Stopping Rules as Experimental Design

Samuel C. Fletcher
Department of Philosophy
University of Minnesota, Twin Cities

1 Introduction

A "stopping rule" in a sequential experiment is a rule or procedure for determining when the experiment should end.[1] For example, consider a pair of experiments designed to obtain evidence about the proportion of fruit flies in a given population with red eyes [Savage, 1962, pp. 17–8]. In both experiments, flies are caught, observed, and released sequentially and fairly, reporting in the end the number of red-eyed flies. In the first, the experiment is designed to stop after observing 100 flies, while the second is designed to stop after observing 6 red-eyed flies. In general the data from these experiments could be very different, but it is also possible that they be the same: in this case, 100 total flies would be observed in both experiments, of which 6 (including the last) would have red eyes. Is the evidence that each of the two would then provide for or against an hypothesis about the proportion of red-eyed flies the same?

The stopping rule principle (SRP) states that this is so:

Stopping Rule Principle: The evidential relationship between the data from a completed sequential experiment and a statistical hypothesis does not ever depend on the experiment's stopping rule.[2]

[1] Despite their name, sequential experiments need not involve any robust experimenter control or manipulation.

[2] Technically, this is restricted to *non-informative* stopping rules, ones which when learned provide no more information about the hypothesis of interest than the data themselves. All parties are in agreement that the SRP does not apply for informative stopping rules. See Raiffa and Shlaifer [1961, pp. 36–42] and Berger and Wolpert [1988, §4.2.7] for formal definitions, examples, and discussion.

So, according to the SRP, the same data should yield the same evidence, regardless of which stopping rule was used. Adherents of the SRP typically apply it to experiments with complicated or ambiguous stopping rules, analyzing the experiment as if it were based instead on a simpler fixed stopping rule. Thus, according to this strategy, if you accept the SRP, "It is not even necessary that you stop according to a plan. You may stop when tired, when interrupted by your telephone, when you run our of money, when you have the casual impression that you have enough data to prove your point, and so on" [Edwards et al., 1963, p. 239].

As I elaborate in section 2, many Bayesian statistical methods satisfy the SRP insofar as they assess evidence for hypotheses in terms of their prior and posterior probabilities, which are invariant under different stopping rules for experiments producing the same data. On the other hand, classical statistical methods (whether Fisherian or Neyman-Pearsonian) do not, insofar as they rely on test statistics whose values depend on the probability distribution of possible—not just actual—data, and clearly two sequential experiments' possible outcomes need not be the same. Thus the SRP, along with the so-called likelihood principle [Birnbaum, 1962, Berger and Wolpert, 1988], which entails it, is a central point of contention between the two schools of evidence and statistical inference—they disagree about the evidential relevance of modal features of the experimental design and process of data collection.[3]

But as authors defending both the Bayesian [Sprenger, 2009, p. 639] and classical [Mayo, 1996, pp. 348, 351, 357] perspectives acknowledge, arguing for the SRP because it follows from a framework for understanding statistical evidence does little to convince others who have accepted a different framework with different conclusions. Rather, it may be more productive to provide independent arguments for or against the SRP—arguments that do not depend on adopting such a framework, besides having different premises—which in turn provide argumentative support for or against broader statistical frameworks, according to how that framework entails or contradicts the SRP. The main goal of this essay is to contend in sections 3–5 that five independent arguments for the SRP advanced by statisticians and philosophers fail to succeed. Viewing a stopping rule as an integral part of a sequential

[3]I have chosen to focus on the SRP rather than the likelihood principle in this essay because of its concreteness and the arguments found in the literature concerning it in particular.

experiment's *design*, though not necessary to show this, makes certain aspects of my contentions more readily comprehensible. Consequently, these arguments do not support adopting standard versions of Bayesianism over classical statistical methods.

The first class of arguments, considered in section 3, conclude that rejecting the SRP leads to an unacceptable sort of subjectivity in statistical inference; these include the arguments from intentions (section 3.1) and deception (section 3.2). I respond that either the sort of subjectivity charged is entirely acceptable if granted, in the former case, or unfounded, in the latter. The second class, considered in section 4, argues from the practical undesirability of rejecting the SRP (section 4.1), or the practical desirability of the consequences of it (section 4.2). These arguments commit to false premises or argumentative gaps that have little hope of being filled without assuming an evidential framework. The last, more technical argument I consider, seeks to show that rejecting the SRP entails untoward decision-theoretic consequences. I sharpen this argument in section 5, only to show that it highlights a general problem with a using fixed-level tests within Neyman-Pearson testing, not to any generic position that rejects the SRP.

All of these arguments seek to establish the evidential irrelevance of (non-informative) stopping rules, but this thesis (the SRP) does not exhaust debates about them. For instance, even if one does not accept the SRP, one can still develop arguments delimiting when stopping rules matter evidentially, or when one's attitude towards them depend on specific practical rather than evidential purposes. Such arguments are beyond the scope of the present essay, although I return to some possible connections with them in section 6, where I also conclude with some reflection on future investigation of other independent arguments for and against the SRP.

2 The Stopping Rule Principle in Bayesian and Classical Statistics

Recall the two fly-sampling experiments introduced at the beginning of section 1 and let $\theta \in [0, 1]$ be the proportion of white-eyed flies in the fruit fly population of interest. Furthermore, assume each catch is statistically independent of each other, and that the population of flies does not change during the experiment (i.e., no births or deaths). Then the statistical models

for the two experiments may be described as follows:

1. Observe N flies. The probability of observing W_1 white-eyed flies is then
$$P_\theta(W_1) = \binom{N}{W_1}\theta^{W_1}(1-\theta)^{N-W_1}. \tag{1}$$

2. Continue observing until R red-eyed flies have been caught.
$$P_\theta(W_2) = \binom{W_2+R-1}{W_2}\theta^{W_2}(1-\theta)^{R}. \tag{2}$$

Note that the number of white- and red-eyed flies caught in both experiments will be the same if and only if $W_1 = W_2$ and $W_2 + R = N$. In this case, $P_\theta(W_1)/P_\theta(W_2) = \binom{N}{W_1}/\binom{N-1}{W_2}$, which is a constant. In what follows I assume these equalities.

A Bayesian analysis of these experiments assumes a prior probability $P(\theta)$ for the population proportions and sets $P(W_i|\theta) = P_\theta(W_i)$ for $i = 1, 2$. When the likelihoods $P_\theta(W_1)$ and $P_\theta(W_2)$ are proportional, as they are in the case at hand, then from the same prior probabilities $P(\theta)$, each entails by Bayes' theorem the same posterior probabilities, $P(\theta|W_1) = P(\theta|W_2)$. So, following Steel [2003, §4], we may note that any Bayesian measure of evidence for an hypothesis of interest that depends only on the prior and posterior probabilities for that hypothesis must satisfy the SRP.[4] For example, if one understands evidence in terms of confirmation by data W_i [Huber, n.d., §6b], both the log-ratio confirmation measure
$$r(W_i, \theta) = \ln\left(\frac{P(W_i|\theta)}{P(\theta)}\right) \tag{3}$$
and the log-likelihood confirmation measure
$$l(W_i, \theta) = \ln\left(\frac{P(W_i|\theta)}{P(W_i|\neg\theta)}\right) \tag{4}$$

[4]Arguments related to this had been much earlier stated [Edwards et al., 1963, p. 237], its conclusion well-known [Savage, 1962, p. 17], but Steel [2003] was, as far as I know, the first to point out the implicit assumption about the dependence of the evidential measure on only the priors and posteriors. (This is not because, e.g., Savage and others might have been considered to focus more on decision rather than evidence; they just seemed to assume as a matter of course that evidence for a hypothesis provided by data is given by the posterior probability for that hypothesis.) When this assumption does not hold, Bayesian measures of evidence need not satisfy the SRP.

have this property. (To see this in the latter case, note that by Bayes' theorem, $P(W_i|\neg\theta) = P(\neg\theta|W_i)/P(\neg\theta) = (1 - P(\theta|W_i))/(1 - P(\theta))$.)

By contrast, classical statistical methods, whether Fisherian or Neyman-Pearsonian, will not satisfy the SRP, insofar as they rely on data whose values depend on the probability distribution of possible—not just actual—data, and clearly the two sequential experiments' possible outcomes are not the same. Explicitly, if data w_i are recorded, they will calculate for any hypothesis θ the p-value $P_\theta(W_i \geq w_i)$, the probability of measuring data at least as extreme (i.e., unlikely) as the data actually measured. The Fisherian then takes the p-value as a measure of disconfirmation for θ, with smaller values indicating higher disconfirmation [Howson and Urbach, 2006, Ch. 5.b]. Thus, data are evidence against that hypothesis to the extent that the data actually measured were extreme or unlikely.

In Neyman-Pearson testing, one sets a threshold value α, called the significance level or type I error rate of the test, so that if the p-value falls below it, the hypothesis is "rejected" but is "accepted" otherwise [Howson and Urbach, 2006, Ch. 5.c]. Additionally, one must select a test statistic that minimizes the probability of acceptance when the hypothesis is actually false, called the type II error rate of the test. These decisions are supposed to be tied with particular actions, hence do not in general have a substantive epistemic import. Consequently, many scientists practice a hybrid of the two types of testing, according to which "rejection" is interpreted as a type of qualitative disconfirmation, while "acceptance" is only interpreted as neither confirmation nor disconfirmation [Mayo, 1996, Ch. 11]. (For further discussion of classical statistical testing, see, e.g., Romeijn [2017, §3.1.1].)

For concreteness, suppose that we are interested in testing whether white- and red-eyed flies are equally represented ($\theta = 1/2$), and that $N = 12$ for the first experiment while $R = 3$ for the second—i.e., $w_1 = w_2 = 9$.[5] Then the p-values for the two sequential experiments come out as

$$P_{1/2}(W_1 \geq 9) = \sum_{w_1=9}^{12} \binom{12}{w_1} \left(\frac{1}{2}\right)^{w_1} \left(1 - \frac{1}{2}\right)^{12-w_1} \approx 0.07, \tag{5}$$

$$P_{1/2}(W_2 \geq 9) = \sum_{w_2=9}^{\infty} \binom{w_2 + 3 - 1}{w_2} \left(\frac{1}{2}\right)^{w_2} \left(1 - \frac{1}{2}\right)^{3} \approx 0.03 \tag{6}$$

[5]The example is an amalgam of those by Savage [1962, pp. 17–8] and Mayo and Kruse [2001, pp. 387–8].

Therefore a Fisherian test of significance would quantify the evidential value of the two experiments differently. Further, a Neyman-Pearson test of the hypothesis that $\theta = 1/2$ at significance level $\alpha = 0.05$, the most commonly selected value, would lead to its rejection (disconfirmation) with the second experiment but not with the first.

If one were to adopt one or the other of these frameworks, one would commit oneself for or against the SRP. But if one has not yet made such a commitment, what can be said? There are other arguments that have been made for the SRP independent of these frameworks, arguments to which I now turn.

3 The Arguments from Intentions and Deception

In this section and the next, I describe and rebut four sorts of arguments for the SRP. The two arguments in this section—from intentions (section 3.1) and from deception (section 3.2)—both take the form of a modus tollens: a failure to adopt the SRP leads to failure of scientific objectivity, i.e., the freedom of the epistemic import of scientific evidence from the personal biases, conventions, and choices of researchers producing the evidence [Reiss and Sprenger, 2017, §4], with deleterious consequences for the scientific enterprise's ability to self-correct; since scientific objectivity in this sense should be upheld, the SRP should be adopted. For present purposes, I shall grant the second premise, focusing my criticism on the first. Part of my strategy will be to stress what all parties to the debate already acknowledge, that stopping rules are a part of the design of an experiment. In doing so I will often use the following heuristic: replace mentions of "stopping rules" in an argument for the first premise with mentions of "experimental design". This isn't absolutely necessary in order to identify what goes wrong with those arguments, but I have found it helpful nonetheless to demystify stopping rules, and hope the reader does as well.

3.1 The Argument from Intentions

Perhaps the most well-known argument for the SRP is the *argument from intentions*. Savage [1962, p. 76] describes the gist of the argument thus, as he heard it from G. A. Barnard in 1952:

> The design of a sequential experiment is, in the last analysis, what the experimenter actually intended to do. His intention is locked up inside his head and cannot be known to those who have to judge the experiment.[6]

Because of these "hidden intentions," [Hacking, 1965, p. 109], it is therefore charged that violating the SRP introduces worrisome subjectivity about the experimenter's mental states into statistical analysis [Berger and Wolpert, 1988, p. 78]. In particular, classical statistics is supposed to be less objective than its Bayesian alternative because "Classical procedures ... insist that the intentions of the experimenter are crucial to the interpretation of data" [Edwards et al., 1963, p. 239], and unlike with the subjectivity associated with Bayesian priors, it is neither explicit nor is it "the kind of subjectivity that may be 'washed out' by repeated testing" [Steele, 2013, p. 945].

The argument from intentions takes as a premise the unverifiability of an experimenter's state of intention, as a part of their mental state. Because the experiment's stopping rule, ultimately, is a *part* of that intention, it too is unverifiable. But how hidden *are* intentions, really? Advocates of the argument from intentions typically take this to be so obvious as to be without need of supporting argument,[7] but in fact, there is overwhelming support that the relevant sort of intentions are just as verifiable as many other aspects of mental life. I shall present two related lines of support for this: first, in various disciplines—law, linguistics, and psychology—intent to act or behave in a certain way is relevant to their concerns. Consequently, they have methods to establish—that is, verify—ascriptions of intent. Second, in the experimental sciences in particular, there are many ways of verifying intent that require much less sophistication.

In most systems of criminal law, the mode, or level, of culpability for a conventional crime, which determines sentencing guidelines and other degrees of punishment, depends on *mens rea*, or the intentional mental state of the perpetrator [Fletcher, 1998]. Establishing various types of intent with regard to the criteria of criminal action (*actus reus*) is a necessary and routine part of criminal prosecution, usually involving careful assessment of the defendant's claims about their intent as well as the reasonably expected inevitable (or

[6] Savage continues: "Never having been comfortable with that argument, I am not advancing it myself." However, he does shortly thereafter [Edwards et al., 1963]. (See also the discussion by Mayo [1996, p. 346–7].)

[7] Perhaps it's a hangover from radical behaviorism?

even probable) consequences of their observed actions. Thus, even when justice for individuals and societies is at stake and consequences are the highest, intent is not something at all unverifiable or subjective, but can be established through the usual evidence one gathers to establish the mental state of another.

Competent speakers of a particular language use similar everyday methods to infer the meaning of utterances. For example, to understand whether a speaker asking "Can you lift your feet?" either inquires about mobility or issues a request depends on contextual features that any language user with pragmatic competence will recognize—e.g., the speaker's operation of a vacuum cleaner. Moreover, recognizing a speaker's intentions goes beyond syntactic competence and is necessary for fluency in a second language [Koike, 1989]. Far from being unverifiable, correctly ascribing intent is essential for interpersonal communication and coordination of action.

Finally, in cases where more extreme precision is needed, psychologists have developed entire empirical research programs dedicated to describing in fine detail the interrelations between attitude, norms, motivations, intentions, and actions [Fishbein and Azjen, 1975, Azjen and Fishbein, 1980]. One of the most influential of these is the Theory of Planned Behavior [Azjen, 1985, 1991], which also includes perceived behavioral control. This has been successful enough to lead to tools for both prediction and control of behavior, such as in matters of health and public policy [Fishbein and Azjen, 2011]. In all this research, intentions are measured through carefully designed and calibrated questionnaires asking directly about introspected intention, behavior, and other related factors, which, when modeled correctly, have been verified to correlate with actual behavior appropriately cued.[8]

The purpose of this cursory review of three fields' involvement regularly establishing intent is not to endorse deferentially all the variegated details of their involvement, but rather show merely that there is quite substantial evidence against the claim that intentions are unverifiable. In the face of this, skeptics may always question whether these fields can really establish strong enough evidence about intent to make it evidentially relevant. But without any details of such a counterargument provided, it is hard to see it as anything beyond a form of generalized and self-defeating skepticism. If reported and

[8]Perhaps it goes without saying, but this is emphatically not a fringe or speculative research program in psychology: as of the end of September, 2018, when this passage was written, these five works collectively have over 176,000 citations according to Google Scholar.

assessed intent to stop an experiment is not considered verifiable according to the usual means, why not extend the same conclusion to reports of the data itself? I take it that this position has already been ceded by advocates of the SRP.

What implications therefore does this entail for the argument from intentions? Following the heuristic proposed at the beginning of this section leads us to the following contrast:

> [The experimenter's experimental design] is locked up inside his head and cannot be known to those who have to judge the experiment.

This clearly isn't so, as detailed descriptions of experimental procedures, including those that encompass what might have been observed but was not, are routine in experimental science. The point is not merely that stopping rules are typically reported in practice, but that the stopping rule for an experiment can be and usually is conclusively verifiable, which confutes the claim that stopping rules are unverifiable or hidden for experiments. Good experimental practice recognizes both the need to make this possibility an actuality in reporting an experiment, as well as its evidential relevance. Typically an experimenter is verily obligated to report all the details of the design that are relevant for replications thereof, including the stopping rule, as Gillies [1990, p. 95] has emphasized.

Howson and Urbach [2006, p. 160] have objected that, because not all properties of an experiment need be similar in a replication—e.g., the color of the experimenter's shoes—an *independent* reason is needed to understand the stopping rule as an evidentially *relevant* part of the experimental design.[9] A successful independent argument against the SRP could provide such a reason, but such an argument is yet forthcoming, they claim. It's beyond the scope of the present essay to evaluate whether there are such arguments—see section 6 for a reference to one prospect. However, this objection cuts both ways: not having a reason to include the stopping rule as evidentially relevant is not itself a reason to exclude it as evidentially irrelevant. In other words, not having a specific prescription about whether stopping rules are included in the evidentially relevant part of the description of experiments is not the same as a permission to exclude that information as irrelevant.

[9]They also point out that replication is not necessary in some sciences, but this is besides the point: as long as it is a concern in some sciences, it helps block the argument from intentions.

Howson and Urbach [2006, pp. 158–9] present a variation of the argument from intentions that deserves a separate response. They invite the reader to imagine that "two scientists collaborate in a trial [i.e., an experiment], but are privately intent on different stopping rules; by chance, no conflict arises, and the result satisfies both" [Howson and Urbach, 2006, p. 158]. For example, they could be observing the eye color of flies, as in the example from section 2, with one intent on stopping after observing 12 flies, while the other intent on stopping after observing 3 red-eyed flies. In such a situation it is ambiguous what the true stopping rule is—which scientist would prevail if they were to run into conflict about stopping the experiment?—yet it seems, all else being equal, that one ought to be able to interpret the resulting data evidentially. They write in conclusion: "*We suggest that such information about experimenters' subjective intentions* [...] *has no inductive relevance whatever in this context*" [Howson and Urbach, 2006, p. 158, emphs. orig.].

First, I note that on logical grounds alone this argument cannot be a modus tollens, as described at the beginning of this section, for the SRP. Because the SRP has the logical form of a negated existential claim, one cannot argue for it in a modus tollens by providing a counterexample to a universal claim. In a word, producing an example in which the stopping rule does not matter evidentially does not show that it never does.

Second, the argument does not even establish that the stopping rule sometimes does not matter evidentially to a hypothesis. This is because it conflates ambiguity with inscrutability or irrelevance. As it was described, the only definite stopping point for the experiment is when 12 observations have been made, 3 of which (including the last one) were of red eyes. But this does not imply directly the evidential irrelevance of the stopping rule, nor the impossibility of making evidential claims from this experiment. An advocate for the evidential relevance of the stopping rule in this case may still maintain that the data from the experiment bear evidentially, if ambiguously, on the hypothesis of equal representation of white- and red-eyed flies. Such an advocate can analyze the data according to different hypothetical plausible definite stopping rules, and present these analyses together, emphasizing what they have in common. For example, if the scientific collaboration was among peers, then due deference would likely have them stop the experiment when one decides to. If they were not peers, deference in stopping the experiment would go to the stopping rule of the epistemic superior. Such pieces of information can help plausibly remove ambiguity from the evidential evaluation of the experiment, but even without them an advocate for the

evidential relevance of stopping rules does not face an experimental outcome with inscrutable evidence.[10]

3.2 The Argument from Deception

A different argument suggests that the absence of ways to account for intentions blows a breach into the bulwark of the scientific community's methods for self-correction. It is thus a failure not of product objectivity, but process objectivity [Reiss and Sprenger, 2017, §1]. For instance, Sprenger [2009, p. 641] advances the following contrast.

> Using fake data involves considerable risk: if continued replications fail to reproduce the results, our experimenter will lose all her reputation. By contrast, she can never be charged for insincerely reporting her intentions. The crucial point here is ... that the scientific community is unable to *control* whether these intentions have been correctly reported.

In a reversal of what is sometimes charged of methods that obey the SRP [Mayo and Kruse, 2001, Mayo, 1996, Ch. 10.3], the argument from deception advances that abandoning the SRP *leads* to problems with misleading evidence, for experimenters can gain evidential advantage by reporting the stopping rule for their experiment that maximizes (or minimized) the evidential import of the experiment for a chosen hypothesis.

But is this so? Following the replacement heuristic for the experimenter considering insincerity:

> She can never be charged for insincerely reporting her [experimental design]. The crucial point here is ... that the scientific community is unable to *control* whether these [experimental designs] have been correctly reported.

[10] Another possible response, suggested by Livengood [2017], is that in fact *two* experiments were being performed, since sometimes creative intentions can matter to what exists. This would be the case when the design is part of the experiment itself, so that two different designs entail two different experiments. The evidential import of each experiment, then, can be evaluated separately. It is not yet clear to me how one should understand these two experiments with respect to the problem of use-novelty or double-counting of data, so I won't discuss it further.

But this is contravened by examples from actual scientific controversies. For example, Franklin [1994] describes the case of the alleged detection of gravitation waves by Joseph Weber in the 1960s and 1970s. After replication attempts by many other groups, Weber's claims were not substantiated, on which the scientific community came to a consensus. Importantly, Weber did not falsify data, but made his experimental design and data analysis plan available, if not entirely in his publications, then through contacts with other researchers. This design included rules for how to start and stop taking data, and when to exclude outliers. By 1977, "he had lost all credibility as far as gravity wave experiments were concerned" [Franklin, 2010, p. 126], although that loss did not extend to other aspects of his work.[11] So, similar mechanisms for self-correction apply here for design as they do for the data themselves, and the possibility of replication is precisely the most central tool for doing so. If continuing replications using the reported experimental design fail to reproduce relevantly similar result, the original experiment will be discredited.

4 The Arguments from Impracticality and Waste

Another class of arguments against the SRP also take the form of a modus tollens: a failure to adopt the SRP leads to pragmatic difficulties, e.g., undue burdens in performing or analyzing experiments or wasting valuable resources; since these pragmatic difficulties are to be avoided, the SRP must hold. These arguments thus involve two sorts of premises: a factual claim that allowing for the evidential significance of any experiment's stopping rule leads to pragmatic woes, and a conditional claim along the lines that an evidential principle—such as the denial of the SRP—holds only if it does not lead to insuperable practical difficulties.

Although I have just framed this class of arguments in the indicative mood, one can also read them as operating on a sort of meta-level about evidential frameworks. On this version, one implicitly supposes that there are different internally viable frameworks for evidence, some of which satisfy the SRP and some of which do not. One is then interested in the external question of choosing one framework over another, criteria for which can

[11] Although credibility can be considered a property of a scientist qua epistemic agent, it seems more important in scientific endeavors as attached to particular claims; Shapin [2010], for example, defends the particularity of credibility claims in science.

well be practical [Carnap, 1950]. For such versions, the second premise of the argument is normative, something like: one *should adopt* an evidential principle only if it does not lead to insuperable practical difficulties.

In what follows, I consider and criticize two arguments of this form, those from impracticality (section 4.1) and from waste (section 4.2). Their first, descriptive premises are not in general true, while their second, conditional or normative premises are charitably tacit—charitably, because without them, these arguments would simply be fallacious ad consequentiam. Further, for the argument from waste, it is hard to find any hint of how its second premises could be motived without adopting a whole framework for evidential reasoning, such as a version of Bayesianism. Without articulating such a motivation, its doesn't provide *independent* reasons why undesirable pragmatic consequences should lead to the adoption of an evidential or epistemological principle.

4.1 The Argument from Impracticality

Sprenger [2009, p. 642] summarizes the argument from impracticality thus: "Specifying the stopping rule in advance sounds good, but specifying the correct, comprehensive stopping rule (which we need to interpret the results properly) is practically impossible." The difficulty is that "there will often be unforeseen eventualities that crop up in sequential experimentation" [Berger and Wolpert, 1988, p. 77] that the experimenter did not anticipate in the description of the experiment's stopping rule. This descriptive premise is supposed to be the case for any experiment, so rejecting the SRP leads to the infeasibility of the evidential evaluation of any sequential experiment. If one then assumes that "ought" implies "feasible" and that one ought to evaluate the evidential significance of each of our sequential experiments, the conditional or normative premises, one then arrives at a contradiction.[12]

There are actually two versions of the argument from impracticality that vary based on the reading of the first, descriptive premise. The stronger version takes Sprenger's asserted practical "impossibility" literally: what is infeasible is a literally correct and comprehensive description of the conditions under which an experiment would stop. For instance, what is the probability at any time that the experiment will stop due to a meteor striking

[12]Except for insisting on modifications of feasibility to approximate feasibility, I will not challenge this premise further, although one could [Southwood, 2016].

the laboratory building? Since this is admittedly practically impossible, the conditional or normative premises entail that one is not obligated to specify the stopping rule of an experiment. However, rejecting the SRP does not in fact require such comprehensiveness.[13] All that is needed is the *probability* of the experiment stopping at a particular point in the sequences of trials.[14] To illustrate this, consider again the fly-sampling experiments from section 2. Just as it isn't necessary to describe how each fly is physically caught from and released to the population to analyze the data from each, only the probabilities (including probabilistic independences) of the outcomes, so too it isn't necessary to describe just all the ways an experiment could stop, just the probabilities that it does so at various steps in the experiment.

This leads to the weaker version of the argument from impracticality, according to which the practical difficulty comes not from the existence of unforeseen ways in which the experiment could end, but from the infeasibility of determining the probabilities of these ways. But conceding this practical impossibility does not entail the impossibility of taking the evidential bearing of stopping rules into account *within a good approximation*. Indeed, the conditional or normative premise about evaluating the evidential significance of each of our sequential experiments should allow for that evaluation to be approximate.

When one had good reason to believe that the probabilities of unexpected stopping are small, one can *idealize* the stopping rule as something much simpler. For example, in the calculations of likelihood of the eye color experiment, suppose the probability that it will definitely halt due to factors largely independent of the data collected—meteors hitting the building, etc.—is given by ϕ. Then the "fixed-N" sample experiment has a probability of collecting W_1 white-eyed flies and $N - W_1 = R_1$ red-eyed flies given by

$$\begin{aligned} P_\theta(W_1) &= P_{\theta,\phi}(W_1 | R_1 + W_1 = N) P_\phi(R_1 + W_1 = N) \\ &= \binom{N}{W_1} \theta^{W_1} (1-\theta)^{N-W_1} (1-\phi)^N. \end{aligned}$$

[13] The issue of "correctness" is actually orthogonal to the issues here and concerns more whether the statistical model and experimental design for the sequential experiment were misspecified.

[14] Steele [2013, p. 945] suggests that all one needs to do is include the different stopping events in the total outcome space for the experiment, but this does not rebut the argument against the practical impossibility of specifying what these all are, as she seems to suggest.

When ϕ is very small, as it often is, it may be idealized away without substantial misrepresentation of the evidential import of the experiment's result.[15]

The case is entirely analogous to cases of idealization from Newtonian physics. In order to calculate the trajectory of a projectile, one needs to determine all the forces on it at all times, as well as how the projectile deforms in response to forces. But practically small forces on fairly rigid projectiles can be neglected without significant loss. So, too, the tools of idealization turns practical impossibility to practical possibility for statistical models.

Once cases such as this are admitted, the descriptive premise in the argument from impracticality no longer holds. Because it is not generally the case that representing the probability of stopping, at least to a good approximation, is impractical, it cannot be that unfeasible demands to represent this probability lead one to the SRP, i.e., to reject the stopping rule's evidential relevance in all cases. Of course, there are many examples of sequential experiments with multiple complex stopping mechanisms, such as when a clinical trial shows excess harm or benefit [Whitehead, 1997]. But the existence of these examples is insufficient, logically, to prove any claim about the universal infeasibility of taking into account stopping rules, which is needed for the form of the modus tollens argument. Moreover, it is doubtful that they are as infeasible as sometimes complained: there is a voluminous literature, initiated for the most part by Wald [1947], on how to model them properly and understand their evidential relevance—see, e.g., Siegmund [1985] and Whitehead [1997]. Claimants of impracticality or unfeasibility ignore about 75 years of developments at their peril.

One could admit these objections but then treat the argument from impracticality as a consideration in favor of an evidential framework that requires fewer demands on researchers. Although taking into account the evidential significance of stopping rules is feasible, it can still be subtle and time-consuming. This version of the argument shades into a version of the argument from waste, to which I turn attention presently.

4.2 The Argument from Waste

Not conforming our notions of evidence to the SRP, its advocates sometimes contend, leads to wasted resources:

[15] One can also treat more complicated stopping rules [Raiffa and Shlaifer, 1961, pp. 39–40].

> if we believe in the evidential, postexperimental relevance of the stopping rule, then we have to be silent on the meaning of data where the stopping rule is unavailable. But if we throw the data into the trash bin, we give away a great deal of what reality tells us. [Sprenger, 2009, p. 642]

In the most extreme case, "If the experimenter forgot to record the stopping rule and then died, it is unappealing to have to guess his stopping rule in order to conduct the [data] analysis" [Berger and Wolpert, 1988, p. 78]. By contrast, for those who adopt the SRP,

> This irrelevance of stopping rules to statistical inference restores a simplicity and freedom to experimental design that had been lost by classical emphasis on significance levels ... Many experimenters would like to feel free to collect data until they have either conclusively proved their point, conclusively disproved it, or run out of time, money, or patience. [Edwards et al., 1963, p. 239]

The second, tacit premise seems to involve the concept of dominance. Understood as a pure (non-normative) conditional, it might read: if an evidential principle entails that, for any sequential experiment, its analysis according to that principle would use or waste no more resources (and uses less or wastes less on at least one occasion) than its analysis according to that principle's denial, then that principle is true.

This (non-normative) conditional premise is highly implausible. Its only motivation, as far as I can see, would be to save the argument from waste from committing an ad consequentiam fallacy. In general just because data analysis would be simplified or resources less wasted by following the SRP doesn't make it true.[16] Thus it is more charitable to interpret the second, tacit premise normatively: if a framework for evaluating evidential entails that, for any sequential experiment, its analysis within that framework would use or waste no more resources (and uses less or wastes less on at least one occasion) than its analysis according to some other framework, then one should adopt that framework. Though much more plausible, this version of the premise also has its problems, to which I will return after discussing the first, descriptive premise.

[16] See also the discussion in Mayo [1996, p. 350–1].

Like with the argument from impracticality, there are two versions of the argument from waste that vary based on the reading of this first premise. The stronger version takes literally Sprenger's assertion that declining to evaluate data from a sequential experiment whose stopping rule is unknown is a kind of *evidential waste*. Cohen [2010, p. 234, emph. orig.] usefully analyzes waste as "any process wherein something useful becomes less useful and that produces less benefit than is lost—*where benefit and usefulness are understood with reference to the same metric*". Here, the data is supposed to be wasted because its evidential use, e.g., in confirming or disconfirming hypotheses, is lost unless it is analyzed. However, in order for declining to analyze the data to count as an evidential loss, it must be evaluated with the *same* framework according to which analyzing the data would count as an evidential gain. But any such framework *cannot* be one according to which the stopping rule is evidentially relevant, for within such a framework, "reality" doesn't "tell us" anything definite through a data set without taking into account the stopping rule. Thus, this strong version of the first, descriptive premise begs the question for the SRP.

Another problem with the stronger version of this premise is that it seems to assume that experimenters working in an evidential framework in which stopping rules are relevant never have recourse to analyzing an experiment if the stopping rule is not known. However, one can often contact the experimenter or their colleagues or students by various means, complete a couple data analyses conditional on different plausible stopping rules, or simply attempt to replicate an experiment.[17] Admittedly, these techniques require more work and may involve more vagueness in justified conclusions than those used within an evidential framework that accepts the SRP. This leads to the weaker version of the first, descriptive premise, suggested more from the quotations by Berger and Wolpert and Edwards et al.. According to this version, adopting an evidential framework that makes stopping rules evidentially relevant does not waste the evidential value of data, but rather requires more *cognitive resources* to evaluate the data from any sequential experiment

[17]Data sets that are so intractable as to be insusceptible even to analyses conditional on various plausible stopping rules, have, on our substitution heuristic, a substantial part of their experimental design unarticulated. In such cases I do not see any necessity to analyze them. But even if there were, and even setting aside the other problems with the stronger version of the descriptive premise, establishing that there are some such intractable examples is logically insufficient to establish the SRP, for the same reasons as for the argument from impracticality discussed in section 4.1.

than from within a framework that accepts the SRP.

I do not contest that adopting the SRP simplifies data analysis or that cognitive effort is a relevant external criterion for adopting an evidential framework. Indeed, it seems to me that this advantage, hence this version of the argument from waste, is most convincing to applied researchers, among the arguments I consider in this paper. But for a framework that adopts the SRP to be truly considered less wasteful of cognitive resources than one that does not—and here I return to the normative version of the second premise—the two frameworks must be compared along more dimensions than just their use of (cognitive) resources. As evidential frameworks, they must be apt, or at least comparable in their aptness, for the same goals.

To illustrate, consider the following anarchist evidential framework: draw whichever conclusions one wishes immediately from an experiment. Because it requires essentially no cognitive resources at all, the normative version of the dominance premise entails that one should adopt this evidential anarchism over any otherwise reasonable alternative. One is driven to this absurdity—I take it that all sides of the debates around the SRP are motivated, if only implicitly, by a rejection of evidential anarchism—only because that premise does not take into account other essential criteria for comparing frameworks, such as their aptness for the particular tasks for which evidence is fashioned.

This raises the pertinent question of whether two arbitrary evidential frameworks, about which we have assumed only that they differ on the SRP, must be fashioned for the same (external) tasks. Surprisingly, advocates of the argument from waste seem to assume that they are, but surely this is theft over honest toil. As Kelly [2016] describes, concepts of evidence serve multiple roles, some complementary and some in tension: evidence can be that which justifies belief, or according to which rational folk apportion belief or determine action; it can also be that which confirms or corroborates a hypothesis, as a guide to truth; and it can be whatever objectively (or intersubjectively) arbitrates between competing hypotheses. Assuming that any two evidential frameworks are arranged for the same myriad roles is implausible, and I know of no reasons for it.

The upshot of these observations is that without further such reasons, the argument from waste cannot succeed in establishing the SRP—without, at least, abandoning the goal to give an argument for (or against) the SRP that does not adopt any particular evidential framework. Typically, proponents of this argument—Edwards et al. [1963], Berger and Wolpert [1988], and

Sprenger [2009] included—aim to compare not arbitrary evidential frameworks but ones using concepts from Bayesian and classical statistics, respectively. This at least allows more concrete traction on the problem of comparison, but gives up on the framework-independent arguments that have been my focus in this essay. Nevertheless, these proponents too seem to assume that these two feuding families of ideas also have the same goals. But as Mayo [1996, Ch. 10] has forcefully argued, these goals are in fact quite different: "the underlying rationale of a number of methodological rules [in classical statistics] is the aim of reliability or severity in the sense I have been advocating, yet that aim runs counter to the aim reflected in Bayesian principles" [Mayo, 1996, p. 320]. Thus, if Mayo is right, without already assuming the methodological aims of an evidential framework, the argument from waste does not pronounce any clear verdict, even between Bayesian and classical frameworks for statistical inference and evidence.

5 The Decision-Theoretic Argument

Sprenger [2009, pp. 645–7] has put forth an argument for the SRP quite different from the previous ones, based on decision theory. Although the full technical statement of the proposition used in the argument is quite involved, the basic idea is simple: roughly, if one does not adopt the SRP in certain circumstances, one is led to incoherence in the sense of having inconsistent or irrational preferences.

In a bit more detail, Sprenger [2009, p. 645] assumes the framework of Neyman-Pearson testing in which decisions about the acceptance and rejection of an hypothesis are made on the basis of the p-value of a statistic chosen to minimize the type II error rate given a fixed type I error rate (significance level). He further assumes that either type of error is worse, in terms of its utility as a confirmational or inferential decision, than the corresponding correct decision for a fixed truth or falsity of the hypothesis under test. Next, he compares two decision rules for a given sequential experiment: one, δ_S, which evaluates the Neyman-Pearson test for a specific hypothesis H_0 against a specific alternative H_1 according to the stopping rule actually used in the experiment, and another, δ_τ, which does the same except according to a fixed stopping rule τ. (For example, τ may indicate analyzing the data as if it were collected according to a predetermined fixed sample size, regardless of what the stopping rule actually was; by contrast, S indicates analyzing

the data according to the actually used stopping rule.) Once one fixes the data from the sequential experiment—but not the stopping rule—then either $\delta_\tau = 0$ ("accept H_0 and reject H_1") or $\delta_\tau = 1$ ("accept H_1 and reject H_0"). Denoting $R(\theta, \delta)$ as the risk, or expected loss, of making decision δ if θ, a specific state of the world, is true, he proves that for all values of θ, either $R(\theta, 0) < R(\theta, \delta_S)$ or $R(\theta, 1) < R(\theta, \delta_S)$. The important corollary is that "Preferring δ_S over $\delta_\tau = 0$ and $\delta_\tau = 1$ leads to *incoherence* for any value of $[\theta]$, in the sense that a Dutch book (namely, a sure loss) can be constructed against these preferences" [Sprenger, 2009, p. 646].

But how successful is this as an argument independent of a framework for statistics? Despite Sprenger's insistence that he has demonstrated that classical statisticians are thus "beaten in their own game" [Sprenger, 2009, p. 645], classical statistical testing, Neyman-Pearson or otherwise, just isn't in the business of providing decision rules that are coherent in a Bayesian sense—their own "game" involves procedures for minimizing error probability, not updating coherent personal probability assignment to states of the world or making bets or forecasts that minimize Bayes risk. One would accept Bayesian coherence as a desideratum if one already accepts the Bayesian framework, but then if one is using typical methods in that framework, they would already satisfy the SRP; why would a Bayesian be using Neyman-Pearson testing?

Instead of dismissing Sprenger's argument for committing a non sequitur or begging the question, one can instead reconfigure it with more neutral assumptions and inferences using the recent work of Malinksy [2015], who provides a way of transforming Bayesian coherence arguments against Neyman-Pearson testing to ones whose conclusions are entirely stated within the Neyman-Pearson framework. To do so, he introduces the concept of *combined risk*. Consider a collection of n experiments labeled by an index $a_i \in I_n = \{a_1, \ldots, a_n\}$ and concerning a common (parameter) space of disjoint hypotheses Θ, and let δ be a decision rule defined over the possible outcomes of all of these experiments. Letting $R_{a_i}(\theta, \delta)$ be the risk of experiment a_i when the hypothesis (state of the world) $\theta \in \Theta$ obtains, the combined risk of the collection I_n is then defined as $R_{I_n}(\theta, \delta) = \sum_{i=1}^{n} R_{a_i}(\theta, \delta)$.

Next he formulates a version of the dominance principle in classical decision theory naturally extended to the case of combined risk.

Dominance: Given a collection of n experiments labeled by an index in the set I, if $R_I(\theta, \delta_1) \leq R_I(\theta, \delta_2)$ for all $\theta \in \Theta$, and the inequality is strict

for some θ, prefer δ_1 over δ_2.

Finally, he shows how in these terms, one can interpret a key result of Schervish et al. [2002] as implying that, for any Neyman-Pearson decision rule, there exists a pair of other decision rules and a pair of experiments such that performing the other decision rules respectively on the two experiments has a lower combined risk than performing the Neyman-Pearson decision rule on both. Thus, by dominance, preferring the Neyman-Pearson rule leads to a contradiction.

In applying these ideas to decision rules involving stopping rules, it suffices for the purposes at hand to do so for the specific case of Sprenger's argument. The key observation[18] is that $\frac{1}{2}R(\theta, 0) + \frac{1}{2}R(\theta, 1) < R(\theta, \delta_S)$, hence $R(\theta, 0) + R(\theta, 1) < 2R(\theta, \delta_S)$. Using Malinsky's strategy, we can interpret this as stating that given two identical experiments, the combined risk of performing Neyman-Pearson testing according to the actual stopping rule used on both is higher than that of performing the same testing according to a fixed stopping rule entailing rejection on the one and acceptance on the other. Thus, preferring to evaluate the test using the stopping rule actually used, as opposed to one fixed stopping rule for the first experiment and another for the second, conflict with the dominance principle.

This is an improvement on Sprenger's argument, for it does not presuppose the relevance of decision-theoretic concepts (such as coherence) foreign to classical statistical testing. But even in this form, it has somewhat limited scope. Unlike the arguments considered in sections 3 and 4, it can at most establish that there are problems with denying the SRP within Neyman-Pearson testing; it's entirely compatible with the SRP being false in general. Nevertheless, if one can establish that a major evidential framework that rejects the SRP has internal problems, it might provide some (eliminative) inductive support to evidential frameworks that accept the SRP, such as certain versions of Bayesianism.

However, the decision-theoretic argument is not even successful at this task. In the first place, it is well-acknowledged that Neyman-Pearson testing is better interpreted as a procedure for making decisions rather than determining evidence for hypotheses [Mayo, 1996, Ch. 11], so proponents of classical statistical methods would view it as straw man for criticizing evidential procedures that violate the SRP. Moreover, the problems that the decision-theoretic argument captures for Neyman-Pearson testing are actually *inde-*

[18] Sprenger [2009, p. 647] attributes it to Teddy Seidenfeld.

pendent of assumptions about the SRP. To see this, recall that argument encoded the evidential relevance of the stopping rule for Neyman-Pearson testing in the decision rule δ_S, which indicates a decision to analyze the data according to the stopping rule actually used. The only features of δ_S used in the proof by Sprenger [2009, p. 646] is that if $\delta_\tau = 0$ and H_0 is true, then there is a probability strictly between zero and one that $\delta_S = 0$—i.e., that the test correctly accepts H_0—and similarity if $\delta_\tau = 1$ and H_1 is true. Thus the proof goes through unchanged if we replace δ_S by any decision rule with this property, including one that satisfies the SRP. For example, consider the rule δ_R according to which H_0 is accepted (and H_1 rejected) simpliciter with probability $p \in (0, 1)$ and rejected (with H_1 accepted) with probability $1 - p$. Though not a very useful version of Neyman-Pearson testing, it satisfies the SRP insofar as the stopping rule is irrelevant to the decision to accepted or reject a hypothesis. Moreover, it conflicts with the dominance principle just as δ_S does. Thus, while the decision-theoretic argument reveals a problem with Neyman-Pearson testing, that problem does not follow from any assumptions about the evidential significance of stopping rules.

Perhaps this should not be so surprising, for the method used in the proof involves the fact, known at least since the work of Cox [1958], that one can always find mixed tests that dominate a given Neyman-Pearson test, simply because it arises from a fixed-level testing procedure—the test is chosen to minimize type II error *given* a fixed type I error. This is certainly a problem for Neyman-Pearson testing, but arguably that procedure's assumption of fixed-level testing is not essential to the broader framework of classical statistics. One can use the whole apparatus of testing while letting the significance level vary according to certain features of the data. For example, Berry and Viele [2008] construct a coherent decision rule—one not susceptible to the type of argument under discussion—for the case of data from normal distributions. In this case, the type I error rate (significance level) drops monotonically as a function of sample size—see Malinksy [2015] for further discussion. In sum, the problems that the decision-theoretic argument reveals arise from fixed-level testing, not (in)attention to stopping rule, which is not an essential feature of classical statistical testing.

6 Conclusions and Future Work

I have rebutted five independent arguments for the SRP: the arguments from intentions and deception (section 3), from impracticality and waste (section 4), and a decision-theoretic argument (section 5). In three of these (sections 3.1, 3.2, and 4.1), I employed a heuristic substitution of "experimental design" for "'stopping rule" to reveal some of those arguments' flaws. For the last argument, originally due to Sprenger [2009], I suggested how it could be improved using ideas from Malinksy [2015], but in the end it was still unsuccessful in establishing the SRP. Thus, these arguments do not provide independent support to typical Bayesian frameworks for statistical analysis.

These are certainly not the only independent arguments for the SRP. I have not considered a different decision-theoretic argument by Berger and Wolpert [1988, pp. 83–5], nor arguments that seek to establish the SRP through arguments for the likelihood principle [Birnbaum, 1962, Berger and Wolpert, 1988], which entails it.[19] Nor have I considered arguments against the SRP, such as the foregone conclusions argument [Mayo and Kruse, 2001, Mayo, 1996, pp. 351–7] suggested by Armitage already in 1959 [Savage, 1962, p. 72], much less attempts to deflect it [Backe, 1999] or explain it away [Steele, 2013, Gandenberger, 2015]. Since the foregone conclusions argument in particular is typically formulated within the framework of classical statistics, a reformulation in framework-independent terms would be helpful to assess it fairly. Here, the substitution heuristic may be useful, as it forces one to see the stopping rule as an element of experimental design. Only once these arguments have been made more precise can we be in a position to adequately assess their strength against the SRP, as I have done here for the arguments for it, freed from question-begging background assumptions. The result can then provide support in debates about the proper framework or frameworks for statistical evidence. If those frameworks permit the evidential relevance of stopping rules, one can also turn to further practical questions, e.g., about how much of a difference they make and the most economical way, in the sense of the scientific cognitive economy, to bring nonideal practice closer to ideal evidential principles. These tasks will be taken up in future work.

[19] As described in these references, the likelihood principle is equivalent to two other principles of sufficiently and conditionality, respectively. Some arguments for the conditionality principle are similar to those for the SRP.

Acknowledgements

Thanks to Greg Gandenberger, Kasey Genin, Jonathan Livengood, Dan Malinksy, Conor Mayo-Wilson, Jan Sprenger, and an anonymous referee for comments on a previous version, and audiences at Minnesota, Munich, Bologna (SILFS2017), Edinburgh (BSPS2017), and Exeter (EPSA2017) for their insightful comments. Part of this work was completed with the support of a European Commission Marie Curie Fellowship (PIIF-GA-2013-628533).

References

I Azjen and M Fishbein. *Understanding attitudes and predicting social behavior*. Prentice-Hall, Englewood Cliffs, NJ, 1980.

Icek Azjen. From intentions to actions: A theory of planned behavior. In J Kuhl and J Beckmann, editors, *Action Control*, pages 11–39. Springer, Berlin, 1985.

Icek Azjen. The theory of planned behavior. *Organizational behavior and human decision processes*, 50(2):179–211, 1991.

Andrew Backe. The likelihood principle and the reliability of experiments. *Philosophy of Science*, 66(Proceedings):S354–S361, 1999.

James O Berger and Robert L Wolpert. *The Likelihood Principle*. Institute of Mathematical Statistics, Hayward, CA, 2nd edition, 1988.

Scott Berry and Kert Viele. A note on hypothesis testing with random sample sizes and its relationship with Bayes factors. *Journal of Data Science*, 6: 75–87, 2008.

Allan Birnbaum. On the foundations of statistical inference. *Journal of the American Statistical Association*, 57(298):269–306, 1962.

Rudolf Carnap. Empiricism, semantics, and ontology. *Revue Internationale de Philosophie*, 4(11):20–40, 1950.

Andrew Jason Cohen. A conceptual and (preliminary) normative exploration of waste. *Social Philosophy and Policy*, 27(2):233–273, 2010.

David R Cox. Some problems connected with statistical inference. *Annals of Mathematical Statistics*, 29:357–372, 1958.

Ward Edwards, Harold Lindman, and Leonard J Savage. Bayesian statistical inference for psychological research. *Psychological Review*, 70(3):193–242, 1963.

M Fishbein and I Azjen. *Belief, attitude, intention, and behavior: An introduction theory and research*. Addison-Wesley, Reading, MA, 1975.

M Fishbein and I Azjen. *Predicting and changing behavior: The reasoned action approach*. Psychology Press, New York, 2011.

George P Fletcher. *Basic Concepts of Criminal Law*. Oxford University Press, New York, 1998.

Allan Franklin. How to avoid the experimenters' regress. *Studies in History and Philosophy of Science*, 25(3):463–491, 1994.

Allan Franklin. Gravity waves and neutrinos: The later work of Joseph Weber. *Perspectives on Science*, 18(2):119–151, 2010.

Greg Gandenberger. Differences among noninformative stopping rules are often relevant to Bayesian decisions. 2015. arXiv:1707.00214.

Donald Gillies. Bayesianism versus falsificationism. *Ratio (New Series)*, III (1):82–98, 1990.

Ian Hacking. *The Logic of Statistical Inference*. Cambridge University Press, Cambridge, 1965.

Colin Howson and Peter Urbach. *Scientific Reasoning: The Bayesian Approach*. Open Court, Chicago, 3rd edition, 2006.

Franz Huber. Confirmation theory. In *The Internet Encyclopedia of Philosophy*. n.d. Accessed 30 Mar. 2018.

Thomas Kelly. Evidence. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, winter 2016 edition, 2016.

Dale April Koike. Pragmatic competence and adult L2 acquisition: Speech acts in interlanguage. *The Modern Language Journal*, 73(3):279–289, 1989.

Jonathan Livengood. Counting experiments. *Philosophical Studies*, 2017.

Daniel Malinksy. Hypothesis testing, "Dutch book" arguments, and risk. *Philosophy of Science*, 82(5):917–929, 2015.

Deborah G Mayo. *Error and the Growth of Experimental Knowledge*. University of Chicago Press, Chicago, 1996.

Deborah G Mayo and Michael Kruse. Principles of inference and their consequences. In David Corfield and Jon Williamson, editors, *Foundations of Bayesianism*, pages 381–403. Kluwer, Dordrecht, 2001.

Howard Raiffa and Robert Shlaifer. *Applied Statistical Decision Theory*. Division of Research, Graduate School of Business Administration, Harvard University, Boston, 1961.

Julian Reiss and Jan Sprenger. Scientific objectivity. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, winter 2017 edition, 2017.

Jan-Willem Romeijn. Philosophy of statistics. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, spring 2017 edition, 2017.

Leonard J Savage. *The Foundations of Statistical Inference: A Discussion*. Methuen, London, 1962.

Mark J Schervish, Teddy Seidenfeld, and Joeseph B Kadane. A rate of incoherence applied to fixed-level testing. *Philosophy of Science*, 69 (Proceedings):S248–S264, 2002.

Steven Shapin. *Never Pure: Historical Studies of Science as if It Was Produced by People with Bodies, Situated in Time, Space, Culture, and Society, and Struggling for Credibility and Authority*. Johns Hopkins University Press, Baltimore, 2010.

David Siegmund. *Sequential Analysis: Tests and Confidence Intervals*. Springer, New York, 1985.

Nicholas Southwood. Does "ought" imply "feasible"? *Philosophy and Public Affairs*, 44(1):7–45, 2016.

Jan Sprenger. Evidence and experimental design in sequential trials. *Philosophy of Science*, 76(5):637–649, 2009.

Daniel Steel. A Bayesian way to make stopping rules matter. *Synthese*, 58: 213–227, 2003.

Katie Steele. Persistent experimenters, stopping rules, and statistical inference. *Erkenntnis*, 78:937–961, 2013.

Abraham Wald. *Sequential Analysis*. Wiley, New York, 1947.

John Whitehead. *The Design and Analysis of Sequential Clinical Trials*. Wiley, New York, rev. 2nd edition, 1997.