicity, we just say that a procedure *generalizes* if it generalizes equally well with the 'vanilla' procedure that just samples n participants.

Then one concludes

Theorem 3. *All sample procedures generalize equally well.*

Proof. Let C be the condition in the sampling procedure p. For any $n \in \mathbb{N}$ and non-empty data set $x \in \Omega^n$, we write $C(x)$ as the event that the condition is true on x, but not on any subset of x. That is, if the procedure p reports a data set x, that means that the first n samples were x and $C(x)$ is true , formally $[p(X) = x] = [(X^n = x) \cap C(x)]$. Since $C(x)$ is computed determinstically from x,

$$[X|(p(X) = x)] = [X|(C \cap (X^n = x))] = [X|(X^n = x)] \tag{2}$$

which is the distribution of X when knowing a fixed sample of size n. □

Note that the distribution of X given a statistic $S(x)$ instead of x itself may be different under two sampling procedures; that is, if an article reports only non-sufficient

statistics, then the stopping rule may make a difference. This includes cases where S is a subsample of x, potentially chosen in a selective way. However, if S is a sufficient statistic for a parameter θ, then all relevant information is reported for this parameter; we say that a sampling procedure generalizes with respect to S and θ, formally

Definition 4. Let X as before and p_1 and p_2 be two sampling procedures. We say that p_1 and p_2 *generalize equally well with respect to a statistic S and a parameter θ* if for any value s of the statistic, $n \in \mathbb{N}$, $n \geq 1$, and $x = (x_1, ..., x_n) \in \Omega^n$,

$$[\theta|(S(p_1(X)) = s)] = [\theta|(S(p_2(X)) = s)] \tag{3}$$

Again, we say a procedure *generalizes with respect to S and θ* if it generlizes equally well with a fixed-size sample of n draws.

The following corollary states that if we report S from a study with a stopping rule, and S is sufficient for a parameter θ, then each stopping rule gives the same information about θ:

Corollary 5. *If S is a sufficient statistic for a parameter θ, then all sample procedures p generalize with respect to S and θ.*

Proof. As S is sufficient, the distribution of θ given $S(x)$ is independent of the data x, that is, $[\theta|S(x)] = [\theta|S(x), X^n = x] = [\theta|X^n = x]$ (the latter equality since $S(x)$ is computed deterministically from x). Again, let C be the condition in the sampling procedure p and $C(x)$ the event that the condition is true on x, but not on any subset of x. We have

$$\begin{aligned} [\theta|(p(X) = x)] &= [\theta|(C(x) \cap X^n = x)] &(4)\\ &= [\theta|(X^n = x \cap (S(x) = s)] = [\theta|(S(p(X)) = s)] &(5) \end{aligned}$$

□

In empirical research, researchers are encouraged to (1) make their whole data set available, (2) report average and sample covariance matrix of a set of variables they assume

to be normally distributed, or (3) report the maximum likelihood estimate of the parameters they estimate in a model, most frequently a Structural Equation Model. While (1) falls under Theorem 3 and therefore always generalizes, (2) and (3) are covered by Corollary 5 when the parameters of the normal distribution, or the parameters of the model, are researched. So in practically relevant instances, the sampling procedure is irrelevant for the generalizability of the results. Whether the researcher 'takes a peek' at the first few participants to decide whether to continue sampling, or sample until they reach a favorable result, or any other stopping rule, when reported the results are equally valid to draw inference on.

What Generalizability Does Not Imply

Note the following: Firstly, if S is *not* sufficient for a parameter of interest, than the sampling procedure does matter. This is for example the case if S just sub-samples some of the data, that is, if a researcher discards some participants because they are not favorable, or if the sampling procedure ignores a data set. In particular, even though reported sufficient statistics generalize equally well independent of the sampling method used, the 'file drawer effect' still exists. The file drawer effect describes the bias from ignoring some studies when combining results (e.g., in a meta-analysis or when investigating two dependent hypotheses in two different studies) because these studies have not been reported (e.g., because of p values that were not, by some criterion, "significant"). The file drawer problem is independent of the problem of sampling methods.

Secondly, the fact that any particular result allows the same conclusions about the population does not mean that specificity or sensitivity of the test are not changed; in fact, often specificity or sensitivity do change. For example, assume a sampling procedure takes a fixed n of samples and only reports them if $p < 0.05$, otherwise ignores the sample. Then for zero true effect size, their α error is of course inflated to 100%, and the β error is reduced to zero. Nevertheless, if they do report a specific outcome, then this outcome is as generalizable as a report would be from a researcher that reached that outcome on his first

try with a fixed sample. This holds both if the whole data set is reported as well as if a sufficient statistic is reported, as for example the average and sample variance of a normally distributed variable. That is, these findings are not in contradiction of the literature about α-inflation from early stopping rules; it is rather that the very fact that the stopping rule has been met precisely cancels this α inflation's effect on generalizability.

Finally, equal generalizability means that we can expect the same future outcome when the same data has been reported. If that is not the case, for example, if one sampling procedure reports fewer participants than another, the interpretation of the result will also differ; the precision of the lower-n estimate will naturally be lower than that of the higher-n estimate.

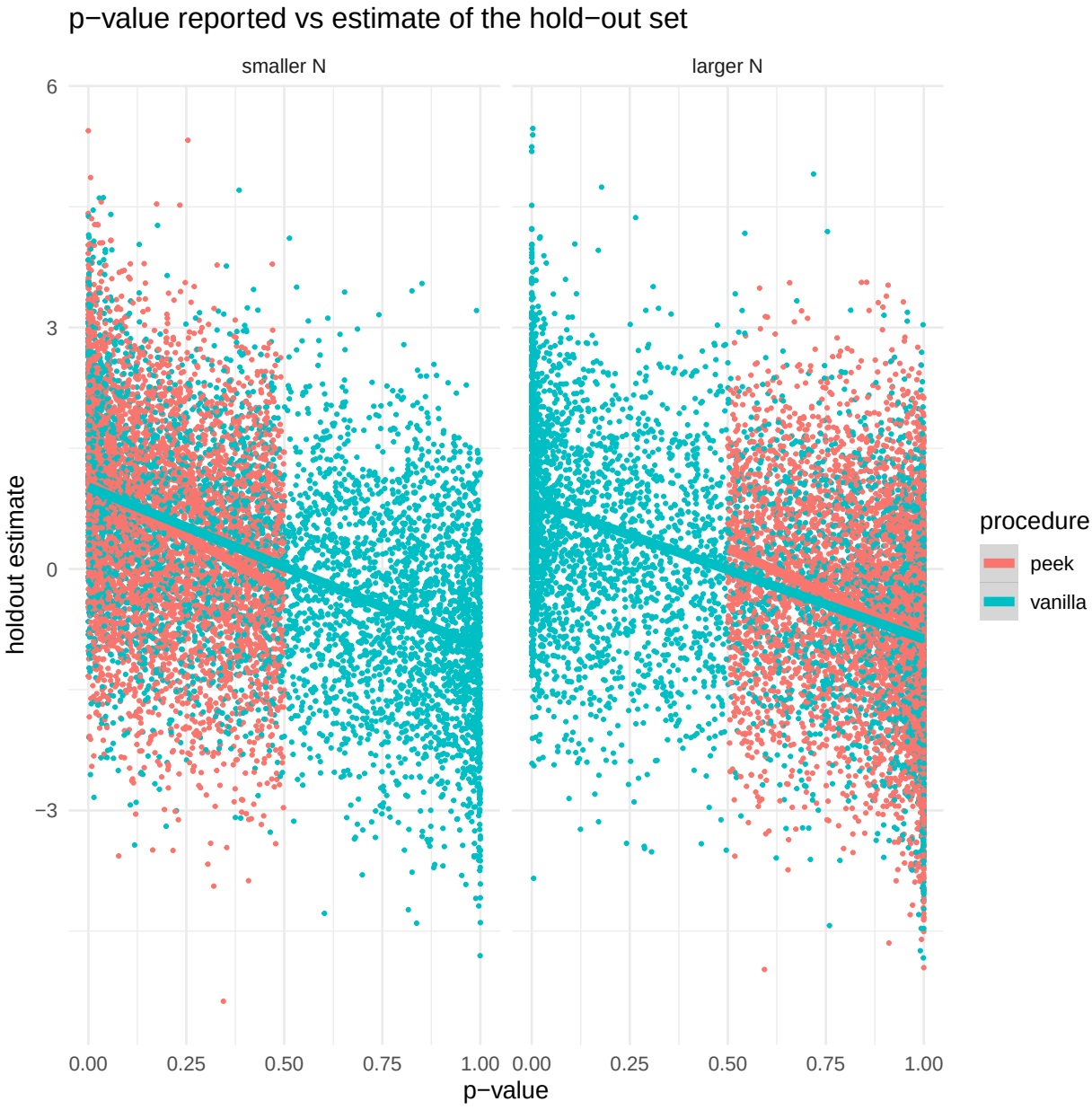
Validity Simulation

To increase the accessibility of the mathematical results, we first set up a small simulation just to demonstrate the results. In this simulation, a single observed variable is investigated either with a peek preview sampling strategy or with a classical fixed-N sampling strategy. In the peek preview strategy, half of the data set is sampled; if the average of the observed data points is above zero, the result is reported. Otherwise, the second half is sampled. In the classical strategy, data sets of the same sizes (either half or the full data set) is simulated. In both cases, a hold-out set is simulated afterwards. The true effect size for each of 10,000 simulation trials is chosen from a standard normal distribution.

```
library(ggplot2)
library(knitr)

result <- data.frame (procedure = factor(character(), levels = c("peek", "vanilla")), N=

trials <- 10000
```

```r
for (trial in 1:trials) {
  effect = rnorm(1, mean = 0, sd = 1)

  # point by peek preview sampling procedure
  x1 <- rnorm(1, mean = effect, sd = 1)
  x2 <- rnorm(1, mean = effect, sd = 1)
  y <- rnorm(1, mean = effect, sd = 1)

  if (x1 > 0) {
    p = 1-pnorm(x1, mean = 0, sd = 1)
    result <- rbind(result, data.frame(procedure = "peek", N = 1, p = p, repEst = y))
  } else {
    p = 1-pnorm(x1, mean = 0, sd = 1/sqrt(1+1))
    result <- rbind(result, data.frame(procedure = "peek", N = 2, p = p, repEst = y))
  }

  # study vanilla with N = 1 and N = 2
  x1 <- rnorm(1, mean = effect, sd = 1)
  x2 <- rnorm(1, mean = effect, sd = 1)
  y <- rnorm(1, mean = effect, sd = 1)

  if (trial %% 2 == 0) {
    p = 1-pnorm(x1, mean = 0, sd = 1)
    result <- rbind(result, data.frame(procedure = "vanilla", N = 1, p = p, repEst = y))
  } else {
    p = 1-pnorm(x1, mean = 0, sd = 1/sqrt(1+1))
    result <- rbind(result, data.frame(procedure = "vanilla", N = 2, p = p, repEst = y))
```

```
  }
}


ggplot(result, aes(x = p, y = repEst, color = procedure)) +
  geom_point(size = 0.5) +
  geom_smooth(method = "lm", se = TRUE, size = 2) +
  facet_wrap(~ N, labeller = as_labeller(c("1" = "smaller N", "2" = "larger N"))) +
  labs(title = "p-value reported vs estimate of the hold-out set",
       x = "p-value",
       y = "holdout estimate",
       color = "procedure") +
  theme_minimal()

## Warning:  Using ‘size‘ aesthetic for lines was deprecated in ggplot2
3.4.0.
      ## i Please use ‘linewidth‘ instead.
      ## This warning is displayed once every 8 hours.
      ## Call ‘lifecycle::last_lifecycle_warnings()‘ to see where this
warning was
      ## generated.
## ‘geom_smooth()‘ using formula = ’y ~ x’
```



Longitudinal Simulation

Simulation Population

We demonstrate the generalizability of sampling procedures by a simulation. For an illustrative example, assume a researcher investigates the decline of symptoms during a cognitive therapy. Symptoms are assumed to be normally distributed and are measured before the therapy, after half the sessions, and after the therapy. The mean effect is

assumed to be linear, so the data will be analyzed by a LGCM with latent intercept and slope predicting the three measurements within each person. The model has two mean parameter and six variance parameter (three measurement errors and the 2×2 covariance matrix of the latent variables).

For data generation, we must necessarily draw population parameters from some distribution. We assume that this distribution represents the long-run probability distribution of effect sizes in hypotheses selected and studies designed by the researchers. That is, these probabilities describe the range of true and zero effects that would result from studies by these researchers over a theoretically infinite career. In order to also describe the frequentist operating characteristics of sampling methods, we define the distribution of population parameters in a simplified way. The intercept of the symptoms is set to 10 with measurement error variance of 0.1. As the parameter of interest, the slope has a 50% chance to be precisely zero (that is, the null hypothesis of no slope is precisely true). If the slope is not zero, it is selected from a normal distribution with mean -1 and variance .25.

For every population generated in this way, data is generated for each sampling procedure outlined below until that procedure reports data or it is certain that the procedure will not report a data set. To see the different cases, we assume that the researcher either reports (1) the full data set, (2) the mean and covariance matrix of the latent variables, or (3) the maximum likelihood estimates of the model parameter. For all procedures, a replication sample is then drawn which is assumed to be so large that the true simulation parameters are recovered.

In this illustrative example, five researchers with different sampling procedures are considered:

Researcher 0 performs a simple random sample of $N = 20$ participants; they are the baseline for the comparison of the others. Researcher 1 samples at least 5 participants until the mean is significantly smaller than zero, in which case he reports the results, or $N = 20$

participants are reached.

Researcher 2 samples at least 5 participants until the mean is significantly smaller than zero or $N = 20$, using a Bonferoni correction on the α level that ensures an α of 0.05 for an omnibus-test using the current number of tests done so far .

Researcher 3 samples at least 5 participants until the estimated mean slope is below -0.5, in which case he reports, or $N = 10$ participants are reached. Researcher 4 samples two participants; if those two both show a negative total trend (i.e., the third measurement of symptoms is below the first), then they continue to a total of $N = 20$ participants and report those. Otherwise, Researcher 4 ignores the sample.

Researcher 5 samples $N = 20$ participants, then checks if the mean is significantly smaller than zero in a χ^2 test. If so, they report the data, otherwise, they ignore it.

Researcher 6 samples like Researcher 0, but then computes the reported statistic only on those participants that benefited from the therapy, unless there were three or less, then he does not report.

Note that Researcher 6 does at first not apply a different sampling procedure, but use the same procedure as the baseline Researcher 0. Only after that they report onl non-sufficient statistics, which is, only a part or even none of their data. So while their primary sampling procedure is covered by the definition, we don't expect them to produce results that generalize equally because the report is not sufficient. Therefore, following the math, we expect all Researchers to generalize equally well with exception of Researcher 6; here, we expect that for the same outcome reported, the mean slope of a replication study will show less decline than for the others.

Computation of the Difference in the Population Expectation

To reiterate our goal, in our illustrative example, we are interested whether we can expect the same effect of the therapy if we get the same report by two researchers using different sampling strategies. So for instance, assume Researcher 0 pre-registered to sample N instances and then does exactly that, while Researcher 1 sampled participants until

reaching a significant result. Now both researcher report exactly the same result in two different articles (say, on two different therapies). Does that mean that the therapy investigated by Researcher 0 should be expected to be more effective than that of Researcher 1? In other words, if we would do replication studies for both therapies with infinitely many participants, would the therapy of Researcher 0 turn out to have a stronger decline of symptoms?

In quantitative terms, this question translates to whether the true mean of the mean slope parameter in the LGCM is different between Researcher 0 and Researcher 2, under the condition that they produce the same report. The outcome of interest in this simulation is hence the difference between both expectations.

Computing this difference needs a little bit of a mathematical trick. Of course, in theory we could simulate a very large number of trials for each sampling strategy and wait until two reports are identical, for instance the two data sets are equal (even though they may have been simulated with different slope means). Then, we could compare the two slope means, and if that happens often enough, we can compute the average distance. However, since the space is 60-dimensional when the report is the full data set on our maximal $N = 20$ participants, this approach is infeasible.

There is an alternative way since for Researcher 0 (the basic fixed-size sampling procedure), we can conclude backwards from the data to the population distribution of the slope mean parameter. This distribution is proportional to the integral over all possible generating slope means from the simulation procedure, multiplied by the likelihood to generate the specified data (or data covariance matrix and mean vector) from this slope mean value. This integral is proportional to the Bayesian posterior for this data representation and the population simulation scheme as prior - which is, fixed values for all parameters but the slope mean, and the above described normal distribution with mean -1 and 0 for the slope mean.

So for every report one of the simulated researchers 1 to 6 submits, we do one

Bayesian analysis and compute the mean of the posterior distribution of the slope mean parameter. This is the expected value of the mean slope for Researcher 0 if they would have reported the same data or covariance matrix, respectively. The difference of this value to the true population mean slope (the mean slope with which the data for Researcher 1 to 6 was initially simulated) is then stored. If the expectation is the same for all possible reports, then this difference should have an expectation of zero over all simulation trials. Therefore, the result of our simulation is this expected difference for each Researcher. Of course, if two researchers both create the same expectation of the real world as does Researcher 0, then in particular they are equally generalizable as well. Note that strictly speaking, we only test a necessary condition; in theory the difference could be positive on some areas and negative on others of the report space and cancel out perfectly when integrating over all points; however, this would seem a very unlikely incidence, so we only report the necessary condition here.

Simulation Results

From the mathematical theory, we expect that Researcher 1 to Researcher 5, when reporting the same result, will all allow the same conclusions about the treatment effectiveness, which in particular is also the same as we would conclude from the same result by Researcher 0. For Researcher 6, in contrast, we expect that the treatment may be less effective given the same reported results, considering that Researcher 6 reports no sufficient statistic for the slope mean wrt. their data.

Figure 1 shows the result of the simulation. Each bar shows the mean difference between the expected efficiency of the therapy for the Researchers 1 to 6 as outlined above and Researcher 0 who uses a standard sampling of the same number of participants. The error bars show the variances of the difference between the simulation trials. As can be seen, there is no difference in the expected performance of the therapy when the research report is identical, no matter which sampling strategy has been used: Sampling with uncorrected p as stopping rule, with correct stopping rule, with estimate-based stopping,

with sneak-previewing some of the participants, or with drawering non-significant results. Only with Researcher 6, when reporting the same result as a researcher who pre-registered the same N, will find the therapy to be weaker in reality than Researcher 0, and therefore also to all others. With more than a third than the mean effect of the therapy in the simulations, this deviation not surprisingly is fairly large.

Explanation of Seemingly Paradox Outcomes

The result that all sampling routines allow the same conclusion to the future, even though some are strongly uniformly criticized in the literature, may seem surprising. There are a number of apparent paradoxes that need closer inspection.

Why does α inflation not imply less generalizability?

If everything else is constant, then a difference in the α error implies a difference in the posterior distribution of the null hypothesis. This is a simple consequence of the Bayes Theorem[2]; if H_0 is the event that the null hypothesis is true and T the event that a test is positive, then

$$P(H_0|T) = P(T|H_0) \cdot \frac{P(H_0)}{P(T)}$$

where $P(T|H_0)$ is the α error. So an increased α error, everything else constant, implies a higher probability of the H_0 given a positive test result.

However, using a sampling procedure that includes a condition does not leave everything else constant. If the sampling procedure somehow favors a positive result, then $P(H_0|C) < P(H_0)$; so adding the condition does reduce the numerator. That is why α inflation does not necessarily mean that one cannot draw the same conclusions from the result if the increased *alpha* is 'paid' for by having to satisfy a condition. Both effects cancel out perfectly, as shown in Theorem 3.

---

[2] Sometimes people assume that frequentistic analysis somehow makes Bayes Theorem invalid, or unusable; that is not the case obviously, the math is not impressed by our choice of terminology framework.

What if we assume that the effect is fixed to zero?

The classical convincing argument why a frequentistic test allows a conclusion is the reasoning that the α error is controlled, that is, if there is zero effect, than we know that errors can only happen with low probability. However, that argument is flawed by the Bayes Fallacy: There always is a prior distribution over the effect, at least due to the observers lack of knowledge about the true effect [3]. In fact, if the true effect would be fixed to zero, than of course all sampling methods[4] will generalize equally, because the replication distribution is known and the same for any kind of test. So assuming that the effect is truly zero makes Theorem 3 trivially true.

What if a sampling procedure only stops on positive effects?

Assume the true effect size, for instance a population mean, is zero, and a sampling procedure samples until the average is positive. For simplicity, assume the distribution is 1 or -1 with equal probabilities. Obviously, no reported average (which is a sufficient statistic of the mean) is lower or equal zero, which gives the impression that the procedure does not generalize. Observe that the event that the method reports with an even number of participants is impossible, because if more than half of the results were 1, then the method would have reported earlier. On odd N, all combinations of outcomes have exactly one more '1' than '0' if they are reported, and have not been

Discussion

In frequentistic thinking, it is usually assumed that something that increases a researcher's chances to get a significant result on a fixed effect size is 'bad practice'. Usually, data obtained in that way is assumed to generalize less, often without a clear definition of what generalization means. This article showed that this notion is not generally true. By the definition that two sampling procedures generalize equally well if, on

[3] You sometimes hear that an effect is 'fixed but unknown'; that is, strictly speaking, a contradiction in itself, because something that is unknown has an uncertainty and is therefore not fixed.

[4] In fact, all methods; even a psychic.

the same report, we can expect the same outcome for future samples, we showed that all sampling procedures generalize equally well when all collected information relevant for the outcomes of interest is reported. Such results are hence equally valid to predict future outcomes no matter whether they were sampled in a standard way or by sampling procedures seen as 'problematic'. In addition to the mathematical proof, a simulation demonstrated this for the accepted as well as for some of the most notorious sampling procedures: pre-registering a sample size and sampling exactly that, sampling until a result is significant (without or with correction), sneak-previewing a section of the data and continuing sampling only if they show descriptively favorable results, reporting data only if it is significant, or sampling until a descriptively positive result shows in the parameter estimates. All these ways to obtain data, and any other that reports all outcomes necessary for the parameters of interest, allow the same conclusions from the data to future outcomes. However, if information important for the research hypothesis is not reported, then the results do not generalize. This in particular includes situations where only partial data is reported selectively.

If no data is reported, no wrong conclusions can be made from that, but meta-analysis of reported data on the same hypothesis while ignoring non-reported data does of course have a biasing effect. So although reported data has the same value even if other data has not been reported, it is still important to report the posterior distribution of the parameters involved. Preregistration contributes to motivate researchers to publish data, no matter what the results are. However, a not significant result without reporting posteriors usually does not justify a scientific article. Therefore, a pre-registered study that revealed non significant result tempts researchers to interpret the null result, or at least to carefully suggest that the null hypothesis might be strengthened, which it is not. Also, significance can be missed by underpowering or not paying attention to minimize measurement error, which is motivated when a null result is enough to publish a scientific article. If the result does not seem strong enough, it would be preferable to sample more

participants until a better call about the research hypothesis can be made, whether that was pre-registered or not. Of course, this sampling procedure should be reported for the sake of meta analyses.

The wrong expectation that α inflation reflects in generalizability is fostered by the use of frequentistic thinking. From the frequentistic perspective, there exists a fixed, non-random effect size. With that image in mind, it obviously doesn't matter what results we find for future outcomes, as the population is fixed anyway. That is why the notion of 'generalizability' remained vague in the field: With strict frequentistic thinking, even not doing any study generalizes as well as any possible sampling procedure. So it needs some degree of dialectic thinking to discuss how well a result generalizes and to stay in the frequentistic framework at the same time, leading to rules on how to sample 'cleanly' that come across as more moralistic than scientific. It needs Bayesian thinking (not necessarily Bayesian methods) to realize that the fact that there is a lack of knowledge about the hypothesis means that there are possible different effects, and that data helps us to reduce our lack of knowledge about that. From that perspective, it is easy to see that a sampling condition which is determinstically obtained from the data does not reduce the interpretability. At the same time, it motivates researchers to find what in frequentistic settings would be 'null results', as these still reduce our uncertainty about the effect in the same way, regardless of any cutting point (e.g., whether the effect is larger than zero) we might be interested in. It also motivates to minimize measurement error, as that improves the reduction of uncertainty, again regardless of where the best estimate of the effect ends up.

Note that the choice between Bayesian and frequentism is a choice of language; the mathematical content stays the same. Although the frequentistic language, as we saw in this article, is confusing and leads us to wrong ideas about the world, it is possible to do correct frequentistic statistics. It happens to be called Bayesian statistics.

References

Armitage, P., McPherson, C., & Rowe, B. (1969). Repeated significance tests on accumulating data. *Journal of the Royal Statistical Society: Series A (General)*, *132*(2), 235–244.

Chow, S.-C., & Chang, M. (2008). Adaptive design methods in clinical trials–a review. *Orphanet journal of rare diseases*, *3*, 1–13.

Fanelli, D. (2009). How many scientists fabricate and falsify research? a systematic review and meta-analysis of survey data. *PloS one*, *4*(5), e5738.

Grieve, A. P. (2024). Probability of success and group sequential designs. *pharmaceutical statistics*, *23*(2), 185–203.

Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS medicine*, *2*(8), e124.

Nosek, B. A., & Lakens, D. (2014). Registered reports.

O'Brien, P. C., & Fleming, T. R. (1979). A multiple testing procedure for clinical trials. *Biometrics*, 549–556.

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

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological bulletin*, *86*(3), 638.

Shaffer, J. P. (1995). Multiple hypothesis testing. *Annual review of psychology*, *46*(1), 561–584.

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