Brit. J. Phil. Sci. **59** (2008), 857–879

How to Discount Double-Counting When It Counts: Some Clarifications

Deborah G. Mayo

ABSTRACT

The issues of double-counting, use-constructing, and selection effects have long been the subject of debate in the philosophical as well as statistical literature. I have argued that it is the severity, stringency, or probativeness of the test—or lack of it—that should determine if a double-use of data is admissible. Hitchcock and Sober ([2004]) question whether this 'severity criterion' can perform its intended job. I argue that their criticisms stem from a flawed interpretation of the severity criterion. Taking their criticism as a springboard, I elucidate some of the central examples that have long been controversial, and clarify how the severity criterion is properly applied to them.

doi:10.1093/bjps/axn034

A common intuition about evidence is that if data **x** have been used to construct a hypothesis $H(\mathbf{x})$, then **x** should not be used again as evidence in support of $H(\mathbf{x})$: The fact that $H(\mathbf{x})$ was constructed to fit or accommodate the data, many think, prevents the data from also counting as a good test of, or reliable evidence for, $H(\mathbf{x})$. This 'no double-counting' requirement captures a general type of prohibition against data mining, hunting for significance, tuning on the signal, and *ad hoc* hypotheses, in favor of requiring predesignated hypotheses and novel predictions. Whether (and when) inferences should be discounted, if not disallowed, when double-counting has occurred, has long been the source of disagreement and debate both in philosophical and statistical literatures. It is well known that certain types of double-counting can lead to unreliable inferences. For example, if one is allowed to search through several factors and selectively report just those that show (apparently) impressive correlations, there is a high probability of erroneously inferring that a correlation is real. However, it is equally clear that there are reliable procedures for using data both to identify and test hypotheses: the use of a DNA match to identify a criminal, radiointerferometry data to estimate the deflection of light, and in such homely examples as of using a ruler to measure the length of a kitchen table. Here, although the inferences (about the criminal, the deflection effect, the table length) were constructed to fit the data, they were deliberately constrained to reflect what is correct, at least approximately. It is the severity, stringency, or probativeness of the test—or lack of it—therefore that should determine if a double use of data is admissible—or so I have argued (Mayo [1991], [1996]; Mayo and Kruse [2001]; Mayo and Cox [2006]).

Hitchcock and Sober ([2004]) question whether the severity criterion can perform this job: if it is interpreted so as to sanction the latter cases of reliable double-counting, they think, then it appears also to countenance the former cases of unreliable selection effects. How then can I claim it correctly discriminates between legitimate and illegitimate cases? The answer to the puzzle is that they have misinterpreted or misapplied the severity criterion so as to inadvertently fall into the very fallacy it was designed to avoid. I show that their interpretations of the severity criterion lead to a false dilemma: either double-counting always leads to minimally severe tests, or else it has no influence on a severity assessment. This discussion focuses just on this puzzle about double-counting. The general issue is at the heart of ongoing controversy, particularly in noisy sciences where controlled experiments are impossible, and is especially relevant with today's explosion of data-dependent algorithms for model search. I will begin with four main theses and positions relating double-counting to the severity criterion, as I construe it.

1 Severity and Use-Constructing: Four Points (and Some Clarificatory Notes)

Inferences involving double-counting may be characterized by means of a rule R by which data **x** are used to construct or select hypothesis $H(\mathbf{x})$[1] so that the resulting $H(\mathbf{x})$ fits **x**; and then used 'again' as evidence to warrant $H(\mathbf{x})$. (Mayo [1996]) called the latter a 'use-constructed' test procedure because it corresponds to the description of $H(\mathbf{x})$ as violating 'use-novelty' (Musgrave [1974]; Worrall [1978], [1989]). The bare bones of *a use-constructed test procedure* is to output $H(\mathbf{x})$ as supported, well tested, indicated, or the like by data **x**, where **x** has been used to construct or select for testing $H(\mathbf{x})$ in such a way as to fit, or pass, or be in accordance with, **x**. For simplicity, we can allow the same characterization to cover cases where **x** is used in arriving at a disagreement or lack of fit with a given claim C. Here the constructed hypothesis $H(\mathbf{x})$ could take the form of asserting that C is false, a discrepancy from C exists, or a rival to C is true.

In many ways the term 'double-counting' is preferable to 'use-constructing': not only is it more general but it emphasizes that the question is not whether it matters that **x** was used in constructing $H(\mathbf{x})$, but whether *re*using the same data alters the evidential import of the observed fit between data **x** and hypothesis $H(\mathbf{x})$. Nevertheless, it is so much less cumbersome to allude to a 'use-constructed' hypothesis than to a 'hypothesis arrived at by double-counting' that I will frequently use the former term. So 'use-constructing' in this discussion will always refer to double-counting, and not to cases where the hypothesis constructed at one stage is later tested on distinct data (and where it is that distinct data that is the basis for the inference in question).

1.1 Point 1: Getting beyond 'all or nothing' standpoints

A central aim of my analysis has been to get beyond two extreme positions: the first, that double-counting, use-constructing and the like *never* count, that 'what the data have to say' about a hypothesis is solely a function of given data **x**, and a hypothesis *H*; the second, that it *always* does: that the strength of an inference to a use-constructed hypothesis will always count less than if **x** were not used in the construction or selection of *H*. Those who hold the second position require (or prefer) the data to be 'use-novel' (Mayo [1991], [1996]; Musgrave [1974]; Worrall [1978]). However, some construction procedures, while guaranteed to output some $H(\mathbf{x})$ *or other*, whatever the data, nevertheless ensure that false or erroneous outputs are rare or even impossible. This claim about a procedure

[1] We could further generalize this to include using the data to determine when data collection stops for a predesignated *H*, but we restrict the definition to the cases of interest in our current discussion.

'rarely erring' is an example of an *error probability*. The supposition that double-counting never counts against an inference results in overlooking the special error probability adjustments that many cases require if misleading inferences are to be avoided. The supposition that it should always be eschewed, on the other hand, indiscriminately disparages all procedures that involve elements of double-counting, thereby disparaging certain perfectly reliable cases of model validation and searching, as well as many data-dependent inferences.

1.2 Point 2: The rationale for prohibiting double-counting is the requirement that tests be severe

The second thesis concerns the epistemological rationale for prohibiting, downgrading or 'discounting' inferences based on use-constructions. Identifying the rationale has long been problematic. The intuition behind condemning double-counting, I argue, echoing Worrall ([1989]), is essentially the familiar Popperian warning that agreements between data and hypotheses only count when they result from stringent attempts to find flaws:

> Observations or experiments can be accepted as supporting a theory (or a hypothesis, or a scientific assertion) only if these observations or experiments are severe tests of the theory—or in other words, only if they result from serious attempts to refute the theory, and especially from trying to find faults where these might be expected in the light of all our knowledge. (Popper, [1994], p. 89)

Although Popper's methodology never adequately captured this intuition, there is clearly something right-headed in condemning the 'too easy' corroborations that we know can be 'rigged' so as to accommodate data while protecting pet hypotheses rather than subjecting them to scrutiny. The most important upshot of requiring that a hypothesis pass a severe test before a fit between H and $\mathbf{x}$ is counted as genuine evidence for H is the recognition that how well H has been tested can be altered by the particular way the observed 'fit' came about.

As members of the Popper–Lakatos school were well aware, this recognition introduced various context-dependencies into an account of confirmation or testing in order to capture 'anti-adhocness' (Lakatos [1970], p. 124). Here is where subtleties enter: If one holds evidential appraisal to be simply a matter of the logical or formal relationship between data $\mathbf{x}$ and hypothesis H, then one is free from the amorphous business of explaining how the manner in which hypotheses are constructed or selected can matter to how well they are tested or warranted. However, there is no way to avoid this task, if one seeks an account that captures what many take to be an essential intuition about evidence—an intuition with which we agree. As difficult as our task is, we need not conclude that construction and selection effects boil down to the

experimenter's intentions that are 'locked up in his head' (Savage [1962], p. 76), as some allege. Such effects may be located within the objective properties that determine the severity of a test, much as with the effects of sample size and experimental design (Mayo and Kruse [2001]).[2]

1.2.1 Severity

I begin with a bare skeletal formulation of the severity requirement:

Hypothesis H passes a severe test T with **x** if,

(i) **x** agrees with or 'fits' H (for a suitable notion of fit[3]), and

(ii) with very high probability, test T would have produced a worse fit with H (or no fit at all), if H were false or incorrect.

Equivalently, (ii) can be written as

(ii) with very low probability, test T would have produced a result that fits H as well as (or better than) **x** does, if H were false or incorrect.

We may abbreviate 'the severity with which H passes test T with **x**' by SEV(T,H,**x**) to emphasize that it involves three arguments. (See Mayo [2005]; Mayo and Spanos [2006]).

This rather bland formulation serves as a brief capsule of the much more vivid context-specific arguments that flesh out the severity criterion when it is clearly satisfied, or flagrantly violated. When severity is clearly satisfied we might argue that the test was genuinely probative—that the occurrence of so good a fit between data **x** and H would be very difficult, extraordinary, virtually impossible (or the like) were H incorrect in what it asserts regarding the phenomenon or question of interest. When it is seriously violated we may voice such complaints as: 'This test permits practically any data to be interpreted so as to count as fitting H rather than giving H's faults a chance to show up by means of clashes with data'. A favorite example of severity being satisfied, promoted by Popper, is the successful fit between Einstein's predicted light deflection H, and the 1919 observed eclipse data **x**; a familiar example of a severity violation is the ease of arranging a successful fit with a vague prediction of an astrological hypothesis. The main points to keep in mind are these:

- When the severity requirement is satisfied, it is because the falsity of H would render the fit (between H and **x**) extraordinary.
- When severity is violated, it is because the falsity of H *fails* to render the fit extraordinary—i.e., so good a fit could easily occur even if H were false.

[2] Members of the Popper–Lakatos school looked to characteristics of the 'research program' for objective ways to characterize the needed context-dependencies, e.g., Musgrave's 'historical' account of confirmation.

[3] Minimally, this requires P(**x**;H) to exceed P(**x**; not-H).

Clearly, if the occurrence of an observed fit between **x** and H was impossible unless H was true, then the inference from **x** to H is fully warranted without threat of error. But inductive inference contains uncertainties and the most we may be able to do is control these errors, and report them at least approximately correctly. Although severity is a matter of degree, for our purposes here, we can keep to qualitative assessments, e.g., highly severe or highly *in*severe tests.

1.2.2 Some clarificatory notes

My goal is to effectively get to the heart of an issue that is often shrouded in technical complexity, keeping tangential issues to a minimum. I see no other way to get philosophers of science from various backgrounds on board, and I am convinced that there is a real need for philosophical clarification here. To that end, I will strive to keep notions as general as possible so as to allow substituting one's preferred way of talking about evidence, and will keep symbols and qualifications to a minimum. Listing some notes about the terminology and assumptions that I will employ should suffice to avoid confusions without getting bogged down. Each will be illustrated and illuminated as we proceed:

(1) *Hypotheses.* A 'hypothesis' H while it can take many forms, should be regarded as a claim about some aspect of the process generating data **x**, or the population from which **x** is sampled (both of which are generally given by an associated model M). Speaking of H being 'true' does not presuppose a realist position, and may be construed in different ways. 'H is true' might mean: H correctly describes the process generating **x**, with respect to the particular aspect under test (as modeled in M)—where this 'aspect' can vary. In some contexts, H might assert that some quantity, say μ, is in a given range, e.g., $\mu = \mu_0 \pm \varepsilon$. Correspondingly, '$H$ is false' has to be specified as a particular denial of what H asserts, e.g., μ differs from μ_0 by more than ε.

(2) *Erroneous inference from an observed fit*. The 'fit' between H and **x** in condition (i) of severity, whatever its form, refers to something observed or recorded in a particular experiment or sample (generally by means of a statistic or summary of the data). It should not be confused with the use of 'fit' as describing H 'fitting' what is actually the case regarding the aspects under inquiry (as in 'H fits the blueprint of the universe'). The latter is what is *inferred* on the basis of an observed fit with a particular dataset **x**. The goal is to infer from observed fits (or misfits) to how well (or poorly) H captures aspects of the process generating the data. 'Errors', or erroneous inferences, here will refer to any mistakes in moving from the observed or recorded fit to an inference about the population or general phenomenon involved (a partial list is in Section 4).[4]

[4] In discussing use-constructing (violations of use-novelty), I had assumed the minimal requirements for an adequate criterion of fit (i) is satisfied, so that, for example, H would not pass the test unless **x** were at least more probable if H is true than if H is false; and presumably others picking up the discussion assume this as well. Thus, I will continue with this assumption here,

(3) *Severity as 1 minus the (appropriate) error probability*. Typically, the successful application of a use-constructed test procedure is identified as outputting a 'fit' so as to satisfy clause (i); thus, attention turns to clause (ii), sometimes called the *severity criterion SC* (that so good a fit be very improbable, were the hypothesis to be inferred false).[5] In a statistical setting the probability in (ii) has to be calculable 'under the assumption' of some statistical hypothesis, such as the 'null hypothesis' of a test, abbreviated as H_0. Here, a fit with hypothesis H typically comes from a disagreement with (or 'rejection' of) the null hypothesis H_0. That is, H_0 is H's denial. There is a *test statistic* D that measures distance between **x** and H_0 in the direction of the alternative H of interest. In the more common, informal setting, requirement (ii) is assessed qualitatively.[6] Accordingly, a severity assessment may be quantitative or qualitative. In either case, the probability in (ii) is understood as in frequentist *error statistics* or (more formally) *sampling theory*.

In null hypothesis testing, the probability in (ii) refers to the probability (or 'sampling') distribution of D calculated under a hypothesis H_0 (representing a denial of H). For example, if H_0 is taken as the (null) hypothesis:

H_0: the mean light deflection is .87″,

and the observed deflection exceeds this by 2 standard deviations (i.e., D = 2-s.d.), a test might infer that this is evidence of a genuine discrepancy from H_0, and evidence for

H: the mean light deflection exceeds .87″.

For a detailed discussion see Mayo ([1996]). Since observing a 2-s.d. difference (or larger) from the true mean (whatever it is) occurs very infrequently (approximately 3% of the time), such an outcome is often said to differ 'statistically significantly' from H_0 (at the .03 level).[7] To infer H, 'there is a genuine discrepancy from the hypothesized mean, H_0' when in fact H_0 is true, is to commit an error (of type 1). If one adopts a rule to infer a discrepancy from H_0 just in the event that D $\geq$ 2-s.d., one ensures the probability of erroneously doing so is approximately .03. This is an *error probability*.[8] This test rule would also be said to correspond to the *observed significance level* (or *p-value*) of .03. The severity associated with the inference is .97 (i.e., 1 – .03), so long as the

although we must keep in mind that some cases may not even ensure good fits in the weakest sense.

[5] Both (i) and (ii) are part of the severity requirement, but using SC this way allows relating our discussion most simply to that of Hitchcock and Sober.

[6] The probability is not a conditional probability where prior probability assignments to the hypotheses are required. P(. ; H) is always the probability of the event in question 'calculated under the assumption' that H is correct, or incorrect. Note that P(A;H) = 1 – P(not-A;H) for an event A. For detailed qualifications, see Mayo 2006.

[7] .03 is the *p*-value associated with the observed difference D. That is, P(D > 2-s.d.; H_0) = .03. I approximate throughout with 2 s.d. rather than 1.96 s.d.

[8] A common fallacy to avoid is regarding .03 as the probability of H_0 given the data (Prosecutor's Fallacy). Rather, P_{H0} (D > 2-s.d.) = .03.

statistical model assumptions are approximately satisfied (Mayo and Spanos [2004]).

All this pertains to ordinary significance testing, without use-constructions or double-counting. In the context of use-constructing, the probability in (ii) must take into account the use-construction rule, say R, because the relevant error probability may be altered.

(4) *Computed versus 'actual' error probabilities*. It will be useful to take advantage of the statistical terms often used when use-constructing alters the observed significance level: we say the significance level that would be computed were the construction rule ignored—i.e., the 'computed' or 'nominal' significance level—differs from the 'actual' significance level, taking the construction rule into account. This statistical point corresponds to the more familiar fact that certain ways of constructing hypotheses so as to fit observed data **x** can decrease the probativeness of the test, and thereby decrease the warrant that accrues to a hypothesis H by dint of its fitting **x**. *The construction rule may restrict the space of hypotheses that could be inferred to those that fit* **x** *in such a way as to make it easier (more probable) to infer a use-constructed H, even if H is false.* If so, then this must be taken account of in appraising the test's capability of avoiding or alerting us to a change in the stringency or severity of the analysis.

1.3 Point 3: Evaluate severity of a test T by its associated construction rule R

From the start, it was emphasized that in cases of use-constructions, severity must be evaluated by considering the properties of the construction rules. The construction procedure may in some cases be sufficiently *stringent* so that it is highly improbable, or even in some cases impossible, to output a false hypothesis. The construction rule R determines the actual hurdle or 'test criterion' that must be met in order that $H(\mathbf{x})$ is arrived at and inferred. I defined:

> *A Stringent Construction Rule (R-*α*)*: the probability is very small, α, that rule R would output $H(\mathbf{x})$ unless $H(\mathbf{x})$ were true or approximately true of the procedure generating data **x** ([1996], p. 276).

The definition of severity does not change, but the way to compute it does. Since the key feature of statistical theory based on error probabilities (*error statistics*) is the insistence that to evaluate a particular inference requires considering the procedures by which the data and hypotheses were generated, error statistics gives a natural backdrop for clarifying assessments in our context. In order to sidestep the central confusions in this arena, we avail ourselves of familiar notational distinctions between a general use-construction rule R(**X**),

X a random variable, a generic output $H(\mathbf{x})$, and a particular fixed hypothesis $H(\mathbf{x}_0)$.[9] Let us illustrate.

1.3.1 Ordinary confidence interval estimation

Ordinary confidence interval estimation provides examples where double-counting need not preclude severity, and offers insight for clarifying several thorny issues. Consider the vector of n random variables $\mathbf{X} = (X_1, \ldots, X_n)$, with each X_i Normal $(N(\mu,\sigma^2))$, Independent and Identically Distributed (IID), and, for simplicity, assume that the standard deviation is known to be σ. (Bold **x** indicates that it is a vector.) A 95% confidence interval estimation rule is based on the sampling distribution of a statistic $\bar{X}$, the sample mean, with standard deviation $\sigma_{\mathbf{x}} = (\sigma/\sqrt{n})$. It may be seen to output inferences of the form

$$H(\mathbf{x}): \ (\bar{x} - 2\sigma_{\mathbf{x}} \ \leq \ \mu \ < \ \bar{x} + 2\sigma_{\mathbf{x}})$$

where $\bar{x}$ is the observed sample mean. $(\bar{x} - 2\sigma_{\mathbf{x}})$ is the generic lower confidence interval bound, μ_{lower}, and $\bar{x} + 2\sigma_{\mathbf{x}}$ is the generic upper bound, μ_{upper}. A *particular* inferred estimate $H(\mathbf{x}_0)$ results from substituting an actual observed mean in $H(\mathbf{x})$, and asserts:

$$H(\mathbf{x}_0): \ \mu \text{ is in the interval } [\mu_{\text{lower}}, \ \mu_{\text{upper}}]. \tag{1}$$

Another construal might be 'exclude μ values outside the interval'. $H(\mathbf{x}_0)$ may be said to have passed with severity approximately .95, assuming, as always, that the assumptions of the statistical model are met. That is because the sample mean differs from *its* true mean, whatever it is, by more than 2-s.d. only 5% of the time.[10] We may write this:

$$\text{P}((\bar{X} - 2\sigma_{\mathbf{x}} \ \leq \ \mu \ < \ \bar{X} + 2\sigma_{\mathbf{x}}); \ \mu) = .95,$$

Or, equivalently, one may write

$$P_{\mu}(\bar{X} - 2\sigma_{\mathbf{x}} \leq \mu \ < \ \bar{X} + 2\sigma_{\mathbf{x}}) = .95.$$

What is very special about confidence interval procedures, and other procedures like them, is that we can make these probabilistic claims without knowing the true mean. One can equivalently infer (about the confidence interval construction rule R):

$$\text{P}(\text{R}(\mathbf{X}) \text{ outputs } H(\mathbf{x}); \ H(\mathbf{x}) \text{ is false}) = .05 \tag{2}$$

Now, post-data, one has a particular interval $H(\mathbf{x}_0)$: $[\mu_{\text{lower}}, \mu_{\text{upper}}]$. '$H(\mathbf{x}_0)$ is false' asserts μ is *not* in $[\mu_{\text{lower}}, \mu_{\text{upper}}]$. For example, a mean $\mu' = \mu_{\text{upper}} + \text{k}$

[9] Philosophers are more familiar with logical variables than random variables. $H(\mathbf{x})$ might be compared to a propositional function. 'Hypothesis function' $H(\mathbf{x})$ becomes a specific hypothesis only by observing $\mathbf{x}_0$ and following the rule R for selecting or constructing $H(\mathbf{x}_0)$.

[10] The inferred interval is $(\mu_{\text{lower}} \leq \mu < \mu_{\text{upper}})$. Because this confidence interval is two-sided, one adds the two .025 (approximate) error probabilities calculated in the example of Note 7.

(for a positive k) renders $H(\mathbf{x}_0)$ false. One is free to ask probabilistic questions post-data, though it seems more natural to include 'would' in that case.

$$\text{P(R(}\mathbf{X}\text{) would output } H(\mathbf{x}_0);\ H(\mathbf{x}_0) \text{ is false)} \leq .05 \tag{3}$$

assuming the statistical assumptions hold. One may also wish to calculate (3) for particular values of μ outside the given interval, say for $\mu = \mu'$. What is the probability that the falsity of $H(\mathbf{x}_0)$ would have led to an interval (produced by procedure R) excluding μ'? No more than .05. Equivalently,

$$\text{P(R(}\mathbf{X}\text{) would yield an interval including } \mu';\ \mu = \mu') = .95 \tag{4}$$

It is useful to write (4) subscripting the hypothesis 'under which' the calculation is made:

$$P_{\mu=\mu'}(\text{R}(\mathbf{X}) \text{ would yield an interval } \textit{including } \mu') = .95$$

Since, with high probability, rule R(**X**) would have signaled the falsity of $H(\mathbf{x}_0)$ by leading to the construction of an interval including μ', rather than $H(\mathbf{x}_0)$, we reason that the observed $\mathbf{x_0}$ is good evidence against μ'. Numerous other probabilistic assertions could be generated along these lines, referring either to test T or equivalently to rule R; the above examples are merely to illustrate.

Admittedly, these assertions can be extremely slippery, and in fact they have been the subject of famous misinterpretations, some of which continue to this day. (As a full discussion of confidence intervals is inappropriate here, the reader is referred to other sources, e.g., Cox and Hinkley [1974], chap. 7). Most notably, the confidence level .95 is not the probability a specific interval estimate is true: the particular interval $H(\mathbf{x}_0)$ will either be true or false, assuming, as in the classical confidence interval context, that μ is regarded as a fixed parameter. Any probabilistic claims apply to the inference method or test rule, just as with significance testing. We can, however, say that the particular inference has passed with severity .95, because *severity always alludes to the properties of the test or inference method* (as do the more familiar confidence levels).

In ordinary confidence interval estimation, then, the use-constructing of the inferred interval does not prevent high severity (as given by the confidence level). Contrast this now with the following interval using an *optional stopping* rule.

1.3.2 Optional stopping rule' R* in confidence interval estimation

In a classic example, the procedure involves a stopping rule that would continue to collect data until some value of μ, say 0, was excluded from the confidence interval. (It can be proved that in sampling from a normal distribution such a procedure would end with probability 1.) This is sometimes described as a procedure of 'trying and trying again' to achieve a result sufficiently far from 0 so as to report a nonzero effect. Since 0 would be excluded from any interval

that R* outputs, the usual confidence interval reasoning, as we just spelled out, would entitle inferring:

$$H(\mathbf{x}_0): \ \mu \text{ is not } 0.$$

However, with the optional stopping procedure R*, there is a high probability that such an inference is in error:

$$P_{\mu=0}(\text{R}^*(\mathbf{x}) \text{ excludes } 0) \text{ is high.}^{11} \quad (5)$$

How high depends on when it ends. The previous assurance of .05 error probability is clearly vitiated, and unless the severity is adjusted, the inference is misleading. Equivalently, we might say, the 'computed' confidence level is .95 but the 'actual' confidence level—and thus the severity associated with the inference—is lower. For a discussion of optional stopping and its influence on severity, see Mayo ([1996], chap. 10), Mayo and Kruse ([2001]).

1.4 Point 4: The ease of passing vs. ease of erroneous passing: *Statistical vs. 'Definitional' probability*

We have seen that ordinary procedures like confidence interval estimation (without optional stopping) involve double-counting and yet are perfectly reliable. How then to explain the common deprecation, if not prohibition, of all double-counting? In an attempt to diagnose the source of the problem, I suggest (in Mayo [1996]) it might be that two parallel questions are being confused, in assessing a use-construction procedure:

(a) What is the 'probability' that a use-constructed procedure passes (infers, outputs) some hypothesis or other?

(b) What is the probability that a use-constructed procedure passes (infers, outputs) some hypothesis or other, even if *this* or *those* (inferred) hypotheses are false?

The assurance of some fit or other in a double-counting procedure could (rightly) lead one to consider that the answer to (a) is high or even one. However, it is only a high value in answering (b) that is problematic. Contrasting (a) and (b) helps to highlight the error in a tendency to slide from answering two very different questions, and that is why I introduced them. But perhaps I did not say enough to draw a radical distinction between the use of 'probability' in the two. I want to remedy this.

First, consider (a). The 'assurance' of some fit or other in the double-counting procedure is just a report of what is typically meant by a successful application

[11] This may be read: 'the probability that rule R* leads to a confidence interval that excludes $\mu = 0$, even though, in fact, $\mu = 0$, is high'.

of the procedure.[12] In other words, *by definition*, applying the (use-construction) procedure involves outputting some hypothesis or other as 'passing' the test. Whether it is through deliberately building a hypothesis to fit **x**, or searching through a basket of hypotheses for one that fits **x**, or suitably altering the hurdle to count as a fit, or any of the other tactics available, some hypothesis that fits **x** will result (else we say the procedure has not been 'applied'). Thus, the answer to (a) would seem to be 1. Because the 'assurance' in (a), when we have it, is just a matter of the definition of a use-construction procedure, it may be called 'definitional' probability to distinguish it from (b).[13] When it comes to assessing a procedure's stringency or reliability—in answering question (b)—the concern is whether the observed fit could plausibly have come about *even though the hypothesis in question is false*. *Construction rule R, successfully applied, may be guaranteed to pass some* $H(\mathbf{x})$ *or other, but it may be rare for R to output false hypotheses.* At least this is so for stringent use-constructing procedures, as with ordinary confidence intervals. In those cases the probability in (b) would be low, and severity would be *satisfied* (the severity would be 1 – the probability in (b)). In such cases, there is no basis for 'discounting' the warrant for a use-constructed $H(\mathbf{x})$.

2 The False Dilemma: Hitchcock and Sober

2.1 Marsha measures her desk reliably

We are prepared now to quickly get to the bottom of the problem that leads Hitchcock and Sober ([2004]) to deny that the severity criterion distinguishes problematic from unproblematic cases of double-counting. The example Hitchcock and Sober (p. 24) give in raising their criticism is a case of simple measurement: to some extent, it follows the pattern of reliable interval estimation. In construction rule R, Marsha uses a tape measure to arrive at and infer a claim about the length of her desk. The inferred claim is of form $H(\mathbf{x})$: the length of the table is $\mu = \mathbf{x} \pm 1$ cm.

> Here, [R], is the procedure of stretching the tape measure along the side of the desk and noting which number accompanies the slash that is closest to the edge of the desk; [$\mathbf{x}_0$] is the result that the closest slash is adjacent to the number 150; and [$H(\mathbf{x}_0)$] is the claim that the desk is between 149 and 151 centimeters wide. Assume, moreover, that Marsha is very reliable in her use of tape measures; it is very unlikely that her measurement will be off by more than one centimeter. It is clear from Mayo's critique of strong predictivism that she would deem this to be a case in which [$H(\mathbf{x}_0)$] has passed a severe test with [$\mathbf{x}_0$]. (Hitchcock and Sober [2004], p. 24)

[12] It is certainly not part of the definition of severity that the probability in (a) equals 1.

[13] Not all use-construction cases are guaranteed to end, but one could always define 'applying the procedure' as arriving at an $H(\mathbf{x})$.

However, it seems to them that the severity criterion for such cases is not satisfied (although it ought to be). It is not satisfied, they claim, because 'Regardless of the length of the desk, we can be almost certain that Marsha will postulate a desk length that fits the result of her measurement as well as $H(\mathbf{x}_0)$ fits the actual result $[\mathbf{x}_0]$'. (I replace D with $\mathbf{x}_0$, T with $H(\mathbf{x}_0)$ for consistency with my notation.)

Now it is true that her procedure R will output some hypothesized length or other, assuming it is applied successfully; nevertheless, it very rarely outputs false hypotheses, by the stipulations of their own example. Therefore, any application of this use-construction rule yields an inferred length that passes with high severity. Why then do Hitchcock and Sober claim that taking account of R results in 0 severity? The reason is that their construal slips into the 'definitional' probability that gives 0 severity to all use-constructed H (i.e., the answer to question (a) is maximally high, but only a high value to the probability in (b) entails low severity). More specifically, it is the result of their characterization of what they call the 'non-rigid' construal of my severity criterion for use-constructed cases:

> SC_{nr}: There is a very low probability that test procedure T would yield so good a fit with whatever hypothesis would have been proposed, if H is false (for a 'rigid' H). (Hitchcock and Sober, [2004], p. 23—my notation).

In SC_{nr}, they explain, the occurrence of H in 'if H is false' still refers rigidly (p. 23), so their non-rigid construal asserts:

> SC_{nr}: There is a very low probability that test procedure T would yield so good a fit with whatever hypothesis $H(\mathbf{x})$ would have been proposed, if $H(\mathbf{x}_0)$ is false.

Note the use of $\mathbf{x}$ and $\mathbf{x}_0$ in the first and second instances, respectively. SC_{nr} comes out false for Marsha's measurement because *by definition* her measurement procedure always outputs some inferred length. But this is just to say the probability in (a) is maximally high, and this does not entail low, much less 0, severity. Thus, to construe severity as in their SC_{nr} is to commit the fallacy that leads to always prohibiting use-constructions as yielding 0 severity! It is not a statistical claim calculated under the hypothesis that '$H(\mathbf{x}_0)$ is false'. Although their SC_{nr} includes the clause 'if $H(\mathbf{x}_0)$ is false', the problem is that, as fixed, it is doing no work. Just consider an imaginary dialogue between Marsha and Bill:

> *Marsha*: 'I infer from my reliable measurement procedure that my desk is approximately 150 centimeters.'

Here, the true but unknown desk length is playing the role of the unknown parameter μ, and 150 is the observed value $\mathbf{x}_0$. Marsha's inference is: infer $\mu = 150 \pm 1$ cm. For simplicity, write this as: infer $H(150)$.

> *Bill*: 'Well if your desk length μ were some value other than 150 centimeters, if it had been, say, 200 centimeters, then you still would have inferred some length estimate, presumably approximately $H(200)$ (i.e., $\mu = 200 \pm 1$ cm). This counts against your inference to $H(150)$!'

This just makes no sense. That she probably would have reported $H(200)$ were the desk 200, and not 150, centimeters is just what we would want! (More precisely, with high probability she would have observed an $\mathbf{x}$ value that led her to infer $H(200)$.)[14]

It should now be clear what has gone wrong: Assessing severity using their SC_{nr} leads to use-constructed cases being given minimal severity—even if the construction rule R would never output a false hypothesis. This was the very flaw I sought to avoid. The correct application of the severity criterion to hypotheses resulting from rule R—clause (ii) of the definition of severity—is:

> *SEV requirement*: For any sample $\mathbf{X}$, there is a very low probability that test procedure T, with construction rule R, would infer $H(\mathbf{x})$, if $H(\mathbf{x})$ is false.[15]

(Compare this to the confidence interval assertion (4)). Since this is satisfied in Marsha's case, the severity criterion is satisfied. One could also instantiate into $H(\mathbf{x})$ and obtain a claim that is true for Marsha's particular measurement with $\mathbf{x}_0$:

> SEV (instantiating): There is a very low probability that test procedure T, with construction rule R, would infer $H(\mathbf{x}_0)$, if $H(\mathbf{x}_0)$ is false.

What one *cannot* do and still expect a statistically grammatical claim is mix both generic $H(\mathbf{x})$ and specific $H(\mathbf{x}_0)$ in the same claim as Hitchcock and Sober do in SC_{nr}.

2.2 A false dilemma

Finding (correctly) that their non-rigid severity criterion SC_{nr} does not give the plausible answer that Marsha's use-construction is reliable, Hitchcock and Sober infer I am left with the only other construal they offer—they call it the 'rigid' version of the severity criterion SC_r.

[14] Admittedly, these assertions demand a careful understanding of confidence intervals which I do not pretend to here supply. I hope only to convince philosophers that it is worth attaining such an understanding. See references.

[15] One may replace 'infer $H(\mathbf{x})$' with 'infer some $H(\mathbf{x})$ or other' if one prefers. This is already conveyed by the generic $\mathbf{x}$.

> SC_r: There is a very low probability that test procedure T would yield so good a fit with $H(\mathbf{x}_0)$, if $H(\mathbf{x}_0)$ is false.

They define the rigid severity criterion SC_r as the appraisal that would ensue were *H not* the result of constructing or selecting; there is no place in SC_r to take the construction rule R into account. In effect, it refers to the 'computed' error probability. Now SC_r happens to get the right answer when applied to Marsha's measurement just because *in this case* use-constructing does not alter severity ('computed' equals 'actual'). But they assume that the price of getting it right for Marsha means the error statistician must always use SC_r and thus must ignore the construction rule R. This is to ignore the fundamental piece of information required for a proper (error statistical) assessment of severity: namely, taking into account the procedure of hypothesis construction or selection (and adjusting error probabilities when it is warranted to do so).

Let me be clear: It is not the wording of SC_r that is so problematic. An error statistician could well construe this as merely an equivocal statement, directing them to consider, as part of the test procedure T, the properties of the construction rule R that led to outputting $H(\mathbf{x}_0)$, and thereby wind up evaluating severity correctly as in SEV above. (I will come back to the error statistician's construal in Section 3.3.) But this is not Hitchcock and Sober's construal: they intend SC_r literally, as giving no role to the construction or selection rule R that led to output $H(\mathbf{x}_0)$. Unsurprisingly, they discover that SC_r does not take account of double-counting when we want to take it into account! (See the example in Section 3.) SC_r lacks the very niche that an appraisal of test T must include if it is to be sensitive to construction and selection effects.

The upshot is that Hitchcock and Sober present us with a false dilemma: a choice between a criterion that would always regard use-constructed inferences as violating severity (SC_{nr}), and one where use-constructing can never alter the severity assessment (SC_r) but treat the case just as if the use-constructed *H* were not use-constructed. Clearly, this was not their intention; it is the consequence of misconstruing the severity criterion developed in the error statistical approach, as it applies to our context of double-counting. Granted, it can be tricky to work with terms such as 'hypotheses that could have been outputted by a construction procedure R'. Such considerations, however, are standard fare in frequentist error statistics, and this entire discussion could be greatly simplified by appealing to this apparatus. I do not mean to suggest that there are no controversial cases regarding double-counting in formal error statistics. There certainly are, and there is considerable scope for philosophers to help disentangle when and why double-counting and selection should matter in practice. My goal here will be to minimize the technical apparatus while setting the groundwork for fostering much-needed attention by philosophers

of science. By explicating some of the more extreme cases, the goal is to set the stage for philosophical scrutiny into the cases still under dispute in practice.

3 Canonical Errors of Inference

3.1 How construction rules may alter the error-probing performance of tests

The most straightforward and least circuitous way to evaluate the stringency of the use-construction rule is first to identify a cluster of cases where use-constructing and selection effects may enter; and second to consider the kind of error or flaw that may arise to prevent the given type of inference from being well-warranted.[16] The following list will serve our current purposes; more could well be added.

The data **x** may be used in selecting or constructing hypotheses (generally with respect to some model) when:

1. Estimating a parameter; measuring a quantity,
2. Inferring an aspect of the source or cause of a known effect,
3. Accounting for a result that is anomalous for some theory, claim, or model H (e.g., by means of an auxiliary $\mathrm{A}(\mathbf{x})$, see 3.2),
4. Inferring the existence (or non-existence) of genuine effects, e.g., statistically significant differences, regularities,
5. Testing the validity of model assumptions: e.g., IID in statistical models.

Whether the test is in a formally modeled context or not, we can identify central errors and threats to validity that arise in constructing and making inferences about the corresponding $H(\mathbf{x})$ in each case. For instance, in Marsha's case the error of concern is reporting a length that differs from the actual length μ (beyond the margins of reported error).

3.2 Rules for accounting for anomalies

Consider #3, sometimes called 'exception incorporation' and is often the butt of criticisms by those who eschew double-counting. We all know of classic cases where pet theories and models are spared from disconfirmation by using the data to invent 'just so stories' and the like. The concern is that this will be done even if the hypothesis or theory is false and was not probed in the slightest by the test in question. Use-constructing can increase the latitude for such

[16] More generally, we would also need to consider whether it is a behavioristic context or one of scientific inference, but here we are focusing on just the latter. There are cases where the construction rule alters the assessment for the former context and not the latter.

flaws. Suppose H is being tested and data $\mathbf{x}$ are anomalous for H.[17] Let rule R′ account for any anomaly by constructing or selecting some auxiliary hypothesis A($\mathbf{x}$) that allows one to restore consistency with data $\mathbf{x}$ while retaining H. We may further suppose that ample auxiliaries are logically available. Applying R′ yields outputs of form

$$H(\mathbf{x}): \ H \ \& \ \mathrm{A}(\mathbf{x}).$$

Now any rule R delimits the kinds of hypotheses that it outputs, and we need to consider the range of outputs to evaluate severity. In this case, for each possible outcome $\mathbf{x}'$, $\mathbf{x}''$, etc. rule R′ yields

$$H'(\mathbf{x}'): \ H \ \& \ \mathrm{A}'(\mathbf{x}')$$
$$H''(\mathbf{x}''): \ H \ \& \ \mathrm{A}''(\mathbf{x}'')$$
$$H'''(\mathbf{x}'''): H \ \& \ \mathrm{A}''' \ (\mathbf{x}'''), \ldots$$

etc.

The corresponding test, T, confronted with an anomaly for H_i, applies R′ and outputs the inference $H_i(\mathbf{x}_i)$, as warranted by the data. Since, the space of outputs never includes not-H, the test fails to probe H. Now suppose R′ is applied and the particular dataset $\mathbf{x_0}$ yields a specific inference:

$$H(\mathbf{x}_0): \ H \ \& \ A(\mathbf{x}_0).$$

Since, test T with rule R′ scarcely guards against the threat of erroneously retaining H, we would say, *of this particular* $H(\mathbf{x}_0)$, that the observed fit between $H(\mathbf{x}_0)$ and data $\mathbf{x}_0$ is not good evidence for the truth of $H(\mathbf{x}_0)$ because

$$\mathrm{P}_{H(\mathbf{x}) \text{ is false}}(\text{Test T, with rule R, fits } H(\mathbf{x})) = \text{high, if not maximal}$$

(where $H(\mathbf{x})$ alludes to some saving conjunction or other). Thus $H(\mathbf{x}_0)$ fails to satisfy the severity criterion for a use-constructed hypothesis in this context, and this is shown by referring to the general procedure R′ that constructed the *particular* $H(\mathbf{x}_0)$.

It is much simpler, I think, to focus on the error of concern—erroneously saving H from anomaly, and the associated error probability. Does rule R′ work hard to avoid this flaw? Scarcely. The error probability with Rule R′ is high or maximal. The corresponding test lacks stringency; so we are unable to warrant $H(\mathbf{x_0})$ with severity, at least by dint of the properties of this test. Of course, a defender of the particular inference may show, or try to show, that she has answered the threat in the case at hand by some *other* means. Having brought

[17] Here it is typical for H, which we may imagine is the currently accepted hypothesis, to play the role of the null hypothesis. Anomalies for H 'fit' various rival claims, and the auxiliaries A are use-constructed in order to retain H.

out the threat to severity that any defender needs to get around serves a valuable role: it points to what an improved argument or experiment or analysis would need to accomplish.

In principle, the 'same' $H(\mathbf{x}_0)$ outputted by the insevere rule R′ could have been inferred on the basis of some other rule that was highly stringent. But, at least for the error statistical way of thinking, the inference to $H(\mathbf{x}_0)$ must be evaluated in relation to the construction rule actually used. It follows that we cannot evaluate how well warranted a particular $H(\mathbf{x}_0)$ is without considering the properties of the rule by which it was constructed: our appraisal of this apparently 'same' hypothesis need not be the same. This is in marked contrast to evidential appraisals that take into account only how well evidence $\mathbf{x}_0$ 'fits' a hypothesis H, for any measure of fit.

Two further points: First, the above rule R′ is certainly not the only pattern for using data to save theories. My purpose here is only to explicate the severity criterion in use-construction cases, and get beyond some of the flawed interpretations. To this end, I deliberately consider some extreme cases. If we cannot make out the needed distinctions with the clear-cut cases, we will be hard pressed to do so for the more equivocal cases. Second, there is no suggestion that one can always calculate the severity associated with a use-construction rule, although the inability to do so might itself be grounds to question it: the onus is on the tester to show the hypothesis in question has been well tested. As a final illustration, consider one of the most classic cases where 'selection effects' demand altering error probabilities, and thus severity.

3.3 Hunting for statistically significant differences

An exemplary class of cases where selection effects alter the capacity of a test to control erroneous inferences is when inferring effects are genuine and not 'due to chance' (#4 in the list). Suppose that in comparing randomly selected samples of adolescents from groups who had and had not been breast-fed as infants, 20 different null hypotheses $H_{0,i}$, $i = 1, 2, \ldots, 20$ are tested, each asserting 0-difference on some factor, e.g., incidence of cancer before the age of 15, high blood pressure, juvenile diabetes, multiple sclerosis, heart disease, maturity levels, psychological problems, and so on. Suppose that only the smallest p-value attained is reported, let us say it is approximately .05, and that all of the other nineteen insignificant results are ignored.[18] Imagine that

[18] The null hypothesis H_0 asserts there is 'no effect' (in the population of interest) on the given factor (e.g., psychological problems) associated with the two types of infant feeding. In familiar tests, to find evidence against H_0 is to find evidence for alternative H_1: there is a genuine effect (in one or both directions). The test statistic D might be the observed difference between the effect rates (in treated and untreated groups) and 0—since 0 is predicted under H_0. P_{H0} ($D > D_{obs}$) = the p-value associated with D_{obs}. It is also called the observed significance level or significance probability. (See Mayo and Cox [2006].)

the one null hypothesis that may thereby be rejected is $H_{0,13}$ (e.g., there is no difference in the incidence of psychological problems in breast-fed versus bottle-fed children), and it is inferred that there is a difference on this factor. That is, rule R—'hunting for significance'—derives from the observed correlation an inference about the general population of breast- versus bottle-fed adolescents, for example:

> $H_{1,13}(\mathbf{x}_0)$: there is a lower incidence of psychological problems in breast-fed versus bottle-fed children.

The resulting test procedure is very different from the case in which $H_{0,13}$ is preset as the single hypothesis to test. In the predesignated case, the possible outcomes are the different significance levels attained by this one factor; in the case of 'hunting', by contrast, the possible results are the possible statistically significant factors that might be found to show a 'nominally' significant departure from the null hypothesis. Hence, the type 1 error probability is the probability of finding at least one such significant difference out of 20, even though the global null is true (i.e., all 20 observed differences are due to chance). The probability that this procedure yields erroneous rejections differs from, and will be much greater than, .05 (and is approximately .64). There are different, and indeed many more, ways one can err in this example than when one null is prespecified, and this is reflected in the adjusted p-value.[19] The severity associated with the hypothesis $H_{1,13}$ that the correlation is real is correspondingly low (about .36). Notice we are evaluating the warrant for this one hypothesis $H_{1,13}$, but we do so by considering the overall probability of erroneously rejecting the chance explanation. By contrast, consider appraising severity by Hitchcock and Sober's 'rigid' rule, SC_r. As they point out, this would treat the inference from the hunting expedition just as if $H_{0,13}$ had been predesignated and fixed from the start. This would thereby fail to properly discount the diminished force of the evidence due to hunting. This conflicts with the goal of the severity criterion I set out. The correct construal of the severity criterion, we have seen, differs from both their rigid and non-rigid formulations.

3.3.1 An anticipated criticism

A critic may still persist that I am using a different definition of severity when $H_{1,13}$ is inferred from a hunting procedure, and when it is inferred by rejecting the single prespecified null $H_{0,13}$. It is true that the severity associated with

[19] Another way to capture this type of hunting procedure is to regard the test statistic being used as the $\min(p_1, p_2, \ldots, p_{20})$, the smallest 'nominal' p-value of the 20 null hypotheses considered (Cox and Hinkley [1974]; Cox [2006]). The probability that the smallest of 20 p-values is .05 can be as high as .65, even if all 20 nulls are true, and one should adjust the actual p-value associated with the inference accordingly. An adjustment that may be used in this type of case is the Bonferroni adjustment.

an inference in the former case must take account of the possible impact of double-counting on the relevant error probability, but, as I have been trying to drive home, the error statistician must *always* take account of factors that alter error probabilities—double-counting is not the only one. One could consider each factor as associated with a different definition, but it would be more correct to simply say: in assessing severity, calculate your error probabilities correctly, and do not confuse 'computed' error probabilities with 'actual' ones.

But the critic may continue to express perplexity. If the severity definition remains constant, she may ask, then how can we be assessing the severity associated with the *one hypothesis* $H_{1,13}$ if it is computed differently in the cases of hunting for significance (where the global null is considered), and a prespecified null hypothesis, respectively? The reply is that the ability of the former test to have avoided the relevant error as regards this (one and the same) hypothesis differs from that of the latter. Granted, one may reject a concern with error probabilities and severity, but that is besides the point just now.

Error-statistical calculations *always* involve considering outcomes other than the one actually attained (whether in using significance levels, confidence levels, or other). In use-construction cases this means considering the corresponding *other hypotheses* that could have been constructed or selected (or other stopping points in the case of stopping rules).[20] We need to look at the range of hypotheses that could be outputted by R, in order to assess the well-testedness of the one hypothesis R happened to have outputted. The concern, at least in the context of scientific inference, is not merely to avoid often announcing genuine effects erroneously in a long-run series[21], the concern is that *this* test performs poorly as a tool for discriminating genuine from chance effects in this particular case. The hypothetical error rates teach us about the test's capacities in the case at hand. Because at least one such impressive departure is common even if all are due to chance, the 'hunting expedition' has scarcely reassured us that it has done a good job of avoiding such a mistake in *this* case. The .05 'computed' *p*-value is invalidated when it comes to the 'actual' value. Even if there are other grounds for believing the genuineness of the one effect that is found, we deny that *this test alone* has supplied such evidence.[22]

[20] If the experimenter searches only for differences in the direction of favoring breast-feeding, yielding a 'one-sided' test, further corrections to the error probability are required.

[21] This would be the concern in a purely behavioristic context. This seems to be the case for multiple testing in microarrays.

[22] Note that if the design is to preset the *p*-value, say to .05, it can happen that none of the 20 factors is selected for testing. One might not be able to determine the probability that at least one is found. Although nothing in our discussion turns on this, if it was thought desirable, the experimenter could define a successful application of this procedure to be limited to cases where

4 Concluding Remarks

The issue of double-counting, use-constructing, and selection effects has long been the subject of debate in the philosophical as well as statistical literature. Rather than taking the position that double-counting always or never matters, we need to consider just when and how it may alter the capacity of the test as a tool for uncovering and avoiding erroneous inferences. I have argued that the severity requirement, properly construed, gives us a platform for judging when we should take double-counting into account, by 'discounting' the evidential weight of an observed data-hypothesis fit. The difficulty Hitchcock and Sober ([2004]) raise, I showed, stems from a flawed interpretation of the severity criterion. Limiting the choice to their 'non-rigid' and 'rigid' formulations, I have argued, leads to a false dilemma—either double-counting always leads to minimally severe tests or else it can have no influence at all on a severity assessment. Well-formed statistical assertions steer us clear from both positions.

A severity assessment concerns a statistical relationship between an event—that test T outputs a fit with H—and a claim that 'H is false' about the underlying data-generating mechanism. We hypothetically consider 'H is false' for purposes of evaluating the stringency, probativeness, or error-detecting capacity of our test. In judging that H passes severely, it is *because* H's falsity would make it so improbable, surprising, or extraordinary to have gotten so good a fit with H. In judging that H does not pass severely, it is *because* the falsity of H fails to adequately constrain the procedure so that it (very probably) would have alerted us to H's falsity (by producing a result discordant with H). Such an *in*severe test procedure fails to provide grounds that the error of concern is being avoided in the particular case. The severity appraisal always depends on features of both the data- and hypotheses-generation procedures. In particular, once the construction rule is applied and a particular $H(\mathbf{x}_0)$ is in front of us, we evaluate the severity with which $H(\mathbf{x}_0)$ has passed by considering the stringency of the rule R by which it was constructed, taking account of the particular data achieved.

As I have emphasized, *the severity criterion remains fixed and does not change*; what changes is how to apply it. One can classify types of applications, if one keeps in mind that the overarching goal is to warrant an inference only to the extent that the errors of interest have been adequately ruled out (as assessed statistically). What matters is not whether H was deliberately constructed to accommodate data $\mathbf{x}$, what matters is how well the data, together with background information, rule out ways in which an inference to H can be in error. The focus in this paper was on some extreme cases because the purpose was

at least one is found; in other cases, one might imagine, the entire study is shelved into a file drawer, leading to what is known as the 'file-drawer' problem.

to get beyond some of the flawed interpretations involving them. A next step would be to take full advantage of the concepts of sampling distributions, statistics, and random variables to get beyond arduous phrasings and ambiguous language. Only then can we determine, for the more complex and interesting cases, just when double-counting creates obstacles to reliable inference.

Acknowledgements

I am grateful to David Cox, Clark Glymour, Aris Spanos, John Worrall, and the anonymous referees for extremely helpful comments on earlier versions of this paper. I would also like to thank C. Hitchcock and E. Sober for useful exchanges concerning the topic of this paper.

Department of Philosophy
Major Williams Hall, Virginia Tech
Blacksburg, VA, USA
mayod@vt.edu

References

Cox, D. R. [2006]: *Principles of Statistical Inference*, Cambridge, UK: Cambridge University Press.

Cox, D. R. and Hinkley, D. V. [1974]: *Theoretical Statistics*, London: Chapman & Hall.

Hitchcock, C. and Sober, E. [2004]: 'Prediction Versus Accommodation and the Risk of Overfitting', *British Journal for the Philosophy of Science*, **55**, pp. 1–34.

Lakatos, I. [1970]: 'Falsification and the Methodology of Scientific Research Programmes', in I. Lakatos and A. Musgrave (*eds*), *Criticism and the Growth of Knowledge*, Cambridge: Cambridge University Press, pp. 91–196.

Mayo, D. G. [1991]: 'Novel Evidence and Severe Tests', *Philosophy of Science*, **58**, 523–52.

Mayo, D. G. [1996]: *Error and the Growth of Experimental Knowledge*, Chicago: The University of Chicago Press.

Mayo, D. G. [2005]: 'Evidence as Passing Severe Tests: Highly Probed vs. Highly Proved', in P. Achinstein (*ed.*), *Scientific Evidence*, Baltimore, MD: Johns Hopkins University Press.

Mayo, D. G. and Cox, D. R. [2006]: 'Frequentist Statistics as a Theory of Inductive Inference', in J. Rojo (*ed.*), *Optimality: The Second Erich L. Lehmann Symposium*, Beachwood, Ohio: Institute of Mathematical Statistics, pp. 77–97.

Mayo, D. G. and Kruse, M. [2001]: 'Principles of Inference and Their Consequences', in D. Cornfield and J. Williamson (*eds*), *Foundations of Bayesianism*, Dordrecht: Kluwer Academic Publishers, pp. 381–403.

Mayo, D. G. and Spanos, A. [2004]: 'Methodology in Practice: Statistical Misspecification Testing', *Philosophy of Science*, **71**, pp. 1007–25.

Mayo, D. G. and Spanos, A. [2006]: 'Severe Testing as a Basic Concept in a Neyman-Pearson Philosophy of Induction', *British Journal for the Philosophy of Science*, **57**, pp. 323–57.

Musgrave, A. [1974]: 'Logical Versus Historical Theories of Confirmation', *British Journal for the Philosophy of Science*, **25**, pp. 1–23.

Popper, K. [1994]: 'The Myth of the Framework', in N. A. Notturno (*ed.*), *Defence of Science and Rationality*, London: Routledge.

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

Worrall, J. [1978]: 'The Ways in Which the Methodology of Scientific Research Programmes Improves on Popper's Methodology', in G. Radnitzky and G. Andersson (*eds*), *Progress and Rationality in Science*, Boston Studies in the Philosophy of Science 58, Dordrecht: D. Reidel, pp. 45–70.

Worrall, J. [1989]: 'Fresnel, Poisson, and the White Spot: The Role of Successful Prediction in the Acceptance of Scientific Theories', in D. Gooding, T. Pinch and S. Schaffer (*eds*), *The Uses of Experiment: Studies in the Natural Sciences*, Cambridge, UK: Cambridge University Press, pp. 135–57.