Don't Forget the Psychology in Analyses of Psychological Data: The Case of Sequential Testing

Elsey, J.W.B.; Filmer, A.I.; Stemerding, L.E.

DOI
10.1525/collabra.24953

Publication date
2021

Document Version
Final published version

Published in
Collabra: Psychology

UNIVERSITY of CALIFORNIA PRESS | Collabra: Psychology

Elsey, J. W. B., Filmer, A. I., & Stemerding, L. E. (2021). Don't Forget the Psychology in Analyses of Psychological Data: The Case of Sequential Testing. *Collabra: Psychology*, *7*(1). https://doi.org/10.1525/collabra.24953

Methodology and Research Practice

Don't Forget the Psychology in Analyses of Psychological Data: The Case of Sequential Testing

James W. B. Elsey[1] [a], Anna I. Filmer[1], Lotte E. Stemerding[1]

[1] Clinical Psychology, University of Amsterdam, Amsterdam, the Netherlands

Keywords: sequential testing, optional stopping, sequential analysis, bayesian statistics, experimenter effects, interim analyses, biases

https://doi.org/10.1525/collabra.24953

Collabra: Psychology

Vol. 7, Issue 1, 2021

Sequential testing enables researchers to monitor and analyze data as it arrives, and decide whether or not to continue data collection depending on the results. Although there are approaches that can mitigate many statistical issues with sequential testing, we suggest that current discussions of the topic are limited by focusing almost entirely on the mathematical underpinnings of analytic approaches. An important but largely neglected assumption of sequential testing is that the data generating process under investigation remains constant across the experimental cycle. Without care, psychological factors may result in violations of this assumption when sequential testing is used: researchers' behavior may be changed by the observation of incoming data, in turn influencing the process under investigation. We argue for the consideration of an 'insulated' sequential testing approach, in which research personnel remain blind to the results of interim analyses. We discuss different ways of achieving this, from automation to collaborative inter-lab approaches. As a practical supplement to the issues we raise, we introduce an evolving resource aimed at helping researchers navigate both the statistical and psychological pitfalls of sequential testing: the Sequential Testing Hub (www.sequentialtesting.com). The site includes a guide for involving an independent analyst in a sequential testing pipeline, an annotated bibliography of relevant articles covering statistical aspects of sequential testing, links to tools and tutorials centered around how to actually implement a sequential analysis in practice, and space for suggestions to help develop this resource further. We aim to show that although unfettered use of sequential testing may raise problems, carefully designed procedures can limit the pitfalls arising from its use, allowing researchers to capitalize on the benefits it provides.

Sequential testing (also known as optional stopping) is the practice of monitoring and repeatedly analyzing data as a study progresses, and deciding whether or not to continue data collection depending on the results. This has major practical and ethical benefits: clinical trials can be stopped as soon as sufficient evidence for benefit, harm, or absence of effects has been convincingly established. Furthermore, resources can be preserved by not running excessive numbers of participants if a conclusion can be reached using a smaller sample size than expected.

As sequential testing involves repeated analyses, careful corrections must be made when using frequentist statistical approaches so as to avoid inflation of type 1 errors (see Lakens, 2014 for an accessible overview; Wald, 1945). In contrast, some authors favoring Bayesian analytic approaches (e.g., stopping testing once a critical Bayes Factor has been reached) argue that sequential testing is 'no problem for Bayesians' (Rouder, 2014), and that the stopping rule is 'irrelevant' from a Bayesian perspective (Edwards et al., 1963; Lindley, 1957; Wagenmakers, 2007). Anscombe (1963, p. 381) asserted that experimenters "should feel entirely uninhibited about continuing or discontinuing" their experiments and even changing the stopping rule along the way, and corrections for sequential testing have more recently been described as "anathema" to Bayesian reasoning (Wagenmakers et al., 2018). The supposition that using the Bayes Factor as a decision criterion "eliminates the problem of optional stopping" (Wagenmakers et al., 2012, p. 636) is considered a major advantage of Bayesian approaches (Wagenmakers et al., 2018).

Researchers continue to debate whether Bayesian approaches advocated by the authors above resolve statistical issues associated with interim analyses (de Heide & Grünwald, 2020; Kruschke, 2014; Yu et al., 2013), and whether this debate points more to misinterpretations of the Bayes Factor than problems with Bayesian approaches *per se*

a j.w.b.elsey@uva.nl

(Rouder, 2014; Rouder & Haaf, 2019). Whatever one's perspective on the relative merits of different analytic approaches, if Bayesian approaches – or indeed appropriately corrected frequentist approaches – do provide a solution to certain statistical pitfalls, can they be said to have eliminated '*the*' problem of sequential testing? The answer, we argue, is no. Potential issues with sequential testing are also psychological: incoming data may affect researchers, in turn causing them to influence the data-generating process under investigation. We are not the only ones to have recognized this, and below we discuss how others have proposed to tackle this problem. While many, including the authors above, would likely acknowledge the possible psychological implications of sequential testing, the tenor of the statements suggests that these issues may be underappreciated, and some researchers reading these discussions of Bayesian statistics in particular may miss that statistical issues reflect only one side of the possible pitfalls.

Several meta-scientific lines of research suggest that researcher expectations and beliefs may influence the results of experiments. The direction of experimental effects can be shifted in line with what researchers are led to believe should happen (Doyen et al., 2012; Gilder & Heerey, 2018; Intons-Peterson, 1983; Klein et al., 2012; Rosenthal, 1994; Rosenthal & Lawson, 1964; Rosenthal & Rubin, 1978). In summarizing several experimenter effects observed across studies involving experimenter interactions with both human and animal subjects, Rosenthal (2002a, p. 5) highlighted that: "When the first few subjects of an experiment tend to respond as they are expected to respond, the behavior of the experimenter appears to change in such a way as to influence the subsequent subjects to respond too often in the direction of the experimenter's hypothesis". Effects of beliefs, expectations, and confidence in an intervention are also apparent in studies of psychotherapy (Dragioti et al., 2015; Leykin & DeRubeis, 2009; Luborsky et al., 1975; Munder et al., 2011, 2012, 2013).

As one key goal of data analysis is to evaluate and update our beliefs about psychological phenomena, it makes sense that beliefs may change with new data. Valid sequential testing relies on the assumption that the data-generating process under investigation remains constant during data collection. If sequential testing is not used carefully, researchers' beliefs and behavior may be changed by the observation of incoming data, in turn influencing the data-generating process. Given the possible underappreciation of such experimenter effects, it is worth briefly highlighting the ways in which they could compromise a psychological experiment, and dispelling some possible misunderstandings of the nature of these effects. To guide the reader's intuition, we will begin with some hypothetical examples that convey how peeks at interim data might influence a researcher in unanticipated ways.

Misconceptions About Researcher Influences

"Blinding experimental conditions prevents the influence of beliefs, or of changing beliefs, because the researcher cannot choose to affect one condition vs. another". Blinding of experimental conditions can nullify certain interpersonal influences, the possibility that researchers are actively favoring or disfavoring one condition over another, or that ratings are affected by knowledge of the experimental condition (Day & Altman, 2000). However, blinding is not always easy. Psychotherapeutic trials may be a case in which the benefits of sequential testing are especially apparent owing to their high cost, difficulty, and desire to get the best treatments out for patients as quickly and with as few 'control' patients used as possible. Yet, blinding is notoriously difficult in such trials: the therapist will almost certainly know what treatment they are delivering, and the patient must actively participate in it (Berger, 2015). Even if blinding can occur, we suggest that changes to beliefs due to sequential testing could still affect one treatment or experimental condition over another, even in within-person experimental designs. For example, beliefs about the presence or absence of an effect could impact a researcher's overall attitude towards an experiment, the diligence with which they execute it, and the enthusiasm with which they interact with participants. This may impact a researcher's ability to engage with their participants, which could be a necessary condition for them to properly participate in the experimental task. A possible effect may thus be nullified to the level of a baseline/control condition, even though the experimenter has not tried to influence a specific condition.

"If beliefs or other such superficial factors affect an experiment, then the effect should not be considered legitimate anyway". In some cases, 'experimenter effects' in an intervention may be an unwanted or non-specific treatment factor. For example, Kaptchuk et al. (2008) found that the warmth of the administering clinician, but not the actual use of real vs. sham acupuncture, was predictive of relief from irritable bowel syndrome (IBS) in acupuncture treatment. This experiment clearly suggested that, in the case of IBS, it is the attitude of the clinician rather than acupuncture itself that produces beneficial effects. Hence, the acupuncture effect in IBS was not 'real' in the sense of being attributable to acupuncture itself. However, in many cases a degree of confidence, warmth, or enthusiasm may be a necessary component of a legitimate mechanism (e.g., the conviction to work with a patient and generate a bond with them in order to tackle distressing issues in therapy). Peeks at interim data may serve to increase or decrease these important treatment components. Even in experimental studies, participants must be sufficiently motivated to properly engage with the task and stimuli at hand. This may be undermined if the experimenter unwittingly telegraphs a blasé attitude about the experiment owing to interim findings.

"We can simply measure experimenter beliefs and behavior to control for or determine their influence". Meta-research on the changing attitudes and beliefs of experimenters is certainly warranted. Indeed, other researchers have suggested that not only the expectations of effects among experimenters, but also among participants, should be used to more fully understand what is driving effects in experimental and clinical manipulations (Boot et al., 2013). However, there are limits to this approach, most notably that we do not know and cannot always measure the many factors that may contribute to experimenter effects. It seems unlikely that asking researchers about their enthusiasm or confidence in a treatment, or confirming adherence to treatment and experimental protocols, can account for the many sub-

tle ways in which people might be affected across the experimental cycle. Studies are also likely to be underpowered to detect such effects, as many are designed to detect an overall treatment effect, not potentially subtle influences upon it. Hence, while we agree that there could be great value in measuring and understanding researchers' beliefs across an experimental cycle, we think it unlikely this would provide a solution to the possible psychological problems raised in sequential analyses.

We are not aware of any empirical tests of the possible psychological effects of using sequential testing, and this may warrant investigation (see 'Acknowledging Limitations and the Unknown' below). Nevertheless, with the brief examples above it can be seen how such effects are at least plausible, and could apply to a wide range of experiments using either frequentist or Bayesian sequential analytic approaches. Means of precluding such effects are therefore desirable.

Solutions to the Psychological Problems of Sequential Testing

The key issue highlighted above is how the transmission of information about the phenomenon under investigation may influence the thoughts and actions of the researcher, in turn compromising the data-generating process. Solutions to this problem revolve around maintaining the benefits of sequential testing, while keeping the researcher blind to the results of interim analyses. We refer to this as *insulated* sequential testing. One recently proposed solution to this is automated sequential testing (Beffara-Bret et al., 2019). For certain types of experiments, Beffara-Bret and colleagues provide a protocol for directly linking data collection to a blinded and automated analysis pipeline. Based upon explicit, predefined stopping rules and analysis plans, the output of this process can tell the researcher whether to continue with data collection or not, without revealing the results. This is a great development for tackling the issues of sequential testing, though there are some drawbacks. The technical/coding skills needed to set up this pipeline – though reportedly modest – may be beyond many researchers, especially those used to handling data with point-and-click software rather than languages such as *R* or *Python*.

Secondly, an automated pipeline may have limited utility in studies where there are ethical concerns that might need expert oversight, and for which all possible worrisome eventualities cannot be determined in advance. For instance, in trials of some psychotherapeutic interventions, there is a possibility of ethically concerning outcomes that may be difficult to define or anticipate in advance – e.g., the average member of a treatment group may improve, while a small minority show dramatically worsened outcomes that do not occur in a control group.[1] In such cases, a stopping requirement based on group mean comparisons may not be triggered, but an informed expert may consider this as grounds to pause the trial. Stopping criteria can of course be put on metrics besides means or group comparisons, but in ethically sensitive experiments one might worry whether all the relevant possible outcomes have been considered. Decisions about whether to stop an experiment or trial on ethical grounds often involve a confluence of different factors being weighed together on a case-by-case basis (Friedman et al., 2015), and where serious ethically relevant outcomes are in play, it is almost certainly irresponsible to think that all eventualities can be coded into a blind pipeline without oversight. An alternative solution in such cases is to have independent or semi-independent analysts performing interim analyses.

The U.S. Food and Drug Administration (FDA) has advised the use of 'data monitoring committees' (DMCs) in some trials (FDA, 2006). These committees are typically composed of domain-experts, ethicists, and statisticians who monitor incoming data to determine if a trial should be stopped early due to excessive risks or conclusive benefits. The FDA advises that the outcomes of group comparisons not be shared with those involved in the study to prevent changes to behavior that may compromise trial integrity. However, these recommendations are predominantly aimed at pharmaceutical and medical device studies, particularly those involving severe risks such as mortality. Lakens (2014) notes that such division between researchers and analysts is rare in psychological science. Assembling full-fledged data monitoring committees may be excessive for most typical psychological experiments, but involving an independent analyst is often viable.

In psychological studies, research personnel could be split such that an ethically responsible team member is given explicit and predefined plans of how to perform interim analyses, as well as the capability to make judgment calls on ethical grounds if unexpected but concerning events occur. This interim analyst would not interact with any participants, and take precautions in communications with other study personnel to inform them only of the decision to continue or not, with more details revealed if discontinuation is the outcome. When such ethical concerns are not an issue, it becomes more feasible to involve interim analysts without domain expertise and who can be more independent of the main study personnel. This may be achieved through 'recruitment' of potential analysts in one's network, or also be an opportunity for 'crowdsourcing' the scientific endeavor (Uhlmann et al., 2019). The Open Science Framework platform *StudySwap* (Chartier et al., 2018) enables researchers to connect and share resources. One unanticipated but approved use case for *StudySwap* is to request the help of independent analysts who may perform interim analyses on one's behalf.[2]

1 A clinician would likely recognize instances of individual patients in their care experiencing concerning symptoms, but in many cases, assessments are performed by the person who did not administer the treatment, or using online data collection for longer-term outcomes. Worrisome effects might therefore be missed by a clinician.

2 We thank Daniel Lakens for bringing our attention to *StudySwap* in an earlier review of this paper

Researchers must also be aware that even with ostensibly blinded sequential analyses, failure to think critically about psychological aspects of the experimental process can more easily result in leakage of 'implied' information about study effects. Based on initial power analyses, study personnel may realize that the mere fact of continuing data collection beyond a certain sample size implies that initial effect size predictions were likely overestimated. Where data collection is time consuming and difficult, this may be compounded by frustration and the dim prospect of continuing with many more participants for relatively little anticipated gain. Intuitively, effects of continuing data collection without statistical information about the effect would have less effect on the experimenter's beliefs than literally seeing evidence accumulating against their hypothesis, but would tend to rise with increasingly drawn-out periods of data collection. Hence, beyond planning for compelling statistical evidence (Schönbrodt & Wagenmakers, 2018), researchers should balance the chances of ending the experiment at a decision boundary (e.g., a critical *p* value or Bayes Factor) with the potential that continuing data collection beyond a certain point may change the experimenters' confidence in the expected effect, and thereby influence the data-generating process. Whatever the choice of minimum and maximum sample size, it would be wise to recognize the minimum as only a best-case scenario: the maximum must be entertained as genuinely possible, and planning a series of experiments on the assumption that each will end at the first opportunity is likely to prove frustrating. While this may seem obvious, it is worth emphasizing – as researchers we are not immune to the sense that luck will be on our side, going through the procedures of a power analysis and highly ambitious sample size that we actually hope and expect to never have to reach.

To facilitate insulated sequential analyses with the aid of independent analysts – either via *StudySwap*, through one's network, or by splitting study personnel – we have developed an evolving resource for researchers: the Sequential Testing Hub (www.sequentialtesting.com). This resource includes a template for information about interim analyses that an independent analyst would need, a section covering some more extensive practical considerations regarding the use of sequential testing, a bibliography of related resources (e.g., for sequential testing power analyses, blinded automation of sequential testing), links to useful software and tutorials, and space for suggestions to expand the coverage of the resource. We encourage researchers with tools and tutorials to contact us so that we can link to these resources, making the process of planning and performing rigorous sequential testing easier. While domain experts may be well aware of the tools and resources we highlight, from our own experience designing sequential analyses, we believe that such an evolving resource will prove useful for those without such an in-depth background who wish to utilize sequential analytic approaches, whether insulated or not.

Acknowledging Limitations and the Unknown

Although existing research gives strong reason to believe that experimenter beliefs can influence study outcomes in a variety of ways, we are not aware of studies that have examined the interplay between interim analyses and experimenter beliefs. In addition, previous research suggests that the impact of expectations can vary across settings, protocols, and outcomes (see Rosenthal & Rubin, 1978 for an early meta-analysis). Some of the most consistent and well-researched expectancy effects are of the manipulated expectations of teachers for student performance in their classes (Rosenthal, 2002b). The extended duration of teacher-student interaction and large scope for interpersonal influence might most closely parallel the possible influence of a therapist over their patient, leading one to expect that therapeutic trials are the cases in which expectancy effects are most probable. Yet, studies also suggest that subtle changes in non-verbal communication in simple experimental studies can have surprising effects (Rosenthal, 2002a). As we noted above, even a possible within-person manipulation in a typical psychological experiment could be affected if researchers develop an attitude that undermines a participants' level of engagement with the experimental task. In addition, we do not yet know the extent to which researchers change their beliefs with incoming data – the reader can likely think of cases in which they feel a certain idea has been disproven long ago and yet is retained with great conviction by a disconcerting number of their peers!

Hence, the magnitude of the potential problems we raise is currently unknown.[3] It is possible to imagine some ways in which one might seek to investigate such effects. One option would be to manipulate researchers' expectations or the data they see as results come in. A key consideration here would be what outcome one cares about: one could measure researcher behavior with stooge 'participants' who are blind to the condition, and who make judgments of how the researcher behaved with them. In this case, we would be looking at tangible researcher behaviors and the feelings researchers elicit in participants, which might be affected by interim checks. We would need to make assumptions about how this could affect the data collected. If we really wanted to know how such manipulations affect processes under investigation, we would need each manipulated researcher to continue collecting full batches of data that are sufficiently powered to detect an effect, and differences in

3 Some anecdotal considerations might be relevant. Firstly, every colleague we have spoken to who has performed interim analyses as part of a research project they care about appears to have shown at least some of the psychological effects we highlighted to some degree, from becoming disillusioned or suddenly invigorated regarding the observed effect, demotivated by the unexpected continuation of data collection, or even wishing to abandon the research altogether after seeing that results do not seem to be coming out as expected at early stages but requiring prolongation of the project. Supervisors of PhD candidates have also mentioned seeing these effects in their supervisees. When researchers hope to find particular results – which applies to most researchers – it is almost inevitable that they will be impacted by seeing these hopes realized or dashed.

that effect, across researchers and the conditions they are in. To get any reasonable number of 'researcher' participants, we would need a huge number of 'participant' participants for each researcher to experiment upon. For this to be of real value, we would also want to understand what types of effects are likely to be affected, and our impression is that one could not cover all or even most bases even in multicenter experiments, as interesting as this might be. Another approach would be to use existing data in which interim analyses have been performed. This may be more feasible, but variation across studies in all sorts of other important factors is likely as great as variation among the ways in which interim analyses were performed (the level of insulation). Moreover, in our experience, studies report relatively little information regarding how interim analyses were performed. To make such research more feasible in the future, we would encourage researchers to go into greater detail about what methods, if any, were employed to try to insulate analyses from such effects.

Given that we do not know the magnitude of such effects, is it right that researchers should feel burdened to take precautions against them? We would stress that we do not wish to impose any unfair or unrealistic burdens on researchers, and it may be argued that taking such precautions could be disproportionately difficult for small research teams or those without a large network. However, readers should also not overestimate the burden of such procedures. In many cases, basic insulated analyses are feasible with the help of just one colleague and, at the risk of sounding facetious, when clear instructions are provided and ethical concerns are not an issue, it might even be possible to recruit analysts on freelance recruitment sites to perform simple checks on anonymized data (remember that one will not be relying on this analyst to perform all of one's final published analyses). In a recent example conducted in our lab, the only change that needed to be made to typical protocols was that the person supervising data collection did not discuss any interim results with the person collecting the data until the stopping point was reached. Truly difficult insulated analyses are more likely in such cases as clinical/psychotherapy trials where there is need for ethical oversight. Such trials are usually already conducted by relatively large research teams, and in this sense the difficulty and importance of insulated analyses might even scale with the size of the research team and resources available. Finally, consider that in many experiments, we do not know the impact that commonplace precautions such as blinding really have, or what would happen without blinding. We nevertheless rightly take quite strenuous precautions because we wish to be able to confidently rule out certain possible confounding influences.

Evidently, there is much that remains unknown in this area. In our opinion, this is more an argument for raising awareness of these possible issues than for ignoring them, as others may be encouraged to explore them. We follow Shadish, Cook, and Campbell (2002) in emphasizing that threats to the validity of an experiment can be both empirical and conceptual. The possibility that sequential testing procedures may result in changes in researcher behavior is at least conceptually plausible and should thus be taken seriously and protected against as far as possible. Having said this, we also recognize that what is deemed possible is going to vary considerably among researchers given different practicalities and assessments of the level of threat such issues pose in their specific experiment. Boot and colleagues (2013) similarly recognized that careful control and assessment of expectancy effects in control/placebo conditions was not easily achieved. The solutions we and others suggest do not always cover all possibilities and can impose a burden on researchers. We do not suggest that non-insulated sequential analyses are invalid, but we do think people should be more aware of the risks. More all-encompassing and easily-applied solutions cannot be expected without broader recognition of the possible problem.

Conclusion

To summarize, our recommendation is to take practically and ethically feasible measures to insulate the research team from information that may influence their attitudes and behavior during the collection of data (see Figure 1). We suggest that researchers report any steps taken to prevent the leakage of information in their study design. When such steps have not or cannot be taken, this of course is not damning, but variables in the dataset indicating at which stage of interim analysis each data point is from should make it easier for meta-researchers to investigate possible influences of interim analyses. We have highlighted how unfettered use of sequential testing may be ill-advised. However, by leveraging the power of automation, crowdsourcing, or other means of involving independent analysts, researchers may be able to perform highly efficient, procedurally rigorous, insulated sequential testing. We also aim to facilitate performance of sequential analyses with an evolving 'hub' highlighting useful tools, approaches, and articles pertinent to the design of sequential analysis – the Sequential Testing Hub – and encourage readers to suggest valuable and useful resources. More broadly, we hope to have emphasized that not all problems arising from the use and misuse of statistics are solely or inherently statistical – a fact that applies beyond sequential testing. It would certainly be a shame if, confident in the power of new statistical and experimental approaches, psychological researchers forget to think like psychologists about the experiments they design.

Contributions

Contributed to conception of paper: 1st, 2nd, and 3rd authors

Contributed to drafting and revising the paper: 1st, 2nd, and 3rd authors

Competing interests

The authors have no conflicts of interest to declare.

Data accessibility statement

There is no data associated with this manuscript.

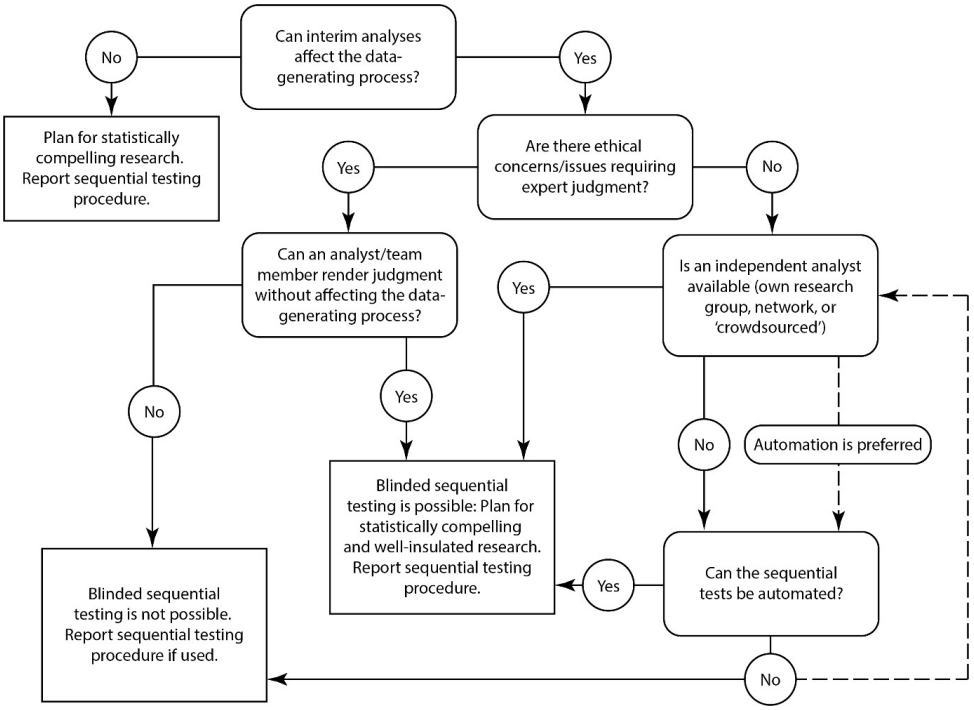

Figure 1. Decision tree for circumstances under which insulated sequential testing might be pursued.

Submitted: February 25, 2021 PDT, Accepted: June 13, 2021 PDT

REFERENCES

Anscombe, F. J. (1963). Sequential Medical Trials. *Journal of the American Statistical Association*, *58*(302), 365–383. https://doi.org/10.1080/01621459.1963.10500851

Beffara-Bret, B., Beffara-Bret, A., & Nalborczyk, L. (2019). A fully automated, transparent, reproducible and blind protocol for sequential analyses. *PsyArXiv*. https://doi.org/10.31234/osf.io/v7xpg

Berger, D. (2015). Double-blinding and bias in medication and cognitive-behavioral therapy trials for major depressive disorder. *F1000Research*, *4*(638). https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4732552/

Boot, W. R., Simons, D. J., Stothart, C., & Stutts, C. (2013). The pervasive problem with placebos in psychology: Why active control groups are not sufficient to rule out placebo effects. *Perspectives on Psychological Science*, *8*(4), 445–454. https://doi.org/10.1177/1745691613491271

Chartier, C. R., Riegelman, A., & McCarthy, R. J. (2018). StudySwap: A Platform for Interlab Replication, Collaboration, and Resource Exchange. *Advances in Methods and Practices in Psychological Science*, *1*(4), 574–579. https://doi.org/10.1177/2515245918808767

Day, S. J., & Altman, D. G. (2000). Blinding in clinical trials and other studies. *Bmj.Com*, *321*(7259), 504–504. https://www.bmj.com/content/321/7259/504?hwoasp=authn:1369647604:4223573:1446970149:0:0:4lLHYzv2ujxaVryjzwQeDA%3D%3D

de Heide, R., & Grünwald, P. D. (2020). Why optional stopping can be a problem for Bayesians. In *Psychonomic Bulletin and Review*. Springer. https://doi.org/10.3758/s13423-020-01803-x

Doyen, S., Klein, O., Pichon, C. L., & Cleeremans, A. (2012). Behavioral priming: It's all in the mind, but whose mind? *PLoS ONE*, *7*(1).

Dragioti, E., Dimoliatis, I., Fountoulakis, K. N., & Evangelou, E. (2015). A systematic appraisal of allegiance effect in randomized controlled trials of psychotherapy. *Annals of General Psychiatry*, *14*(1), 25. https://doi.org/10.1186/s12991-015-0063-1

Edwards, W., Lindman, H., & Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological Review*, *70*(3), 193–242.

FDA. (2006). *Guidance for clinical trial sponsors: Establishment and operation of clinical trial data monitoring committees*.

Friedman, L. M., Furberg, C. D., DeMets, D. L., Reboussin, D. M., & B., G. C. (2015). Statistical Methods Used in Interim Monitoring. In *Fundamentals of Clinical Trials* (5th ed., pp. 373–401). Springer.

Gilder, T. S., & Heerey, E. A. (2018). The role of experimenter belief in social priming. *Psychological Science*, *29*(3), 403–417.

Intons-Peterson, M. J. (1983). Imagery paradigms: How vulnerable are they to experimenters' expectations? *Journal of Experimental Psychology: Human Perception and Performance*, *9*(3), 394–412.

Kaptchuk, T. J., Kelley, J. M., Conboy, L. A., Davis, R. B., Kerr, C. E., Jacobson, E. E., & Park, M. (2008). Components of placebo effect: randomised controlled trial in patients with irritable bowel syndrome. *British Medical Journal*, *336*(7651), 999–1003. https://www.bmj.com/content/336/7651/999/

Klein, O., Doyen, S., Leys, C., Magalhaes, P., Miller, S., Questienne, L., & Cleeremans, A. (2012). Low hopes, high expectations: Expectancy effects and the replicability of behavioral experiments. *Perspectives on Psychological Science*, *7*(6), 572–584. https://doi.org/10.1177/1745691612463704

Kruschke, J. K. (2014). Goals, power, and sample size. In *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stain* (2nd ed., pp. 359–398). Academic Press.

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

Leykin, Y., & DeRubeis, R. J. (2009). Allegiance in psychotherapy outcome research: Separating association from bias. In *Clinical Psychology: Science and Practice* (Vol. 16, Issue 1, pp. 54–65). https://doi.org/10.1111/j.1468-2850.2009.01143.x

Lindley, D. V. (1957). A statistical paradox. *Biometrika*. https://www.jstor.org/stable/2333251

Luborsky, L., Singer, B., & Luborsky, L. (1975). Comparative studies of psychotherapies: is it true that "everybody has won and all must have prizes"? *Archives of General Psychiatry*, *32*(8), 995–1008. http://www.ncbi.nlm.nih.gov/pubmed/772674

Munder, T., Brütsch, O., Leonhart, R., Gerger, H., & Barth, J. (2013). Researcher allegiance in psychotherapy outcome research: an overview of reviews. *Clinical Psychological Science*, *33*(4), 501–511. https://www.sciencedirect.com/science/article/pii/S0272735813000275

Munder, T., Flückiger, C., Gerger, H., Wampold, B. E., & Barth, J. (2012). Is the allegiance effect an epiphenomenon of true efficacy differences between treatments? A meta-analysis. *Journal of Counseling Psychology*, *59*(4), 631–637. https://doi.org/10.1037/a0029571

Munder, T., Gerger, H., Trelle, S., & Barth, J. (2011). Testing the allegiance bias hypothesis: A meta-analysis. *Psychotherapy Research*, *21*(6), 670–684. https://doi.org/10.1080/10503307.2011.602752

Rosenthal, R. (1994). Interpersonal Expectancy Effects: A 30-Year Perspective. *Current Directions in Psychological Science*, *3*(6), 176–179. https://doi.org/10.1111/1467-8721.ep10770698

Rosenthal, R. (2002a). Experimenter and clinician effects in scientific inquiry and clinical practice. *Prevention and Treatment*, *5*(1), 38c. https://psycnet.apa.org/record/2003-04137-005

Rosenthal, R. (2002b). The Pygmalion effect and its mediating mechanisms. In J. Aronson (Ed.), *Improving Academic Achievement* (pp. 25–36). Academic Press. https://www.sciencedirect.com/science/article/pii/B9780120644551500051

Rosenthal, R., & Lawson, R. (1964). A longitudinal study of the effects of experimenter bias on the operant learning of laboratory rats. *Journal of Psychiatric Research*, *2*(2), 61–72.

Rosenthal, R., & Rubin, D. B. (1978). Interpersonal expectancy effects: The first 345 studies. *Behavioral and Brain Sciences*, *1*(3), 377–386. https://www.cambridge.org/core/services/aop-cambridge-core/content/view/S0140525X00075786

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

Rouder, J. N., & Haaf, J. M. (2019). On The Interpretation of Bayes Factors: A Reply to de Heide and Grunwald. *Psyarxiv*. https://files.osf.io/v1/resources/m6dhw/providers/osfstorage/5d40ac7e26ebf5001984ae8d?action=download&version=1&direct

Schönbrodt, F. D., & Wagenmakers, E. J. (2018). Bayes factor design analysis: Planning for compelling evidence. *Psychonomic Bulletin and Review*, *25*(1), 128–142. https://doi.org/10.3758/s13423-017-1230-y

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Houghton-Mifflin.

Uhlmann, E. L., Ebersole, C. R., Chartier, C. R., Errington, T. M., Kidwell, M. C., Lai, C. K., McCarthy, R. J., Riegelman, A., Silberzahn, R., & Nosek, B. A. (2019). Scientific Utopia III: Crowdsourcing Science. *Perspectives on Psychological Science*, *14*(5), 711–733. https://doi.org/10.1177/1745691619850561

Wagenmakers, E.-J. (2007). *Stopping Rules and Their Irrelevance for Bayesian Inference*.

Wagenmakers, E.-J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., Love, J., Selker, R., Gronau, Q. F., Šmíra, M., Epskamp, S., Matzke, D., Rouder, J. N., & Morey, R. D. (2018). Bayesian inference for psychology. Part I: Theoretical advantages and practical ramifications. *Psychonomic Bulletin and Review*, *25*(1), 35–57. https://doi.org/10.3758/s13423-017-1343-3

Wagenmakers, E.-J., Wetzels, R., Borsboom, D., van der Maas, H. L. J., & Kievit, R. A. (2012). An Agenda for Purely Confirmatory Research. *Perspectives on Psychological Science*, *7*(6), 632–638. https://doi.org/10.1177/1745691612463078

Wald, A. (1945). Sequential tests of statistical hypotheses. *The Annals of Mathematical Statistics*, *16*(2), 117–186.

Yu, E. C., Sprenger, A. M., Thomas, R. P., & Dougherty, M. R. (2013). When decision heuristics and science collide HyGene View project Fast and Frugal View project. *Psychonomic Bulletin and Review*, *21*(2), 268–282. https://doi.org/10.3758/s13423-013-0495-z

SUPPLEMENTARY MATERIALS

Peer Review History

Download: https://collabra.scholasticahq.com/article/24953-don-t-forget-the-psychology-in-analyses-of-psychological-data-the-case-of-sequential-testing/attachment/63477.docx?auth_token=FjpuAm-JE7YcgT42MUDX