# The Stopping Rule Principle and Confirmational Reliability

Samuel C. Fletcher*

*University of Minnesota, Twin Cities*

February 2, 2023

**Abstract**

The *stopping rule* for a sequential experiment is the rule or procedure for determining when that experiment should end. Accordingly, the *stopping rule principle* (SRP) states that the evidential relationship between the final data from a sequential experiment and a hypothesis under consideration does not depend on the stopping rule: the same data should yield the same evidence, regardless of which stopping rule was used. I clarify and provide a novel defense of two interpretations of the main argument against the SRP, the foregone conclusions argument. According to the first, the SRP allows for highly confirmationally unreliable experiments, which concept I make precise, to confirm highly. According to the second, it entails the evidential equivalence of experiments differing significantly in their confirmational reliability. I rebut several attempts to deflate or deflect the foregone conclusion argument, drawing connections with replication in science and the likelihood principle.

## 1 Introduction

Many types of scientific and engineering studies involve repeated observations or tests on subjects of the same type in the service of making inferences about that type. For example, lake ecologists selectively net fish to determine the their species distribution; mechanical engineers randomly test widgets coming off of a factory line to infer the reliability of the manufacturing process to implement their design; and medical scientists enroll willing patients meeting selection criteria to test the effectiveness of new procedures and drugs. Within statistics, such studies are known as *sequential experiments*,[1] and a natural question to ask about their design is how the designer ends them. It could be according to some predetermined number of observations; it could instead be determined by the content of the observations; it could

*scfletch@umn.edu

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

even be fixed by some independent random mechanism, or some complicated combination of all of these. This determination is called the sequential experiment's *stopping rule*.

In principle, the stopping rule for a sequential experiment can be *informative*, in the sense that learning it provides information or evidence about the hypotheses under consideration, beyond the information or evidence provided in the data themselves. I will provide a more in-depth discussion of this property in section 4.2, but until then, I assume that the stopping rule under consideration is not informative in this sense. (I define what it means for a stopping rule to be *non*-informative in section 2.1.)

Can the (non-informative) stopping rule of a sequential experiment affect its evidential interpretation? In other words, given two sequential experiments that yield qualitatively the same observations, must those experiments bear evidentially in the same way on statistical hypotheses, even if the experiments had different stopping rules? This question is relevant not just for philosophers of science interested in how data provide evidence for hypotheses, but also for scientific practitioners. An affirmative answer would entail the need for especial care in understanding how sequential experiments end, thus, how they are to be directly replicated (a theme to which I return in sections 3.4 and 6): evidence could depend not just on what was observed, but what could have been observed. A negative answer, meanwhile, would entail a great simplification in the aspects of experimental design relevant for evidential evaluation: data always bear the same evidence, no matter how their collection ended.

This latter negative answer is provided by the Stopping Rule Principle (SRP):

**Stopping Rule Principle (Informal):** The evidential relationship between the data from a completed sequential experiment and a statistical hypothesis does not depend on the experiment's stopping rule.

Adherents of the SRP typically apply it to experiments with complicated stopping rules, analyzing those experiment as if they were based instead on simpler fixed stopping rules. 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 out 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)

The SRP is neutral about the structure of evidence, such as whether it is comparative—i.e., whether it applies to individual hypotheses or is a relation between two of them—and whether it is qualitative or quantitative. Rather, it is a constraint on evidence's *equivalence conditions*.

Discussions of the SRP and related principles are typically couched in terms of evidential equivalence, or the irrelevance of stopping rules to evidence, where the structure and possible polysemy of evidence is otherwise left unspecified. For example, Berger and Wolpert (1988, p. 25) affirm that "We presuppose nothing about what this evidence is; it could (at this stage) be any standard measure of evidence, or something entirely new", including some multidimensional measure. Following Steel (2003), I presuppose further that evidence should be understood as *incremental confirmation*. I do not assume that confirmation is quantitative, but only that its degrees are partially ordered, their order interpreted as "at least as

confirmatory as," and that they have a distinguished element, the neutral degree. Elements larger than the neutral degree are (positive) confirmatory degrees, while elements smaller than the neutral degree are disconfirmatory degrees.[2] In these terms, the SRP is a constraint on the "level sets" of incremental confirmation—which data sets are equally confirmatory of certain hypotheses.[3]

After introducing some terminology and concepts to make the SRP more precise in section 2, I describe how the SRP is typically seen as a central point of contention between Bayesian and classical schools of statistical inference and evidence. But, I point out, in part following Steel (2003) again, that this actually depends on how *confirmation* is formalized in the two. In section 3 I then reformulate in these confirmation-theoretic terms a version of an important argument against the SRP, sometimes known as "reasoning to a foregone conclusion." This argument proceeds by modus tollens: accepting the SRP entails the evidential irrelevance of an extreme failure of reliability, in the sense of "how well an experiment can distinguish a true hypothesis from among alternatives" (Backe, 1999, p. S355). If one demands either a modicum of reliability for confirmation or that equally confirmatory experiments do not differ too much in their reliability, then one must reject the SRP.

Now, the basic idea of this "foregone conclusion" argument has been much discussed in the literature—see, e.g., Feller (1940); Robbins (1952); Anscombe (1954); Savage (1962); Kerridge (1963); Cornfield (1970); Berger and Wolpert (1988); Mayo (1996); Kadane et al. (1996); Backe (1999); Mayo and Kruse (2001); Steele (2013); Gandenberger (2015). However, essentially all of this discussion situates this argument within the conceptual and evidential framework of classical statistics. Those skeptical of this framework thus have had little motivation to take the argument seriously. My reformulation, by contrast, articulates the argument using concepts of evidence, evidential support, or confirmation that most schools of statistics share. This allows me to synthesize, strengthen, and clarify the foregone conclusion argument so as to bypass objections and attempts to deflate it in sections 4 and 5. My reformulation is therefore *not* primarily intended to convince obstinate adherents of the SRP (although I would welcome such conversions!). Rather, it is to show that their grounds for accepting must lie in biting a bullet: their complete rejection of the minimal evidential relevance of confirmational reliability to confirmation. This reframes the debates on statistical methodology away from the confusing clash of grand statistical frameworks and towards the status of a concept well-defined for all.

Readers knowledgeable of debates between classical and Bayesian statistical schools will recognize that their conflict over the SRP (with the qualification about the formalization of confirmation notwithstanding) is really just a special case of their conflict over a wider (and deeper) principle, the Likelihood Principle (LP), which entails the SRP (Birnbaum, 1962; Berger and Wolpert, 1988). I have chosen to focus on the SRP in this essay because of its concreteness and the existing literature surrounding it, but a rejection of it entails by modus

[2]I leave open the possibility that there are degrees of confirmation incomparable with the netural degree. These might be interpreted as degrees that are partially confirmatary in some resepects and partially disconfirmatory in others, although this interpretation may demand a further structural assumption, e.g., that for any degree, it and the neutral degree have a lower bound and an upper bound.

[3]If one did not wish to identify evidential equivalence with equality of incremental confirmation, one could instead postulate that the former, understood as a separate, primitive concept, entails the latter in all cases in which it is defined. Nothing else substantial in the following would then change.

tollens a rejection of the LP, too. I will return to my arguments' implications for the LP in the concluding section 6.

# 2 The Stopping Rule Principle and Experimental Reliability in Bayesian and Classical Statistics

## 2.1 The Formal Stopping Rule Principle

To state the SRP more precisely and illustrate how it interfaces with typical methods in Bayesian and classical statistics, it will be helpful to define mathematically what it means for a stopping rule to be noninformative, for which I roughly follow Raiffa and Schlaifer (1961, pp. 36–42). Consider an experiment whose data is represented by the sequence of random variables $Z = X_1, \ldots, X_i, \ldots$, where $|Z|$ is the length of $Z$. The stopping rule for the experiment should determine the probability that a certain *number* of these random variables are actually observed. Thus I suppose that the probabilities of the experiment's various outcomes depend in general on two parameters: the hypothesis of interest $\theta \in \Theta$ and the stopping parameter $\phi \in \Phi$. The probability of making at least one observation is $P_{\theta,\phi}(|Z| \geq 1)$, while that of making at least one more observation, given that $k$ observations have been made, is $P_{\theta,\phi}(|Z| \geq k+1 | X_1, \ldots, X_k)$.

A *sequential* experiment is one for which each conditional probability $P_{\theta,\phi}(X_i | X_1, \ldots, X_{i-1}, i \leq |Z|)$ does not vary with $\phi$, i.e., once one conditions on past data and the fact that there will be at least one more datum recorded, the stopping parameter no longer makes a difference to the probability distribution. To emphasize this, I drop $\phi$ from this expression, writing $P_\theta(X_i | X_1, \ldots, X_{i-1}, i \leq |Z|)$. (The relevant probability for the first element in the sequence, $X_1$, is $P_\theta(X_1 | 1 \leq |Z|)$. The probability of observing $Z$ with exactly $|Z| = n$ is then just

$$P_{\theta,\phi}(Z, |Z| = n) = \prod_{i=1}^{n} P_{\theta,\phi}(|Z| \geq i | X_0, \ldots, X_{i-1}) P_\theta(X_i | X_0, \ldots, X_{i-1}, i \leq |Z|) \times (1 - P_{\theta,\phi}(|Z| \geq n+1 | X_1, \ldots, X_n)), \quad (1)$$

where $X_0$ is a constant ("dummy") random variable. This likelihood may be factored into two components, $g_\theta(Z)$ and $s_{\theta,\phi}(n, Z)$, representing the "data-generating process" and the "stopping process," respectively:

$$g_\theta(Z) = \prod_{i=1}^{n} P_\theta(X_i | X_0, \ldots, X_{i-1}, i \leq |Z|) \quad (2)$$

$$s_{\theta,\phi}(n, Z) = \prod_{i=1}^{n} P_{\theta,\phi}(|Z| \geq i | X_0, \ldots, X_{i-1}) \times (1 - P_{\theta,\phi}(|Z| \geq n+1 | X_1, \ldots, X_n)), \quad (3)$$

i.e., $P_{\theta,\phi}(Z, |Z| = n) = g_\theta(Z) s_{\theta,\phi}(n, Z)$. The stopping process (or stopping rule) is said to be *proper* when the experiment will stop almost surely, i.e., $P_{\theta,\phi}(|Z| < \infty) = 1$, and *improper* otherwise.

These preliminaries allow for a precise definition of what it means for a stopping rule to be noninformative.

**Noninformative Stopping Rule:** A sequential experiment with data-generating process $g_\theta(Z)$ has a *noninformative stopping rule for* $\theta$ just when

1. its stopping process $s_{\theta,\phi}(n, Z)$ does not depend on $\theta$,
2. $(\theta, \phi)$ can take on any value in $\Theta \times \Phi$, and
3. if $P_{\theta,\phi}(Z)$ is determined according to $P_{\theta,\phi}(Z) = P(Z|\tilde{\Theta} = \theta, \tilde{\Phi} = \phi)$ by the experiment, where $P$ is a probability measure and $\tilde{\Theta}$ and $\tilde{\Phi}$ are random variables, then $\tilde{\Theta}$ and $\tilde{\Phi}$ are independent.

Informally, noninformative stopping rules do not provide information about the parameter (hypothesis) of interest beyond what the data themselves provide, even indirectly—that is, by providing information about a different parameter that is not independent of the one of interest.[4] This includes probabilistic information, which is available typically (but not exclusively) for Bayesian analyses.

The SRP then provides a constraint on how data $z$ from sequential experiments $Z$ may confirm an hypothesis $\theta$,[5] i.e, on the incremental confirmation measure $c(Z = z, \theta, \theta')$. This measure may be comparative or non-comparative. Comparative measures specify the confirmation of $\theta$ over, or relative to, $\theta'$ by data $z$ from experiment $Z$. Non-comparative measures specify the confirmation of $\theta$ simpliciter; for these, $\theta'$ is a dummy variable (i.e., $c$ is constant with respect to $\theta'$).

**Stopping Rule Principle (Formal):** Consider a confirmation measure $c$ and two sequential experiments, with noninformative stopping rules, for the parameter $\theta \in \Theta$, whose outcomes are described by the respective random variables $Z$ and $Z'$. If for particular outcomes $Z = z$ and $Z' = z'$ their data-generating processes are equal for all $\theta \in \Theta$, i.e., $g_\theta(z) = g'_\theta(z')$, then $c(Z = z, \theta, \theta') = c(Z' = z', \theta, \theta')$ for all $\theta, \theta' \in \Theta$.[6]

In a word, the outcomes of sequential experiments which differ only by a noninformative stopping rule are evidentially equivalent.

## 2.2 Illustration in Bayesian and Classical Statistics

The types of confirmation measures that typical Bayesian statistical methods use implicitly satisfy the SRP, while those for typical classical statistical methods do not. To see this, it will be helpful to consider a simple example of two sequential experiments with binary outcomes—say, sampling fruit flies from a population with either red or white eyes to determine the proportion of white-eyed flies, denoted by $\theta$ (Savage, 1962, pp. 17–18). In both experiments,

[4]See, e.g., Raiffa and Schlaifer (1961, pp. 38–40) for examples of directly and indirectly informative stopping rules.

[5]In this essay, aside from the stopping parameters $\Phi$ as described in the foregoing, I am setting aside the possibility of *nuisance* parameters, parameters whose values are necessary to determine the probability distribution for the data but which are not the object of inference or confirmation. While I believe the essential ideas here extend to such cases, I leave demonstrating that to future work.

[6]Some statements of the SRP require, effectively, that $z = z'$ (e.g., Berger and Wolpert, 1988, p. 76). But this needlessly excludes situations where there is a trivial relabeling of the values of the data, or when the order of the data recorded from the sequential experiment is different.

flies are caught, observed, and released sequentially and fairly, with the number of white-eyed flies reported in the end. 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). If the random variable $X_i$ represents the outcome of the $i$th catch, with the values 1 and 0 representing white and red, respectively, then both experiments have the same data-generating process $g_\theta(Z) = \theta^{\sum_i X_i}(1-\theta)^{\sum_i (1-X_i)}$. However, the two experiments have different stopping rules:

1. Observe $N$ flies. The probability of observing $W_1 = \sum_{i=1}^{N} X_i$ white-eyed flies is then

$$P_\theta(W_1) = \binom{N}{W_1}\theta^{W_1}(1-\theta)^{N-W_1}. \tag{4}$$

   Thus $W_1$ has a binomial distribution with $N$ independent trials and "success" probability $\theta$.

2. Continue observing until $R$ red-eyed flies have been caught. The probability of observing $W_2 = \sum_{i=1}^{|Z|-R} X_i$ white-eyed flies is then

$$P_\theta(W_2) = \binom{W_2 + R - 1}{W_2}\theta^{W_2}(1-\theta)^{R}. \tag{5}$$

   Thus $W_2$ has a *negative* binomial distribution with $R$ "successes" needed for stopping and $1-\theta$ the probability of "success".

In both cases, one can show that the stopping process is simply $s_{\theta,\phi}(n, Z) = \delta_{n,N}$, where $\delta_{n,N}$ is a Kronecker delta.[7] Hence, both stopping rules are uninformative for $\theta$ (Raiffa and Schlaifer, 1961, pp. 38–39). 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 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 $i = 1, 2$. To consider compatibility with the SRP, one must select a Bayesian confirmation measure, a variety of which exist (Huber, nd, §6b). Following Steel (2003, §4), we may note that any Bayesian confirmation measure that depends only on the prior and posterior probabilities of a hypothesis of interest satisfies the SRP in cases where the experiments to which it is applied share the same space of hypotheses.[8] For example, both the log-ratio confirmation measure

$$r(Z = z, \theta, \theta') = \ln\left(\frac{P(\theta|Z=z)}{P(\theta)}\right) \tag{6}$$

and the log-likelihood confirmation measure

$$l(Z = z, \theta, \theta') = \ln\left(\frac{P(Z=z|\theta)}{P(Z=z|\neg\theta)}\right) \tag{7}$$

[7] $\delta_{n,N} = 1$ if $n = N$ and vanishes otherwise.

[8] This argument, which proceeds via the Likelihood Principle introduced in section 6, 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 confirmation measure.

satisfy the SRP for the binomial/negative binomial experiments, since these have a common hypothesis space (the success probability). (Note that these are both non-comparative confirmation measures, so $\theta'$ is a dummy variable for each.) To see this in the latter case, note that by Bayes' theorem,[9]

$$\frac{P(Z=z|\theta)}{P(Z=z|\neg\theta)} = \frac{P(\theta|Z=z)P(\neg\theta)}{P(\neg\theta|Z=z)P(\theta)} = \frac{P(\theta|Z=z)(1-P(\theta))}{(1-P(\theta|Z=z))P(\theta)}.$$

By contrast, classical statistical methods (whether Fisherian or Neyman-Pearsonian) will not, 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 $\theta$ 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 data are evidence against that hypothesis to the extent that this probability is low, i.e., the data actually measured were extreme or unlikely.

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., $w_1 = w_2 = 9$.[10] Then the p-values for the two sequential experiments come out as

$$P_{1/2}(W_1 \geq 9) = \sum_{w_1=9}^{12} \binom{12}{w_1} \left(\frac{1}{2}\right)^{w_1} \left(1-\frac{1}{2}\right)^{12-w_1} \approx 0.07, \tag{8}$$

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

Therefore a test of significance at level $\alpha = 0.05$ of the hypothesis that $\theta = 1/2$ would lead to rejection with the second experiment but not with the first. In terms of non-comparative confirmation, one would have $c(w_1 = 9, \theta = 1/2, \theta') < 0$ yet $c(w_2 = 9, \theta = 1/2, \theta') \geq 0$ (taking zero as the neutral element dividing confirmation from disconfirmation). Even if one were taking the p-value as a measure of disconfirmation (à la Fisher), there would still be a difference between the two.

## 2.3 (Dis)Confirmational (Un)Reliability

When it comes to the satisfaction of the SRP, one way of understanding why there is a difference between, on the one hand, Bayesian analyses using (for example) the log-ratio (equation 6) or log-likelihood (equation 7) confirmation measures, and, on the other, classical analyses using p-values, is that the latter are, but the former are not, sensitive to the *(dis)confirmational (un)reliability* of an experiment. Some experiments, by their own

[9] This calculation shows that one should not be misled into thinking that $l$ depends only on the likelihood $P(Z|\theta)$ and not the prior $P(\theta)$, despite its name.

[10] The example is an amalgam of those by Savage (1962, pp. 17–18) and Mayo and Kruse (2001, pp. 387–388).

lights, have a high probability of (dis)confirming a certain hypothesis, at least to some degree, when that hypothesis is false (resp. true) and some alternative hypothesis is actually true (resp. false). Such an experiment may be said to be (dis)confirmationally unreliable for that hypothesis against that alternative to the degree that this probability is high, and (dis)confirmationally reliable to the degree that it is low. (When referring to all of these notions at once hereafter, I will often refer to them simply as "reliability.")

Formally, let $Z$ again be a random variable representing the outcome of an experiment, $\theta, \theta' \in \Theta$ be parameters (hypothesis) determining the probabilities for its potential outcomes, and $c(Z, \theta, \theta')$ be an incremental confirmation measure whose values (or "degrees" of confirmation) are partially ordered and include a unique designated "neutral" value $q_0$. The ordering represents increasing confirmation and $q_0$ denotes the degree that is neither confirmation nor disconfirmation. Note that this is compatible with both quantitative and qualitative confirmation measures; $c(Z, \theta, \theta')$ could, for instance, take on numerical values or simply three basic ordered qualitative values (e.g., "confirm," "neither confirm nor disconfirm," and "disconfirm"). Further, assume hereafter that, if the set of stopping parameters $\Phi$ is nonempty, then the stopping parameter $\phi$ is known.[11] Hence, I will drop reference to $\phi$ from the notation to reduce clutter. Then, according to $c$, the $q$-confirmational unreliability of $Z$ for $\theta$ over $\theta'$ against $\tau \neq \theta$ is[12]

$$CU_c(Z, \theta, \theta', q, \tau) = P_\tau(c(Z, \theta, \theta') > q). \tag{10}$$

This is just the probability of experiment $Z$ confirming $\theta$ over $\theta'$ more than degree $q$ with confirmation measure $c$ when in fact $\tau \neq \theta$ is the case.[13] Similarly, according to $c$, the $q$-disconfirmational unreliability of $Z$ for $\theta$ is

$$DU_c(Z, \theta, \theta', q) = P_\theta(c(Z, \theta, \theta') \leq q). \tag{11}$$

This is just the probability of experiment $Z$ confirming $\theta$ over $\theta'$ no more than degree $q$ with confirmation measure $c$ when in fact $\theta$ is the case. Finally, one may define confirmational and disconfirmational reliability, respectively, as

$$CR_c(Z, \theta, \theta', q, \tau) = 1 - CU_c(Z, \theta, \theta', q, \tau), \tag{12}$$

$$DR_c(Z, \theta, \theta', q) = 1 - DU_c(Z, \theta, \theta', q). \tag{13}$$

So far, I have assumed that the confirmation of hypotheses applies to so-called "point hypotheses," single parameter values $\theta \in \Theta$, each of which determines completely a probability distribution for the experimental outcomes of experiments for them. But one can generalize the definition of (dis)confirmational (un)reliability, equations 10–13, to apply also to so-called "disjunctive" or "composite hypotheses" $H \subseteq \Theta$, which in this context are (non-singleton)

[11] It is typical in an experimental design to specify the stopping parameter completely, or else control it enough that the likelihood function for the experiment is well approximated by one so specified.

[12] If one uses the likelihood ratio as a comparative confirmation measure, then confirmational unreliability (equation 10) is essentially the same as what Royall (2000) call the "probability of misleading evidence." (The only difference is that Royall uses a non-strict inequality.)

[13] One could of course extend the definition of $CU_c$ to the cases in which $\tau = \theta$, but in this case the confirmational unreliability is just equal to the disconfirmational reliability (equation 13), i.e., $CU_c(Z, q, \theta, \theta', \theta) = DR_c(Z, q, \theta, \theta')$.

*sets* of parameter values. Disjunctive hypotheses do not in general uniquely determine a probability distribution for experimental outcomes; rather, each of their elements, which *are* point hypothesis, determines one such distribution. Analogously, a confirmation measure will in general assign a (non-singleton) *set* of confirmation values to a disjunctive hypothesis, given an experimental outcome for them.

If one assumes that the partial order of confirmation values has the greatest-lower-bound property, i.e., that the infimum of a set of values always exists, then one can assign to a disjunctive hypothesis the infimum of the confirmation values:

$$c(Z, H, H') = \inf_{(\theta,\theta')\in H\times H'} c(Z, \theta, \theta') \tag{14}$$

for disjoint $H, H' \subseteq \Theta$. Since one can interpret a disjunctive hypothesis, naturally, as the disjunction of its constituent point hypotheses, it is natural to suppose that it is confirmed no more than its least confirmed constituent. That is what equation 14 formalizes. Then the *maximal* $q$-(dis)confirmational unreliability is just

$$mCU_c(Z, H, H', q, T) = \sup_{\tau\in T} P_\tau(c(Z, H, H') > q), \tag{15}$$

$$mDU_c(Z, H, H', q) = \sup_{\theta\in H} P_\theta(c(Z, H, H') \leq q), \tag{16}$$

and the *minimal* $q$-(dis)confirmational reliability is defined just as with equations 12–13:

$$mCR_c(Z, H, H', q, T) = 1 - mCU_c(Z, H, H', q, T), \tag{17}$$

$$mDR_c(Z, H, H', q) = 1 - mDU_c(Z, H, H', q). \tag{18}$$

The sense in which these are "maximal/minimal" is that they represent the worst cases with respect to the different hypotheses under consideration. If one also has a probability measure on the point hypotheses and the disjunctive hypotheses are measurable with respect to it, then one also define the *expected* $q$-(dis)confirmational (un)reliability by replacing the suprema in equations 15–16 with expectations with respect to this measure. Regardless of which version one chooses, if the disjunctive hypotheses considered are just singletons, then the generalized version $q$-(dis)confirmational (un)reliability becomes effectively equivalent to the versions applicable just to point hypotheses, equations 10–13. These generalizations and the additional properties required of the partial order are not required for the sequel, so I will retain reference only to the point-hypothesis versions of (dis)confirmational (un)reliability there. But one could adopt these generalizations for the sequel at the cost of their concomitant additional assumptions.

Whether one considers disjunctive and point hypotheses or only point hypotheses, (dis)confirmational (un)reliability is well-defined for Bayesian and classical confirmation measures,[14] which are nevertheless differently sensitive to them. If a Bayesian confirmation measure for an hypothesis $\theta$ (perhaps over another, $\theta'$,) depends only on the prior and posterior probabilities for $\theta$ (and perhaps $\theta'$), then its confirmational reliability certainly depends on more in general—note the dependence on $\tau$ in equation 12. One might describe this dependence as being on the "modal" or alethic possibilities that form part of the backdrop to an experiment. In any

[14]For the former, one sets, for any random variable $A$, $P_\theta(A) = P(A|\theta)$.

case, even insofar as a Bayesian confirmation measure does depend on this modal structure, it is not at all clear that, all else being equal, differences in the reliability of two experimental outcomes would necessarily make a difference to the Bayesian confirmation of an hypothesis under test.[15]

In classical statistics, by contrast, the confirmation measure is often a function of the p-value of a statistic of the experimental outcomes under some hypothesis. In Fisherian testing, for example, the extent to which a p-value approaches zero is the extent to which it provides disconfirmation of the tested hypothesis. In Neyman-Pearson testing of an hypothesis, this p-value is mapped to two confirmation values, "accept" and "reject," depending on whether the p-value rises above or below a fixed value, called the size, significance level, or type I error probability of the test, which is also the experiment's disconfirmational unreliability for that hypothesis (taking $q$ to be "reject"). The power of the test for the hypothesis against a certain alternative is then the experiment's confirmational reliability against that alternative, and the type II error probability is its confirmational unreliability. In Mayo's "severe testing" account of classical statistics, the severity of a test is an additional criterion for an experiment to confirm an hypothesis (Mayo, 1996, pp. 179–181).

In the next section, I reconstruct and make precise arguments against the SRP based on the conviction that reliability should somehow matter for confirmation. Before continuing to those arguments, I first must clarify the scope of the disagreement about reliability. *All* parties to the debate about the SRP seem to agree that confirmational reliability is important for and a relevant factor in the *design* of experiments.[16] Regardless of how experimentalists intend to interpret their results, they try to minimize their sequential experiments' costs—whether pecuniary, temporal, material, ethical, or otherwise—which depend in general on the stopping rule. Rather, the debate concerns whether stopping rules have any *epistemic* import, i.e., some bearing on how the results of experiments with them provide evidence that guides us toward truth. Those who accept the SRP would answer negatively, and inversely for those who reject it.

# 3 The Stopping Rule Principle Versus Reliability

## 3.1 Foregone Conclusions

Confirmation measures that satisfy the SRP in general can equate the confirmations for an hypothesis (perhaps over another) provided by two experiments with vastly different reliabilities.[17] This has been observed in the context of the debate around the SRP at least as early as 1959, when Peter Armitage remarked that, under the right circumstances, a researcher adhering to the SRP could justifiably continue a sequential experiment until reaching a *fore-*

[15]It would be interesting and worthwhile to determine more precisely how reliability and any particular Bayesian confirmation measure are interdependent, but that question is beyond the scope of the present work.

[16]See, for example, the brief statements by Edwards et al. (1963, p. 239) and Berger and Wolpert (1988, p. 78), as well as a fuller statement by Backe (1999, p. S358) and decision-theoretic justifications by Sprenger (2009, pp. 644, 648) and Steele (2013, §3).

[17]To avoid needless repetition, I will henceforth leave tacit parentheticals re-affirming that confirmation can be comparative or non-comparative, unless confusion might arise.

*gone conclusion*, an inevitable confirmation to degree $q$ of an hypothesis essentially regardless of its truth (Savage, 1962, p. 72).[18] This is clearly a case of maximal $q$-confirmational unreliability. The basic idea is that the researcher adopts the following stopping rule: continue taking new data until the resultant total would confirm a predetermined hypothesis more than degree $q$. Is such a stopping rule proper? Although the details are subtle, this is indeed possible in the framework of Bayesian statistics using the log-likelihood confirmation measure (equation 7) only if the Bayesian agent is allowed to adopt certain merely finitely additive (i.e., not also countably additive) probability distributions as priors for parameter values. But even without such special priors, much the same conclusions can arise—cf. the discussion of section 3.2. As Mayo (1996, p. 356) observes, advocates of the SRP in print such as Savage (1962) and Berger and Wolpert (1988, p. 83) are "plainly uncomfortable" with this conclusion; they suggest to readers to trust their intuitions in simpler cases in which the SRP (and the LP) should clearly hold. I agree with Mayo and Mayo and Kruse (2001, p. 400) that it is dubious to hold such examples as exotic (especially, again, in light of the discussion of section 3.2).

One way of drawing out the conflict more precisely is to observe that cases such as this show that the SRP can be incompatible with the following property of a confirmation measure $c$, given contextually chosen positive and negative confirmation levels $q_+, q_-$:

**No Foregone Conclusions (NFC):** There is no experiment with identifiable outcomes $Z$, parameters $\Theta$, parameter values $\theta, \theta' \in \Theta$, and confirmation values $q_+ \geq q_0$ and $q_- \leq q_0$ such that either $CU_c(Z, \theta, \theta', q_+, \tau) = 1$ or $DU_c(Z, \theta, \theta', q_-) = 1$.

An experiment with identifiable outcomes $Z$ and parameters $\theta \in \Theta$ is one for which the map $\theta \mapsto P_\theta(Z)$ is injective. In other words, NFC excludes from consideration experiments for which the parameter value and the outcome probability distribution make no difference to each other. With these excluded, a confirmation measure satisfying NFC never allows one to draw foregone conclusions—almost surely misleading evidence for or against an hypothesis.

Backe (1999) suggests in response that Bayesians avoid using such stopping rules before committing to run an experiment, but this prescriptive rule does not help when one is faced with how to analyze an experiment already completed that may have violated the rule—certainly assuming the experimenter was entirely rational in this way would be naive. More productively, however, Kadane et al. (1996), following Savage (1962, pp. 72–3) himself, Kerridge (1963), and Cornfield (1970) show that foregone conclusions are avoided if one uses *countably additive* priors. One can adapt their conclusions to the present context as follows: if such a Bayesian adopts the countably additive prior $P(\theta)$, the non-comparative log-ratio confirmation measure $r$ (equation 6), and attempts to continue observing new data until $r(Z, \theta_0, \theta') > q \geq 1$, then

$$P(|Z| > \infty) < e^{-q},$$
$$P(|Z| > \infty | \neg\theta_0) < (e^{-q} - P(\theta_0))/(1 - P(\theta_0)) < e^{-q}.$$

[18] The example had already been known in the context of debates around the proper analysis of data from experiments with "optional" (i.e., probabilistic) stopping rules in classical statistics (Savage, 1962, p. 18). (In retrospect, the division between deterministic and non-deterministic stopping rules is not in general invariant with respect to a redescription of the experiment, but this won't matter for present purposes.)

In a word, the stopping rule is not proper, so the probability of stopping with misleading confirmation is bounded above, decreasing exponentially in the degree of confirmation. An experiment with such a stopping rule thus has, for increasing degrees of foregone confirmation, significantly decreasing chances of ever reaching completion.

## 3.2 Unreliability

As important and interesting as the results reviewed at the end of section 3.1 are, I agree with Mayo and Kruse (2001, p. 393) that

> the most important consequence of the Armitage example is not so much the extreme cases ... but rather the fact that ignoring stopping rules can lead to a high probability of error, and that this high error probability is not reflected in the interpretation of data

when that interpretation—how data confirm hypotheses—is bound by the SRP. In a word, it is not *maximal* unreliability which is the issue per se, but sufficiently *high* unreliability—or equivalently, sufficiently *low* reliability. This suggests instead to demand of a putative confirmation measure $c(Z, \theta, \theta')$ some property that instantiates the following schema:

**$\varepsilon$-Adequate $(q_+, q_-)$-Reliability for $\mathcal{E}$ ($\varepsilon \mathrm{A} q_\pm \mathrm{R}\mathcal{E}$):** For any experiment in $\mathcal{E}$ with outcomes $Z$ potentially informative for parameters $\theta, \theta', \tau \in \Theta$, $CR_c(Z, \theta, \theta', q_+, \tau) \geq \varepsilon(\theta, \theta')$ and $DR_c(Z, \theta, \theta', q_-) \geq \varepsilon(\theta, \theta')$.

(Suggested pronunciation: "epsilon-acre.") The sense in which $\varepsilon \mathrm{A} q_\pm \mathrm{R}\mathcal{E}$ is a schema, unlike NFC, is that, syntactically, it is an open sentence with variables $\varepsilon$, $q_+$, $q_-$, and $\mathcal{E}$. All else being equal for a given $\mathcal{E}$, choices of $\epsilon : \Theta \times \Theta \to [0, 1]$ that are larger over its arguments, higher values of $q_-$, and lower values of $q_+$ correspond to stronger strictures of reliability on a confirmation measure.

Not all instances of $\varepsilon \mathrm{A} q_\pm \mathrm{R}\mathcal{E}$ plausibly hold. Letting $\mathcal{E}$ be the class of all experiments, $\varepsilon(\theta, \theta') = 1$, and $q_+$ and $q_-$ be arbitrarily high and low confirmation values, respectively, would yield an implausible instance of $\varepsilon \mathrm{A} q_\pm \mathrm{R}\mathcal{E}$ that requires all experiments, almost surely, to maximally confirm an hypothesis if it were true, and maximally disconfirm an hypothesis if it were false. But I do claim that when $\mathcal{E}$ is finite, there will be everywhere positive $\varepsilon(\theta, \theta')$, $q_+ \geq q_0$, and $q_- \leq q_0$ for which $\varepsilon \mathrm{A} q_\pm \mathrm{R}\mathcal{E}$ should hold.[19] Which values these will plausibly be will depend on the confirmation measure under consideration—not all measures take on values in the same partially ordered space—and the experiments $\mathcal{E}$. Although I do not have a recipe to provide for this, I can indicate a few difference-makers. Experiments with more potential data might admit of stronger strictures on an instance of $\varepsilon \mathrm{A} q_\pm \mathrm{R}\mathcal{E}$ that they should satisfy. For confirmation measures related to properties of classical tests, such as size and power, these properties may correlate with stronger strictures similarly. As the power of test varies with $\theta$, relative to a given $\theta'$ (say, taken as the "null" hypothesis), so might $\varepsilon(\theta, \theta')$ vary. The values of $\varepsilon(\theta, \theta')$, $q_+$, and $q_-$ for a given $\mathcal{E}$ may also depend on any probability

[19] I restrict my claim to finite $\mathcal{E}$, since it seems possible to have an infinite collection of experiments, each element $\alpha$ for which a confirmation measure satisfies $\varepsilon \mathrm{A} q_\pm \mathrm{R}\mathcal{E}$ with $\varepsilon_\alpha(\theta, \theta') > 0$, but such that for some $\theta, \theta'$, $\inf_\alpha \epsilon_\alpha(\theta, \theta') = 0$.

assignments to the parameters investigated, as is often the case for Bayesian confirmation measures. One might allow for $\varepsilon(\theta, \theta')$ to vary, for a given $\theta'$, as the prior for $\theta$ varies. Since more might be at stake in some experiments than others, how the confirmation of hypotheses might influence action or policy, broadly interpreted, can also play a role. Stronger versions of $\varepsilon \mathrm{A} q_{\pm} \mathrm{R} \mathcal{E}$ make precise different ways in which this decision-making procedure should be "open-minded" to how the evidence may indicate hypotheses of interest.

$\varepsilon \mathrm{A} q_{\pm} \mathrm{R} \mathcal{E}$ constrains the "absolute" reliability of confirmation measures, but one might consider instead *relative* reliability. As Mayo (2004, p. 105) has suggested, one can instead identify the problem with the SRP as the fact that it "entails the irrelevance of the procedures generating [the data] that do not alter likelihoods even though they can dramatically alter error probabilities." Given an experimental outcome $Z = z$, its likelihood is simply $\mathcal{L}_{Z=z}(\theta) = P_\theta(Z = z)$, a function of the parameter (hypothesis) of interest with the data considered fixed. Likelihoods are said to be *equivalent* when they are proportional—i.e., $\mathcal{L}_{Z=z}(\theta)$ and $\mathcal{L}_{Z'=z'}(\theta)$ are equivalent when there exists some $k > 0$ such that $\mathcal{L}_{Z=z}(\theta) = k\mathcal{L}_{Z'=z'}(\theta)$ for all $\theta$. If the likelihoods from two different experiments are equivalent, they generally arise from observed data that are highly relevantly similar, if not identical. Finally, "error probabilities" in Mayo's sense are (dis)confirmational unreliabilities, with functions of the p-value of a certain statistic taken as the confirmation measure.[20] So, Mayo's alternative complaint about the SRP, more precisely put, is that it doesn't allow for equivalent likelihoods from different experiments with nevertheless *different* reliabilities to engender *different* degrees of confirmation for a given hypothesis.

In the present context, however, experiments with different stopping processes $s_{\theta,\phi}$ and $s_{\theta,\phi'}$ may have different parameters for those processes—$\phi \in \Phi$ but $\phi' \in \Phi'$ with $\Phi \neq \Phi'$—so that the experiments' associated likelihoods do not have the same domain. Consequently, Mayo's alternative complaint must be modified to apply to these cases. Her reference to the "procedures generating the data" gives a hint. Given an experimental outcome $Z = z$, define a *data-generating partial likelihood* as $\mathcal{L}^g_{Z=z}(\theta) = g_\theta(Z = z)$, where $g_\theta$ is a data-generating mechanism for the experiment. Say that two data-generating partial likelihoods $\mathcal{L}^g_{Z=z}(\theta)$ and $\mathcal{L}^g_{Z'=z'}(\theta)$ are equivalent when there exists some $k > 0$ such that $\mathcal{L}^g_{Z=z}(\theta) = k\mathcal{L}^g_{Z'=z'}(\theta)$ for all $\theta$. Then Mayo's revised alternative complain about the SRP is that it doesn't allow for equivalent data-generating partial likelihoods from different experiments with nevertheless different reliabilities to engender different degrees of confirmation for a given hypothesis.

In the interest of arguments against the SRP that do not appeal directly to confirmation thresholds, one can abstract away from them to arrive at the following property schema, instances of which one might demand of a putative confirmation measure $c(Z, \theta, \theta')$.

**$\varepsilon$-Similar Uniform Reliability for Likelihood-Equivalence of $\mathcal{E}$ ($\varepsilon$SURLE$\mathcal{E}$):** For any pair of experiments in $\mathcal{E}$ with potential outcomes $Z$ and $Z'$ for the same parameter (hypothesis) space $\Theta$ and any $\theta, \theta' \in \Theta$, if the outcomes $Z = z$ and $Z' = z'$ are data-generating partial likelihood equivalent and $c(Z = z, \theta, \theta') = c(Z' = z', \theta, \theta') = q$, then $|DR_c(Z, \theta, \theta', q) - DR_c(Z', \theta, \theta', q)| \leq \varepsilon(\theta, \theta')$ and $|CR_c(Z, \theta, \theta', q, \tau) - CR_c(Z', \theta, \theta', q, \tau)| \leq \varepsilon(\theta, \theta')$ for all $\tau \in \Theta$.

[20] Here the confirmation should be taken to be qualitative, rather than quantitative (Mayo, 1996, p. 179n3): confirmation is "passing" a highly severe test.

(Suggested pronunciation: "epsilon-surly.") $\varepsilon$SURLE$\mathcal{E}$ states that at least for experiments whose outcomes are relevantly highly similar, those experiments confirm a hypothesis equally only if they are not too differently (dis)confirmationally reliable. In other words, at least for data-generating partial likelihood equivalent experiments, sufficiently large differences in reliability make a difference to confirmation. All else being equal for a given set of experiments $\mathcal{E}$, *smaller* values of $\varepsilon : \Theta \times \Theta \to [0, 1]$ on any of its arguments correspond to stronger strictures of comparative reliability on a confirmation measure. Unlike $\varepsilon \mathrm{A} q_{\pm} \mathrm{R} \mathcal{E}$, there is no dependence on confirmation values $q_+, q_-$ for $\varepsilon$SURLE$\mathcal{E}$, since it assumes the "difference-making" power of reliability on confirmation doesn't depend on the degree or quality of confirmation.[21] This is the sense in which it concerns "uniform" reliability. But like with $\varepsilon \mathrm{A} q_{\pm} \mathrm{R} \mathcal{E}$, I claim that when $\mathcal{E}$ is finite, there will be some $\varepsilon(\theta, \theta')$ everywhere less than 1 for which $\varepsilon$SURLE$\mathcal{E}$ is a plausible property to demand of a confirmation measure. Also like with $\varepsilon \mathrm{A} q_{\pm} \mathrm{R} \mathcal{E}$, the same contextual considerations apply in deciding what values of $\varepsilon$ would be plausible to demand.

Neither $\varepsilon \mathrm{A} q_{\pm} \mathrm{R} \mathcal{E}$ nor $\varepsilon$SURLE$\mathcal{E}$, as schema, conflict *logically* with SRP—mere schema (open sentences) are not the sorts of things that can so conflict. Moreover, it is not the case that *every* instance of these schema conflicts with the SRP. (Let $\varepsilon = 0$ for $\varepsilon \mathrm{A} q_{\pm} \mathrm{R} \mathcal{E}$ and $\varepsilon = 1$ for $\varepsilon$SURLE$\mathcal{E}$.) The compatibility of particular instances of the schema depends on the class $\mathcal{E}$ of sequential experiments and the values of $\varepsilon$ and (for $\varepsilon \mathrm{A} q_{\pm} \mathrm{R} \mathcal{E}$) $q_+$ and $q_-$.

But conflict is easy to come by for particular choices. Consider again the case of the experiments about the proportion of eye-colors in a fruit fly population in section 2.2, and for simplicity suppose that the only two proportions of white-eyed flies under consideration are $\theta_1 = 0.5$ and $\theta_2 = 0.6$. Furthermore, suppose that $\mathcal{E}$ consists only of an experiment $Z$ which stops either when $l(Z = z, \theta_2, \theta') > 0$—i.e., $\theta_2$ is confirmed according to the log-likelihood confirmation measure (equation 7)—or seven flies have been sampled. One can then calculate that $CR_l(Z, \theta_2, \theta', 0, \theta_1) \approx 0.273$ and $DR_l(Z, \theta_2, \theta', 0, \theta_1) \approx 0.855$.[22] So, if one sets $q_+ = 0$ and $\varepsilon(\theta, \theta') = 1/2$, say, then the log-likelihood confirmation measure $l$ will not satisfy $\varepsilon \mathrm{A} q_{\pm} \mathrm{R} \mathcal{E}$ because its confirmational reliability is too low. In other words, the chances of confirming $\theta_2$ when it is false (and instead $\theta_1$ is true) are much better than even odds.

To see the conflict with $\varepsilon$SURLE$\mathcal{E}$, let $\mathcal{E}$ include as well another experiment $Z'$ observing flies from the population, but which stops when just one fly has been caught. It follows immediately that $CR_l(Z', \theta_2, \theta', 0, \theta_1) = 0.5$ and $DR_l(Z', \theta_2, \theta', 0, \theta_1) = 0.6$. So, supposing now that in the first experiment, $Z$, only one fly was actually caught, one can calculate that $|CR_l(Z, \theta_2, \theta', 0, \theta_1) - CR_l(Z', \theta_2, \theta', 0, \theta_1)| \approx 0.227$ and $|DR_l(Z, \theta_2, \theta', 0, \theta_1) - DR_l(Z', \theta_2, \theta', 0, \theta_1)| \approx 0.255$. Since the antecedent conditions of $\varepsilon$SURLE$\mathcal{E}$ are satisfied for the two experiments—again, $l$ satisfies the SRP—$l$ does not satisfy $\varepsilon$SURLE$\mathcal{E}$ even for $\varepsilon$ with appreciable values, such as $\varepsilon(\theta, \theta') = 1/4$. So, in this example, experiments which are supposed to be confirmationally equivalent according to the SRP can differ in reliability by over 0.25.

The conflict needn't always be the same for both $\varepsilon \mathrm{A} q_{\pm} \mathrm{R} \mathcal{E}$ and $\varepsilon$SURLE$\mathcal{E}$. Consider a third experiment $Z''$ that is exactly like $Z'$ but which stops only when seven flies have been

[21] Here one might make an exception for $q_0$, for that neutral case might intuitively accommodate a variety of experiments with various reliabilities. I do not see how anything that follows in this essay depends on taking a stand on this issue, however.

[22] Here I have adapted the calculations of Steele (2013, p. 957) to the present terminology.

caught. Then one can calculate that $CR_l(Z'', \theta_2, \theta', 0, \theta_1) = 0.5$ and $DR_l(Z'', \theta_2, \theta', 0, \theta_1) \approx 0.710$. So, supposing now that in the first experiment, $Z$, seven flies were also actually caught, and that the number of red- and white-eyed flies in the two experiments was the same, one can calculate that $|CR_l(Z, \theta_2, \theta', 0, \theta_1) - CR_l(Z'', \theta_2, \theta', 0, \theta_1)| \approx 0.227$, which is the same as when comparing $Z$ with $Z'$, but $|DR_l(Z, \theta_2, \theta', 0, \theta_1) - DR_l(Z'', \theta_2, \theta', 0, \theta_1)| \approx 0.145$. Keeping the same $\epsilon = 1/2$ and $q_+ = q_- = 0$ for $\varepsilon\mathrm{A}q_\pm\mathrm{R}\mathcal{E}$ as in the previous comparison would entail that if $\mathcal{E} = \{Z', Z''\}$, $\varepsilon\mathrm{A}q_\pm\mathrm{R}\mathcal{E}$ does not hold of $l$. But doing the same with $\epsilon = 1/4$ for $\varepsilon\mathrm{SURLE}\mathcal{E}$ does allow for it to hold (once we have verified that the analogous calculations hold with $\theta_1$ and $\theta_2$ permuted).

Similar conclusions for both $\varepsilon\mathrm{A}q_\pm\mathrm{R}\mathcal{E}$ and $\varepsilon\mathrm{SURLE}\mathcal{E}$ follow from using the log-ratio confirmation measure $r$ (equation 6). In a word, both $\varepsilon\mathrm{A}q_\pm\mathrm{R}\mathcal{E}$ and $\varepsilon\mathrm{SURLE}\mathcal{E}$ seem to have some satisfiable instances for confirmation measures that satisfy the SRP, but not all plausible instances are satisfiable.

What about confirmation measures that do not satisfy the SRP? If one takes one's confirmation measure $c$ to be the p-value $p_s$ of a statistic $s$ measuring the discrepancy of the data from what is expected under $\theta$, then:

- $DR_{p_s}(Z, \theta, \theta, q_-) = 1 - q_-$, where $q_-$ would be the size of a test of $\theta$ based on $s$—the probability of the test rejecting $\theta$ when it is true.
- $CR_{p_s}(Z, \theta, \theta', q_+, \tau) = POW_s(\theta, \tau, q_+)$, the power of a test of significance, based on $s$, of $\theta$ at $\tau$ with size $q_+$—the probability of the test rejecting $\theta$ when in fact $\tau$ is true.

Good practice for such tests often suggests a size of 0.05 for a test of $\theta$ and a power of 0.80 against the most relevant $\tau \neq \theta$. In these cases, high values of confirmational reliability for $\theta$ and of disconfirmational reliability for $\theta$ against $\tau$ follow. Thus for such $\theta$ and $\tau$, there will be plausible instances of $\varepsilon\mathrm{A}q_\pm\mathrm{R}\mathcal{E}$ that will be satisfied, such as in the above case of testing for the possibility of two proportions of eye-colors in the fruit flies. As for $\varepsilon\mathrm{SURLE}\mathcal{E}$, if its antecedent is satisfied for two experiments using $p_s$ as their confirmation measure, then the two experiments must have identical (dis)confirmational reliabilities. Thus $\varepsilon\mathrm{SURLE}\mathcal{E}$ will be satisfied.

## 3.3 The Reason for the Conflict

Although I have not proved that a general conflict between the SRP, on the one hand, and *any* instantiation of $\varepsilon\mathrm{A}q_\pm\mathrm{R}\mathcal{E}$ or $\varepsilon\mathrm{SURLE}\mathcal{E}$, on the other, is inevitable for an arbitrary confirmation measure, the examples from section 3.2 show that for some seemingly reasonable $\varepsilon(\theta, \theta')$, $q_+$, and $q_-$ there will indeed be conflict for some commonly used Bayesian confirmation measures.[23] (Since for some $\varepsilon(\theta, \theta')$, $q_+$, and $q_-$ these are not reasonable properties to demand of a confirmation measure, the absence of a general conflict for all values is no deficiency.) The reason for this is similar in both cases, although the differences are also worth highlighting. In all cases, one advantage of using $\varepsilon\mathrm{A}q_\pm\mathrm{R}\mathcal{E}$ or $\varepsilon\mathrm{SURLE}\mathcal{E}$ is that these

[23] See also examples 20 and 21 in Berger and Wolpert (1988, pp. 75–76, 80–81) for further instances in which similar conclusions apply.

properties are formulated using concepts that appear to be common to all accounts of confirmation of probabilistic theories; they do not presuppose concepts particular to classical statistics.

What is therefore the framework-neutral source of the conflict between the SRP and $\varepsilon\mathrm{A}q_{\pm}\mathrm{R}\mathcal{E}$? Any confirmation measure which satisfies the latter must have, for any experiment, sufficiently low probabilities of confirming a hypothesis (beyond some specified degree) when it is false and disconfirming it (at least to some specified degree) when it is true. On the one hand, an arbitrarily selected confirmation measure, even if it does not satisfy the SRP, may not have this property. For example, a confirmation measure based on the p-value of a statistic will not in general satisfy the SRP, but it may also not be confirmationally reliable, either: both advocates (e.g., Mayo (1996, Ch. 11)) and critics (e.g., Howson and Urbach (2006, Ch. 5)) of classical testing acknowledge that achieving a high p-value on a test of an hypothesis is not generally good grounds for confirmation (much less acceptance) of that hypothesis. On the other hand, it is not clear whether an arbitrarily selected confirmation measure that *does* satisfy the SRP will *not* satisfy $\varepsilon\mathrm{A}q_{\pm}\mathrm{R}\mathcal{E}$ for reasonable $\varepsilon(\theta, \theta')$, $q_+$, and $q_-$. But two experiments that differ only by a stopping rule will in general differ in their reliabilities ($CR_c$ and $DR_c$) for reasonable choices of confirmation measure $c$. Because a confirmation measure satisfying the SRP will entail that two such experiments with otherwise identical outcomes will be evidentially equivalent, it leaves open the possibility that one experiment falls below the reliability threshold. Consequently, while satisfying the SRP makes it more difficult for a confirmation measure to satisfy $\varepsilon\mathrm{A}q_{\pm}\mathrm{R}\mathcal{E}$ for reasonable values of $\varepsilon$, $q_+$, and $q_-$, the source of the conflict for the Bayesian confirmation measures $l$ and $r$ comes more directly from the fact that the reliability of an experiment does not seem to make a difference (or enough of one) for that experiment's ability to confirm or disconfirm.

When it comes to the conflict between the SRP and $\varepsilon\mathrm{SURLE}\mathcal{E}$, the source is more direct. As I just mentioned, two experiments that differ only by a stopping rule will in general differ in their reliabilities ($CR_c$ and $DR_c$) for reasonable choices of confirmation measure $c$. As long one can find some such pair of experiments whose reliabilities differ to a sufficient degree for some data-generating partial likelihood equivalent outcomes, $\varepsilon\mathrm{SURLE}\mathcal{E}$ will be violated. Such outcomes are data that are very relevantly similar to one another, so in such cases $\varepsilon\mathrm{SURLE}\mathcal{E}$ demands that sufficiently large differences in reliability make some difference to confirmation. The SRP, on the other hand, denies this can ever be the case.

## 3.4 The Justification for Requiring Reliability Criteria for Confirmation Measures

There is a significant literature on axiomatic constraints on confirmation measures and Bayesian versions especially (Crupi, 2016). Yet while reliability as a notion pertaining to individual measuring instruments has been applied in Bayesian epistemology and confirmation theory (Bovens and Hartmann, 2002, 2003), (dis)confirmational reliability has not, or at least not in these terms. (I discuss the connection with the literature on truth-convergence briefly below.) There is nevertheless a simple intuition behind requiring a criterion of this sort: confirmation of a hypothesis $H$ requires "a test that is highly capable of probing the ways in which $H$ can err" (Mayo, 1996, p. 9). This includes the ability to disconfirm $H$ when

it is false, but also confirm $H$ when it is true.

Now, in the context of the above quotation, Mayo is concerned with distinguishing her own "severe testing" account of scientific learning with what she calls the "evidential-relation" view, according to which the relevant properties to be examined are between the data and hypotheses. For severe testing, by contrast, these accrue (at least in part) to the *methods* used (Mayo, 1996, p. 72). But the constraint of reliability can be applied without complete rejection of the form of an evidential relationship.[24] Indeed, very general considerations from reliabilist epistemology buttress the motivation that the best evidence for or against a hypothesis—that is, incremental confirmation or disconfirmation of it—ought to come from a reliable source or method (Goldman and Beddor, 2016).

How this motivates $\varepsilon \mathrm{A} q_{\pm} \mathrm{R} \mathcal{E}$ or $\varepsilon \mathrm{SURLE} \mathcal{E}$ depends on how one makes precise these broad, informal insights. One way is to make sufficient reliability a precondition for a certain degree of confirmation. That is what $\varepsilon \mathrm{A} q_{\pm} \mathrm{R} \mathcal{E}$ implements. On this strong condition, no confirmation can be had without that confirmation also being reliable: the chances of it leading one astray (either for a false hypothesis or against a true one) must be sufficiently low. Another way is merely to require that reliability must *make some difference* to confirmation. At least for data sets that are data-generating partial likelihood equivalent—about as similar as they can get, according to the models of the experiments from which they were produced—if the experiments that produced them have sufficiently different reliabilities, then that fact should make some difference to the evidential bearing those data sets have on the hypotheses for which the experiments were conducted.

Often reliability in epistemology is understood as reliability in the long-run: a method is reliable if it eventually leads one to the truth. In this context, various convergence theorems for Bayesian confirmation can be invoked: e.g., that if hypotheses can be identified (i.e., distinguished) eventually from the data, then eventually data confirm all true hypotheses and disconfirm all false ones almost surely (Huber, nd, §7). These theorems come with the inevitable interpretive objection that they show very little about real data and evidence, for in the long run we're all dead.[25] How are we supposed to understand the reliability of methods when data, and our ability to collect it, are finite? Should not reliability, at least as it applies to the short-term, come in degrees? This is precisely where confirmational reliability criteria can play a role. If the evidence that a finite body of data provides for an hypothesis depends both on the data collected and, to some degree or other, the confirmational reliability of the experiment to produce such data, one has a surer justification for using that data to support hypotheses.

Finally, a further cost to rejecting confirmational reliability criteria is that many scientists have endorsed some form of it, especially in the context of replication. I shall for concreteness focus on the evidence for this in psychology, and because replication has been of greatest concern in that field as of late. Simmons et al. (2011, p. 1359) report various degrees of "flexibility" that researchers might employ—sometimes known as "questionable research

[24] Mayo (1996, Ch. 10.4) seems to acknowledge this in her discussion of Bayesian methods that also assess reliability.

[25] I interpret the illuminating and important literature (mostly in mathematical statistics) on rates of convergence to the truth—both in the finite-dimensional (Le Cam, 1973, 1986) and infinite-dimensional cases (Ghosal et al., 2000; Shen and Wasserman, 2001)—to be offering a kind of palliative for this, extending these long-run guarantees to the medium-term.

practices" (QRPs)—to "to falsely find evidence that an effect exists [rather] than to correctly find evidence that it does not," supporting their contentions with simulation studies. Two features of these studies are of note here. First, the worrisome feature of evidential reports on which they focus is a higher-than-nominally-reported false positive rate, which is a particular form of disconfirmational unreliability. So, researchers in psychology generally take reliability seriously when it comes to the quality of the evidence. Second, one of the degrees of flexibility is using a stopping rule of the following form: collect a certain number of observations, then if a particular (null) hypothesis is not sufficiently disconfirmed, collect a certain number more and report the disconfirmation level.[26] They report that this increased the false-positive rate—the estimated probability of sufficiently strong erroneous disconfirmation—by about 50% over the reported value derived from an experimental protocol using a fixed stopping rule. The point is that if one uses a confirmation measure that satisfies the SRP, then one is unable to adjust the (dis)confirmation a data set entails for an hypothesis by its reliability.

Although Simmons et al. (2011) employ simulations to demonstrate this, there is also evidence that it is a real concern amongst practicing psychologists. Indeed, John et al. (2012) measured the prevalence of QRPs and attitudes towards them among psychologists. Perhaps surprisingly, they are somewhat common (depending on the type of QRP), and researchers were much more likely to think them defensible in their own scientific practices than in those of others. One of these QRPs was "Stopping collecting data earlier than planned because one found the result that one had been looking for"—i.e., the desired level of (dis)confirmation of a particular hypothesis. For confirmation measures that satisfy the SRP this should entail no problem, but this is apparently not so because psychologists take reliability to be important. Yu et al. (2014) continued this line of research by showing that various heuristics that scientists use to decide when to stop collecting data can affect not only the false positive rate but also the false negative rate and estimated effect size—i.e., the confirmational unreliability for a range of hypotheses.[27] This is so even whether one uses classical confirmation measures (based on p-values) or Bayesian ones.[28]

[26]In more detail: for one two-valued independent variable, collect 20 normally distributed observations per value; if sufficient disconfirmation does not occur, collect 10 more observations per value (Simmons et al., 2011, p. 1361).

[27]Bias in estimated effect sizes plays a role in confirmational unreliability because that latter concept relates two hypotheses by the probability of confirming one if the other is true. It also plays a role in the statistical theory of estimation, which is beyond the scope of the current argument, although issues concerning the SRP can well be extended to it. One can then have debates about the demerits of bias analogous to those about reliability that I describe in sections 4 and 5. See, e.g., Savage (1962, p. 18), Berger and Wolpert (1988, pp. 80–82), and Howson and Urbach (2006, pp. 164–166) for arguments dismissive of the importance of bias.

[28]In particular, they examine the use of Bayes Factors, the comparative confirmation analog of $l$ (equation 7). See also Sanborn and Hills (2014), who use simulation studies for a different variant, so-called Bayesian hypothesis testing. They emphasize that while there is no inconsistency in using Bayesian methods, those methods do not take into account something important for confidence in the conclusions of the research of psychological science. (See also Rouder (2014) and Sanborn et al. (2014) for a dialog on this topic.)

# 4 Deflating Reliability?

Because of the novel way I've made precise, in sections 3.2 and 3.3 how the conflict between demanding the SRP hold of a confirmation measure, on the one hand, and properties like $\varepsilon\mathrm{A}q_{\pm}\mathrm{R}\mathcal{E}$ or $\varepsilon\mathrm{SURLE}\mathcal{E}$, on the other, there have not been direct responses to it in the literature. But there have been responses to the more informal version of the foregone conclusion argument I outlined in section 3.1, which requires confirmation to be sensitive in some way to reliability. Some of these are broad enough to apply, potentially, to my own more formal version. One more formal but under-discussed response is that there is something subtly inconsistent about requiring a confirmation measure to be sensitive to reliability properties. I will discuss and critique that response in section 5. But first, I will discuss two responses that instead attempt to deflate the importance of reliability directly or indirectly, so as to mitigate the conflict with the SRP.

## 4.1 Altered Priors

Much of the discussion of the SRP has been against the wider backdrop of debates about different schools of statistical methodology. In particular, as I have alluded in section 2.2, advocacy of the SRP has tended to correlate with espousal of Bayesianism, and denial with adherence to classical statistics. In this context, Berger and Wolpert (1988) consider a two-party scenario: an experimenter running an experiment similar to the binomial experiment from section 2.2,[29] and an observer who determines the evidential value of the data the experimenter produces for a parameter $\theta$:

> A Bayesian conditionalist might not completely ignore a stopping rule ... if he suspects it is being used because the *experimenter* thinks $\theta$ might be zero. The Bayesian might then assign some positive prior probability, $\lambda$, to $\theta$ being equal to zero, in recognition of the experimenter's possible knowledge. (Berger and Wolpert, 1988, p. 81)

Practically, they note, this may have the consequence that foregone conclusions against $\theta = 0$ will be harder to achieve.

Mayo and Kruse (2001, §6.2) suggest that this is an odd response for an advocate of the SRP, for it seems to *concede* the importance of reliability for confirmation measures, against the SRP.[30] That might be so, but it might instead just be a plain reminder of a circumstance under which a Bayesian may not ignore learning a stopping rule—namely, when doing so would cause them to update their priors. (By definition the "Bayesian conditionalist" updates, and ought to update, her beliefs based on observed data.) Such an update might force greater *or less* belief in a particular value of a parameter, such as $\theta = 0$, depending on what the Bayesian learns, and believes about how what she learns weighs on the states of the world.

[29] Their example is that of testing the mean of a normal distribution with known and fixed variance, but the substantive point at issue here is the same because the stopping process for the experiment does not depend on $\theta$.

[30] See also Mayo (1996, p. 356n30) and Steele (2013, p. 955).

However, this is just to acknowledge that *the stopping rule may be informative* for a Bayesian. Recall from section 2.1 the definition of a stopping rule that is noninformative for a parameter $\theta$: its stopping process does not depend on $\theta$, and any further parameters beyond $\theta$ on which the stopping process depends are logically and *probabilistically* independent thereof. Although the example of testing under discussion has a stopping process that does not depend on $\theta$,[31] if the Bayesian (non-trivially) updates her priors for $\theta$ upon learning the stopping rule, then the stopping process depends on parameters that are *not* probabilistically independent of $\theta$. That means that the stopping rule is (indirectly) informative, and the SRP does not apply. Hence such a response as that of Berger and Wolpert quoted above does nothing to deflate the conflict of the SRP with reliability properties for confirmation measures.

## 4.2 Error Theory

Nevertheless, the framework that Berger and Wolpert (1988) consider, of an assessor of the evidence facing data produced by a distinct experimenter, has been influential in other attempts to deflate the importance of reliability criteria for confirmation. Steele (2013, p. 951), writing about a "persistent experimenter" who chooses the (so-called "optional") stopping rule that stops taking data when sufficient confirmation of a particular hypothesis is achieved, notes that

> a stopping rule can be newsworthy in and of itself if the experimenter is distinct from the inference-maker. In the persistent experimenter cases, it is the experimenter's initial choice of the optional stopping rule that is informative, because it reveals something of the experimenter's motivations or attitudes; these revealed attitudes have a bearing (in Bayesian terms) on the truth of the hypothesis under comparison, *before the experiment has even begun*, in a manner that produces the optional stopping intuition. (Steele, 2013, p. 951)
>
> [...]
>
> In fact, in all of these cases where the experimenter differs from the inference-maker, and where the experimenter's choice of test/stopping rule depends on her prior probabilities/utilities, this choice may be informative for the inference-maker. (Steele, 2013, p. 954)

This is precisely the interpretation of the proposal by Berger and Wolpert (1988) at which I arrived by the end of section 4.1: a stopping rule can be (indirectly) informative for a parameter $\theta$ when the possible stopping rules depend on other parameters not probabilistically independent of $\theta$.

If Steele left it at that, she would also be open to the same criticism, that this is a nonresponse to the conflict between the SRP and reliability properties for confirmation measures. But she attempts to go further by arguing that the only situations in which learning the stopping rule of an experiment *intuitively* makes a difference to the evidential value of that experiment is *precisely* those in which the stopping rule is informative. The

[31] See also footnote 29.

error of the opponent of the SRP is to extrapolate those intuitions to *all* cases.[32]

To do so, Steele (2013, p. 952) proposes to examine a different example in which she claims no such intuition arises. It has the following features:

- The experimenter and evaluator of the evidence are the same.
- One is testing two point hypotheses against one another in a sequential experiment like that of section 2.2.
- One's confirmation measure is a comparative version of $l$, the Bayes Factor.
- There is a fixed cost for an erroneous inference, i.e., confirming a false hypothesis or disconfirming a true one, at the end of the experiment.
- The cost of an observation is positive but constant.
- Once should act to minimize expected costs.

It turns out then that the experimental design satisfying these criteria is one which does not use a fixed stopping rule, but rather stops when one of the two hypotheses is sufficiently confirmed above the other. Steele then claims that no intuition concerning the relevance of reliability criteria arises nor enters into the evaluation of the evidence from this experiment.

However, even setting aside the appeal to an intuition which the present author does not share, there are at least three significant flaws with this argument. In the first place, it doesn't even have the right logical form to establish its conclusion. A single example is insufficient to establish that intuitions about the relevance of reliability criteria for confirmation never arise for sequential experiments without informative stopping rules. Perhaps the case is intended as the basis for an (implausible) inductive argument to a universal generalization, but then some motivation for why *all* cases are analogous to it is yet forthcoming. Any exception to the generalization blocks the force of the argument.

The second flaw is that whether optional stopping is more or less reliable or pragmatically costly than a fixed stopping rule for a proposed sequential experiment is not actually relevant to the question at hand. That question is whether, given two experiments that produced identical data, differences in the experiments' stopping rules make a difference to the evidence each experiment yields. The example is compatible with both affirming and denying this difference-making. In particular, affirming that reliability makes a difference for confirmation does not entail that an optional stopping rules sometimes is the better design for an experiment, because by itself that affirmation makes no prescription about experimental design at all. It makes only a claim about dependence, that however a sequential experiment (dis)confirms an hypothesis depends not just on the values of the data collected, but the reliability of that collection method.

[32]Steele (2013) also considers two other "error theories" that she rejects: the first is that the intuition for reliability concerns arises from a conflation with considerations of experimental design, as Backe (1999) had urged. The second involves a conflation with experiments in which only a summary statistic of the experimental outcomes is available. Steele (2013, pp. 949–950) rightly rejects these, in my view, as inaptly responding to the issue at hand. However, as I discuss in the sequel, this is also my view of her preferred error theory.

These previous observations are related to the third flaw: despite Steele's assertions to the contrary, the example is not actually relevantly analogous to the example used in the foregone conclusions argument. The most important reason for this is that there is no chance of a foregone conclusion, or one nearly so: ceteris paribus, the stopping rule does not make it likely, e.g., that one hypothesis will be confirmed even if it is false, at least without making any explicit calculations. Relevantly analogous examples are ones in which it is entirely obvious either that the (dis)confirmational unreliability of the confirmation measure used in the experiment is unacceptably high (cf. $\varepsilon \mathrm{A} q_{\pm} \mathrm{R} \mathcal{E}$), or that there are two different possible stopping rules for the data produced which differ significantly in their (dis)confirmational reliability (cf. $\varepsilon \mathrm{SURLE} \mathcal{E}$). Since it is not obvious in this case, if the calculations of the confirmational reliability of the proposed confirmation measure showed that it violated $\varepsilon \mathrm{A} q_{\pm} \mathrm{R} \mathcal{E}$ or $\varepsilon \mathrm{SURLE} \mathcal{E}$ for reasonable $\varepsilon(\theta, \theta')$, $q_+$, and $q_-$, I would doubt that no questions about their importance would ever be raised.

The origin of the last two flaws might be in Steele's misconception of the circumstances under which (indirectly) informative stopping rules can arise: "The argument here is that the optional stopping intuition arises from our association of optional stopping tests with a particular sort of shady experimenter" (Steele, 2013, p. 955).[33] But it just ain't so!

> Why, after all, should we think the experimenter is using [an optional stopping rule] to *deceive* you? Why not regard his determination to demonstrate evidence against the null hypothesis as a sign that the null is *false*? Perhaps he is using [the optional stopping rule] only because he *knows* that [it is false] and he is trying to *convince* you of the truth! (Mayo and Kruse, 2001, p. 398)

Such an experimenter, ever the hopeful idealist, does not intend to mislead, but his experimental methods might still be unreliable. Indeed, in the discussion of their empirical results regarding psychologists' tendency to employ optional stopping rules, Yu et al. (2014, p. 279) write that

> One need not postulate that decisions to terminate sampling prematurely are an act of deception, since responsible researchers can easily fall prey to the allure of small sample results [in which small samples are believed to be representative of a population.] ... [Also,] scientists may be intrinsically more motivated to find effects than to indicate that effects do not exist. ... These tendencies may predispose individuals to engage in optional stopping.

The intuitions that reliability matters to confirmation come not just from the extreme cases of maximal unreliability and scientific fraud, but also from moderate cases of unreliability, workaday earnest scientific motivations, and mundane cognitive heuristics.

# 5 The Decision-Theoretic Argument

The attempts, outlined in section 4, to accept the intuitiveness of reliability's import for evidence while deflating its substantive impact were not successful. But as I mentioned at

[33] See also Gandenberger (2015, p. 13), who calls them "disingenuous experimenters."

the beginning of that section, there is another response to the foregone conclusions argument and rejections of the SRP more generally, due to Berger and Wolpert (1988, pp. 83–5), that argues that rejection of the SRP is inconsistent with other properties acknowledged to be important for reliable confirmation. In particular, this response observes that denying the SRP leads in certain cases to following *inadmissible decision rules*.

Explaining this response requires introducing some further notions from decision theory. Suppose that, faced with the conclusion of a given sequential experiment, one must decide how to act. This could include deciding merely what to believe (with its concomitant if not entirely foreseeable effects on future concrete action), but also how to report and interpret the evidential value of the concluded experiment, or adopting a particular policy for action. In any case, let $\mathcal{A}$ be the set of possible actions, $\mathcal{Z}$ the set of possible outcomes of the experiment, and $\Theta$ (as before) the set of parameter values (i.e., states of the world) determining the probabilities (or the probability densities) for these outcomes. A decision rule for $(\mathcal{Z}, \mathcal{A})$ is then a plan for making an action depending on the outcome of the experiment. Formally, it is represented by a function $\delta : \mathcal{Z} \to \mathcal{A}$. Depending on the data, for example, one could announce the (dis)confirmation of an hypothesis to a particular degree.

Actions, hence decision rules, do not in general affect the world indifferently. Each action may entail a cost, or *loss*, as measured on some definite scale, which is assumed to be represented by a positive real number. (Negative numbers then represent gains.) Moreover, that loss may also depend on the state of the world. We may represent these dependencies by a loss function, $L : \mathcal{A} \times \Theta \to \mathbb{R}$. If one's actions are governed by a decision rule $\delta$, and the state of the world is $\theta$, then observing outcome $z$ would lead to a loss $L(\delta(z), \theta)$.

Because we are only assuming that the state of the world determines the probabilities (or the probability densities) of various outcomes of the experiment, various outcomes could potentially be observed even if the state of the world and the decision rule were fixed. Thus different losses can potentially be incurred depending on how the data lead to different actions according to $\delta$. The *risk* $R$ for a given decision rule $\delta$ when the state of the world is $\theta$ is then defined as the expected loss,

$$R(\theta, \delta) = \mathbb{E}_\theta L(\delta(Z), \theta), \tag{19}$$

where $\mathbb{E}_\theta$ is the expectation operator under the probability distribution for $Z$ determined by $\theta$.

With these notions in hand, Berger and Wolpert (1988, pp. 83–85) consider two experiments $Z_1$ and $Z_2$ that differ only by their stopping rule—e.g., the two considered in section 2.2—and assume that some action will be taken after one or the other of them is complete. That is, they assume that there has been selected respective decision rules $\delta_1$ and $\delta_2$ for the two experiments. They then make the following four assumptions about these experiments:[34]

**Convexity of Actions (CA)** The set of actions has the structure of a real vector space, on which the loss function is strictly convex. I.e., for any $a, a' \in \mathcal{A}$, $\alpha \in (0, 1)$, and $\theta \in \Theta$, $L(\alpha a + (1 - \alpha)a', \theta) < \alpha L(a, \theta) + (1 - \alpha)L(a', \theta)$.[35] (For instance, the action

[34] At least, what follows are the assumptions I have reconstructed from their exposition. Except for the Weak Conditionality Principle, what follows are my own names for these assumptions.

[35] Berger and Wolpert (1988, p. 84) remark that "More general loss functions can be handled also" but it is difficult to see how their proof strategy could be substantively generalized without this assumption.

could be to report the best estimate $\hat{\theta}$ of $\theta$ based on the evidence from the experiment, with a loss function quadratic in the difference: $L(\hat{\theta}, \theta) = |\hat{\theta} - \theta|^2$.)

**Weak Conditionality Principle (WCP)** Consider the "mixed" experiment $Z^* = (J, Z_J)$, in which either $J = 1$ or $J = 2$ is observed, each with probability 0.5 and independent of the true value of $\theta$, after which $Z_J$ is performed. Then, given that one has committed to use a confirmation measure $c$, $c((j, z_j), \theta, \theta') = c(z_j, \theta, \theta')$ for all $\theta, \theta' \in \Theta$.

**Decision Inequivalence (DI)** There exists some potential outcome $z$ of *either* $Z_1$ and $Z_2$ such that $\delta_1(z) \neq \delta_2(z)$.

**Decisions Supervene on Confirmation (DSC)** Given that one has committed to use a confirmation measure $c$, if the results $z$ and $z'$ from any of the experiments $Z_1$, $Z_2$, or $Z^*$ are such that $c(z, \theta, \theta') = c(z', \theta, \theta')$ for all $\theta, \theta' \in \Theta$, then for any decision rule $\delta$, $\delta(z) = \delta(z')$.

I shall return to the interpretation of these conditions presently, but first I complete the statement of their argument.

Let $\delta$ now be any decision rule for $Z^*$ (satisfying in particular DSC). Berger and Wolpert (1988, p. 84–85) consider the following distinct decision rule for $Z^*$:

$$\delta^*((j, z^*)) = \begin{cases} \frac{1}{2}\delta((1, z)) + \frac{1}{2}\delta((2, z)) & \text{if } z^* = z, \\ \delta((j, z^*)) & \text{otherwise.} \end{cases} \tag{20}$$

This decision rule is exactly like $\delta$, except for when the data are equal to the value mentioned in DI, for which the decision rules for the individual experiments $Z_1$ and $Z_2$ are different. In that case, $\delta^*$ averages over what those decision rules would entail. Berger and Wolpert then prove that $R(\theta, \delta^*) < R(\theta, \delta)$ for all $\theta \in \Theta$. Thus, $\delta$ is an *inadmissible* decision rule because it is dominated by $\delta^*$.[36]

How do Berger and Wolpert interpret this result so as to argue that some sort of stronger reliability property in conflict with the SRP should not be required of a confirmation measure $c$? I reconstruct their reasoning as follows. In the first place, DI and DSC imply by modus tollens that $c(Z_1 = z, \theta, \theta') \neq c(Z_2 = z, \theta, \theta')$ for some $\theta, \theta' \in \Theta$. This is just to deny that the SRP holds in this case. By WCP, $c(Z^* = (j, z), \theta, \theta') = c(Z_j = z, \theta, \theta')$ for $j = 1, 2$ and all $\theta, \theta' \in \Theta$, so by DSC, $\delta((j, z)) = \delta_j(z)$ for $j = 1, 2$. That's just to say that, at least in the case of data $z$, $\delta$ implements the decisions that $\delta_1$ would if in fact $j = 1$, and those that $\delta_2$ would if in fact $j = 2$. These are the decision rules resulting from using a confirmation measure that does not obey the SRP in this case, and indeed, by DI, $\delta((1, z)) \neq \delta((2, z))$. So, in this case, adopting a confirmation measure that does not obey the SRP, as strong reliability properties would sometimes require, leads to a kind of irrationality, which inadmissible decision rules represent.

Another way of casting the conclusion of this argument is that because there is in some cases a conflict between the SRP and instances of $\varepsilon \mathrm{A} q_{\pm} \mathrm{R} \mathcal{E}$ or $\varepsilon \mathrm{SURLE} \mathcal{E}$, such instances must

[36] One decision rule $\Delta$ is said to *dominate* another, $\Delta'$, just when $R(\theta, \Delta) \leq R(\theta, \Delta')$ for all $\theta \in \Theta$, and the inequality is strict for at least one value of $\theta \in \Theta$. A decision rule is *admissible* if and only if it is not dominated by another decision rule.

conflict with the conjunction of the above premises and the demand to act only according to admissible decision rules. Can some of these be separated as relatively innocuous background assumptions? In other words, is there a clear source for the conflict with these strong reliability properties?

In the first place, the above argument uses the framework of standard decision theory centered around maximizing expected utility, which entails that actions should follow admissible decision rules. Although this framework has been criticized as a general framework for rational decision-making (Steele and Stefánsson, 2016, §5), it is difficult to hold as being inapplicable in *every* decision problem. All that is needed for the argument is that there is *some* decision problem involving sequential experiments of the sort described. This, therefore, does not seem to be the source of the conflict. I suggest that it can be taken as a background assumption.

To an even greater extent, both CA and DSC are highly specialized assumptions that apply only to a narrower class of decision problems. CA requires the set of actions and the loss function to have very particular properties not found in general decision problems. DSC mandates that the decision rule used is essentially a function of the confirmation measure used. This is also somewhat unusual because the confirmation measure is for incremental confirmation, and does not describe one's total evidence for the state of the world, which ought rather to be determinate of one's decisions. Nevertheless, there are plausibly decision problems in which they do apply, such as a situation in which there is no other evidence for the value of $\theta \in \Theta$ that obtains except for that produced by the experiment, and the action to be taken is reporting an estimate $\hat{\theta}$ for $\theta$, with (say) a loss function quadratic in the difference between them. Thus, I suggest that these assumptions can also be taken as a part of the background to the conflict.

By contrast, DI just formulates one way of assuming that the SRP does not hold in some situation or other. It supposes that there is some common potential outcome of two sequential experiments that would result in two different actions. As I showed above, against the backdrop of DSC this entails that the confirmation provided to some $\theta \in \Theta$ would be different for the two experiments, despite the common outcome, and this is just a denial that the SRP hold for that confirmation measure in this case. Since this is entailed by some instances of $\varepsilon \mathrm{A} q_{\pm} \mathrm{R} \mathcal{E}$ and $\varepsilon \mathrm{SURLE} \mathcal{E}$, this is one of the sources of the conflict.

The other source of the conflict must then be the remaining premise which I have yet to discuss, namely WCP. Indeed, even outside the context of Berger and Wolpert's argument, one can already identify it as being in conflict with some reliability properties one could require for a confirmation measure. To see this, one need only observe that the outcome of a mixed experiment in general does not have the same (dis)confirmational reliability as its "component" experiment performed. This is because the probability of a mixed experiment achieving a certain (dis)confirmation of an hypothesis is going to be the average of that probability for each of its two "components." If these two probabilities are distinct, as will in general be the case for experiments with different stopping rules, then this average will lie strictly in between the two. To assume WCP is to assume already that differences in (dis)confirmational reliability do not bear on confirmation, at least for experiments that have a form like that of $Z^*$.

In sum, one can understand Berger and Wolpert's argument as proceeding in the context of standard decision theory, concerning decision problems where the loss function is strictly

convex on the set of actions (CA) and where decisions supervene on incremental confirmation (DSC). The argument then shows that, in this context, WCP is in conflict with DI, which is implied by certain plausible instances of $\varepsilon \mathrm{A} q_{\pm} \mathrm{R} \mathcal{E}$ and $\varepsilon \mathrm{SURLE} \mathcal{E}$. Thus WCP is the source of this conflict with these strong reliability properties. But on consideration of WCP, it becomes clear it directly conflicts with these properties regardless of the decision-theoretic context. This is not to the detriment of Berger and Wolpert's argument, as these properties are logically stronger than DI in this decision-theoretic context. But because my present goal is to isolate the source of the conflict with those properties, it is notable that one can extricate the conflict from the decision theoretic background, as WCP does not invoke any decision-theoretic concepts.

Now, of course Berger and Wolpert urge one to reject DI in the face of their reductio argument. But in light of the foregoing discussion, and the motivations for demanding confirmational reliability properties like instances of $\varepsilon \mathrm{A} q_{\pm} \mathrm{R} \mathcal{E}$ or $\varepsilon \mathrm{SURLE} \mathcal{E}$ canvassed in section 3.4, it is worthwhile to compare the intuition behind WCP. The underlying intuition being tapped seems to be that the component experiment that was actually performed of the mixed experiment is "really" just the same as performing that component without mixing. The initial "mixing" process to choose which component experiment to perform should not matter because it is independent of $\theta$—it is, technically speaking, *ancillary* to $\theta$. However, (dis)confirmational reliability is a modal property of a confirmation measure applied to an experiment; one must imagine repeated experiments in order to determine an experiment's reliability. In particular, one must be careful not to conflate imagining just the component that was performed repeated with imagining the mixture experiment itself being repeated. The two are not equivalent, and that should, according to reliability criteria, be reflected as a difference in their ability to (dis)confirm hypotheses.

As a technical complement to this insight, one might observe that just because a statistic is ancillary to $\theta$ does *not* mean that it is necessarily independent of any *sufficient* statistic for $\theta$, one that "captures all the information about $\theta$" (Berger and Casella, 2002, p. 272).[37] That a statistic is ancillary to $\theta$ does not mean that it is not indirectly informative (Berger and Casella, 2002, §2.4.2–2.4.3). The literature on ancillary statistics and their interpretation is large—see, e.g., Ghosh et al. (2010) and references therein—so further examination of WCP along these lines and its conflict with reliability criteria for confirmation must await another occasion. The important point here is that one should not extrapolate the irrelevance of the mixing mechanism for $\theta$ to its irrelevance for estimators of $\theta$.

# 6 Conclusions and Implications

Ultimately, when one makes the foregone conclusions argument more precise and nuanced using $\varepsilon \mathrm{A} q_{\pm} \mathrm{R} \mathcal{E}$ and $\varepsilon \mathrm{SURLE} \mathcal{E}$, it remains successful in exhibiting a conflict between SRP and (dis)confirmational reliability criteria for confirmation. For reasonable $\varepsilon(\theta, \theta')$, $q_+$, and $q_-$, conflict arises for simple sets of experiments $\mathcal{E}$. Besides having the virtue that it can be formulated without presupposing a classical statistical framework already inimical to the SRP, it also avoids the rebuttal, discussed in section 3.1, that it attempts to establish a

[37] A statistic $T$ of the data $X$ from an experiment for a parameter $\theta$ is said to be sufficient for $\theta$ when the probability distribution of $X$ conditioned on $T(X)$ does not depend on $\theta$.

conclusion that is too strong to be true—that the SRP must conflict with the NFC property. Moreover, as I described in section 4, attempts to deflate the importance of reliability, albeit directed at the original foregone conclusions argument, are unsuccessful.

The final counterargument I considered, in section 5, attempts to use decision theory to reveal a subtle inconsistency in rejecting the SRP, at least for specialized examples. That attempt hinges on accepting WCP, which I showed is already in conflict with reliability criteria for confirmation. Hence it is the source of the conflict, regardless of whether one situates it in a decision-theoretic context. This bears on the much larger debate about the LP, for Birnbaum (1962) famously proved that the conjunction of WCP with the Weak Sufficiency Principle is *equivalent* to the LP for any confirmation measure $c$:[38]

**Weak Sufficiency Principle (WSP)** Given any experiment $X$ for $\theta \in \Theta$, a sufficient statistic[39] $T$ of $X$ for $\theta$ and $x_1, x_2$ two possible outcomes for $X$, if $T(x_1) = T(x_2)$, then $c(x_1, \theta, \theta') = c(x_2, \theta, \theta')$ for all $\theta, \theta' \in \Theta$.

**Likelihood Principle (LP)** Given any two experiments $X_1$ and $X_2$ for $\theta \in \Theta$ and respective possible outcomes $x_1$ and $x_2$, if $X_1 = x_1$ and $X_2 = x_2$ are likelihood equivalent for all $\theta \in \Theta$, then $c(x_1, \theta, \theta') = c(x_2, \theta, \theta')$ for all $\theta, \theta' \in \Theta$.

Thus, while a rejection of the SRP entail a rejection of the LP, by Birnbaum's theorem it raises the question of which of the WSP or WCP must be rejected. The arguments of section 5 show that at least the WCP must be rejected if one is to accept strong reliability properties for confirmation.

Finally, as I discussed briefly at the ends of sections 3.4 and 4.2, stopping rules are alleged to play an important role in assessing the replicability of scientific results, especially in psychology. Following reporting methods that would be legitimized under the SRP has been deemed a QRP by some (Simmons et al., 2011; John et al., 2012; Yu et al., 2014; Sanborn and Hills, 2014); although there have been some defenses along Bayesian lines (Rouder, 2014), a fuller assessment of the evidential role of stopping and reliability in the context of these debates would be fruitful. Old foundational debates on the nature of evidence and confirmation may find new life in that wellspring.

# References

Anscombe, F. J. (1954). Fixed-sample-size analysis of sequential observations. *Biometrics*, 10:89–100.

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

Berger, J. O. and Wolpert, R. L. (1988). *The Likelihood Principle*. Institute of Mathematical Statistics, Hayward, CA, 2nd edition.

[38] Actually, Birnbaum used slightly different principles; this version follows Berger and Wolpert (1988).

[39] See footnote 37 for a definition of sufficiency.

Berger, R. L. and Casella, G. (2002). *Statistical Inference*. Thomson Learning, Pacific Grove, CA, 2nd edition.

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

Bovens, L. and Hartmann, S. (2002). Bayesian networks and the problem of unreliable instruments. *Philosophy of Science*, 69(1):29–72.

Bovens, L. and Hartmann, S. (2003). *Bayesian epistemology*. Oxford University Press, Oxford.

Cornfield, J. (1970). The frequency theory of probability, Bayes' theorem, and sequential clinical trials. In Meyer, D. L. and Collier, Jr, R. O., editors, *Bayesian Statistics*, pages 1–28. Peacock Publishing, Itasca, IL.

Crupi, V. (2016). Confirmation. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, winter 2016 edition.

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

Feller, W. (1940). Statistical aspects of E.S.P. *Journal of Parapsychology*, 4:271–298.

Gandenberger, G. (2015). Differences among noninformative stopping rules are often relevant to Bayesian decisions. arXiv:1707.00214.

Ghosal, S., Ghosh, J. K., and van der Vaart, A. W. (2000). Convergence rates of posterior distributions. *Annals of Statistics*, 28:500–531.

Ghosh, M., Reid, N., and Fraser, D. A. S. (2010). Ancillary statistics: A review. *Statistica Sinica*, 20(4):1309–1332.

Goldman, A. and Beddor, B. (2016). Reliabilist epistemology. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, winter 2016 edition.

Howson, C. and Urbach, P. (2006). *Scientific Reasoning: The Bayesian Approach*. Open Court, Chicago, 3rd edition.

Huber, F. (n.d.). Confirmation theory. In *The Internet Encyclopedia of Philosophy*. Accessed 30 Mar. 2018.

John, L. K., Loewenstein, G., and Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science*, 23(5):524–532.

Kadane, J. B., Schervish, M. J., and Seidenfeld, T. (1996). Reasoning to a foregone conclusion. *Journal of the American Statistical Association*, 91(435):1228–1235.

Kerridge, D. (1963). Bounds for the frequency of misleading Bayes inferences. *The Annals of Mathematical Statistics*, 34(3):1109–1110.

Le Cam, L. M. (1973). Convergence of estimates under dimensionality restrictions. *Annals of Statistics*, 1:38–53.

Le Cam, L. M. (1986). *Asymptotic Methods in Statistical Decision Theory*. Springer, New York.

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

Mayo, D. G. (2004). Rejoinder. In Lele, S. R. and Taper, M. L., editors, *The Nature of Scientific Evidence: Statistical, Philosophical, and Empirical Considerations*, pages 101–118. University of Chicago Press, Chicago.

Mayo, D. G. and Kruse, M. (2001). Principles of inference and their consequences. In Corfield, D. and Williamson, J., editors, *Foundations of Bayesianism*, pages 381–403. Kluwer, Dordrecht.

Raiffa, H. and Schlaifer, R. (1961). *Applied Statistical Decision Theory*. Division of Research, Graduate School of Business Administration, Harvard University, Boston.

Robbins, H. (1952). Some aspects of sequential design of experiments. *Bulletin of the American Mathematical Society*, 58:527–535.

Rouder, J. N. (2014). Optional stopping: No problem for Bayesians. *Psychonomic Bulletin and Review*, 21(2):301–308.

Royall, R. (2000). On the probability of observing misleading statistical evidence. *Journal of the American Statistical Association*, 95(451):760–768.

Sanborn, A. N. and Hills, T. T. (2014). The frequentist implications of optional stopping on Bayesian hypothesis tests. *Psychonomic Bulletin and Review*, 21(2):283–300.

Sanborn, A. N., Hills, T. T., Dougherty, M. R., Thomas, R. P., Yu, E. C., and Sprenger, A. M. (2014). Reply to Rouder (2014): Good frequentist properties raise confidence. *Psychonomic Bulletin and Review*, 21(2):283–300.

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

Shen, X. and Wasserman, L. (2001). Rates of convergence of posterior distributions. *Annals of Statistics*, 29:687–714.

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

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

Steel, D. (2003). A Bayesian way to make stopping rules matter. *Erkenntnis*, 58:213–227.

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

Steele, K. and Stefánsson, H. O. (2016). Decision theory. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, winter 2016 edition.

Yu, E. C., Sprenger, A. M., Thomas, R. P., and Dougherty, M. R. (2014). When decision heuristics and science collide. *Psychonomic Bulletin and Review*, 21(2):268–282.