Naturalism & Objectivity: Methods and Meta-methods

Jean Anne Miller

Dissertation submitted to the faculty of the Virginia Polytechnic Institute and State University in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
In
Science and Technology Studies

Deborah G. Mayo (Chair)
Richard M. Burian
Ellsworth R. Fuhrman
Aris Spanos

August 20, 2008
Blacksburg, Virginia

Keywords: epistemology, error statistics, methods, meta-methods, micro-sociology of science, minimal a priorism, naturalism, new experimentalism, normative naturalism, objectivity, philosophy of experiment, philosophy of science, relativism, social epistemologies, reflexivity, relativism

Naturalism & Objectivity: Methods and Meta-methods

Jean Anne Miller

ABSTRACT

The error statistical account provides a basic account of evidence and inference. Formally, the approach is a re-interpretation of standard frequentist (Fisherian, Neyman-Pearson) statistics. Informally, it gives an account of inductive inference based on arguing from error, an analog of frequentist statistics, which keeps the concept of error probabilities central to the evaluation of inferences and evidence. Error statistical work at present tends to remain distinct from other approaches of naturalism and social epistemology in philosophy of science and, more generally, Science and Technology Studies (STS). My goal is to employ the error statistical program in order to address a number of problems to approaches in philosophy of science, which fall under two broad headings: (1) naturalistic philosophy of science and (2) social epistemology. The naturalistic approaches that I am interested in looking at seek to provide us with an account of scientific and meta-scientific methodologies that will avoid extreme skepticism, relativism and subjectivity and claim to teach us something about scientific inferences and evidence produced by experiments (broadly construed). I argue that these accounts fail to identify a satisfactory program for achieving those goals and; moreover, to the extent that they succeed it is by latching on to the more general principles and arguments from error statistics. In sum, I will apply the basic ideas from error statistics and use them to examine (and improve upon) an area to which they have not yet been applied, namely in assessing and pushing forward these interdisciplinary pursuits involving naturalistic philosophies of science that appeal to cognitive science, psychology, the scientific record and a variety of social epistemologies.

To my parents Edmund P. Miller and Jean C. Waller Miller who raised us to believe that an education was both intrinsically and extrinsically valuable—you were right! And to my brother, Edmund "Bud" Miller who figured this out first and lead the way.

Acknowledgments

I want to thank my committee for their generous help and patience on this project—Dick Burian for his extensive comments on my defense draft; Skip Fuhrman for always asking the questions that made me think beyond the dissertation; and Aris Spanos, who helped me to understand not only the technical side of error statistics and misspecification testing but their philosophical importance as well. Most of all, I want to thank my committee chair, Deborah Mayo, for her invaluable help, encouragement and patience at all stages of this project and for pushing my analytical skills much further than I thought they (or I) could ever go.

I thank my fellow graduate students Emrah Aktunc and Tanya Hall for reading and helping me edit several chapters when they were really rough and my friend and colleague Mary Cato for slugging through an entire early draft of this dissertation. I, of course, take full responsibility for any remaining mistakes.

I want to acknowledge a grant from the National Center for Ecological Analysis and Synthesis (NCEAS: PIs: M. Taper and S. Lele), which enabled me to learn about and think through the Little Rock Lake acidification experiments, and the late Tom Frost, who introduced and helped guide me through them.

Without the patience, support and encouragement from my family and friends, I would not have been able to overcome all the obstacles to completing this project—thank you one and all. I especially want to thank Mary Ellen Jones and Voula Saridakis, who led the way with humor and conviction in completing our STS degrees.

Most of all, I want to thank my best friend, Steven Jacobson, who kept me grounded throughout and always gave me the best advice—both practical and intellectual.

Naturalism & Objectivity: Methods and Meta-methods
Table of Contents

Chapter 1: Normative Naturalism & the New Experimentalism

The central problem of epistemology is to understand the epistemic quality of human cognitive performance, and to specify strategies through whose use human beings can improve their cognitive states. (Kitcher 1992: 74)

A better strategy is to regard naturalism as a set of methodological rules developing a consistent and testable picture of the world. (Giere 2006: 60)

Rules for learning from experiment are empirical and may be evaluated empirically. (Laudan 1987: 27)

The picture corresponding to error statistics is one of an activist learner in the midst of an inquiry with the goal to find something out. (Mayo & Spanos 2010: 19)

1.0 Introduction

I have three goals for this chapter. First, I want to introduce some of the terms and concepts that I will be using as well as sketch in some background to what we can call the 'new naturalism' and 'new experimentalism' in philosophy of science.[1] I will indicate the main tenets of these rather loosely defined schools of thought specifically as they are used to study scientific methodology, and some of the challenges facing them therein. The key feature of what I am calling the "new experimentalism" strand in naturalism is the view that looking to experimental practices holds the key to objectivity for warranting scientific inferences based on empirical data. But how this is to be done is still an open question in philosophy of experiment.

The background discussions here set the stage for the chapters that follow. In them, I will scrutinize several different naturalist approaches to grounding objectivity via experimental practices. I will discuss the lessons that I think can be learned from some of the failures and successes of these accounts both methodologically and reflexively meta-methodologically. A meta-method is simply a method; we can call it a second order method. A meta-method is used

[1] I am following Philip Kitcher (1992) and D. Mayo (1996) who also traced it to Robert Ackermann (1989) respectively in the use of these terms.

to study object level or first order methods (e.g., a philosophical method that is used to study scientific methods would be an example of a meta-method).

1.0.1 Evidence & objectivity: my position. My second and main goal for this chapter, which is woven throughout, is to introduce the position that I endorse, which is Deborah Mayo's philosophy of Error Statistics (ES). Her philosophy embraces a strong concept of objectivity. Fundamentally, objectivity for an error statistician means one's method of inquiry minimally upholds what Mayo has identified as the "weak severity" principle:

> **Weak Severity Principle:** Data $\mathbf{x_0}$ do *not* provide good evidence for hypothesis *H* *if* $\mathbf{x_0}$ result from a test procedure with a very low probability or capacity of having uncovered the falsity of *H* (even if *H* is incorrect). Mayo & Spanos (2010: 21 italics added)

Mayo and Spanos explain the reasoning behind this principle: "We intuitively deny that data $\mathbf{x_0}$ are evidence for *H* if the inferential procedure has very little chance of providing evidence against *H*, even if *H* is false (ibid.)." The basic notion at work here is that for data to be considered evidence for a hypothesis requires more than that it merely fits the hypothesis; the data must also be the result of a test or method that with high probability will detect the falsity of the hypothesis under consideration if it is false (but not otherwise). This notion is the foundation of the error statistical approach to reliable empirical inference.

An important contrast to traditional philosophies of science becomes immediately apparent in that on the ES account, evidence e is never accorded the status of a 'given.' In fact, this account is based on understanding how the procedures and methods used to generate and model data *into* evidence for/against a specific hypothesis provides for (or underlies) claims for scientific objectivity. On the severity interpretation of evidence, evidential status is always context dependent—a function of the specific hypothesis under test, the data generating procedure(s) used, and the actual data produced. This can be encapsulated in the severity

function: SEV(Test T, outcome **X**, inference H) (see Mayo & Spanos 2010). Granted, objectivity is a tough concept to pin down, especially in a way that is not circular, or too weak or simply too negative. (e.g.., defined only negativity solely in contrast to subjectivity).

1.0.2 The strength of weak severity: Notice that weak severity is not that weak. Indeed, it is much stronger than many of the standard notions of objectivity proposed by philosophers including Pierre Duhem and Karl Popper. There are at least two implicit components to such an appraisal, which can be found in Mayo (1983, 1996). Some researchers locate objectivity in the attitude of the researcher. For Duhem, the final choice between hypotheses was determined not by logic or experimental facts but by the physicists innate sense of 'good sense' (bons sens):

> These motives which do not proceed from logic and yet direct our choices, these 'reasons which reason does not know' and which speak to the ample 'mind of finesse' but not to the 'geometric mind,' constitute what is appropriately called good sense. (1954: 217)

Now good sense is not an epistemic trait but a moral virtue for Duhem and could be cultivated by practicing scientists. It would also would be found in differing degrees among them: What is the virtue?

> The sound experimental criticism of a hypothesis is subordinated to certain moral conditions; in order to estimate correctly the agreement of a physical theory with facts, it is not enough to be a god mathematician and skillful experimenter; one must also be an impartial and faithful judge. (ibid. 218)

However, it is humanly impossible to be totally impartial and unbiased, especially as there are compelling reasons to suspect that all humans harbor biases of which they are unconscious. Moreover, as Duhem points out, "it is impossible to leave outside the door, the theory that we wish to test…(182)." I think it would be much better and actually more honest to adhere to the principle of minimal severity and admit the existence of biases and the influence of theory and take active steps to detect, and control or eradicate errors arising from their influence on the

experimental results. Given the definition of weak severity, it should be clear that an appraisal is not merely the result of the whims, desires or idiosyncrasies of the one holding (or denying) the claim, no matter how virtuous. In chapter 4, I will take up this topic more fully when I discuss Helen Longino's method of multiple perspectives which can be seen as an attempt to replace Duhem's moral method of cultivating impartial observers.

Popper rejected Duhem's notion that objectivity lied in the moral judgments or good sense of individual scientists. An objective appraisal should also be inter-subjectively accessible. Popper states that "the objectivity of scientific statements lies in the fact that they can be *intersubjectively tested"***.** (Emphasis added (1959: 44) Further, 'observational' statements are accepted, (always provisionally, of course, for Popper[2]) ,"which are distinguishable by the fact that there exists at the time a 'relevant technique' such that 'anyone who has learned it' will be able to decide that the statement is 'acceptable'. (Lakatos 1970:106). Popper attempted to ground objectivity by the use of inter-subjectively available technologies and methods along with willingness, indeed fervor, to subject claims to attempted falsifications. (Though he never really could spell out what constituted a sincere or severe attempt at falsification.) *However, a lack of bias along with inter-subjective scrutiny is not enough to ensure that an inquiry is good. One can imagine an inquiry that is totally unaffected by human prejudices and uses only well-known techniques that none-the-less suffers from egregious design flaws.*

Another constraint, as Charles Saunders Pierce suggested, is that our methods probe that that we are orienting ourselves correctly in this world. That is, it is not enough that our inferences are unbiased; they must also really describe what the case is. Objectivity for the error statistician is achieved by using methods that can unearth and control the errors that impede

[2] As a strict falsificationist, Popper never accepted any statement as confirmed, which would require an inductive logic. At most it was corroborated, which meant a sincere attempt to falsify it had been attempted and yet it had not yet been refuted, i.e., falsified.

learning, whether they arise from our own subjectivity or because of the fact that data is incomplete and laden with noise, various experimental and observational errors, mistaken background assumptions, and so on. Learning from error is the watchword of this school of thought.

1.0.3 Firing up meta-methodological debates Can we meet this strong sense of objectivity? I claim we can, at least some of the time. Others in Science and Technology Studies (STS), as seen in the recent "science wars" disagree. They think experimental methods fail to provide objective warrant for claims in science and deny the possibility of objectivity in their accounts of the experimental appraisal of scientific claims. This debate over methodology and objectivity is the primary subject of this dissertation.

My third goal for this chapter is to explain why I think revisiting methodological and meta-methodological debates is important for philosophers and more generally researchers in STS, and I will suggest a list of criteria a naturalist methodology should meet if it is to fulfill the normative goals that many of us still hold, at least implicitly.

1.1 Logical Empiricist Background: What I am calling the new naturalism, whether in philosophy, history or sociology, may be seen as the result of a rebellion against logical positivism/empiricism (see Alan Musgrave (1974), Philip Kitcher (1992), Ronald Giere (2007), David Bloor (1976), Harry Collins (1985, 2004), and Helen Longino (1990, 2002)). The logical empiricists (and Karl Popper, too) following Gottleib Frege held that logic, not psychology, was relevant to studying how scientists produced and justified scientific knowledge. They held that the justification and warrant for scientific claims were to be grounded on *a priori* principles of induction, or in Popper's case, deduction. Traditionally, philosophies of science took the form of one or another logic of confirmation. Again, Popper was an exception and attempted to ground

the logic of science deductively as falsification, as 'conjectures and refutations.' These analytic philosophers of science sought to develop a universal logic that would systematically relate any given piece of evidence to any given hypothesis. Once this logic was formulated, scientists could bring their hypothesis (h) and evidence (e) to the philosopher of science, who would then plug this information (h and e) into an algorithm to determine to what degree the evidence confirmed or disconfirmed that hypothesis. These attempts failed for a variety of reasons including: (1) it was impossible to develop a sufficiently rich language that could capture theories and evidence of interest and (2) all attempts suffered from paradoxes (e.g., the raven paradox) or other problems as was internally admitted.[3]

While the new naturalists compare themselves favorably with logical empiricists, the naturalist turn was inspired not so much by these failures but by the work of Thomas Kuhn.[4]

1.1.1 Kuhn's rebellion—naturalism. Thomas Kuhn, a historian of science, 'revolutionized' the entire meta-methodological project when he charged that abstract logics of confirmation or falsification did not and could not explain how *real* scientists (historical or current) actually generate or warrant knowledge claims. To understand that, Kuhn claimed, would require looking at and describing the actual practices scientists engage in as they create knowledge. He felt this would require an intrinsically sociological or psychological explanation (see Kuhn 1970: 238). He based these claims on his historical case study of the Copernican Revolution and many other important episodes in the history of (mainly physical) sciences.

[3] See Suppe, Mayo (1996), Musgrave (1974), Salmon for a fuller discussion.

[4] Writing around the same time as Kuhn, the philosopher W.V.O. Quine was very influential in naturalizing epistemology, though much less visible than Kuhn in STS. He is perhaps best known for writing several influential essays, including "Two Dogmas of Empiricism," claiming that the analytic-synthetic distinction employed by philosophers did not exist and that, on close inspection, all knowledge is synthetic. This is just a technical way to say that no knowledge is *a priori* (i.e., independent of experience) but gained and, more important from the point of my project, justified *a posteriori* or empirically. Knowledge for Quine is a web of belief, and what we are psychologically prepared to accept as "new knowledge" is highly dependent on what we already accept as knowledge—knowledge is thus intrinsically contingent, a historical artifact in many respects. Thus, according to Quine, the *a priori* logics so beloved by philosophers were not only irrelevant to science but to all of epistemology.

Popper (1970: 54) disputes Kuhn's interpretation of these episodes as well as some of his historical facts.[5] However, whether or not one agrees or disagrees with Kuhn's analysis of history, and whether or not one accepts Kuhn's sociological conclusions, as Philip Kitcher points out:

> philosophers in light of Kuhn seemed to be faced with a methodological dilemma: either insist philosophers know *a priori* the principles of confirmation and evidence and must conclude that the actual reasoning of scientists is cognitively deficient or else abandon the *a priori* status of methodological claims and use the performance of past and present scientists as a guide to formulating a fallible theory of confirmation and evidence (1992: 73).

While one can debate the merits of Kuhn's arguments and assertions as well as the particulars of his conclusions, the upshot for naturalist philosophers was that one must reject the *a priori* in epistemology, even if one does not accept the turn from epistemology to psychology or sociology.

1.2 The New Experimentalists: I will regard *Naturalism* in the domain of meta-analysis of science as the view that philosophers and others should justify their claims empirically. The *new experimentalists* form a subset of naturalists in philosophy who have turned away from theory-dominated philosophies of science (representations) to focus on experimental practices (interventions).[6] Unfortunately, just as they were getting started, the whole movement seemed to fizzle away leaving a legacy of case studies but not much else. So yet another goal I hold is to revive the new experimentalism as holding the key to solving problems in philosophy of science.

The key feature of what I am calling the new experimentalist strand in naturalism is the view that looking to experimental practices holds the key to objectivity in the sciences and to

[5] One of the things we will look at in chapter 4 is the role history of science often plays as empirical evidence for claims made in naturalist philosophy of science.
[6] This distinction is from Hacking (1986).

solving philosophical problems (see Mayo 1996: chapter 3). Mayo has identified three broad themes that unite the new experimentalist literature:

1. Understanding the role of experiment is the key to circumventing doubts about the objectivity of observation.
2. Experiment has a life of its own apart from high level theorizing (pointing to a local yet crucially important type of progress).
3. The cornerstone of experimental knowledge is the ability to discriminate backgrounds: signal from noise, real effect from artifact, and so on. (Mayo 1996: 63).

As Mayo points out, the hallmark of the new experimentalists is their attention to the daily nitty-gritty details of local, particular experimental practices and technologies. In order to understand how data forms an "objective constraint" and serves as an "adjudicator" in science (1996:60), we need to scrutinize experimental practices and arguments.

1.2.1 A life of its own: Ian Hacking (1983) was one of the first philosophers to claim that experiments have a life of their own. What this means is that experimenters can develop and test their own local or topical hypotheses independently of the larger global theories (or paradigms) floating around. These topical hypotheses, as Mayo points out, often revolve around discriminating errors, for example distinguishing a real effect from an artifact. Often, this can be done in a way that provides theory-independent warrant for data.

A canonical example of how experiment can live a life of its own independent of larger theories is Hacking's well-known example of the use of three microscopes, each based on a different physical theory (optical, fluorescent and electron) for providing evidence that a dense body seen in a cell was a 'real' body and not an artifact of any one of the technologies used to observe it. As Hacking points out, it would be a miracle for all

three different technologies to display the exact same artifact in the exact same position if it really were an artifact. Hence, the method of using three different types of microscopes provides us with good grounds for claiming that we have observed a real effect, and so can rule out the artifact explanation of our observations. The ability to make these sorts of arguments supports claims that experimental appraisals are objective in so far as they can be theory independent.[7] While the new experimentalist literature is filled with examples of these types of methods or more generally arguments for detecting and weeding out errors, Mayo is unique in making this focus on "arguing from error" the cornerstone of her experimental epistemology.

1.2.2 Arguing from Error Mayo argues that when we turn and look at the "actual experimental processes and reasoning that are used to arrive at data" (1996: 60), we find a general pattern of reasoning, which she calls an "argument from error" that provides a systematic rationale underlying experimental arguments.

> *Argument from error:* it is learned that an error is absent when (and only to the extent that) a procedure of inquiry (which may include several tests) having a high probability of detecting the error if (and only if) it exists nevertheless fails to do so, but instead produces results that accord well with the *absence* of the error. (Mayo 1996: 64)

She notes with approval (1996: 63) Peter Galison's point that "the task of removing backgrounds is not ancillary to identifying the foreground—*the two tasks are one and the same* (1987: 256)." The ability to sustain an argument from error (removing backgrounds) is the ability to sustain objective inferences about phenomena (foreground). After all, that was the whole point of experimenting—to get objective control of errors in finding answers to our questions about

[7] For excellent examples and discussions on the epistemological significance of independence arguments for resolving problems with theory-laden observations see also Kosso and Shapere.

nature. (Even the exploratory aspects of experimentation—just poking around—requires an acute awareness of ruling out errors.)

1.2.3 Canonical Models of Error Ruling out errors is obviously important to any robust sense of objectivity but how do to so and how to do so systematically is not obvious for there appears to be innumerable errors. However, the situation is not as dire as one may think. Mayo identifies a handful of general error types, around which we can devise strategies for error detection and control:

6 Canonical Error Types

1. Mistaking spurious for genuine correlations,
2. Mistaken directions of effects,
3. Mistaken values of parameters,
4. Mistakes about causal factors,
5. Mistaken assumptions of statistical models,
6. Mistakes in linking statistical inferences to substantive scientific hypotheses. (Mayo & Spanos: 2010: 19)

As errors run to type, we may be able to, and in fact, often can, partition the space of potential errors into a manageable framework given the hypothesis under test, the methods available for testing, and the data produced. I illustrate this using a fairly complex example of such a partitioning in chapter 3 in my discussion of a BACI (Before-After-Control-Intervention) experiment conducted at Little Rock Lake to test the effects of acid rain on a lake ecosystem.

The new experimentalist narratives provide a rich source of examples of how scientists ruled out specific errors of the types above in order to make reliable inferences. Many excellent examples can be found in the work of Ian Hacking, Nancy Cartwright, Alan Franklin, and Deborah Mayo. The counterparts to the new experimentalists in philosophy are the micro-sociologists and anthropologists in the sociology of science, who investigate individual laboratories and specific research groups or projects (e.g., Harry Collins, Bruno Latour, and

Andrew Pickering). Another rich source of the details of experimental practices can be found in the work of historians like Peter Galison and feminists such as Donna Haraway and Helen Longino. However, the methods detailed in these studies can and should be assessed based on their properties for detecting and controlling errors in an inquiry and this is the very thing that is often overlooked or taken for granted in these studies, if not dismissed out of hand. This leads to a lot of confusion about what is really "doing the work" for objectivity in experimental situations and I hope to clarify and resolve some of this confusion.

1.3 Naturalist Theses Currently, there are two main approaches to naturalism outlined in Kitcher 1992 and similarly identified by the anti-naturalist Worrall 1999. In one trend, researchers appeal to historical or contemporary case studies to provide empirical evidence for claims about science. In the other naturalist approach, researchers look to a specific science or to particular scientific theories and findings (e.g., from psychology, evolutionary biology, cognitive science, sociology, etc.) to underwrite their claims. There are several themes that all naturalists share. To flesh out the experimental conception that I will build on, let us consider *four* broad *epistemological theses* that Philip Kitcher has identified as shared by naturalists:

1. "The central problem of epistemology is to understand the epistemic quality of human cognitive performance, and to specify strategies through whose use human beings can improve their cognitive states.
2. The epistemic status of a state is dependent on the processes that generate and sustain it. [For him, these include history, psychology, cognitive science, history and sociology.]
3. The central epistemological project is to be carried out by describing processes that are reliable, in the sense that they would have a high frequency of generating epistemic virtuous states in human beings in or world.
4. Virtually nothing is knowable *a priori*, and, in particular, no epistemological principle is knowable *a priori* (1992: 74-76)."

Cognitive performance, processes for producing epistemic states and reliability are all terms that are somewhat vague and ambiguous, and hence diverse readings and interpretations of these theses are available. The differences between naturalists arise from differences in how best to

achieve these goals or in how they interpret the implications or proper application of these theses. For example, in thesis two above, Naturalists interpret 'processes' as referring variously to physical, methodological or social processes (e.g., physical brain processes (chapter 2), experimental methods (chapter 3, 4), or group dynamics (chapters 5, 6)). I will explore these different interpretations in subsequent chapters. One thing we should note is that 'cognition' is specifically a scientific term for thought. As such, it carries its own implicit set of assumptions that needs to be recognized and acknowledged as John Worrall (below) does.

1.3.1 Naturalist assumptions Worrall, (a minimal a priorist in his own words) after a thorough perusal of the naturalist literature provides a list of *shared assumptions* made by naturalists (and which underlie, and I believe, clarify wonderfully Kitcher's list of theses above.) After each assumption, I indicate in brackets the Kitcher thesis with which it goes.

1. Humans are part of the ordinary causal order of the world [K1];
2. Any account of what constitutes scientific knowledge must be consistent with a descriptive account of how real cognitive agents arrive at such knowledge [K2];
3. Accounts must also fit how human cognitive machinery works to acquire & accredit theories themselves [K3];
4. Methods have improved [K3]; and
5. Background knowledge constrains what we take to be a good theory. [K4] (Worrall 1999: 340)

Worrall's assumptions almost map directly one on one onto Kitcher's theses. Because cognition is the scientific term for thought it implicitly assumes, as Worrall points out above, that human thinking is a part of the casual order of the world (e.g., not independent of it as say a Cartesian immaterial mind could be). Thus strategies for improving cognition require actual strategies that real agents (not ideal agents) could us to make improvements. Worrall, however, is no naturalist or at least not a through-going one. While he accepts and even applauds the naturalist's assumptions above, the difference between his position and the naturalists is that Worrall thinks

these assumptions cannot be justified naturalistically without entirely begging the question or ending up in a pernicious relativism. This is because, as he understands it, naturalism attempts to provide a purely *descriptive* account of scientific practices, but researchers do not just passively observe and describe, they must first actively decide and select what to observe and describe.

1.3.2 The case for circularity and relativism In the case of naturalist philosophy of science, one must make choices about which sciences and scientists to draw upon to use in their accounts (e.g., Kuhn favored Gestalt psychology, Giere vision science), or which case studies and which aspects to study (e.g., why Newton's principia but not his alchemy? Why chose Newton to begin with?). The problem here is that assumption 2 above requires deciding which cognitive agents and which scientific case studies to look at in order to make sure we are describing the ones who are producing solid scientific knowledge, not at ones who may be blocking progress or delving into pseudo-science.

But these choices, he claims, already require us to have some selection criteria to demarcate shoddy science from science that is well done. Thus, while he commends many of the naturalist theses, he argues that a thorough-going naturalism is an untenable position to hold. These approaches, Worrall claims "promise more than they can deliver: freedom from *a priori* while still underwriting the epistemological specialness of science (339)." Another way Worrall poses the problem is: how, if we are not presupposing that science is special, do we know what to describe to show it is special?

1.3.2.1. "Just ask the scientists": Ron Giere argues that we must 'ask the scientists' and that naturalists cannot go beyond what scientists claim constitutes the 'specialness' of science. But Worrall raises a good point pertinent to Giere's stance, to wit, which scientists do we ask? How do we make that choice unless we have already implicitly devised a set of criteria for

picking who is and who is not doing good science? That is, before naturalists start describing practices, it seems to Worrall that they must already have a pretty good idea about which sciences and scientists to look at. But this knowledge is not part of their naturalist toolkit—we need to remember that they are "just asking the scientists about good practices." So Worrall concludes they must be relying implicitly on some *a priori* evaluative principles or else they are being viciously circular.

Finally, Worrall argues that if we do not have criteria for making this choice, then our choice of whom to look at or which cases to study will be the result of our own personal and social idiosyncrasies. If two individuals made different selections then that might lead them to make conflicting inferences about science and leave them with no non-arbitrary way to settle the dispute between their findings on scientific practices. This leads us to holding a strong relativist position. While some naturalists, such as Harry Collins, may not mind being labeled relativists, most philosophers, including Giere, claim they can avoid that label.

1.3.2.2 Un-judged judges: This view of naturalism as purely descriptive (shared by many naturalists, e.g., Giere), along with the acknowledgement that choices need to be made to get the descriptive project off the ground leads Worrall to claim that some sort of *a priori* evaluative criteria must be drawn upon, either implicitly or explicitly, by both types of naturalists to get their respective games off the ground. He claims that to avoid circularity and relativism while still completing the evaluative and normative tasks philosophers set for themselves, requires postulating at least a small set of *a priori* evaluative principles to provide a standard for making the choices required to fill out the above naturalist assumptions. These criteria he calls the "un-judged judges."

The ability to simply stipulate that some basic assumption or reason is *a priori* only seems to shield it from scrutiny, which seems to me to be the last thing we would want in making the choices required to fill in the naturalist assumptions. Thus, Worrall's shift to a minimal a priorism seems to me equally, if not more, dangerous than any threat posed by circularity in the naturalist position that he indicates. If we accept for the sake of argument his position, that naturalists are unable to free their meta-methodologies from circularity, then it seems to me that a scrutinized circularity would seem to better meet the goals of objectivity than un-judged judges do. Better, however, would be to discover how to break the chains of such perceived circular reasoning.

1.3.3 Knocking down the circularity tower. Kitcher simply dismisses problems of circular reasoning in naturalism as "synchronous skeptical games," that naturalists ought to refuse to play with the skeptics. Mayo, I think, provides the key for how to break out of them. An assumption left over from pre-Kuhnian classical philosophy of science is that hypotheses or theories are only as secure (or reliable) as the raw data below them. Mayo calls this the "tower view" of knowledge. She has argued instead that the (often highly) modeled data and hypotheses higher up in the inferential chain are actually much more secure and reliable than the raw data below. Her simple example is in measuring the length of a table, taking several measurements and averaging them provides a much more reliable estimate of the length of the table than any of the individual measurements. This view is captured by the old carpenter's adage to "measure twice, cut once!" It is in demolishing the tower view of knowledge that one breaks the chains of circular reasoning. This is the task of separating signal from noise, one of the key themes of the new experimentalism and can be accomplished as Mayo argues by using methods (e.g., averaging) to avoid errors (e.g., the value of a parameter, length). If we understand the empirical

properties methods have for accomplishing these tasks, then it seems we have a naturalistic way to break out of the circularity that Worrall is concerned about without having to just simply stipulate that some set of methods are a priori justified. Instead, what is needed is an understanding of how using the properties of methods to detect and eradicate errors allow us to move up the inferential chain reliably. Can this sort of low level, localized type of reasoning provide normative guidance for science and meta-science in general? I believe the answer is yes. So let me turn to another problem raised against naturalism, it's supposed lack of normativity, and see how such a move would work.

1.4 Laudan's Normative Naturalism Perhaps the most serious problem for many philosophers when considering the merits of a naturalistic approach is the often apparent lack of normativity in such accounts. Larry Laudan is perhaps the best known proponent of a position we can identify as 'Normative Naturalism' in philosophy of science.[8] This is the view that meta-appraisals of scientific methods are themselves best served by following an empirical methodology (or meta-methodology), and that the results of such a study done well carry normative implications. Laudan captures the essence of the normative naturalist view:

> Normative Aim: "[T]he chief aim of the methodological enterprise is to discover the most effective strategies for investigating the natural world (Laudan 1987, 27)."
> Naturalistic Methodology: "Rules for learning from experiment are empirical and may be evaluated empirically" (ibid.).

The above sounds similar to Kitcher's naturalist theses previously given. An important difference here is the locus of analysis. Kitcher's naturalism will focus on both individual psychological (i.e., cognitive) apparatus and larger sociological forces for

[8] Latour is perhaps best known outside of academia for drawing normative lessons based on the "Machiavellian" implications of his theory. Surprisingly, Collins, a relativist, has explicitly drawn norms from his sociological studies. This suggests to me that philosophers are not the only ones who see normativity as an important component in meta-theories of science.

affecting scientific epistemology—trying to get at the empirical 'whys' behind theory choice. Laudan's approach is a means-ends analysis, where individual methods or rules provide the means to achieve epistemic ends such as reliable inferences, predicative ability, etc. As such it is a pragmatic brand of instrumental rationality focusing on the 'how' of theory choice. My discussion below draws extensively on Mayo & Miller (2008).

1.4.1 Pragmatic Norms Normative rules, for Laudan, take the form of hypothetical imperatives linking means and ends (e.g., if you want theories with high predicative reliability (end e), reject ad hoc hypotheses (means m). The soundness of these rules depends on empirical claims about the connection between means and ends. So in order to determine the correctness of any rules postulated will depend crucially on empirical information about the relative frequencies with which various epistemic means will likely promote sundry epistemic ends. Laudan looks to historical case studies (his empirical basis) to supply the needed empirical information about the correlation between various means and ends. This method for testing methods (i.e., using means/ends correlations) already requires Laudan to use some rule of appriaisal and therefore assume a method (his) works, in which case his project is open to Worrall's charges of circularity.[9] Laudan acknowledges this point but argues because some form of enumerative induction (arguing from instances) is ubiquitous to theories of methodology then his use of an enumerative inductive rule to get his meta-methodology up and running is innocuous. Worrall would obviously disagree, but for

[9] See Mayo & Miller (2008: 308) for a similar discussion

now, let us put this problem to the side and examine Laudan's method for assessing rules.[10]

The rule that Laudan pins his meta-methodology to takes the following form:

> **Rule (R1):** If actions of a particular sort, m, have consistently promoted certain cognitive ends, e, in the past, and rival actions, n, have failed to do so, then assume that future actions following the rule 'if you aim is e, you ought to do m' are more likely to promote those ends than actions based on the rule 'if your aim is e, you ought to do n.' (Laudan 1987: 25).

Using R1 to determine which methodology is the better one to use, he claims, requires that "we simply inquire about which methods have promoted, or failed to promote, which sorts of cognitive ends in the past (1987:27). The results of such an inquiry then provide the normative force in Laudan's naturalism. That is, one ought to choose method m if it has been found to be correlated with end e using R1 and not use n.

On the surface, it looks like Laudan's method achieves the twin ends for which it was chosen—it is thoroughly naturalistic, an empirical method of correlations and frequencies working on an empirical data base, the historical record; and it is normative insofar as the correlations found become rules to guide future actions. *One question remains, however, and it is this: is Laudan's method itself a good method? Does it promote reliable inferences?*[11]

1.4.2 The unreliability of Rule (R1): In Mayo and Miller 2008, we argued that Laudan's method R1 is a highly unreliable method and further that ES was a better, more reliable method to use in order to achieve Laudan's goals. I summarize our arguments below. One thing to note when reading the summary of our criticisms is that

[10] We can put it aside for now because the difference in their approaches to rule R(1) below is that Worrall accepts it by fiat—a priori.

[11] Reflexively, someone immersed in Laudan's method could pose this meta-question thusly: For assessing the reliability of methods to achieve (end e), is Laudan's method (M) more successful than alternative methods, e.g., error statistics (N)?

while Laudan's method demands an alternative method for his means/end comparison—he is a dyed in the wool comparistivist after all (e.g., see Laudan 1997 and Mayo's reply 1997)—our critique of his method is not itself comparative. That is, our criticism would stand even if there were no alternative method for comparative purposes. This is because there is no need to postulate an alternative method in order to assess the empirical properties of his method as will become apparent below. Let us take a walk in Laudan's world wearing our ES glasses.

1.4.*2.1 Warranting the antecedent in R1:* To apply Laudan's method, we need to determine if the antecedent of R1 is warranted. Mayo & Miller (2008) points out that: "evidence for the antecedent will consist of finding that method M promotes end e more often than rival methods *in some sample of cases.*" Hence, as Laudan realizes, the antecedent is a statistical hypothesis about the effectiveness of M (and N) to achieve e. M does not have to win out in every single case, only the majority of them. This however, leads to the first question we must ask ourselves if we are using his method, which is: how do we know that we have found even one instance where applying M promoted reliability? Let me stress here that the problem we were raising has nothing to do with Hume's traditional problem of induction about inferring that methods, which worked in the past will work in the future. The problem we were raising is the problem of using a sample of means/ends correlations to determine if the method even worked in the first place! So to begin to apply Laudan's method we need to ascertain:

1. Has a given method M actually been applied?
2. Was it the application of method M that promoted the end of reliability?

The second question is by far the more important and worrisome for getting Laudan's project off the ground. We summed up the problems with warranting the antecedent of R1 as follows:

> More generally, finding an instance in which method M is satisfied (violated) and cognitive end e reached (not reached) fails to tell us that reaching (failing to reach) the end was due to satisfying (violating) the rule. (Mayo & Miller 2008: 308)

1.4.2.2 False Negatives: We argued that false negatives pose a major threat to Laudan's project. This is because the historical record, which he relies on, is "a highly unrepresentative sample of applications of methodological rules with regard to the questions of reliability (ibid.)." There are several reasons for this. First, it is unlikely that a failed method would make it into the public record because no one will publish their negative results or failures as there is no perceived need much less any incentive to do so. Scientific journals publish and awards are given for success, not failures of method. Coming up with or using a method that doesn't work (provides unreliable data) does not garner awards or recognition. The upshot then for Laudan's empirical historical database, as we put it in Mayo and Miller (2008) is that "there will be scant evidence in the historical record when a method or criterion was applied and the goal was not achieved."

So, as we stated (paraphrased from ibid: 308), the type of information that the meta-methodologist would be looking for, cases where violating a given rule creates an obstacle to achieving the goal of reliability, would simply not show up. It is just that there would be little evidence of cases where violating a rule resulted in an unreliable theory because if the theory were not rejected and purged from the record, then its

survival would have been due to being shored up using other methods. This brings us to a second problem with Laudan's meta-methodology.

1.4.2.3 A normative bind: On page 309, we pointed out that even if we did have the type of negative evidence required for a means/ends assessment, we would still be in a bind normatively. This is because the advice scientists are looking for and that philosopher's desire to offer is not negative advice like 'don't use Method m if you want reliable inferences,' but positive advice of the type: 'use method m to achieve reliability'. We summed up the problem with Laudan's reliance on R1 thus:

> Without knowing whether satisfying a rule is the, or even a contributory cause of the goodness of a resulting theory, and conversely, whether violating the rule is the reason that a theory fails to be reliable, information about correlations even where obtainable, does not help. A given theory may be seen as resulting from applying any number of methods. (Mayo & Miller 2008: 309)

If we think back to the definition of minimal severity, we quickly realize that there is almost zero chance of finding out that a method is the cause of the success or failure to achieve end e using R1. Through the lens of minimal severity, then we reject R1 as being a highly unreliable meta-method for assessing the reliability of methods.

1.4.2.4 Calling all capacities: What would be needed for Laudan's meta-methodology to succeed in its goal of assessing and hence recommending (or rejecting) methods in an effort to achieve reliable inferences?

> Laudan's meta-methodology could achieve what he wants it to only if it were supplemented with a way to determine whether the method was causally efficacious in the context in question. This demands a separate method for determining the capacities of methods. (ibid.)

We then pointed out that solving his problem simultaneously rings a death knell for his meta-methodology:

> But if one had such a separate method for determining whether and how rules work to achieve ends, then one would not need to collect correlational evidence in the first place." (Mayo & Miller 309).

Circularity is not the real challenge to Laudan's enterprise. The real problem with his brand of normative naturalism is that the mere fact that a method was used (often in conjunction with many other factors) is insufficient to warrant the claim that it was the use of or failure to use a particular method that explained the success or failure at hand. His method is an instantiation of the fourth type of canonical error in my list above, specifically the error of mistaking a correlation for a cause. What is really wanted is a way to determine the causal efficacy of methods for reaching the ends that Laudan desires, e.g., reliable inferences. If we had that information, then the causal claim could be sustained and we would not need to use means/ends correlations to begin with.

1.4.3 The causal mechanisms of methodological rules: What would a causal mechanism in a methodological rule look like? Let us start by looking at an example. Imagine you are a scientist and you want to prove your drug X is a more effective cure for disease D than the current treatment T is. So the experimental problem is to show your drug X is more effective than T. You have a group of people with D, how will you allot them for you experiment? If you are like most people, you would have a tendency to assign the healthiest people to the group that will take X and place the unhealthiest of the lot into group T. And if you were a champion of T, you tend to do the reverse. Wouldn't it be much better, that is more reliable, to divide the people into the two groups randomly? One could put all the names in a hat and after mixing it thoroughly, the first half pulled out go into the group receiving X and the remainder will be assigned to T. That method of assignment does not allow the deck to be stacked in favor of one treatment over the others. But randomization does more than stop one from stacking the deck as

it were based on observable factors. It also ensures that hidden variables between subjects are also randomly and hence equitably dispersed.

The rationale behind randomization should be becoming clear. As a method it forces the distribution of other factors that could potentially bias the results between the two groups to be equally distributed between them and hence they ought to cross cancel out leaving only the different treatments to cause differences between the two groups. Here we see the reason a method is chosen is based on its having the property to eradicate an error, in this case the possibility that a factor other than the one under consideration (drug X) is effecting (confounding) the outcome.

Ideally, in testing a causal claim, such as the efficacy of a drug to cure a disease for example, one would want to be able to control all factors that could affect the outcome that are not due to the drug under test but are the result of another causal factor. These other factors that could confuse a scientist as to the cause of a result are known as confounding factors or confounders for short. While literal control in most, if not all cases, is virtually impossible, randomization is one method that allows one to argue as it they had literal control.

1.4.4 Localization of the severity assessment: Meeting the standard of testing required by Mayo's principle of minimal severity is based on viewing testing as "probative." Severe testing is not a passive activity—for data to *become* evidence requires that the testing (or more generally data generating procedures) are such that the hypothesis is genuinely "probed" for ways in which it would be a mistake or error to assert it. Different hypotheses will lend themselves to different inferential errors, and hence the same data may carry very different evidential import for different hypotheses (e.g., data from a method that may constitute strong

evidence for a correlation may pose weak evidence for a stronger causal claim).[12] The main gain of such an approach (i.e., localization using the severity function discussed previously) is to make error detection and management in empirical inquiry, including an ability to check and test background assumptions, manageable. This is crucial for an account of objectivity, and what has been missed in other naturalist accounts. This approach to methodology (more fully developed in Mayo 1996, Mayo & Cox 2007 and Mayo & Spanos 2005; 2008) exemplifies a central feature of the new experimentalism: the need to "go small" and look at local experimental testing methods and practices. Of course, working out how to link up all the more local studies as well as how to design tests of larger phenomena remains a current topic of research (see Mayo & Spanos 2010; Staley (2008), Mayo & Miller 2008). Nonetheless, Mayo's error statistics has successfully captured and extended a general form of inductive reasoning—severity or arguing from error—used not only in science but in the law and even common sense. (I will cover this approach more thoroughly in chapters 3 and 4.)

1.5 Reasoning from error & Meta-methodology: This same reasoning from error, I claim, should guide our meta-methodological discussions and prescriptions similar to the way Laudan's meta-methodology was assessed above, and in Mayo & Miller (2008). In the ensuing chapters, I will apply this general approach to some of the more popular "naturalist" meta-methodologies and rules in STS and philosophy of science specifically to evaluate their empirical properties for uncovering errors.

1.5.1 Testing meta-methodological claims: I shall argue that in theory it is relatively straightforward for naturalists making claims about scientific methodology and the reliability/stability of inferences based on empirical evidence to resolve their differences.

[12] This is in direct contrast to other statistical accounts (e.g., Liklihoodists and Bayesians) where evidence is evidence and carries the same weight for an inference independent of how the data became evidence; that is, the procedures for generating it, including stopping rules, etc. see Mayo 1996, 2004.

Naturalist meta-methodologies like the methods in science should be open to scrutiny to assess their empirical strengths and weaknesses for meeting their self-appointed tasks. That is, since naturalists claim to base their accounts on empirical evidence, their proposed accounts of empirical evidence should be subject to empirical scrutiny and corrections. This reflexivity blurs the line between method and meta-method, which simplifies the discussion a good deal. For example, if one claims that the only way to get objectivity is by bringing multiple and conflicting views to bear on an inquiry, then one's proposed method for doing this should have the capacity (empirical properties) for securing the stated goal.

Given the central role and supposed importance of testing in accounts of science, it is surprising that most naturalists in STS and its allied disciplines are content to justify their accounts merely by showing that they "fit" some self-selected and opportunistic set of data.

1.5.2 Broader definition of experiment. Let me be clear here--when I talk about 'testing' I am using Mayo's much broader concept of experiment here:

> I understand "experiment," …far more broadly than those who take it to require literal control or manipulation. Any planned inquiry in which there is a deliberate and reliable argument from error may be said to be experimental (1996; 7)."

Neither objectivity nor arguing from error is unique to experimental practice. Reasoning from error is a general type of inductive reasoning used not only in science but in the law and common sense (see Mayo 1996; 2006) and can be successfully employed by athletes, musicians, gardeners and cooks. From figuring out how many miles per gallon a car really gets to which combination of ingredients made (or ruined) dinner, most of us use this type of reasoning every day. The difference is one of degree.

1.5.3 The error statistical assumption: Scientists have developed very sophisticated and often quantitative methods for making these types of arguments. The general principle of

reasoning being advocated here is based on the assumption that warranted 'experimental'/'experiential' arguments demand severe or reliable probes into error. Every account ends up, as Kitcher points out, starting at some assumption. The ES account does better, I would argue, because this assumption can be evaluated by any human being based on their own experience of making and correcting mistakes in everyday life. Furthermore, this assumption is, or at least can be, inter-subjectively evaluated every step of the way in a well-designed experiment. Whether or not a method has the empirical properties for detecting, controlling or eradicating an error is context dependent. The context, which is composed of the specific hypothesis under test, the method used to test it as well as the data produced, carries all the information that is needed to determine whether or not the test can meet the standards of minimal severity or not.

But what does this mean for testing a meta-methodology? Using the above definition, what it means is that similar to empirically assessing methods based on their properties to detect, control or eliminate errors, we need to look at and scrutinize the properties of meta-methods for detecting errors in the inference(s) we wish to sustain when we use them. In this, we can be guided by following one of Mayo's versions of her severity principle:

> **SP**: If a method has no chance to detect an error, then the fact that it has not detected the error is not grounds for inferring that the error is absent.

Thus, if we use a method that has no chance of detecting a particular error, then the fact that we don't detect an error is not evidence it is not there. Referring back to the medical drug test trial, if randomization was not used, then we would be unable to rule out the possibility that one or more confounding factor were responsible for any differences between the two groups (X and D).

One of the goals of meta-methodological accounts is explaining the objectivity or stability of science even given dramatic shifts in large-scale theory change. The differences

between various philosophers and sociologists of science looking at scientific methodologies arise not only from comparing their starting assumptions and methods or their normative injunctions to scientists, but can also be traced to holding different concepts of objectivity/stability.

1.6 Explaining Objectivity or Stability in the Sciences If our aim is to explain the objectivity or stability of inferences based upon evidence in the sciences, then we should be able to assess/test empirically whether or not a method has the properties necessary and sufficient for the claims being made about its use. I bring up for consideration that for those who are wary of "objectivity," they may without too much of a stretch substitute "stability."[13] (This tactic is also available to those who eschew epistemological talk in favor of a metaphysical starting point (e.g., Latour). As I have already stated, the ES account that I endorse holds a strong version of objectivity.

Some naturalists, like Giere, attempt to justify that we are objectively orienting ourselves to the world based on the fact that human biological capacities and technologies have evolved. Evolutionary theory for him provides a link, however partial or tenuous, of a direct, though unconscious human access to reality as evinced by our survival and technological success. Others like Philip Kitcher point out that evolutionary survival is not strong enough to support such a claim. At most, we can infer that humans do not have too high a rate of false positives in their inferential practices (i.e., that something does not pose a risk when in reality it does pose one). Even Giere admits reproductive success is too crude a measure for scientific objectivity, though he still uses it.

Kitcher constructs a decision theoretic naturalist epistemology of science at both the individual and institutional levels. However, the key to objectivity for him is provided by the

[13] Stability can apply to phenomena or theories or laws.

methodological rule of explanatory unification. Following this rule, he claims, provides objective constraints on both individual and socially-based decisions about theories. This is because we do not just seek truth but significant truth. More recently there has been a push to locate objectivity in social norms and practices. Both Longino and Collins argue that social methods are needed to ensure objectivity. These social approaches to objectivity can take a variety of forms. Most involve some form of disciplinary consensus as the standard of objectivity or else a more complicated dialectical or interactional form of objectivity derived from anthropology. In this case, while there is still room for the subjectivity of the knower, external reality, composed of not just things but other knowers, constrains inferences.

1.6.1 Empirical Rationales not Algorithms Most of the attempts discussed above fail egregiously in two ways. One, after claiming that logical empiricism is dead and there is no single algorithm for evidence and inference, they take that to mean that reasons cannot play a decisive role. But the lack of an algorithm does not mean that reasons cannot and do not play the deciding role. This confusion is also apparent in their treatment of methods as algorithms and their failure to take into account that methods, including their own, have empirical properties that enable the detection, manipulation, and correction or eradication of real empirical errors. Errors exist—they are not merely the product of reason but can also mislead reason.

A second problem arises because many of the rules or methods that philosophers and researchers in STS champion are hasty generalizations that come from the fact that they are picking up on some rule or element that in specific situations functions as an error probe and so facilitates objectivity in those contexts. Their success in those contexts can be error statistically justified but this is a very different justification from that which is often mistakenly given. This is a case of getting the right answer but for all the wrong reasons, and the situation is undesirable

from a prescriptive point of view. Thus, for example, I do not deny that social methods play a role in ascertaining and ensuring objectivity in scientific inference. However, where these methods play a role is not in consensus formation. Instead, these methods work, where they do, when the kinds of errors that they can detect occur but not otherwise. In a way, I am proposing the more radical interpretation of social methods—that they play the same role as other empirical methods and should be assessed like those other methods, in terms of whether their properties facilitate objective error control. This is also the case for methodological rules like unification and Hacking style-technological manipulations (e.g., electron guns).

1.6.2 Explanatory unification: Not every type of explanatory unification will work for objectivity. Unification works for objectivity to the extent that it serves as an error probe or check on inferences, similar reasoning lies behind the use of technological success. For example, Hacking's example of building an electron gun to shoot electrons provided objective warrant for claiming that electrons objectively exist and have certain properties but clearly it does not exhaust all there is to learn and know about electrons. We may have all sort of misguided and incorrect notions about their nature that would not be discovered in the course of building PEGGY II (the electron gun). Indeed, tracing technological innovations and improvements is one way of seeing how we learn more and correct our knowledge. This is also why the lack of unification or technological success is not necessarily an indication of failure.

The problem I see in these accounts is that they have picked up on methods that work on some specific occasions and then try to extend them to work in all situations. These methods only work to the extent that they can detect or control specific types of errors in very specific types of inferences given local inferential situations (i.e., processes for generating and interpreting data), but this does not make them universally valid. What can be stated generally is

that objectivity is achieved by employing methods that have the requisite properties for overcoming errors in specific local circumstances. Just as there are many roads to get to the top of the mountain, so too there are many methods for achieving reliable inferences in a variety of contexts and goals. But the rationale underlying the use of all of them remains the same—to be able to control error probabilities objectively for the purposes at hand. I will develop this view in chapters 3 and 4.

1.6.3 A radically new approach to naturalism: Mayo's error statistics provides for a radically different approach to normative naturalism, drawing on the insights (and failures) of Laudan and the new experimentalists. And though she and I both share Laudan's goals and aims for a normative naturalism, this project is not based on a method of means/ends correlations as his is. Indeed, purposely going against the "correlational" tide in STS, this project uses and extends the ES framework and purposely attempts to find the "mechanisms," that is, the empirical properties of methods or rules to detect, deflect, and eliminate errors, and so to uncover the causal properties of methods. The benefit of such an approach is to develop methods not for rational reconstructions, but forward looking methods for learning. I investigate and more fully develop and justify this approach to methodology by using error statistics as a lens to scrutinize and learn from other meta-methodologies to find the good and ditch the bad.

1.6.4 Generalizations, norms & naturalism: Many in STS apparently have abdicated taking any normative stance at least insofar as methodological prescription goes. Contrary to Worrall, Collins[14] and others, the fact that there is no one algorithm for assessing methods or making inferences based on empirical evidence does not mean reasons and arguments do not

[14] Collins is founder of the Empirical Program of Relativism (EPOR), which claims while reasons and arguments have a role to play in scientific theory choice, they are not decisive, rather social factors, such as interests determine when and how experiments end.

play a decisive role in scientific inferences and even in choosing methods and determining "observations."

While I agree that one of the best insights we have gained from naturalist approaches that is embodied in the new experimentalism (in both its philosophical and sociological guises) is that we "must go small and local;" nonetheless, I agree with Mayo that this "localization" or "contextualization" does not preclude us from saying anything general or making normative claims (see chapter 4). Finding the appropriate localization; however, is not obvious and oftentimes is quite tricky. But it is the general concern with error, or arguing from error/severe testing that Mayo has identified as the general style of inductive reasoning that works as a methodological norm not only to underwrite the reliability of (some) scientific inferences, but to provide (along with the lines of the piecemeal approach developed by Mayo) and formalized in the severity triad (Mayo & Spanos 2006) for the *appropriate localization.* By homing in on how error or a concern with error can work as an aid to learning rather than merely create an obstacle, Mayo has provided a key to unlocking how methods work when they do and why they do not work in other cases. This in turn suggests criteria that a normative naturalist approach to methods should meet.

1.6.5 Criteria for assessing naturalist accounts of science. There is always some rationale or other behind the use of specific methods in particular situations. This is the case also for meta-methods. The fact that we can give a rationale, however, does not mean we are justifying the use of a method *a priori*. Instead, we want to assess whether or not the method has the empirical properties or characteristics that enable it to fulfill the rationale behind its use. In short, we must be able to explain how methods work. Below are the key criteria a philosophy or

other type of meta-analysis of scientific methods must meet if it is to fulfill the traditional goals of philosophy of science and yet be vigorously naturalistic:

1. Provide a means to evaluate and justify methods empirically.
2. The assessments in (1) should produce norms for how to proceed or not proceed methodologically. (That such norms will be context specific is to be expected.)
3. The methods and methodological norms promulgated should be compatible with human capacities (both strengths and weaknesses.)

Ideally, after meeting the first three basic goals above, from a philosophical point of view, a superior meta-methodology would afford one the ability to:

4. Provides a demarcation between (better and worse) science & pseudo- or non-science, not simply assume one exists (or fails to exist).
5. Account for the, if not real, at least apparent, local and pluralistic nature of methods.
6. Can solve (resolve) or provide insight into philosophical problems (e.g., skepticism, relativism, Duhem's problem, and so on).

Most accounts fall short of meeting these goals or even their own goals; and where they do get it right, it seems they are relying on ES-like criteria. In other cases, their claims go far beyond what it seems they would be justified to infer given their evidence and methods.

1.7 Conclusion: Why Meta-Methodological Discussions Matter If we philosophers of science, and more generally naturalist epistemologists and other STS researchers, expect to be taken seriously and play a normative role in the practice of science, policy or education, then we need to be able to justify our claims about science or technology with specific evidence. *Thus we need to be able to sustain the even stronger claim that we are engaged in 'normative naturalism,' that is, that our claims about science not only describe (successful) practices (in hindsight) but can extract forward looking norms to 'prescribe' or 'guide' practices as well.* The focus of this dissertation is that naturalists can sustain this stronger claim, learn from the mistakes and errors made in previous naturalist attempts, and avoid the pitfalls and criticisms faced by previous attempts. Such a project is important, not only for solving problems in object-level sciences (e.g.,

replication debates in ecology), but for providing warrant for our own meta-methodologies and underwriting the reliability of our normative advice and activism.

In addition, while it is commonly thought that the distinguishing mark of science lies in its methodology, which underwrites the reliability of scientific inferences, meta-level studies of how science is actually conducted or practiced have, in general, failed to provide insight into how methods underwrite this reliability, if indeed they do. However, this type of insight is desperately needed especially in areas of research where complex large-scale phenomena and observational rather than experimental procedures are the norm (e.g., ecology, climatology, cosmology, sociology, evolution etc.) And also, by extending Mayo's general account of error reasoning explicitly into meta-methodology, we can improve our work in our own fields and have some shared criteria, a touchstone, if you will, for collaboration and critique.

Chapter 2: Philosophical "Science of Science" Approaches: Giere & Kitcher

Drawing on the deliverances of the sciences, naturalists view members of our species as highly fallible cognitive systems, products of a lengthy evolutionary process. How could our psychological and biological capacities and limitations fail to be relevant to the study of human knowledge? (Kitcher 1992: 58)

I would suggest that evolutionary theory, together with recent work in cognitive science and the neurosciences, provides a basis for such understanding….The study of science must itself be a science. The only viable philosophy of science is a naturalized philosophy of science. (Giere 1999:160, 173)

2.0 Introduction. A popular approach to philosophical naturalism is to look to evolutionary theory and the various human sciences to inform, if not entirely replace, the philosophy of science and (more generally) epistemology. The underlying tenet behind these approaches is that we should look to our most advanced scientific theories—"draw on the deliverances of the sciences" as Kitcher puts it above—from psychology, cognitive science, biology, sociology, and other cognate fields to fulfill the traditional tasks of the philosopher of science: warranting scientific practices and knowledge, justifying methodological principles, demarcating science from pseudo-science, and so on. Even stronger claims are forwarded to the effect that epistemology will ultimately reduce entirely into one or more of these sciences. Paul Churchland is an outspoken proponent for neuroscience assuming this role. Another variation views sociology as providing the best way for science to know itself (e.g., Bloor). Social approaches are covered in chapters 5 and 6.

While Churchland and Bloor are absolute reductionists, most naturalists in philosophy of science take a less reductionist and more eclectic approach, borrowing from many sciences, as do, for example, Ron Giere and Philip Kitcher. While authors disagree about the extent to which we should draw on the sciences, the primary thesis behind all approaches in the 'science of science' style of naturalism is that the only way to warrant human thinking about science is by

generating scientific knowledge about human thinking. A question I want to explore is how looking to either psychology, sociology or other cognate sciences can yield objective knowledge about knowledge? I focus on Giere's general argument that naturalism is the best method for human beings to gain objective knowledge. Kitcher introduces explanatory unification as providing an objective constraint on scientific knowledge that works at both the individual psychological level as well as sociologically at the group level.

2.1 Giere's Radical Methodological Naturalism Giere's early work (1985) took a decision theoretic approach to explaining theory choice in science, though as he emphasizes, his turn to decision theory was not to provide an "account of rational choice" but to give a descriptive account—specifically as a specialized part of ordinary belief-desire psychology (p. 347) to which he is still sympathetic.[15] However, his approach since the 1990s emphasizes appealing directly to scientific knowledge, (e.g., results and theories from vision science, cognitive science and social science, see for example Giere 1990, 2001, 2003a, 2003b and 2006). Doing so, he thinks is the way to understand science, to champion it as our best form of knowledge and to resolve philosophical problems about it. Following Giere 2001, we can label his current view "Critical Hypothetical Evolutionary Naturalism (CHEN). His 2001 paper in honor of Don Campbell, an evolutionary epistemologist, provides a concise yet full overview of Giere's naturalist position and I will follow it closely in my discussion below.

2.2 Evolutionary naturalism A primary theme running through philosophical naturalism revolves around taking evolution seriously in understanding epistemology. For Giere, all that is required for "[a]n evolutionary naturalism would be, at a minimum, a naturalism taking cognizance of and compatible with man's status as a product of evolution (2001: 54)." This view is captured in three assumptions that Giere (1990: 339) explicitly lays out:

[15] Personal conversation (2005) at Virginia Tech in Blacksburg VA.

ASSUMPTION 1: Human perceptual and cognitive capacities have evolved along with human bodies.
ASSUMPTION 2: … [T]hese capacities are fairly well adapted to the environment in which they evolved.
ASSUMPTION 3: Our technology and ability to manipulate the surrounding environment show that in relevant and significant ways, scientific knowledge provides an adequate model of the world.

What an evolutionary account does, Giere claims, is allow us to get away from our "subjective experience or intuitions, which stymied both empiricists and rationalists (1999: 161)." That is, we do not have to justify our knowledge based on either the accuracy of our sense impressions or one or more *a priori* principles. Instead what we have learned from an evolutionary perspective and modern neurobiology, according to Giere, is that "[i]n fact, we possess built-in mechanisms for quite direct interaction with aspects of our environment. The operations of these mechanisms largely bypass our conscious experience and linguistic or conceptual abilities." (*ibid.*161) [16] As Giere sees it, the directness of this interaction or link, however partial, between us and aspects of our environment provides the foundation of objectivity.

Now while we are unconscious of these mechanisms, this does not mean that we can never conceive of or talk about them. After all, if this were the case, he would have nothing more to say. Instead, Giere wants us to look to modern cognitive science, psychology, and other "human" sciences (e.g., vision science and sociology) to explain the direct connections between human minds and the many parts of reality, including social realities, with which they interact. It is these interactions that compose the human world. Naturalists of Giere's stripe believe that if we can only learn how those direct connections work, then we would be able to underwrite the objectivity of inferences based on them. I would add my caveat that this is only provided that we can distinguish when they are working correctly and when they are not. For a Gierean naturalist, to answer problems in philosophy of science will require scientific rather than philosophical

[16] In footnote 6, he acknowledges a debt to Paul Churchland for this line of thinking.

research.[17] Thus, cognitive science, psychology and sociology will (ultimately) replace and answer the questions in epistemology; vision science will solve philosophical conundrums about observation, and so on. Giere concedes, "There is nothing particularly evolutionary about my understanding of that part of scientific epistemology concerned with the empirical testing of scientific hypotheses" though he thinks such testing is itself the product of social evolution. Rather for him, "the nature of historical change in science" is more directly evolutionary on his view.

2.2.1. Evolution of research groups & their models Following Kuhn, Giere sees "the evolution of research groups within research specialties" as the engine of scientific change. Unlike Kuhn, Giere sees science as progressive albeit in a historically contingent manner. I return to this point in the next section. Using an evolutionary analogy, he will understand scientific models, a concept he prefers to theories, to which individuals are committed as "traits" of the individuals. Historical changes in models (e.g., from Newtonian mechanics to Einsteinian, Static earth to dynamic earth (plate tectonics), etc.) come about because the relative distribution of these models (i.e., traits) evolves through time. In layman's terms, my model or theory wins because it is popular and its use spreads throughout the scientific community. [18]For Giere, then, the "central issue is: What are the mechanisms underlying the variation, selection and transmission of these intellectual traits within research groups. " He thinks these mechanisms are mainly cognitive or social. (61). He wants us to understand these social and psychological mechanisms as providing the "genetics" for an evolutionary model of science (62.) Thus to understand how scientific knowledge, cast in the form of models, changes, requires an

[17] One of Giere's favorite science for answering philosophical questions about science is the science of color vision (e.g., 2006a,b; 2001).
[18] In Latourian terms, my network is the longest.

understanding of psychological and social mechanisms, which in turn will require us to use psychological and sociological knowledge to uncover and understand them.

Giere states that "to maintain any sort of realist view of science, one of these mechanisms must be the outcomes of experimental tests of various models." (61). But he emphasizes that experimental outcomes are "but one mechanism relevant to the differential growth of one research group over others." Other mechanisms he lists include: material resources to support graduate and post-graduate students, inspiring teachers and enterprising academic organizers (62)". This evolutionary model provides a way to organize our understanding of science but provides "little basis for understanding the details of any changes in science." This understanding will require "appropriate cognitive and social models (2001: 62). Now Giere is less inclined to provide the actual details or theories from these sciences that are required for this task. Early on, he did cite Nadler et al (1978) to make the claim that his modeling concept of scientific knowledge was supported by cognitive science because human animals thought in models. Of course, 'rational reconstructions' to support almost any philosophical concept of science can be made, even if the concept is false. Thus, such reconstructions do not provide evidence for the position. Currently, Giere who feels assured that the sciences will ultimately provide the answers philosophers are looking for, is more concerned to argue that the general naturalist approach he advocates is the one all philosophers should follow.

2.3 Giere on naturalistic explanations Giere explains that providing a negative description of naturalism is easy—it simply requires ruling out any supernatural explanations (especially religious) and stands against any appeals to *a priori* intuitions or justifications to defend claims. Providing a positive account of naturalism is much more difficult, he claims. To begin constructing a positive account, Giere defines naturalism as: "the position that all aspects of the

world can be given a naturalistic explanation" (2001: 55). Natural explanations include "most obviously scientific explanations" but also "historical explanations and even every-day concepts" (ibid.).

Giere asks: "What, one might reasonably ask, constitutes a scientific explanation?" (ibid 54) His reply is that the ***best*** answer a naturalist can give is:

> **Scientific explanation:** A scientific explanation is an explanation sanctioned by a recognized science. To say more is to risk going beyond the bounds of naturalism. Giere (2001: 55)

Similarly, I guess to determine what counts as a historical explanation, one looks for explanations that are sanctioned by historians, and for commonsense we should look at "what people in common would agree on: that which they "sense" as their common natural understanding…." (Wikipedia).[19] Interestingly enough, Wikipedia goes on to state: Common-sense ideas tend to relate to events within human experience…and commensurate with human scale…. Often ideas that may be considered to be true by common sense are in fact false." This is not an auspicious recommendation for the truth of naturalist explanations! Worse, Giere has offed a very unsatisfactory and circular definition of a scientific explanation—scientific explanations are explanations accepted by scientists! What most philosophers are really interested in understanding is why do scientists accept this particular explanation(s) and why did they reject this other one (or more)? Looking again at his definition, two questions immediately leap to mind:(1) What does Giere mean by "sanction" and (2) how is a science "recognized" and by whom?[20]

---

[19] http://en.wikipedia.org/wiki/Commonsense

[20] If Giere thinks he is just trying to give a descriptive social definition, he would still, indeed even more so answer these two questions to provide an adequate social science definition.

In regards to the first question, does sanction mean the explanation is considered true in some way? Must it be part of scientific canon and included in textbooks? Or does it suffice for sanctioning if the explanation is merely entertained or tolerated or used in some way by the scientific community or some members of it?[21] Remembering the parable of the King's new clothes, is it enough that the rulers in this case the scientific elite, accept an explanation? Swept under the carpet in Giere's definition above is the fact that the term "sanction" implies that some sort of criteria have been met in order to warrant the act of sanctioning an explanation. But this is precisely the locus of debate in philosophy of science--what are these criteria, if there are any?

Giere's definition, which only qualifies "sanctioned" with "by a recognized science," implies that our only criteria for what constitutes an 'acceptable' (rather than accepted) scientific explanation comes about either by way of an appeal to authority (scientific authority) or to popularity (the majority of scientists) both of which are fallacies and for good reason—both appeals are demonstrably unreliable methods for reasoning and can easily lead to the acceptance of false claims. This would surely violate one of the tenets (listed in chapter 1) of the error statistical approach, which is based on trying to find and articulate methods for ensuring that false claims do not go undetected.

From the perspective of error statistics, an explanation would be sanctioned only to the extent that it had passed a severe test. Remember the idea behind a severe test is not only does a hypothesis (or explanation) fit the data but the test or method used to generate the data, which is "testing" the hypothesis, is one that such a close fit would be

[21] Giere ultimately will go with the latter idea that an explanation fits the world in some way that is useful to some members of a scientific community at some time.

highly improbable if the hypothesis (explanation) were false.[22] Notice on this account, the reason an explanation is sanctioned is because it has been put through the wringer and survived. Now if Giere offered criteria for sanctioning that would be one thing, but he says there are none. All he requires is that the sanctioning of the explanation is done by a "recognized science." This would suggest that we have some way of recognizing legitimate science from, minimally, non-science. But again, in order to be able to recognize sciences, implies that we have some sort of criteria for what it means to be a science or on a smaller scale, a scientific practice. So Giere must have some criteria of goodness or legitimacy in order to recognize science.

Giere goes on to claim in the same paper that:

> At the most general level, not being willing to appeal to essences, naturalists cannot attempt to solve the demarcation problem by providing a definition that separates science from non-science. At a less abstract level, naturalists know that what counts as a scientific explanation changes over time….Ultimately, naturalists can do no more than follow such historical developments." Giere (2001: 54)

But do we have to have captured the "essence" of science in order to, say, separate out egregiously flawed scientific practices from exemplary ones? Granted we may not have an exhaustive definition to separate science from non-science, but surely we must have some ways to roughly distinguish between the two or how can we even have recognized sciences? On the previous page, I discussed how on the error statistical account, this demarcation is made based on the concept of severe testing. This demarcation does not require one to have captured the essence of science or any one science. It does the work, however, of sorting out and distinguishing good tests (experimental practices) from flawed tests (bad practices).

[22] Obviously for any complex sort of explanation, a series of severe tests will be required to probe the various ways that an explanation could be false by partitioning parts of it for testing—see Chapter 1.

Now when he says naturalists are not willing to appeal to essences above, to be generous, I can only infer that Giere is saying that naturalists do not have an exhaustive definition of science, one that is neither too broad nor too narrow but instead captures fully the "essence" of what it means to be a "science". Given that science is a very broad term and as a field of human knowledge is growing, this is not surprising but neither is it particularly troubling I think for naturalists. Following the scientists, as Giere wants us to, suggests that it is their methods that underwrite the success of their claims and this seems a good place to begin to try and "recognize" scientific practices.

The error statistician just denies it is so difficult to point to methods that would exemplify the extreme cases of clearly unsatisfactory and unscientific means for learning about the world, nor to identify others that are well esteemed in science and very useful for finding things out about the world. As I showed in chapter one, we are able to look at properties of methods (scientific ones as well as others) and determine if a method is able or unable to promote reliable learning. But notice, I did not require or even suggested that there was an essential property that all methods share that identifies them as a scientific method. Instead, I am suggesting, as spelled out in chapter one, that methods have many different properties and that scrutinizing those properties in specific contexts to see whether using the method would promote reliable learning or not, by ruling out specific errors, etc. is essential for assessing whether or not the data produced using that method provide a severe test of a hypothesis, i.e., is evidence for or against it in a specific context. This is very different from Giere's naturalist approach.

Giere relies on weak and purely sociological criteria of group self-identification and recognition for identifying a scientific explanation and a recognized science. Giere's guiding evaluative principle appears to be: go with whatever the scientific community accepts. But would

we want to? The real danger in holding a "sanctioning attitude" of relying solely or heavily on recognized sciences/scientists as the only source of evaluation of science and scientific explanations is that this attitude could simply shield the scientific community from criticism. The history of science is littered with examples of how shielding scientific explanations from criticism leads to epistemic disaster. The Piltdown man fraud and the n-rays fiasco are but two examples of the unreliability of scientific evidence bereft of criticism until exposed by other scientific and non-scientific communities. And to give up on criticism of science would indicate that Giere has given up on objectivity.

2.2.2 Whose science, whose scientist? Science is not a monolithic enterprise—there is almost always dissent, especially at the frontiers of research. John Worrall (1992) points out, deciding which scientists to ask presupposes that one has already been able to identify which scientists or group of scientists have the correct explanation (to use Giere's vocabulary) or are doing things correctly, however we may want to cash that out later. But then to choose which scientist or research group to look at requires we have some extra-scientific criteria or principles to make this determination. But the reason we are looking to and simply reporting on what scientists say is because philosophers are supposedly unable to come up with any criteria for such determinations to begin with so the door to relativism is wide open. You have your favorite scientist (science) and I have mine!

Worrall further points out that it seems odd to say that "philosophers have criteria for evaluating the correctness of 'naturalised', descriptive accounts of science but no such standards for object-level science exist" (Worrall 1992: 342). This is because in choosing which scientists or sciences to look at presupposed some evaluative principles for making such a determination. On the other hand, the fact that philosophers are simply out describing scientists suggests that he

must think that scientists have criteria for evaluating evidence and inferences and philosophers do not. But as Giere wants philosophers to be more scientific then they would need access to these scientific criteria in order to achieve Giere's goal of naturalizing philosophy so there seems to be some tension in his position here. This leads me to question whether even Giere believes that the only naturalist tool we have to assess the status of scientific explanations is whether or not they are sanctioned by scientists.

2.2.3 Changing types of scientific explanations Giere suggests naturalists can do no better than to follow the scientists because when we look at scientific explanations, for example:

> …naturalists know that what counts as a scientific explanation changes over time. For most of the seventeenth century, for example, mechanical explanations appealing to action at a distance would have been rejected. In the eighteenth century, after the impact of Newton's *Principia*, such explanations became common place. Ultimately, naturalists can do no more than follow such historical developments (ibid:55).

Now Giere just blithely asserts that because scientific explanations change over time that naturalists can do no more than act like journalists and record the history of science as told them by scientists (or historians). Why is this suddenly thrown into a discussion where he is trying to claim that we cannot separate science from non-science? This seems to me to be a red herring. Just because the types of explanations change and evolve overtime, by itself does not mean that it cannot be determined that there are better, more objective methods to direct these changes, as well as patently wrong ways to assess how to change explanations. For example, given the predicative ability of Newton's laws, one may be willing to overlook the problem of apparently occult action at a distance, at least to begin with. As Giere himself points out in his footnote to Darwin as an exemplar of explanations, the fact that the mechanics of inheritance weren't entirely worked out while acknowledged as a problem, did not mean that other parts of the theory could not be reliably tested. Again, if we take a piecemeal approach to science, then we

do not accept the whole theory, only parts that have passed severe tests. This is because passing a severe test provides strong evidence that at least about the phenomenon tested, that part of the theory has gotten it correct.

Let us also note how his discussion above sounds as if there was just a sudden gestalt shift between the two types of explanations as if there were no discussion nor experimentation going on between these two centuries that made such acceptance and change in explanation types reasonable or justified. Nothing could be further from the truth.

The fact is that concerns were raised about the apparent action at a distance required by Newton's law of gravitation. These were raised by Cartesians, who claimed such actions were occult and instead they championed a theory of vortexes (think little tornado twisters or water going down a drain) to explain gravity and magnetism and debated in the scientific (natural philosophy) journals of the time.

Now just because explanations change over time, this does not mean that we cannot distinguish between good and bad explanations or better and worse practices. In fact, one would suppose that as practices became better, explanations should change over time to reflect our better ability to test and refine them. And as explanations changed and allowed for new ways to manipulate or intervene in nature, (e.g., using inventions based on new theories such as Newton's), then we would expect both scientific practices to improve and scientific explanations to improve as well. And as scientific inventions, with or without a backing theory, were played with, we again would expect to find knowledge expanding and changing. These sorts of feedback loops and changes in knowledge and technology do not mean that we cannot distinguish good and bad practices at particular times rather they seem to be one of the reasons for the ability to make such discernments.

Let us note that scientists were uncomfortable with action at a distance--and still are hence the search for a carrier, either gravity waves or gravitons that continues to this day! Nonetheless, Newton's theory has stood up well to testing (beginning with Newton himself as well as Cavendish 1776, through to today where launching rockets to the moon are routine tests of his three laws of motion.) Parts of Newton's theory have broken down, and it is now considered a special case of Einstein's more general theory of relativity. However, the fact that we now have a better more general theory that handles very fast and very gravitationally dense objects better than Newton's suggests that our methods are allowing us to learn more and improve and expand our knowledge—it does not suggest that we cannot evaluate those methods and whether changes in explanation types are warranted or not, *in specific cases*.

2.3 Naturalist Criticism Giere Style. Giere is promulgating a critical naturalism:

> This does not mean that naturalists cannot criticize scientific practices. Such criticism, however, can only be based on common sense or on a critical understanding of *other* scientific practices, there being no extra-scientific basis for any other sort of appeal (ibid. 55).

As happy as I am to see a fellow naturalist enjoining criticism, his sources for it, unless filled out seem rather odd. What does he mean by common sense? Giere's "common sense" is a very vague and ill-defined term. Which other scientific practices should we call upon and when? And, given we are criticizing a scientific practice to begin with what warrant do we have for assuming the other practices, those on which we are basing criticism, are not themselves equally open to the same or similar criticisms?

2.3.1 Criticism from other sciences, some concerns. Condemning work on the grounds of not meeting standards of objectivity is not unusual in the sciences. Condemnation may be for using unreliable methods or for allowing personal biases and preferences to color one's work or for other reasons. The existence of (actual and potential) bad practices is particularly obvious

where scientific evidence is required for public policy purposes (for numerous examples see MacGarity & Wagner (2010); Michaels (2008).) But such condemnation is not necessarily based on an understanding of other sciences—for example; I do not need a well-developed psychological theory to criticize specific cases of industrial (e.g., diet pills, tobacco) research because it is biased. While, psychologists tells us that all scientists as human beings are biased, such knowledge does not tell us anything about whether or not, in the specific case at hand, if bias has entered in such a way into the inference or test procedures so as to cause a mistake. (After all if bias is present but does not affect the inference or only serves to strengthen it, then it is not a problem, so no need to worry about it.)

Now if we understand Giere correctly, then naturalists would want to call upon 'real' scientists or scientific research rather than mere pop psychology in applying psychology to philosophical or scientific problems. So, should I call upon a Freudian or a Jungian or a cognitive or a behavioral psychologist—all of whom disagree vehemently with one another's theories? How do we choose a scientist or theory for our philosophizing? As Worrall discusses (ibid. 343):

> She [cognitive scientist] develops cognitive psychological theories about her scientist-subjects—theories she regards as well-supported by the evidence. Suppose now there is a rival cognitive psychologist who holds that different theories are supported by the (same) data. Must we simply record this difference of opinions, there being no way of judging one cognitive theory as better justified by the evidence than the other? …Naturalism entails relativism.

For Worrall, naturalists who claim to be simply "following the scientists" or "describing science" in choosing which science or scientists to describe or follow are engaged in a philosophical exercise that ends up in either to vicious circularity or relativism.

To make matters worse, the sciences he calls upon—psychology, cognitive, social—are perhaps the most roundly criticized of all the sciences, especially for having little, if any, consensus, poor testing practices, weak (if any) evidence for their claims, and almost no

predicative ability (Popper 1957, 1980). So, it seems rather odd then to use them as a source of criticism for other, more established sciences like physics or chemistry. Perhaps Giere will want to say that he means to draw on common sense psychology (like his earlier ordinary belief psychology), which brings the legitimate source of naturalistic criticism to common sense.

2.3.2 Common Sense criticisms—some concerns Giere suggests that common sense provides a legitimate source of criticism to scientific explanations but this raises several concerns. First, he never tells us what common sense is or how it works to help secure objective inference. I guess it is to be left to our common sense? Second, it is almost a truism that common sense tends to be at odds with many scientific explanations—that the earth is spinning rapidly about its axis does not fit with common sense, that thoughts are biochemical electrical stimulations doesn't fit common sense experience, and these conflicts are just in the macro world. Moving into relativistic and quantum mechanical worlds as my undergraduate physics teacher told the class—it is often best to leave common sense at the door.

Third, common sense will often counsel us to listen to and follow authorities, but this will hardly lead to a robust sense of criticism. Now, Giere may claim that common sense tells us that one should not trust people who have a vested interest in the outcome (or conclusion). But, all scientists to one degree or another have a vested interest in their outcome. So is that generalization really useful? No, unless we can figure out how their interests have or could enter into, and, affect their tests or data as evidence[23] for and so negatively influence the inference in question. Maybe common sense will counsel us to listen to the opposing side. But, those taking the opposite side often have an equally vested interest in the opposite conclusion and hence could be equally untrustworthy or biased. The problem with Giere's advice to use common sense

[2323] There are many parts/facets to an experiment and at every level there is a potential for bias and other human errors to creep in unnoticed unless we have the tools to keep them out or to alert us to their presence or in some manner account for them. See Mayo 1996: chapter 7 n the hierarchy of experiment.

is that while common sense may raise flags and suggest caution in general, it seems incapable of operating in any reliable way to indicate specific problems or errors in scientific reasoning.

So while psychology and sociology may be useful scientific sources for learning about "how' human beings as individuals and groups 'reason' (the mechanics of thinking), they do not distinguish good reasoning from bad (the contents of reasoning). This it seems to me is the role philosophers have traditionally fulfilled.

2.4 Giere's methodological turn Giere thinks one of the main challenges to his positive characterization of naturalism is that "it is difficult to defend against the charge of simply begging the question against all those who would appeal to the supernatural, or, more likely, to a priori principle." This seems to be the worst sort of straw man opponent. Skeptics are the real philosophical challengers demanding justification for naturalism in philosophy of science. He suggests that many would-be naturalists fall into the "self-defeating trap of trying to provide a transcendental argument for naturalism" (ibid.: 55). To avoid this dilemma, as Giere sees no way to provide a naturalistic justification for his naturalism without begging the question, he takes what he calls a methodological turn—to practice naturalism and defend it "as a method, not a doctrine" (ibid.:55). His general formulation of this turn is:

> **Giere's Methodological Turn:** For any aspect of the world, seek a naturalistic rather than a supernaturalistic explanation.

To him, the virtue of taking such a stance is that it does not require one provide a transcendental justification. Actually for Giere because it is a methodological injunction or suggestion and not a thesis, it does not require justification at all. (Commands and advice are not true or false, so don't need justification.)[24] But this is a really odd stance especially given his proto-scientific approach

[24] Commands and advice can, however, be criticized but not in reference to veracity, only effectiveness, or sincerity. And, indeed, Giere only criticizes Voodoo, prayer and other "supernatural" methods based on what he claims is their ineffectiveness compared with Western medicine. But unless we can explain why they are so ineffective (and why

to philosophy. Scientists are always asked to justify their choice of methodology—why this cut-off and not that, to use a blind or a double blind trial, commitments to replication and randomization and criticisms for not implementing those methods, and the list go on and on. Giere's idea that somehow turning all of one's theses and positions into methodological injunctions or suggestions allows one to avoid justifying their choice of method is just plain wrongheaded. Moreover, for someone who is all fired up against the focus on language found in traditional philosophy of science, he seems to have mastered the linguistic games he despises by disguising theses as methodological commands to avoid justification. That is a sleight of hand if I ever saw one.

Now Giere does say that commitment to his naturalist method can be "somewhat" justified based on past successes such explanations have had. Okay, but what will count as a success for him? Elsewhere, Giere has stated that the success of science is obvious. He also asserts that no one can argue with 300 years of success referring to science. But Worrall, rightly, points out that just because something is obvious does not mean it does not have to be defended. "In asserting that it is 'obvious' that scientific ways of proceeding are superior to alternatives, Giere is in effect simply assuming what he is intending to demonstrate, while pretending that it is no sort of assumption at all" (Worrall (1992: 348)). Thus, Giere is engaged in circular reasoning unless he can argue that his method (naturalism, which is based on science) has some properties

---

Western medicine is effective), then it could simply be the case that we are not correctly following the method of prayer, voodoo, etc. or that our sample of cases as we (he) are coming from a western perspective are biased—that prayer and voodoo are wonderfully effective, so much so that those who practice these methods right very rarely, if ever, even have to step foot into a doctor's office or hospital. The successes simply go unrecorded. Further, and more curiously, we do have naturalistic explanations for (some) successes of prayer, e.g., optimistic people recover, mind over matter, that is psychological states do come into play in some aspects of disease outcomes. (This is one reason for the need of placebos and double blinding in medical trials.) I, for one, think it is better to examine the properties of methods and criticize them based on how those properties help or hinder the achievement of the goals for which they are being used, rather than to hide behind methodological injunctions to avoid justifying our choice of methods as Giere does.

that make it superior than others for reaching specific goals. But that will require him to analyze his method rather than who use it.

Finally, Giere claims taking his naturalistic methodological turn allows naturalists to criticize non-naturalists even though he admits that it is entirely biased in favor of naturalistic explanations. Though he doesn't attempt to justify this stance, he does try to explain why he thinks this is non-problematic when he talks about his Naturalistic Priority (next section.) Of course non-naturalists can criticize right back at the naturalists, though if naturalists have taken Giere's methodological turn above, they apparently can simply ignore them!

Giere does not mention whether such a stand will allow different schools of naturalism to criticize one another, which would seem crucial for a critical naturalism if it were to be self-correcting. Instead, his naturalistic injunctions for method seem quite feeble. First, it looks like he is merely setting up an apparatus that will drive out those championing a theological explanation for the world but not give any guidance for judging the variety of competing naturalist schools already out there. Importantly, his rules also shield naturalism from criticism—not just the hokey kind but sincere criticism whether it is from other naturalists or those following an a priori path.

Let us move on for now and get acquainted with some of the specific maxims he has formulated for his methodological approach to naturalism.

> **Naturalistic Priority:** "The availability of a *naturalistic* explanation of a recognized phenomenon renders unnecessary any non-naturalistic explanation" (57).

He admits this maxim is not neutral between naturalistic and non-naturalistic explanations nor does he want it to be. It is "part of the stance for developing a thoroughgoing naturalist approach to understanding the world" (57). He notes that the above refers to 'available' scientific

explanations. Here he is not referring simply to logically possible explanations, as that would be too weak, and on the other hand, he is not demanding that the available explanations be fully developed ones as that would be too strong a requirement for scientific explanations to have to meet.[25] Instead he defines an 'available' explanation as "one that, relative to the science accepted at the time, is plausible enough that something not too different is likely eventually to prove correct" (ibid.)[26] He admits it is vague but again, he re-iterates his claim that this is about the best one can do. This is the best we can do?!

Just as for explanations, once again for philosophers the real question of interest is *why* did a sociological group accept an explanation as plausible *and* were they justified, i.e., had good reasons, for doing so? Because Giere believes that naturalists should accept what they see scientists accepting, he is probably satisfied that the entertainment of an explanation by the scientific community provides sufficient naturalistic evidence for plausibility. But, we must ask ourselves—do scientists really appeal to this type of criteria? How would a scientist apply this rule—*accept explanations that are plausible insofar as something not too different from it will prove to be true*? Surely, this could only be truly warranted in cases where we've circumscribed the area in question very well, for example, the inverse square law of electrical attraction[27]. Given the big changes in explanation types Giere has documented, this rule could not work in the past. And even if we think the theory or model in hand toady is close, then which part of it needs adjusting? Giere seems to be demanding that scientists are fortune tellers and know what the true future theory will look like so that they can judge which of the available explanations are close to it.

[25] He points out that Darwin's evolutionary theory of natural selection is an exemplar of a scientific explanation even though having huge holes prior to genetics being available to provide a mechanism of transmission.
[26] Apparently, there is no "punctuated equilibrium" in the evolution of scientific explanations!
[27] See Miller 1998. This local (or laboratory) law has remained & been measured under multiple changes in large scale electrical theory (from Henry Cavendish 1776 up until the present day).

Does the method Giere is recommending above have a good track record? Is it a good method? No, for one thing following it would seem to rule out the possibility of scientific revolutions. For example, Einsteinian physics was hardly plausible to the Newtonian physics of his time, and ditto quantum mechanics, which Einstein did not see as plausible at all. So quite clearly Giere's method will not work, for one would be hard pressed to find a more respected scientist than Einstein but following Giere's method, then quantum mechanics would not be able to get off the ground.

I think a better candidate for determining plausibility lies in a more narrow range. This is the view that a new explanation must be able to account for well-known and severely tested experimental effects and laws. For example, the inverse square law of electrical attraction and repulsion has remained throughout 234 years and three paradigm shifts in electrical theory.[28] If a theory cannot handle well tested experimental effects, then we would say it is not plausible. But otherwise, I think all sorts of theories are plausible. New theories tend to be implausible early on—e.g., early string theory, punctuated equilibrium, plate tectonics— all considered radical given the paradigm of the day. (It would be more correct to say new theories are not judged on their reliability early on for their proponents probably consider them quite plausible.) If scientists were using Giere's method above, then they would not have considered these explanations as candidates for being scientific explanations. So, this does not seem to be a very good methodological rule. Nor does it seem to qualify as a naturalistic one, that is if we are going to follow the scientists as Giere suggests, for clearly it is one they do not follow, unless we rewrite plausibility as I suggested above and ditch Giere's vague idea of plausibility as being an

[28]First measured by Henry Cavendish 1776, the inverse square law remains part of electrical science today having been incorporated under 1 and 2 fluid theories, particle and field theories of electricity (see Miller 1996).

explanation that is "something not too different [from currently held beliefs] is likely eventually to prove correct" (ibid.).

This brings up the issue of whom and what determines plausibility for Giere? He claims it is the research group that will determine whether an explanation is plausible or not. But which research group? If we go back to his 'evolutionary' explanation, then it's the one that gets either the most converts or the best press or what have you. The one thing Giere does not require is that the group has rigorous testing standards that they require any future plausible explanations to survive. If they did and they tested an explanation, then it would seem to carry weight. But this is only because it was well tested, not because of some vague notion of plausibility to future truth. But just as in the argument above, the method Giere is espousing here by forcing explanatory imagination to stay close to current theory would seem to stifle inquiry rather than promote it.

Fortunately, in the second half of his CHEN article, Giere shifts gears away from his discussion of explanation and plausibility, which seem hopelessly vague to me, to some more concrete naturalistic epistemological norms. Here he takes a Laudanesque turn giving it his own modeling flair. Let us see if this later discussion provides a better framework for naturalism.

2.5 Conditional norms & means/ends naturalists. Giere states that naturalists cannot provide categorical norms—norms that proscribe actions unconditionally.[29] But he does not take that to mean naturalists cannot be normative. He, like Laudan, champions naturalists' ability to put forward conditional norms of the type:

Giereian* C*onditional norm: "If you want to achieve [goal] G, take [action] A."[30]

[29]Note he says naturalists do not require categorical norms of "thou shalt do science" but yet, his naturalistic priority says that one shall always choose a naturalistic explanation and as for him scientific explanations are the premiere type, this priority pretty much with replacement would be reads as "thou shalt do science". Circularity appears to be rampant in the norms he formulates for his naturalism.

[30] Norms following the pattern above, he believes will allow him to cash out the empirical success of science. We may also try to see this as a better, more objective way to implement his notion of plausibility... maybe.

Before he embarks on normative epistemology, Giere is clear about the types of epistemic things to which he thinks these conditional norms apply. He wants to get away from the view that scientists deal with "linguistic entities where the connection between statements and world is captured in notions of reference and truth (58)." Instead, he will focus on models where the "connection between models and world is not truth but similarity, or fit" (ibid).

> By fit, Giere suggests a Campbellian reading where:
>
> Individual models then fit the world in something like the way individual organisms fit their environment. This analogy at least makes it clear that the fit of individual models to the world is multidimensional and that the fit of an overall theory is ever changing and never perfect" (ibid.: 59).

He states he is not claiming anything more for the analogy but if we go back to his statements about sociology and research groups, I think he is. In particular the analogy would be just as an organism survives its environment and reproduces, so a model must survive its research group and reproduces itself into other research groups—spreading like bunnies, or cancer.[31] But is this survival the only check of a misfit between Model M and the world?

2.5.1 Giere's comparativism Giere wants his discussion of comparative fit to allow him to cash out and explain his take on "the problem of providing a naturalist justification for epistemic norms." He does this by way of an example—a crucial experiment between two rival models, he calls it. The experiment (T) is assumed to have an observable output (**x**), which falls into a one dimensional range R of numerical values.[32] The models and experimental set-up are related thusly:

> (i) If the model M_1 provides a good fit to the real world, then it is very probable that the experiment will yield an outcome in the range R_1 and very improbable that it will yield an outcome in the range R_2.

[31]He does not claim that fit here leads to reproductive success for example. Indeed elsewhere he says an argument can be made that science and its models decrease reproductive success (e.g., invention of atomic bombs, etc.).
[32] This is a standard comparativist test, e.g., as defended by Laudan (see for example Laudan (1997).

(ii) If the model M_2 provides a good fit to the real world, then it is very probable that the experiment will yield an outcome in the range R_2 and very improbable it yields an outcome in range R_1 (ibid).

This leads him to develop the following decision rule:

> **Comparative decision rule:** If the setup yields a reading on the range R_1 chose M_1 as best fitting. If the setup yields outcomes in R_2, choose M_2 as best fitting (ibid.)

He translates this experiment into the following epistemological norm:

> **Giere's Model testing norm:** If one wishes to decide empirically which of two rival models better fits the world, design an experiment satisfying the conditions stated above (ibid.)

The justification for the usefulness of this norm, according to Giere, is that an experiment satisfying these conditions will provide a basis for a reliable decision between the two models. Will it?

Let us ask ourselves if this is a reliable or severe test. We first note that his two models are not required to exhaust the model space of possibilities.[33] Thus the test may or may not be informative, whether or not either of the models passes or fails it. (See chapter 3 for a fuller explanation.) Now he admits that there is always the possibility that neither model fits the world very well, in which case the experiment is inconclusive. He does not tell us, nor does his set-up above suggest how we will know if the experiment is inconclusive? But what guarantee does his norm and method above give us that such a lack of fit will become apparent to us? What caveats, if any, does he put on his comparison rule about accepting the better fitted model? Below, I provide a counter-example to show just how weak and unreliable, (insevere) his 'crucial model test' is. The problem here has been identified by Mayo as a general problem with other

[33] Hence it is open to Duhemian under-determination—another known or unknown model may fit better than either one. This may be why Giere seemed so comfortable with historical contingency of explanations—going another historical route, we may have found another model instead (see my discussion later in this chapter)S. See Mayo ( ) for why comparativist tests of this sort are prone to many errors, especially of the Duhemain type.

versions of comparative testing (e.g., Laudan) is that the comparativists' best tested model does not mean the model has been well tested at all (1997).

2.5.2 Best tested does not mean well tested. There are two ways to understand this claim. First, the comparativist test may simply be a bad test and neither model has been probed well for flaws. Here we can imagine a drug trial in which one drug A does better than another drug B at curing a disease but neither one does better than a placebo. So drug A compared to drug B is better—but that is not to say much. Neither is very good as a placebo beat them both! Secondly, the model that is comparatively well tested may not have been well probed for flaws. Mayo often refers to General Relativity, which many philosophers, including Laudan, would like to accept *tout en court* on the grounds that it is comparatively best tested of all rivals. But, Mayo (1996, 1997, 2010) argues, so long as there are areas in which it has not been well tested,[34] (probed for errors), then accepting the entire theory is premature and unwarranted. Turning again to a drug trial example, while drug A may have a much higher efficacy rate at say curing cancer than drug B, it could have potentially devastating side effects (like death) that have not yet been tested. This is a flaw in drug A that a comparativist test would not have to probe before according Drug A the status of best tested. But, tests about one part of the model need not rub off onto other parts, e.g., the drug's efficacy may be great but that does not mean that the drug's safety is equally great. Yet, that is what comparativists are assuming when they want to accept the entire model before it has been severely tested in its entirety. Laudan and others disparagingly call this piecemeal approach to accepting only parts of a theory that have survived severe tests as "balkanization" (e.g., Laudan 1997). Mayo, on the other hand, emphasizes that it is precisely the scientific attitude of continuing to test both un-probed areas as well as refining and making more esoteric (pace Kuhn) tests of previously tested areas that is really the engine behind scientific

[34]This would also include areas that have not even been tested at all.

progress—of learning about the world and our models, rather than merely resting on our laurels and rushing to accept some large-scale theory.

The example she uses to illustrate this piecemeal approach to testing is Eddington's eclipse experiments between Newton's (N) and Einstein's (E) theories of gravitation. Using photographic plates and a fortunate solar eclipse, Eddington was able to measure the deflection of light from nearby stars around the sun. Here is a case that seems to match up nicely to Giere's layout above. N predicted deflection effect to be R_n and E predicted it would fall in R_e. 2 out of 3 locations reported R_e and only one set of plates (Sorbal) reported R_n. According to Giere, Einstein's model wins, has the best fit, so we are done, right? Notice, if we take Giere (and other comparativists) seriously, then we really should have accepted all of Einstein's GTR after Eddington. Mayo argues that what has passed a severe test is the angle of deflection of light, not the entire theory. Certainly if we follow Giere, scientists should not have been constructing straw-metric models in the 1970's to probe it for how GTR could be wrong, given we've already accepted it as true or well fitted, etc. (see Will, Mayo). Once a large scale theory has passed a local test about one or maybe two of its predictions, the comparativists are ready to accept the *entire* theory as passing. And, then, it would seem natural to no longer work on it. But we can see this violates our minimal severity criteria for it is a very unreliable method for assessing theories. The history of science is littered with examples of how initially successful theories—and indeed longstanding successes—when investigated in esoteric depth were found to be false and unreliable in a variety of contexts (e.g., Newton at super speeds). If the goal is learning about the world, then cutting off inquiry prematurely is not the way to go.

This continual pressing forward and learning model is reflected in the actual practices of scientists—you know the ones we are supposed to be following. After Eddington, those killjoys

only let us accept the Einstein is a better model for deflection and that any competing model will also have to get the same (or close to it) measurement for light deflection. Also, see Mayo (1996), they were thinking up errors and other effects that could save Newton (immediately after the Eddington experiments) and later, thinking up other ways to prod and probe Einstein even if it means thinking up a whole zoo of alternative theories/models (See Mayo, Will).

2.5.3 External Validity This brings up the question of whether or not Giere's decision rule above is a reliable one. And here it is important to emphasize that his norm is: If one wishes to decide empirically which of two rival models better fits *the world*, design an experiment satisfying the conditions state above [e.g., (i) and (ii)]. It is also crucial to note that when Giere talks about a model fitting, he means to the world, not to the experimental data. He is making a radical induction in extrapolating from fitting the data from a single experiment to fitting the world!

The point is the best tested comparatively gives us no assurance that a model has been tested, or put through the wringer, at all. This is because we need more than "fit" criteria to ensure that even one hypothesis has been reliably or severely tested by which I mean probed for how it would be a mistake to take it fitting some outcome as evidence that the hypothesis has correctly captured (some aspect of) the phenomena generating that outcome. For a theory to be well or severely tested requires that the test can probe the theories (or models) for flaws or errors. That is, we want information not only that the hypothesis fits the data but how probable would that fit be if the hypothesis were false. If we do not assume that the test qualifies as Mayo-severe, then there is not even an implicit restriction in considering what level of model is being tested.

For example, is the model the Newtonian model of the universe or the Einsteinen model or are the models simply the predicted light deflection that are being tested? That levels of

hypothesis or model are important for determining severity and hence for what (or how much) can and cannot be inferred in comparing two models, as we saw in the previous discussion of Eddington's experiment.

2.6 Giere and the comfort of naturalism. Giere rules out supernatural explanations but what reason does he give for choosing scientists as having better explanations of the world than other groups of people? If he cannot provide an epistemological rationale or principle of evaluation for how scientists reason and why their approach is superior or underwrites reliability, then such choice is either supernatural or arbitrary. Giere assumes naturalist explanations of the type given by the sciences are the best and then he simply begs the question in defending his position that philosophers should take this naturalist route too. Indeed, rather than seeking a naturalized philosophy of science, he seeks a philosophy derived from science. *Giere is really promulgating a turn to scientism not scientific philosophy or naturalism.*[35]

We must also note that his approach doesn't seem able to give us forward looking norms or methods, which is what we really want. At most his account is an evolutionary version of rational reconstruction but he has not given us a reason to believe that scientific methods or knowledge are responsible for our survival.

Giere concludes that naturalism is a viable and all encompassing project—for it includes all of nature including culture and the enterprise of science itself. Of course, he could not conclude anything else given his naturalistic method and his naturalist's priority both which declare by fiat naturalism is the best method for philosophizing. However, he says one must be careful not to present it as theses which would require justification and end up in "self-defeating

[35] See the 2007 interview where he talks about not recommending that his university hire a fresh PhD in philosophy of language because by the time that person is ready to retire, that entire field will have been taken over by psychologists/cognitive scientists and not even be a philosophical field anymore. He regards this as a positive outcome of the turn to naturalism.

attempts at transcendental arguments in its favor. Vigilance is required to keep arguments for naturalism within naturalistic limitations." So he says the best way he sees to do this is to regards it as set of methodological rules for developing a consistent naturalistic picture of the world. Labeling naturalist theses as "methodological injunctions" does little to belie the circularity of his position. However, for Giere, "[s]uccess in applying these rules gives comfort to those pursuing the program and encourages other to join." To couch this in the formulation of a means-ends norm that he enjoins naturalists to use, then:

> If your goal is to take comfort in your pursuit of naturalism and get recruits, then you should apply Giere's rules for being a naturalist.

Positively that is the most we can hope for from his critical hypothetical evolutionary naturalism. But is this really the most naturalists can hope for?

I agree whole heartedly with Giere's advice to become a methodological naturalist. However, I see the shift to a focus on methodology not to provide a reason to avoid justification but instead to provide empirical justification or warrant for accepting (parts) of theories and for understanding the objectivity and reliability of (some) scientific knowledge. Both objectivity and reliability are based on cultivating methods for the detection and elimination of errors, Kitcher, although he has much in common with Giere, also proposes a criterion for objectivity that transcends historical contexts though it is rooted in the nature of human beings.

2.7 Kitcher's plea for a return to traditional naturalism Kitcher also wants to re-introduce psychology into philosophy without sacrificing the traditional normative role held dear by philosophers of science. Like Giere, Kitcher emphasizes that the new naturalists *deny* that epistemology is to be based on logic and *a priori* principles:

> Both the reintroduction of psychology into epistemology and the suspicion of the a priori are well supported and…[there is] an important connection between them. Specifically for knowledge claims (aka justified true knowledge) not only true but

> the knower knows it's true for the right reasons, this requires: perceptual knowledge depends on the right kind of relation between the knower and the facts known…generalized references to the characteristics of the psychological mechanisms of subjects (1992: 61).

That is, naturalists argue that epistemology, if it is to be relevant to how real humans learn about the real world, must capture an empirical relationship between knower and fact. He announces that "the need for psychology is hidden when we detach subjects from decision-making contexts" (*ibid.* 87). One psychological fact that is at the heart of Kitcher's philosophy of science is the Quinean view that what we accept as knowledge is based on what we have previously accepted as knowledge. Knowledge is a historical entity. Also, in concert with Kuhn, Kitcher feels we must "use performances of past and present scientists as a guide to formulating a fallible theory of confirmation and evidence." So far this seems quite similar to Giere's position

2.7.1. Kitcher & the eliminativists Kitcher, unlike Giere, does *not* think that eliminativist projects are needed when introducing cognitive factors because "the goal of naturalistic epistemology and philosophy of science is to understand and improve our most sophisticated performances, and about this, eliminativists have presently very little to say (*ibid.* 87)." Kitcher notes that cognitive psychology can treat the use of images and tactic knowledge while retaining traditional epistemological categories, so we should try to enrich rather than eliminate our vocabularies. Kitcher wants to encourage naturalists to return to traditional naturalism—one that re-introduces psychology (and sociology) into epistemology and recognizes knowledge as fundamentally historical and socially embedded, but still seeks to fulfill traditional normative functions and uncover stable (rather than merely relative) principles of method and epistemic appraisal.[36] How is this to be accomplished?

[36] See Roth (1999: 91) for a good discussion on radical versus traditional naturalism.

Psychology and sociology in philosophy of science take the form of decision theory for Kitcher in his own work. For Kitcher, what provides objective constraints on the judgment and choice of explanations by both individuals and groups is the shared goal of explanatory unification. Regardless of which decision theory one implements, Kitcher's, Goldman et al's, Subjective Bayes, Bayesian Information Criteria (BIC), Akaike Information Criteria (AIC), and so on, objectivity is achieved to the extent that the theory used facilitates explanatory unification.

2.8 Reliabilists' projects Kitcher finds reliablists' projects congenial to his version of traditional naturalism:

> The most prominent contemporary versions of naturalism formulate the meliorative epistemological project in terms of enhancing the reliability of the cognitive processes we employ…a process that confers justification is reliable in the sense of belonging to a type that generates true beliefs with high frequency. (1992: 65).

Kitcher admits this is a vague standard. Goldman, *et al* focus on the reliability of "belief generating processes where frequency concepts report on the property of more frequently generating true beliefs over false" (*ibid.*). How do we know that a process has generated a true belief in the first place and how often, we may ask, has it produced false ones? Wouldn't Goldman first need a method or test to determine which beliefs are true and which false? Wouldn't Goldman also need a way to assess the reliability of these belief generating methods for the goal of producing true belief independent of the frequency of (supposedly) true beliefs? The reliable way to do that, it seems to me, would be to subject their beliefs reformulated as specific local hypotheses to severe tests. At least that way they can try to weed out hypotheses/beliefs that are false. This is not the route Kitcher will take.

For Kitcher, a virtuous state is truth, but not any truth, only significant truth. As a naturalist, he feels he is in a good position to articulate an objective standard of finding

significant (to humans) truth. Explanatory unification is the key to finding significant truth, Kitcher claims. It is the key to unlocking the nature of self-correction in science both for individuals and institutions.

2.9 A method for objectivity: explanatory unification Kitcher's objective standard is:

> Given the nature of the world, of the beings in question, and the kind of representations that are sought, there will be determinate answers to questions about how it is best to proceed and hence an objective epistemological standard." (Kitcher (1992: 101))

He does not accept mere pragmatic goals (after all, this not what scientific knowledge is all about according to him) including "fit" as Giere desires. Kitcher wants not only prediction or control but understanding. He says this can be interpreted as realism or causal explanation but also (and clearly he prefers to interpret it so) as *explanatory unification*. Though all of these goals or aims, etc., are different, he claims "they agree in taking as the goal of inquiry the production of a certain type of structured account." He says, in short, this structured account can be put into Aristotelian terms as finding the "order of being" (*ibid.*106). How do we know we have truth?

2.9.1 Explanatory unification for individuals In *The Advancement of Science*, Kitcher attempts to show that underdetermination and relativism as made out by the likes of Shapin and Schaffer, and Collins are unwarranted. He does this by looking at some of the specific examples they examined and argues that they have not made their case. His main point is that relativists can only make their case if internal consistency is the only standard to which we hold an explanation. But this is simply a false view of how humans produce knowledge according to Kitcher.

For Kitcher, "contemporary scientists acquire their languages and conceptions of explanatory dependencies from their predecessors" (1993: 24). Explanations must take this shared background into account, internal consistency is not enough (*ibid.* 294-297). Explanatory

unification determines that new knowledge unifies existing knowledge (perhaps by re-interpreting or replacing parts of explanations) with new knowledge (e.g., Newton as a special case of Einstein.) The historical web of knowledge (*pace* Quine) provides a shared background of knowledge against which new theories and explanations are assessed. Kitcher points out several "local-experimental" ways to break out of explanatory deadlocks. One is to "establish the reliability of an instrument (or technique) by connecting its performance to procedures that can be carried out *independently;* one shows the dependence of particular variations in the performance of the instrument on changes in the design or manufacture which can be understood by deploying independently accepted schemata" (*ibid.* 296). These are the sort of low-level hypotheses, often functioning as informal error checks, from which Mayo builds her account. However, it is because these techniques and instruments have survived severe tests that they have been accepted into background knowledge, and thus can be used of their reliability to achieve specific ends. It is not because they are accepted background knowledge that makes them reliable. That is to get the relationship backwards and to miss the real import of these methods.

Although Kitcher offers local means of assessment for particular claims, he feels they are insufficient (remember his Duhem discussion) to provide objective constraints across the board. At any time in history, scientists are working within a web of belief, accepted theories, methods, etc. Their shared aim is to produce a structured account of the world and it is this goal of explanatory unification that allows scientists in different camps to objectively compare and hence choose theories.

This goal is constrained by accepted knowledge, though this fund of knowledge varies from individual to individual. This variation doesn't threaten objectivity for Kitcher because for

an explanation to be successful (and accepted) it must be able to unify current knowledge, which is historically contingent. The goal of explanatory unification requires and so forces coherence in individual believers. Admittedly, there will be times when two theories are at a standoff, but Kitcher feels there has not yet been a case where such a standoff remained or could not be resolved by one or another theory's ability to unify old and new phenomena. It is the compelling desire for unified explanation that provides an objective standard for assessing theories for individual scientists but also for institutions. Kitcher also sees social structures as playing a fundamental role in producing scientific knowledge.

2.9.2 Explanatory unification in consensus formation For Kitcher, science as an institution determines the content of knowledge by consensus formation with the aim of forming progressive consensus practices—by which he means practices that produce better unifying explanations. Kitcher attempts to map out the optimum organization of cognitive labor as he calls the coordination, cooperation and competition among scientists for distributing effort within the community (Kitcher (1993: 303-4)). He claims that mapping out the space of epistemically important characteristics of the community can show whether or not our institutions "do a good job of coordinating the efforts of individuals" (*ibid.* 306.) He argues that "there are advantages for a scientific community in cognitive diversity. Intuitively a community that is prepared to hedge its bets when the situation is unclear is likely to do better than a community that moves quickly to a state of uniform opinion." (*ibid.* 344)[37] Now the normative job is to figure out "what kinds of social arrangements might foster welcome diversity" (ibid.).

[37] He uses a decision theory matrix to calculate this as well. There are many different types of mathematical frameworks available for decision theories to use, Bayesian, Akaike,….. Regardless of which decision theory one wants to appeal to, Kitcher is arguing that the real objective constraint across individuals and institutions in science comes from this goal or rule to unify all explanations in both individual scientific domains and within the larger domain of science. So this rule is my focus here.

Interesting in its own right, at least to some people, Kitcher attempts to put Merton's norms into a mathematical framework, both at the level of the individual (decision theoretic) and at the institutional level. Social structures can be evaluated and assessed: "Given the actual social structures present in scientific communities, the input from asocial nature is sufficiently strong to keep consensus practice on track" (ibid. 165). What is interesting to me in Kitchers work is not his decision theory. There are many forms out there as I noted at the beginning of my discussion of him. Rather, it is the role of explanatory unification as objective constraint that intrigues me. The goal of unifying explanations forces consensus practices.

2.10 Explanatory unification—an alternative view Regardless of how sophisticated the mathematical structure Kitcher has devised, a problem arises in that Kitcher is not quantifying the right stuff. Indeed, we could apply his calculations and his decision theoretic approach to many other cultural practices—religions, fraternities and sororities, political organization, etc.,—that attempt to provide and even do give a structured and unified explanation of the world that Kitcher would surely deny are "objective" in the same sense as scientific knowledge is. (Indeed, one of his main goals is to distinguish scientific theories of evolution (objective) from pseudo-scientific theories of creationism (subjective). A variety of structured accounts of diverse phenomena abounds and has throughout history. To put this another way, as a practicing scientist, rather than a Christian theologian, how is one to make sense of Kitcher's method? Clearly not any type of explanatory unification will work. So what types would work as an objective constraint on explanations? I think appealing to the error statistical theory can help us here.

2.10.1 Argument from a miracle. Explanatory unification of diverse phenomena can be seen as a powerful qualitative argument from error to the extent that it seems highly improbable,

indeed almost a miracle, that a theory could unify such a wide variety of phenomena under its explanatory umbrella, if it were false. Surely, one of the many phenomena would disagree with its predications in that case, (i.e., if the explanation were false.) It would be a miracle that a false theory would get it right about so many, diverse phenomena. Each phenomenon about which it makes a prediction can be seen as an individual error probe. Note for this type of argument to go through, the unification cannot have been the result of *ad hoc* results (because then it would not be much of a miracle, but a set of constructed 'fixes'). Here objective constraint is a achieved based not simply on unification but on how unlikely such a successful unification would be if the theory was getting it totally incorrect, was in fact composed of false claims about those phenomena. It is by this type of reasoning that explanatory unification can be seen as a *test* of explanations but only to the extent that the individual phenomena succeed as error probes of the theory in question.

2.10.2 Phenomena as error probes. Unification doesn't count if it is too easily achieved because then it would not provide a test of the theory. As Popper points out, confirmations or fits can be easily achieved and should only count if they are sincere tests of a theory. Mayo interprets Popper's sincere tests as severe tests, in which it would be highly improbable that the data would fit the theory or explanation, (even qualitatively speaking), if the theory or explanation were false.

So, why does diversity count for unification? If the phenomena are diverse, this suggests that different parts/aspects of our explanation are being probed and connected by the individual phenomenon. A widely discussed example of this is that agreement in calculations of Avagardo's number N on thirteen different phenomena is considered evidence that the molecular-kinetic theory as a whole is getting it correct. As Mayo (1996) puts it, this agreement "effectively ruled

out the worry that that extrapolations from one phenomenon to another would not hold up (249).” By ruling out this concern, the unification can be seen as evidence that the molecular-kinetic theory as a whole is getting it correct not only about Brownian motion but also about gases, radiation, etc.

Unification is an effective error probe because, by connecting phenomena, it provides an effective way to check for different errors using the other phenomena. Thus, as Hacking (1983: chapter 11) suggested, different types of microscopes can serve as error checks on one another; similarly we can see the several phenomena as serving as different “instruments” for checking via Avogadro’s number the global kinetic-molecular theory. But in the error statistical explanation above, unification is desirable not in itself but because it provides an effective check on specific types of error—e.g., extrapolating errors, etc.

So while Kitcher is getting it right about the desire and usefulness of explanatory unification, he is getting it right for the wrong reasons. And, on his account, and most others, that is not a very reliable way to go about explaining methods or meta-methods. Though I should note that the key point for Kitcher may well be that by unifying domains, a fundamental set of explanatory patterns can be developed and deployed to cover new domains (and in the process provide new cross checks on what has been done so far). On this view, we have two objectives—unification as a method of discovery or exploratory research, and unification as an explanatory error-probe.[38]

2.11 But what is the real role of naturalism? It seems to me the real attraction of naturalistic approaches is the hope they hold out for understanding how scientific practices underwrite scientific knowledge and for explaining (not merely assuming) the success of science. The problem with science of science approaches is that they tend to end up as Worrall points out in

[38] I thank Dick Burian for pointing this out to me.

either pernicious relativism or circularity. Looking at both Giere and Kitcher's accounts, a similar naturalistic story could be told for almost any other form of human knowledge production. Further, these accounts offer little if any prescriptive advice for how we ought to proceed nor do they offer tools or technique for assessing and evaluating disagreements between 'tribes' of knowledge gatherers without begging the question or surrendering to relativism. While they enjoin us to turn to our species most sophisticated practices, they inevitably lead us into psychology and sociology, which do not seem to be well developed enough for the role being thrust upon them. More to the point, they seem inappropriate for the task. Even if I could compare the brain images of, say, a Newtonian and an Einsteinian, how would that information tell me which theory was correct?

Are all such naturalist accounts doomed to fail? According to John Worrall, the answer is yes. According to me, the answer is a resounding no! So let us take Kitcher at his word that" the goal of naturalistic epistemology and philosophy of science is to understand and improve our most sophisticated performances "and turn now to a conglomeration of tools and techniques from standard statistical practice, experimental design, and so on that we can group loosely under the umbrella of error statistics.

Chapter 3: Objectivity & Frequentist statistical testing

> *The rationale of statistical methods and models is found in their capacity to systematize strategies for learning from data, and thereby for furthering the growth of experimental knowledge (Mayo 1996: xi).*

3.0 Introduction. In this chapter, I take up Mayo's concept of severe testing, which provides the formal, quantitative apparatus of induction in the error statistical philosophy of experiment. In her reformulation/extension of standard (i.e., Fisher-Neyman-Pearson) frequentist statistics, she keeps central the role of error probabilities—hence her calling her philosophy of experiment, error statistics. This formal epistemology is also congenial with the Gierean/Kitcher view that naturalists should look to the sciences themselves (in this case statistical science) for formulating a naturalistic philosophy of science. Severe testing is at the heart of the error statistical approach to objectivity.

3.0.1 Objectivity and errors. As discussed in chapter 1, objectivity on the error statistical account has two components. First, in contrast to subjectivity, objective inferences should not be erroneously influenced by the personal biases or idiosyncrasies of the individuals or groups making them. Second, following C.S. Peirce, the goal of objective inferences is to orient ourselves correctly to the world. So, objective claims must latch onto the world in the right ways. We can define 'errors' as the obstacles that stand in the way of assessing claims objectively, whether from human biases, from the incomplete nature of the data, or malfunctioning equipment, etc.

Errors in claims arise because (1) scientific claims to be genuinely ampliative go beyond the results of specific experiments or observational data, and (2), as Mayo points out,

> "…the [i]ncompleteness of observations, inaccuracies of measurements, and the general effects of environmental perturbations and "noise," cause observations to deviate from testable predictions, even when the scientific claims from which they

> are derived are approximately true descriptions of some aspect of a phenomenon" (Mayo 1983: 27).

How can we get around these disturbing factors? This is where statistical practice is most often called upon—for objectively assessing claims based on observationally or experimentally derived data. But we cannot simply turn to statistics to save the day philosophically because there are multiple and conflicting statistical theories and methods available. Besides which, as epistemologists, we want to understand why methods work when they do and why they fail to work when they do.[39]

Choosing which, if any, statistical theory to use and the methods within them requires that we understand (1) the rationale for their use, and (2) if their "machinery" can provide the means for achieving the type of objective scrutiny of claims wanted in the sciences. In specific cases, we need to understand clearly what the particular hypothesis we are learning about is claiming and how the statistical tools being used can/cannot help us in assessing that claim. The debates between Bayesians and frequentists in statistics is a formal analog of similar debates on objectivity in STS and philosophy over how best to deal with the role played by human subjectivity—biases, desires, etc.—in assessing scientific claims based on evidence and of the new experimentalists' debates on how data achieves or is assigned the status of evidence to begin with. Thus, looking at these formal debates can provide clarity to our own more roughly hewn qualitative debates.

3.0.2. Brief overview. I begin with the frequentist approach, first explaining how it applies probabilistic ideas, as they are used to quantify the relative frequencies of events, in order

[39] See *Statistical Science* 18, 2003 for just one example of the debates and controversies surrounding the foundations of statistics. Also, in that issue, a more technical discussion of the ES view, of both the formal debates and objective Bayesian attempts to bridge the two schools can be found in Mayo's "Could Fisher, Jeffreys and Neyman Have Agreed? Commentary on J. Berger's Fisher Address": 19-24.

to bridge the gap between a claim and the incomplete and error-prone data scientists have. This is done by first embedding the material experiment into a statistical model, and then using the probabilistic structure of the latter to assess the relevant error probabilities associated with the inference in question. For example, in the case of testing a hypothesis, one assesses the relevant error probabilities associated with a certain test statistic, often in the form of a *distance* (provided by one or another measure of fit) between the output of a *hypothesized* data generating mechanism (dgm) with known distribution (i.e., what is expected under the null/test hypothesis) and the output (observed data) of the *actual* data generating mechanism that produced the observed data.

Second, I review some common criticisms of this general approach and how they can be circumvented by adhering to Mayo's severity principle. This principle formally lays out our intuitions about what is required for data to play the role of evidence in testing a hypothesis. So even though frequentist statistics often seems to resemble a hodge-podge of strategies, severity provides a general underlying principle of evidence behind error probabilistic thinking that ties these various strategies together and can be used 'meta-statistically' to evaluate their effectiveness for objective inference (Mayo 1996, 2004, 2005.) A common criticism regarding the objectivity of scientific inferences is the reliance on and use of background assumptions in testing. Many in STS and philosophy of science, following Pierre Duhem, regard this reliance on background assumptions as an intractable problem. Therefore, I also provide a brief glimpse at misspecification testing, which is the key for testing the background assumptions (inductive premises) used in standard statistical tests.

Third, I briefly discuss the main competitor to the frequentist approach for assessing scientific claims based on evidence, which is to be found in the various Bayesian schools of

statistics. In these approaches, probability is used to quantify and update an individual agent's degree of belief or confidence in claims and to continue updating it based on a similar assessment of evidence[40] in a rational and coherent manner. I will indicate why this approach does not meet the standards of objectivity laid out above (particularly #2), even given the new push to supplant subjective degrees of belief with "objective" priors.

3.1 Frequentist approach. The noisiness and variability found in the real world objects and phenomena that are of interest to scientists and the data (samples drawn from a population of these things) are well known. Both noise and variability in the data that scientists can gather pose a real challenge for testing hypotheses using it. The question facing scientists, statisticians, philosophers and others is: How do we pull the signal out of the welter of noise, including sources that are unknown to us, so as to be able to claim to have evidence for or against a claim? Mayo supplies the frequentist answer: "by being able to objectively control error frequencies, error statistics (ES)[41] is able to objectively evaluate what has or has not been learned from the result of a statistical test" (1983: 297). In this chapter, we will unpack this statement to get at the underlying rationale for objectivity behind these commonly used procedures.

3.1.1 Frequentist Goals & Elements: The goal of statistical testing for frequentists is viewed: "as a means of *learning* about variable phenomena on the basis of limited data" (Mayo 1983: 298, emphasis in original). The goal is achieved using various methods to "detect discrepancies between (approximately) correct models of a phenomenon and hypothesized ones" (ibid. 299). So to begin, we need to connect up statistical tests to the real world phenomena about which we are interested in learning.

[40] Note, here the fundamental point is that probability is attached to an individual's state of mind about a proposition, which may be a proposition about the world (a hypothesis), or about the evidence so produced. The experiment is assessed only insofar as its characteristics would affect my degree of belief in the evidence, or so I assume.

[41] I replace her NPT* which was an early version of the now full blown and more aptly named Error Statistics.

The key move for frequentists in using statistical tests for scientific claims requires that we frame the substantive hypothesis[42] of interest in terms of the unknown parameters of an appropriate statistical model specified in terms of the random variable(s) underlying the data. This is done by expressing the hypothesis as a restriction on the range of values of the model's parameters. That is, statistical hypotheses are claims about the underlying Data Generating Process or Mechanism (DGM); which sub-model generated the data in question.

Next, we choose a test rule (statistic) that maps possible values of the parameter together with the space of potential observations, called the sample space, onto a subset of the real line. An integral aspect of this approach is to ensure that we choose our test hypotheses about the value of the parameter(s) in such a way as to expose any and all alternatives to it. This is done by carefully choosing our test hypothesis, also known as the null hypothesis, in such a way as to partition the space of alternatives exhaustively. Often the null is a point hypothesis (i.e., a specific numerical measure, say 0), and the alternative is defined as greater than, or less than that number. This way of formulating the test hypothesis allows us to exhaust the space of alternatives. In one-sided tests, the alternative hypothesis only needs to cover greater than or less than the null hypothesis.

For example, in testing a drug for side effects, making our null hypothesis:

H_0: No increased risk ($\delta=0$)

then the alternative would be:

H_1: risk is positive ($\delta > 0$)

In this example, we are not testing whether the drug decreases risks (e.g., like where aspirin lowers the risk of a heart attack), but are only looking at whether its use carries increased risks

[42] Once we appropriately circumscribe or narrow our hypothesis using the piece-meal approach to be explained in the next chapter. See Mayo 1996, chapter 5.

over not using it (or, perhaps, over some standard accepted treatment). Our chosen test rule then maps each hypothesis from the parameter space onto the sample space (e.g., possible observations/values). Thus, the null hypothesis divides the parameter space into two areas (no risk and increased risk) and simultaneously divides the sample space also into two regions—one with results that would accord with the null (no increased risk), the other which would discord with the null. Now the null does not have to be zero or no effect, as I will illustrate in the next section; it can take any value, or range of values within the parameter space.

3.1.2 Going Fishing. In order to avail ourselves of the use of statistical tools, we need to first translate or model our scientific questions into statistical ones. As Mayo and Spanos discuss in depth, we can do this insofar as "statistical methods can be used to connect questions about the phenomenon of interest to questions about distributions of observable random variables that model the data-generating mechanism or the 'population' in question (Mayo & Spanos ES: 10).

For example, suppose above-average fish length is an indicator of the presence of an environmental hazard, say pollution, and we know based on other information that the average fish length in our lake should be 12 inches if all is well and greater than 12 inches if the hazard is present. One way to determine if pollution were occurring would be to determine what the average fish length is in my local fish population. If I find that some fish are greater than 12 inches, is this fact alone evidence that the lake is polluted? No, because some fish will be longer than 12 inches just by chance or natural variation and others will be shorter. This variation allows us to model average fish length as a random variable, and statistical theory assures us that there is a naturally occurring and calculable distribution to this variation *because* the data is seen as being generated by a random variable.

Sometimes, it is appropriate to invoke the results of the Law of Large Numbers (LLN) and the Central Limit Theorem (CLT),[43] with a view to consider our measurements of fish lengths as a sample coming from the normal distribution. (This will be most important, as we will see.) In plain English, these results guarantee that catching extremely short and extremely large fish will be rare events, while it will be common to catch fish whose lengths will fall in the middle of these two extremes. Why is this important? Because this allows us to frame our scientific question: "Are fish getting longer?" as a statistical question about "What is the average (mean) length of fish in our lake?" This property of the population (of lake fish) is called a parameter and in frequentist statistics it is fixed. By fixed, I mean that there is a "correct" answer/fish length and other values are incorrect. So, hypotheses about the value of the mean are either right or wrong.[44]

Uncertainty enters because even though the parameter is fixed, we expect there to be variations either in the population of natural objects (fish lengths) or in our measurements (and probably both) that show up even when the null hypothesis is correct. Modeling our experimental results as a random variable provides a distribution to assess the distance between the results observed and those expected under the null. This distance indicates which departures are probably due to natural variation and which indicate the hypothesis is false. We can divide the sample space neatly into two regions of telling evidence.

H_0: $\mu = 12$, where μ denotes the mean,

H_1 : $\mu > 12$.

[43]Mayo, (1996: 171) explains succinctly: "CLT shows us that regardless of the real underlying distribution of the population, the sample mean $\bar{X}$ is approximately normally distributed. As the sample size n increases to infinity, the sampling distribution of $\bar{X}$ approaches a Normal distribution with mean equal the mean of X itself and standard deviation the standard deviation of X divided by the square root of n. And this holds no matter how the random variable is really distributed, as long as the sample is IID."

[44] This is in direct contrast to Bayesian methodology where the parameter is not fixed and where probability attaches to the hypothesis itself.

To re-iterate: the key step taken above was to frame the (or part of the) 'substantive' question as a statistical hypothesis (about means). This step allows investigators to gather data that can be modeled as a random variable—called the test statistic (e.g., gather samples of fish lengths and take the average). The samples are the data and the average constitutes the basis of the test statistic in response to that initial question, posed as a statistical hypothesis.

3.1.3 Severity interpretation. Following Mayo 1983, 1996 and elsewhere, let us illustrate how the severity interpretation of significance reasoning works using numbers. We want to learn about the population using a sample. Of course, most readers are probably familiar with the problems arising from enumerative induction—of extending properties from a sample straight to claims about the larger population or the next observation (e.g., the two fish in my sample are 12 inches, therefore all the fish in the lake are 12 inches or even to the next fish will be 12 inches). Obviously going from two fish to make an inference about all the fish in the lake is a poor (weak) induction, by which I mean it would be a very unreliable method for inferring the hypothesis that they are all 12 inches long is correct. This would constitute a classic example of the Fallacy of Hasty Generalization—a fallacy because often one would be lead to infer a false conclusion (the average lake fish length is 12 inches) even when all the premises (e.g., 2 fish are 12 inches long) are true.

Philosophers, notably J.S. Mill, have set out a variety of "rules of thumb" about securing large-enough sample sizes, having varied samples, etc. But can we have a better, more sophisticated type of inductive rule to guide inferences from a sample to the larger population? What do all these hodge-podge of rules have in common? What their rules of thumb attempt to

help us avoid and what we want to know is how frequently our sample will mislead us provided our assumptions are met.[45]

A key contribution of Neyman-Pearson testing is to have delineated the two main errors we are open to when assessing these discrepancies between our hypothesis and the data in hand. First, we could be led to reject our hypothesis even though it is correct (and noise or variability was to blame). This mistake is called a type I error. The second error is that we could fail to reject our hypothesis even though it is false and the alternative is correct. This is known as a type II error. Being able to calculate the frequency of making these two errors using standard testing procedures provides a measure of the reliability (or severity) of those procedures (tests) for assessing specific inferences. Mayo emphasizes that it is this knowledge that allows for the objective assessment of what we can and cannot learn about a claim using data produced by a chosen procedure/test.

Most scientific claims will require several tests to probe them for errors and hence warrant them, a process that could extend over years, even decades (see Mayo 1996, 2010). However, the general approach is the same throughout (see Mayo 1983,1996, Spanos 1999)—link substantive scientific claims to statistical claims about a population of interest, then model the observations as a sample from that population, and link what can be learned about the population by the sample using statistical testing rules (e.g.., significance testing, confidence intervals, etc.). Below, I discuss the basic elements required for this connection.[46] To see how this general approach to testing works, we will look at a common and commonly misunderstood frequentist test procedure, significance testing.

---

[45] Here we are assuming that our samples of fish lengths are NIID—normal, independent and identically distributed. Remember our discussion in the last chapter—the problem with the BACI samples was that they were not independent but instead showed Markov dependencies as they were sampled through time.

[46] A fourth step, misspecification testing, can be used to test that the assumptions of the statistical test have been adequately met, which I will briefly introduce towards the end of this chapter. (See Spanos 1986, 1999)

3.1.4 Significance Testing. Say we are interested in a parameter μ, which represents the mean value of some quantity (e.g., fish lengths). The statistic for learning about μ is the sample mean $\bar{X}$, known to be a 'good' estimator of μ. The significance question about our particular results can be stated as: "If the null hypothesis is true, how probable is it that we observed the results we got?" If the null is true, then the majority of the results should fall within two standard deviations of the predicted or expected value under the null (in our example μ=12) 97% of the time. Any result falling outside of that 2 standard deviation range would be very rare. If we were to generate such a rare result, that would suggest that we should reject our null hypothesis (something funny is going on) and accept the non-null or alternative (e.g., average fish size is larger than the null). This type of reasoning provides an objective assessment for what our experimental results do and do not indicate about the null hypothesis and the alternative, as I will explain shortly. But first, how do we calculate this value?

Significance testing allows us to calculate what the mean length of our fish would have to be in order to reject our null hypothesis that the average is 12 inches and all is well in the lake.[47] We are sampling from a population of fish and each fish length can be represented by the variable X. A random sample of size n is taken, $\mathbf{X} = (X_1, X_2 \ldots X_n)$ where each X_i is distributed normally with unknown mean μ and known standard deviation (sd)=.2. Our toxic fish test is:

H_0: $\mu=\mu_0=12$

H_1: $\mu > \mu_0$

Notice H_0 is a simple (1 value) hypothesis while H_1 is a composite hypothesis because it is composed of many possible values for length (e.g., 12. 1, 12.6, 13, …) in fact, all values greater than 12.

[47] Remember, in our example fish lengths larger than 12 inches indicates pollution in the lake.

N-P tests provide a test rule to tell us for each possible outcome $\mathbf{x} = (x_1,\ldots,x_n)$ whether we should reject or accept H_0. The rule (to accept or reject) is given in terms of a test statistic or distance measure:

$$D(\mathbf{X}) = (\bar{X} - \mu_0) / [\sigma / \sqrt{n}]$$

Where $\bar{X}$ is the sample mean, whose sampling distribution is normal with mean μ and standard deviation $[\sigma / \sqrt{n}]$.

To test the two hypotheses above, we observe randomly *n* fish and average their lengths to get $\bar{X}$. We are doing a one-sided test because we ignore those under 12 inches (all values lower than that predicted by the null). Critics complain that because a scientist "chooses" to only look at fish 12 and higher that NP testing is just as subjective as, say, Bayesian testing, only frequentists insidiously sweep their "decisions" under the carpet. So let us note here that one way objectivity enters in, is that our calculations are not locked up in our heads, (e.g., like subjective degrees of beliefs, or expert opinions even), but are instead (at least potentially) publicly accessible and checkable.[48] Now it could be the case that toxic chemicals also make fish grow tinier and clearly this test would not pick up on that. However, that fact does not affect the reliability of *this* test for warranting the inference it is licensing, any more than being majorly underweight (e.g., anorexic) for indicating poor health in humans (e.g., anorexia) affects the reliability of using obesity as an indicator of poor health.

What this test does claim to do is to measure larger fish as pollution indicator and not considering fish below average do not mitigate its effectiveness in achieving that specific aim. Though note, as we are taking an average, if fish are consistently falling

[48] Another scientist need only peer inside my lake, not my head, to replicate or check my data.

below it, this would be identified by our sample mean. If our null is true, then $\bar{X}$ is normal $(\mu, \sigma^2/n)$ and μ=12. If the alternative is correct, then $\bar{X}$ is normal $(\mu, \sigma^2/n)$, where μ > 12.

3.1.5 Evidence & the Cutoff point. What data would be required to reliably reject our null hypothesis, or in the words of Mayo, Popper's successor, 'severely test' it? And if we do not reject, does that mean we accept it? NP testing rules have us calculate a cutoff point at which we agree that our data is so inconsistent with the null that we decide to reject the null hypothesis. That is, data beyond the cutoff point is evidence against the null and for (some) value in the alternative, while data below the cutoff point is evidence against the alternative and towards (but not for) the null.

Here, we need to be careful. There is a distinction between accept/reject here because the null is a point and the alternative is a composite. Data above the cutoff is evidence for some value in the alternative, but data below the cutoff, while it rejects the alternative, is only evidence for some value under the cutoff, not precisely the null. (This is the old chestnut that no evidence against is not evidence for. Later, we'll see how a post data evaluation can demystify this and put a halt to abuses.)

3.1.6 Toxic fish, the cutoff and error probabilities. This is where error probabilities come into play. The cutoff is determined by calculating the distance required so that one would only mistakenly reject the null (a type 1 error) a certain amount of the time. For example, returning to our toxic fish—we really don't want to reject our null that all is well more than say 3% or 1% of the time if it is true; where .03 and .01 are significance levels, e.g., the probability of committing a Type 1 error. By setting the test rule so that the difference between the observed mean and the mean

expected assuming H_0 is true to corresponds to a specified small statistical significance level—for example, .03 or .01 are commonly used— we can ensure that we do not reject our null when it is true more than some small percentage of the time. (When set out beforehand, this is known as the size of the test.)

The cutoff point, which divides the sample space, can be determined for each test. For example, if our sample mean is 3 standard deviations above the null, then a significance of .001 is attained. ($\bar{X}$ observed is greater than or equal to 12 plus 3 standard deviations, in our example 12.6 inches.) For reaching a significance level of .03, we make our cutoff 12 + 2 standard deviations or 12.4 inches for the case above. This is because we know that under the standard normal distribution, that larger fish beyond two standard deviations would only occur by chance 3% of the time. However, we also want to avoid the other type of error, which is failing to reject the null when the alternative hypothesis is true.

One can only minimize one type of error at a time. On NP testing, one chooses to set the most worrisome error up as the type 1, fixed error. While one's choice/judgment is used as to which is the worst error and one sets up the test based on that decision, this information is made explicit and regardless of the choice made, what will follow once it is made is again apparent and hence open to scrutiny and criticism by others. The best test goes to one with a low type 1 error that at the same time has the smallest type 2 error.

Under NP testing, the power of the test (to minimize type 2 error) is always calculated using the cut-off point. Here we also calculate a distance but substitute the alternative H for H_0. For example, if our fish test with significance level 03, n=100, the cutoff is 12.4 (2 standard deviations), and given the standard deviation is 2, what would

be the type 2 error for H if μ is 12.2? To be clear here, the question is: What is the probability that our test accepts H_0 (12) when in fact H (12.4) is true? Here we calculate D= (H) 12.4 – (obs)12.2 = .2, which is one standard deviation. The area to the left of 1 on the standard Normal curve is approximately .84, so β, the probability of committing a type 2 error is .84, which means that the power of the test (to reject H_0=12) and accept H (12.4) is 1-β, which is .16, very low.

What if H is 12.6? (Remember the alternative is a composite.). Then D =12.4 - 12.6 =-.2, which is -1 sd. The area to the left of -1 on the standard Normal curve =.5 – (the area between 0 and 1) =.5 -.34 =.16. (Note: here is a case where H is greater than the cutoff point.) The probability of a type 2 error is only.16 and so the power of the test is .84. Note that the power of a test is always measured in relation to the pre-specified cut-off point (12.4 in our example above). Post-data, that is the actual value of the ensuing result, whether it greatly exceeds or just barely misses the cut-off point is not seen to provide any more information beyond the accept-reject decision. This decision can be seen as a form of statistical induction or pace Neyman, as a guide to inductive behavior (Neyman 1955).

3.1.7 Severity calculations as measures of evidence: Mayo's severity principle provides a formal evidential construal of standard testing. This is accomplished because it links the relative frequencies of the two types of errors and uses them to calculate the probativeness of a result from a test to a specific hypothesis being entertained. Because severity is a measure of the probativess—that is the properties of the test for detecting these errors—we can see how its measure can indicate the evidential status of the actual

data produced by that test for telling about a specific hypothesis being entertained (e.g., the null, or one or another of the alternatives).

3.1.8 Severity Principle. The general rationale behind frequentist testing in general and error statistical reasoning specifically has been captured by Mayo in her **Severity Principle**,[49] which is a general principle of inference—it is not a measurement itself or a test. We have seen this principle before, but here it is again, first in its most minimal form:

> ***Severity Principle (SP):*** Data **x** (produced by process G) do NOT provide good evidence for hypothesis *H* if **x** results from a test procedure with a very low probability or capacity of having uncovered the falsity of *H* (even if *H* is incorrect).

What does this mean in plain English? It means if your test procedure has no chance of discovering that H is false when it is, then it cannot produce "good" evidence for H. On a more positive note:

> ***Severity Principle (full)***. Data $\mathbf{x}_0$ (produced by process G) provides a good indication of or evidence for hypothesis *H* (just) to the extent that test T *severely* passes *H* with $\mathbf{x}_0$.

Severity in the definition above requires that (1) data $\mathbf{x}_0$ fit *H* (for some accepted measure of fit) ***and*** (2) such a close fit would be highly improbable if *H* were false.

In plain English, this means that a test procedure that with high probability can detect the falsity of H and yet does *not* detect that H is false, provides good evidence for H. Let us flesh this out quantitatively.

In the case of a type I error, severity is 1—the probability of a type I error, e.g., 1-α or 1-the significance level of the test. So a test of size .03 would have a corresponding

[49] The severity insight here can be generalized and hence applied as a guiding rule even in qualitative cases.

severity assessment of 1-.03 = a test with 97% severity for the null hypothesis. The more severe a test, the more reliable, or perhaps more aptly put, better probed, the ensuing inference to be sustained is based on the data produced by it. Let us be clear here—the null hypothesis would either be true or false—the 97% here refers to a characteristic of the test where a more severe test provides better evidence for (or against) a hypothesis. [50]

3.1.9 Severity vs. power: Mayo's post-data twist. Mayo's post-data twist for calculating severity for the alternative is to replace the pre-specified cut-off used to calculate power with the observed mean and the ensuing calculation is the severity calculation. So while NP power is *always* calculated at the cutoff for rejection, severity can take into account that the observed values provide evidence for or against the specific hypotheses, which compose the (composite) alternative hypothesis. This is because severity can be used to calculate post-data the distances that can be ruled out. So for example, while we said 'do not reject,' this is not to accept the null as true. We want to know what evidence our test has provided for ruling out specific discrepancies from the null severely. The reasoning being employed here, as Mayo has argued, is the same as that in grading papers or other school tests. A failing mark on a test does not indicate that the student doesn't know anything about the material. Further, it is clear that a failing mark of 58% suggests that the student knows more than were s/he to have gotten only a 23%. In a similar manner, we reason that rejection farther from the cutoff provides stronger evidence against the null than evidence that just makes the cutoff for rejection. Let us now turn to some criticisms of frequentist approaches, some of which the post-data severity interpretation briefly discussed above, resolves.

3.2. Criticisms of the frequentist approach.

[50] See Mayo (2005) for an in depth discussion of the evidential warrant between "highly probed" versus "highly probable" hypotheses.

There are and have been many criticisms of the frequentist approach since its inception. These criticisms tend to be variations on a small set of themes, some of which we'll briefly examine here. Most result either through misunderstanding or misapplying frequentist tests. Others are apt but under Mayo's re-interpretation lose their bite. Let us look at some examples.

3.2.1 Subjectivity is hidden: I have already pointed out that a common criticism is that the frequentist approach requires scientists to make decisions about many aspects of the testing process—choosing cut–off points, sample sizes, choice of the null hypothesis and alternative, etc.—and this, so the critic contends, introduces a subjective element that is swept under the carpet. It is true that these and other decisions are part and parcel of this approach; however, the choices made are (or should be) reflected in the error probabilities and need not affect what the test does or does not say about the particular hypothesis and data at hand.

For example, the fact we ran a one-sided test in testing fish lengths and did not look at smaller than average lengths does not affect the results or our inferences about whether the size of fish were increasing or not in our lake. This is because the claim under test dealt with the increase in fish lengths. That we used it to indicate toxicity was a background assumption but not itself under test. However, that assumption would have been explicitly stated, but assessing it is a separate though obviously key question.

We can clarify the point here using a previous analogy. Given that one measure of health is a person's weight, the fact that being grossly underweight (e.g. starvation) is also very unhealthy does not affect the results of studies on the effect of obesity on human health. This underscores why it is so important to clearly identify both the null

and the alternative hypotheses—to ensure that they exhaust the space of alternatives. Once made, these decisions are public and open to scrutiny and criticism. Thus, these decisions are open to objective appraisal. The severity assessment will reflect the reliability of the test procedures these decisions lead to for making specific inferences given the data generated. This can be illustrated by a common fallacy committed in rejecting a null hypothesis.

3.2.2. The large n problem & the fallacy of rejection. In calculating standard deviations for our fish above, 1 standard deviation (sd) was equal to .2 inches. So a 12.2 inch fish was 1 sd from 12; 12.4 was 2 sd (just at our cutoff for rejection), 12.6 was 3 sd, and so on. This was based on a sample size of 100 fish lengths. However, critics point out that with a large enough sample size, I can reject any hypothesis I choose. This is known as the "large-*n* problem." This problem arises because the standard deviation is a function of the sample size and as sample size n increases then the size of a standard deviation decreases. I said that two standard deviations for a sample size of 100 was .2 because we divide by the square root of n to find the deviation required. But what if the sample size were increased to 10,000?

With a sample of 10,000, then a fish that measured 12.02 would be 1 sd, 12.04 would be 2 sd, 12.06 equals 3 sd and so on. But wait, those values are much smaller than our original 2 sd cut-off of 12.4! On our original test, these measures of 2, 3, or more standard deviations don't even make the original cut-off for rejection. And to reach our original cut-off point would require a distance measure of 10 standard deviations! What is going on? According to critics, especially Bayesian ones, the apparent paradox above simply shows that frequentist testing is deeply flawed.

According to Mayo and other error statisticians this simply shows that a hypothesis cannot be assessed independent of the test, and why evidence should never be taken as simply given. Instead, for data to be construed as evidence for or against anything, we must know the conditions of its generation including sample size, the use of stopping rules, etc. In the case above, a test with a sample size of 10,000, we need to take into account the sensitivity of the test and look not just at standard deviations in isolation, which may be meaningless, but also what measure (e.g., .4 or .04 inches), a standard deviation reflects and whether or not that quantity is measuring anything significant (i.e., important).

3.2.3 Substantive versus statistical significance. The large *n* problem above results from the fact that as the sample size increases, the test becomes more sensitive. Remember, the test here is measuring the distance between the mean fish length as expected under the null and the observed sample mean. The problem here is that the test is too sensitive; it is picking up on inconsequential differences. The large-*n* problem causes confusion when we conflate statistical significance with substantive significance. What I mean is that just because something is statistically significant doesn't mean it is important.

Again, to cash this out analogically, we can group people in a variety of ways—for instance by age. Now if we have for our sample 100 people, the standard deviation between them will probably be measured in years but if we increase our sample to a million people, we can have our standard deviation in months or days and so on, until we are down to minutes. However, depending on the type of inference we want to entertain, say, tracking voting behavior based on age—while months may calculate out to be

statistically significant, they may be totally silent about voting habits; whereas, differences in years or decades (e.g., indices of generation gaps) may be quite telling evidentially for assessing voter trends.

My point is that increasing our sample size will make our test more sensitive but may very well make it too sensitive so that the real effect or phenomena that we are testing for gets swamped out in minutia. This is the statistical equivalent of missing the forest for the trees! But does it mean frequentist tests fail to be objective? No, and the fact that increasing sample size to get an "effect" (e.g., 2 standard deviations) is explicit in the calculations makes disentangling substantive from statistical significance visible and tractable. Mayo often likens this to having a smoke alarm that goes off even when toast is burning—it is a good indication that the toast is done but not that the house is ablaze.

3.2.4 Behavioral versus evidential construal. A common complaint, starting with R. A. Fisher (1955), has been that while frequentist testing is appropriate for industrial applications, where we are concerned with capping the percent of defective items in a batch, while maximizing profits; it is not appropriate for the sciences, which deal with truth, with knowledge and evidence. As Mayo discusses in many places, the behaviorist approach wherein to reject is correlated with one action and to accept with another is primarily Neyman's project. If we follow E. Pearson (1955), we find hints about the evidential construal that Mayo developed.

In the evidential construal of frequentist statistics, what is important about error probabilities is not what will be the case in the long run, though of course for justificatory purposes that tactic is available. Rather, a test's error probabilities are important for gauging the reliability of this test and what it indicates now. An intuitive example is

using a pregnancy test. The reason I want one with good error probabilities attached to it is not because in the long run, if I keep using it I'll know my status! Instead, those probabilities suggest that it is a reliable method for detecting my pregnancy status now.

We need to be careful here, because again, the probability attaches to the test, the procedures or tools—not the hypothesis. And this makes perfect sense, for either I am or am not pregnant, I cannot be 99.9% pregnant in any sensible construal of the term. Here again, the probability measure characterizes the properties of the test for reliably indicating which of the two conditions is the case (e.g., the reliability of the test to correctly identify my condition is highly probable, say, using a standard test, and could be quite poor if my test was for being two days late in getting my period or about analyzing sheep innards.) On an evidential construal, a test's error probabilities are used to assess the reliability of the test as a tool in the specific case of the instance of its use—exactly analogous to how we assess the reliability of other tools, methods and even cars.

3.2.5 Too coarse grained & too mechanical an appraisal. Following Fisher (1954), many charge that NP accounts of testing are simply mechanized procedures for accept/reject decisions similar to those used in manufacturing plants. Another complaint is that they are too coarse grained. From an evidential perspective, it would seem important information to know whether a rejection was due to a result that just made the cut off or whether it was the result of data that was 2, 3 or more standard deviations away from that predicted by the null. Once again, let us remember our classroom testing example to clarify this point. The analogy here is that getting a grade of 59% is less telling about failure (about your lack of knowledge of the topic) than say getting a 44%, and much less than earning a 23% or 2% on the same test. Let us also note, returning to

the an earlier criticism about the supposed hidden subjectivity of test decisions such as cutoff points, that while choosing what grade constitutes a failing grade here may be quite subjective (e.g., in some classes it is <60%, in others <40%, while in some graduate courses <80% is failing), the information conveyed by the grade is not. That is, while in one class a 60% may be passing and in another it is a failing grade, the fact that a student who earns a 60% has demonstrated s/he knows 60% of the material on the given test remains the same, whether it is construed as a passing or failing grade.

3.2.6 Criticisms preliminary wrap-up. In sum, most of the critics' complaints arise because people misuse tests. Some may choose large sample sizes to get a significant effect about nothing of substantial interest. Others fail to understand that making choices/decisions about test specifications do not make assessment of the results of a test a subjective matter even though some choices may make for a poor test,. This is because once the test is laid out, what it does and does not say about a hypothesis is not prone to anyone's decision and can be calculated independent of your or my beliefs about it. Hence these assessments are objective on both criteria I first laid out. That is, what tests say are open to inter-subjective scrutiny, and they hook into the world like other tools, e.g., microscopes and shovels, do. Just as we can assess whether a shovel or a spoon will work better for a specific task (e.g., digging a ditch or uncovering a delicate fossil, respectively), so too can we assess tests on this account (e.g., to parallel the spoon above, we may want to use a large sample to detect a slight effect, perhaps a risk increase).

A major criticism against the objectivity of any type of testing, statistical or otherwise, that appears in the statistical, philosophical and more generally STS literature

is the role played by background assumptions. Some would even say that background assumptions close the door on objectivity. However, while the validity of the tests and procedures that I discussed above crucially depend on their background assumptions either being (at least approximately) met, or being able to argue in individual cases that their results are robust against violations,[51] we do have methods for testing that they hold in specific cases. Let us turn to this topic now.

3.2.7 Testing the assumptions. I talked above about hypothesis testing and error probabilities, but the validity of these tests depends on their assumptions being met. In statistical testing, the question or hypothesis is modeled so that the test statistic comes from a known distribution (e.g., normal distribution). In a way, we are replacing a real world unknown distribution with a similar but known distribution (e.g., Fisher's (1942) famous lady tasting tea example was modeled so that the guessing hypothesis was modeled as a Bernoulli (coin flipping) distribution. *Statistical adequacy of the specified statistical model ensures that there are only two types of errors one can commit in testing hypotheses and these error probabilities are ascertainable given the statistical model.* The adequacy of these models rests on their assumptions being met given the particular data in hand. Statistical models can be characterized by three features—their distribution (e.g., normal), whether members are independent of each other, and if they are identically distributed or not (Spanos 1986, 1999).

For example, in using significance test reasoning for judging the efficacy of a drug against a null hypothesis that there is no difference between those taking the drug and those taking a placebo, an underlying assumption is that patients have been randomly assigned to either the control (placebo) or test (drug) group. If instead, the sickest individuals are assigned to the placebo group, then this assignment would artificially make a drug appear more effective

[51] See Mayo 1996 for a case where a violation actually made a test more severe.

than it really is. This is because at least part of the distance measured is due to the biased group assignment. In an example that I will talk more about in chapter 5, ecologists attempted to use significance testing on samples taken between a control and an acidified basin in a lake. However, the samples used were not independent samples, but were samples taken through time. Philosophers and others in STS often decry the reliance of scientists, experimental tests and outcomes on a variety of assumptions, but much less time is spent investigating how scientists argue and test that test assumptions are adequately met, or even if violated that (sometimes) their results are robust against the violation. In chapter 5, I discuss this further using an ecological example.

3.2.8 How to test assumptions—Mis-Specification (M-S) testing.

How do we test various assumptions? Pre-data this is done through experimental design—e.g., randomization, replication of experimental and control units, etc. Post-data, once the data are in, we can test if the assumptions were violated and if such violations are problematic by asking if our data came from a universe where the assumptions underlying its production hold. We can answer these questions using a variety of techniques—graphical methods such as turning the plot sideways as a visual check on normality, runs tests for independence, and also more formal methods are all available (Spanos 1999). But don't these tests of assumptions also have assumptions? The answer is yes but only the original model assumptions are involved in assessing the relevant error probabilities associated with M-S testing. In evaluating the type I error of a M-S test the null is always the same:

H: All the model assumptions are valid for the data in question

In evaluating the alternative ~*H*, one considers departures from the specific assumption being tested and retains the rest of the model assumptions. Hence, to protect the reliability of such

evaluations one often uses joint M-S tests, which are robust to departures from the retained assumptions.

The idea behind M-S testing is to divide and conquer by testing the assumptions one (or a subset) at a time by using other methods that do not require the assumption under scrutiny to be used. Here again, statistics provides a formal model that we can look at to organize analogous informal and qualitative testing of background assumptions instead of merely reciting a litany of problems about assumptions.

Unlike N-P tests which operate and are evaluated within a specified model, with M-S testing the space of alternatives (models and model assumptions) is infinite. One is truly testing outside of any model. However, as pointed out in the last section, the assumptions being tested can be grouped into three general categories, e.g., one of the most common and well-known distributions is the "normal" characterized by: distribution normal (bell curve), and the samples are independent and identically distributed (NIID). In biological cases, the samples are rarely independent, but are "Markov" dependent, which means each sample is dependent on the one preceding it, however, the dependencies are short lived. While always working with a framework of assumptions, (remember, unlike N-P testing there is nowhere to stand outside of the model), still one can be clever and test one (or more) assumption(s) at a time, by cleverly partitioning the set of possible models. Then one can "hold" two of the characteristics (e.g., assume they are correct for the moment) and use the data to test if the third assumption has been violated.

For example, if the independence assumption is being violated—a fairly often occurrence in large-scale ecosystem experiments, then one can look at the data to see if there are any temporal trends one can pick up—there generally are! A common test of independence is to conduct a run's test, a series of + and – to reflect if the result is up or down. If the process

generating the data is really random, then one should not see patterns in the series of +'s and –'s. Often times, at least in the biological sciences, one can run a Monte Carlo simulation to see what effect an error in the assumptions would have if it were occurring in one's results (see Miller & Frost). Aris Spanos has developed a comprehensive and extensive approach to misspecification testing from specification (laying out the assumptpions of the NP model), to misspecification (testing those assumppptions) to respecification (coming up with a replacement model based on the M-S tests), and crucially, he stresses the importance of carrying out misspecification testing of the respecified model (Spanos 1986, 1999). This last step is crucial because without it, one is simply fitting a model to the data, without checking if such a fit could occur even if the model is egregiously false/inadequate.

Many of the criticisms in section 3.2 above originate from a competing school of statistical thought, Bayesianism. Members of this school use probability in a fundamentally different and incompatible manner to Frequentists. Bayesians also deny that many of the assumptions, which misspecification testing above check, are needed for objective statistical inference and indeed see attempts at meeting them (e.g., randomization) as a source of (hidden) subjectivity rather than as restraints on it. Let us turn to them briefly.

3.3 A Brief word on the Bayesian alternative. The Bayesian school of statistics takes an entirely different view of applying probability and statistics to scientific inferences, or more accurately to human reasoning. Objectivity on this account is to be pursued and subjectivity constrained by making the agents (e.g., scientists) think rationally. This is done by forcing their subjective opinions and beliefs through Bayes' theorem:

Bayes' Theorem: $$P(T/e) = \frac{P(e/T)P(T)}{P(e/T)P(T) + P(e/:\ T)P(:\ T)}$$

Bayesians start with an individual's prior degree of belief (however gotten) in a hypothesis (known as priors) and input varying degrees of belief about the evidence and all other possible hypotheses, even those not yet thought of, (known as the catch-all) into the formula above, to calculate a posterior degree of belief. The idea is that the use of this formal mathematical framework will allow the evidence to shape and mold any individual beliefs, idiosyncrasies and biases in a consistent and coherent manner and thus into a more rational belief to which the whole community is to converge upon eventually. (There are a variety of "washout" theorems that attempt to prove that the individual differences at a starting point on this road will "wash-out" after enough evidence/time has gone through the Bayesian mill.) Mayo and others have provided extensive criticisms of the subjective Bayesian approach, notably chapter 3 in Mayo 1996.

To illustrate differences between the two approaches, frequentists, as Mayo points out, are at pains to try not to let subjective beliefs influence data/evidence and thus focus their efforts on devising methods and taking measures to control and subtract out their influence. Bayesians instead start with subjective beliefs and their end product traditionally has been subjective beliefs, and they appeal to evidence to correct their beliefs.

3.3.1 Bayesianism & evidence. The second major difference, and what makes this approach pretty much a non-starter from my point of view, Bayesians take the evidence as given. How data becomes evidence does not fall under their purview, as they

clearly state (e.g., Howson & Urbach, 1989:272).[52] Instead, their approach aims only to make one's beliefs coherent and consistent. This of course fails ignominiously the second criterion for objectivity with which I started this chapter. Nor can we simply say, well, that is okay, their machinery simply kicks in later in the inference game. This is because their perspective leads them to eschew rules for observation and experiment that are not only standard in scientific practice and are required for the standard statistical tests that we have been discussing; but moreover, whose rationales must be met in order to have good evidence (e.g., randomization, stopping rules, replication, etc., see Mayo (1996), Miller & Frost). The Bayesian approach not only allows for such violations but indeed they are vociferous about not requiring them.

Many ecologists find Bayesianism attractive because the phenomena they are working with are quite large-scale and so do not lend themselves easily to significance testing. For example, the underlying assumptions of replication and randomization are violated in whole ecosystem experiments as there will only be one treated and one control unit. But, as we will see in the next chapter and as mentioned in chapter 1, simply avoiding NP tests and substituting other methods does not avoid the errors those tests and the assumptions underlying them were designed to detect and control. And hence, these errors are still present; however, they will not be accounted for much less eliminated using Bayesian statistics. Turning a blind eye does not make errors go away.

3.3.2. Why the Null is the key. Other ecologists complain about standard testing procedures because as they point out, everyone knows the null is (probably) false, so why waste valuable time and resources to reject it? Why not use Bayesian approaches which allow us to model and attach numbers to any number of interesting hypotheses, ones that

[52] See Mayo 1996: 86-6 for a fuller discussion of this point.

scientists think may be true? The quick answer is because the null provides a milepost for measuring the distance between what is expected under it (the hypothetical or expected result) and what is produced by the experiment or observations (the actual result). While indirect, this type of learning provides invaluable information about the alternative, which of course in many cases is the real hypothesis of interest.

The reason to focus on the null and why cleverly formulating it are so key, is that if well done, like a skeleton key, it unlocks a wealth of information because we know how to model it, while the true state of nature is unknown to us. (That is after all what we are trying to find out about.) While the Bayesians allow us to measure our beliefs in whatever hypothesis excites us or that we want to know about, we must be clear that we are measuring our beliefs, not the true state of nature and this includes the evidence—for that, too, appears and is quantified by Bayesiains based on our beliefs about it, and those beliefs can be entirely detached from the methods used to produce the evidence.

3.3.3. Objective priors---O'Bayes. We cannot forget that the grinder of Bayesian analysis typically works by applying probability to subjective degrees of beliefs, which are either determined introspectively or elicited externally by adducing betting odds, etc. However, there is a new movement to replace these subjective beliefs with objective measures. Often this is done by using an "ignorance" prior such as starting off by assigning a .5 (chance) probability to the null hypothesis. But this assignment actually seems rather high, because to be coherent, our probability space needs to sum to 1. For Bayesians, as probability attaches to the hypothesis (the parameter), this means that we have to divide up the parameter space amongst all possible hypotheses,[53] so to give 50% to just one of them seems rather too generous. More so, if we remember that a major

[53] This would include even those not yet thought of, the so-called Bayesian catch-all factor.

complaint about null hypotheses is that they are often false. So the "ignorance" prior seems more biased than ignorant.

Other major problems with ignorance priors are that the same prior that expresses ignorance under one parameterization, contains information under another parameterization. Thus, these abstract ignorance priors are actually quite context sensitive, which negates their supposed neutrality to context and subjectivity.

3.3.4. A new twist--Conventional Priors. In fact, many current Objective Bayesians (O'Bayes) are not even trying for ignorance priors but instead are trying to develop what are regarded as "conventional priors," that is the prior is simply an agreed upon tool for finding reference posteriors, in particular ones that will match frequentist results.

What these O'Bayes are searching for are objective priors—ones with good frequentist properties. It is important to understand that for them 'objective' here *only* means not subjective. That is, priors that do not reflect the investigators personal subjective beliefs in a hypothesis. In some ways, their quest reminds me of a trick question I used to ask my classes at the beginning of the semester: "Do you want your essays and other work graded objectively?" And, of course, the unanimous response was "Yes!" Then, I would say, "Very well, this class like most classes would tend to end up with a normal curve describing the way the grades will fall. So, I'll pass a hat along, and everyone take a tile. On the tile is your grade for the term." Not unexpectedly, they were aghast and cried out 'unfair.' To which I responded, "What's wrong, this way of grading is entirely objective, my personal biases, idiosyncrasies, etc. will in no way play a role." To which the students would inevitably respond, "But the grade doesn't reflect my ability

or how well I could do in the course!" And this seems to me to be the problem with the O'Bayesian measures.

If conventional priors are not degrees of belief, nor the probability of an event, then what are they? That is, what does the prior represent? This question is especially pertinent, as another problem with them is that they can sometimes sum to more than one—so called improper priors. At best guess, they seem to function as mathematical *deus ex machina,* which drops down to mysteriously to replace beliefs with a *tabula rasa* for the data to etch itself upon. The problem here, among many, is that even if we grant that these conventional priors are 'impersonal,' there is no necessary reason to think that they connect up with the real world in relevant ways.[54] For empirical sciences that will not do. Ignorance is bliss, and may be a prior, but it does not attach to the world.

To be generous, the objective Bayesians can be seen as using either ignorance or conventional priors as blank slates upon which the data can write its story. By eliminating subjective degrees of belief as the basis of their inferential calculations, O'Bayesians appear to be moving closer to meeting the first demand of scientific objectivity by finding a way to hold at bay individual biases and idiosyncrasies. But if we want the data to be dominant in some sense—surely we need to ask penetrating questions about how it was gathered and produced? After all, we can all think of cases where data are weak, terrible, or themselves biased, and thus, when they speak, they only speak lies. To give data dominance without inquiring whether it deserves that dominance or not, as error statisticians do by evaluating error probabilities, seems to me the height of folly—exactly what scientists do not want. Technical brilliance aside, the "GIGO" (garbage in, garbage out) rule applies to these calculations as to others.

[54] Figuratively speaking, a blank slate can also make a formidable wall.

3.3.5 Hindsight is 20/20. The problem here is that while O'Bayesians can point out with pride how their approach allows them to match frequentist answers, they seem to be getting the right answer for all the wrong reasons. It is easy to get the winning lottery numbers once they have been posted, but that hardly seems a good or reliable method for winning the lottery. Matching frequentist results but ignoring the frequentist methods, which provided the results seems to me to be the same thing.

While there are many technical problems with these new objectivist approaches, the main one I see is simply, if you are looking at matching up your final posterior with frequentist measures and outcomes (e.g., results from significance tests, etc.), then why not just use the frequentist methods to begin with? Why go through all sorts of odd contortions to get there? Especially as that reduces your method to a purely reconstructive one and scientists need forward looking methods.

What this new trend towards matching Bayesian outputs with frequentist results suggests to me is that they are still not asking the right questions for meeting the second objectivity requirement of connecting with the world. Questions such as: what data would be relevant for probing this or that feature of the world? Is the data a reliable indicator of the true state of nature? And so on. These are the questions for which frequentists offer us the tools to answer while Bayesians are engaged in Monday night quarterbacking. Indeed, I have even argued that attempting to apply frequentist tests, even where their use is invalid can provide fruitful insights into what one's experiment and data can and cannot reveal about various hypotheses under test (Miller & Frost, Miller).

3.3.6 Minor original insight. One point of departure regarding the formal side of ES that I make from Mayo, Spanos, and others, perhaps, can be found in Miller & Frost. Briefly, I think that *attempting* to apply formal statistical tools, *even where they are unwarranted and invalid* (e.g., the lake example in the next chapter), can serve a useful purpose. The purpose these attempts serve is to indicate and underscore those parts of an experimental design or observational protocol *just* where they make the use of these methods invalid because they violate statistical assumptions. And it is locating these violations, which opens a door to the various errors that undermine one's data's ability to function as evidence for a hypothesis that is the key for improving our procedures, or for suggesting how to weaken our claims. The attempt to apply formal statistical models will highlight the errors and thus suggest what more needs to be done to clean up the test.

Simply substituting graphical or other methods, which though valid as they do not require the same assumptions be met as formal methods do, often only serve to sweep the errors under the carpet. Substituting weaker methods does not mitigate but only hides them. The attempt to use statistical methods acts to shine a light on our procedures. While in some cases the procedures will need to be modified or in the end will be unusable, by pointing out potential problems to avoid, such attempts can make for a more objective assessment of what can and cannot be claimed, even and perhaps most especially in those cases where we could not use the tests validly. (See the next chapter for more on this point.)

3.4 Conclusions. My goal in this chapter has been to introduce the more formal aspects of the error statistical approach. The same type of reasoning and the same rationale behind testing methods underlies qualitative experimental reasoning as much as does the

enormous amount of work that must to be done to reliably claim one's data provides evidence for or against a claim. The rationale both for the use and for the attempted use of statistical tests and misspecification testing remains the same—for data to count as evidence for our inference, it must provide a severe test of or probe for ways that the inference could be false or be a poor model. This ability to probe for, detect and communicate potential and actual errors/lack of errors is the key to objectivity in producing and using empirical evidence for/against scientific claims. Most of this discussion and further elaborations on these themes can be found in the papers by Cox, Mayo and Spanos (both individual and joint) listed in the bibliography.

3.4.1 Anachronism or not? I imagine that many scholars in HPS and STS will object to the error statistical account of scientific inference as providing an account of scientific induction based on the ground that it is anachronistic. How can it claim to capture scientific reasoning when experimental reasoning has been around for much, much longer than statistical reasoning, which is primarily a product of the 20th century? This is a good question. Spanos (2009) uses Lord Rayleigh's work in the 1890's on the discovery of Argon to argue and illustrate how formal statistical tools can be used retrospectively to formalize (*not reconstruct*) the highly sophisticated though informal inductive reasoning used by earlier scientists (using their methods and data) to assess the reliability (or unreliability) of their conclusions. Thus, another use for modern statistics is as a lens to understand why often unfamiliar or archaic procedures worked or failed to work. The power of modern statistics is that it generalizes and formalizes a much more general style of inductive reasoning, which Mayo calls 'reasoning from error,' that

antedates not only statistics, but science as well. This is the topic of the next chapter, so let us turn there now.

Chapter 4: Naturalistically appraising methods.

The picture corresponding to error statistics is one of an activist learner in the midst of an inquiry with the goal to find something out. (Mayo & Spanos (2010:19))

Philosophy of science without history of science is empty; history of science without philosophy of science is blind. (Lakatos (1978:102)).

4.0 Introduction: Objectivity is often defined in contrast to subjectivity and requires that: "All claims are to be open to being tested and criticized by impartial criteria independent of the whims, desires, and prejudices of individuals" (Mayo (1983: 197))." Furthermore, when discussing objectivity in the sciences, as the claims being made are about the "empirical world," objectivity would seem to further require that claims are not only rendered free of human biases and idiosyncrasies but that "impartial criteria" are provided to evaluate whether they hook onto the world in the right ways.[55] By "the right ways" I mean they must also adequately model or capture aspects of the phenomena under scrutiny (Cox & Mayo (2006)). Thus, impartiality or neutrality on the part of scientists, even if it were possible, would be insufficient for scientific objectivity; the claims one makes must also reflect the underlying phenomena by fitting or adequately modeling the world in specified ways.

Given the requirements for objectivity above, then to claim to have objective evidence for a scientific inference must meet two requirements as well. First, to be evidence data must be intersubjectively accessible and unbiased (or biased in a well understood way so as to be usable nonetheless) to meet the first requirement above. Second, not any data can serve as evidence for a hypothesis, because the data needs to reveal how successfully or to what extent the hypothesis under scrutiny reflects the underlying phenomena.

[55]The important difference between deduction and induction is that the truth of the conclusion in the latter are contingent that is they depend crucially on the way the world is and further, these conclusions are ampliative, meaning they go beyond their premises. Thus, evaluating these arguments depends on our ability to assess and evaluate how well they help hook into and anchor the conclusion to the world. As Cox & Mayo point out, this is achieved using methods that "constrain…by evidence and checks of error (2010: 276)."

Problems arise because (1) for scientific claims to be genuinely ampliative they must go beyond the results of specific experiments or observational data, and (2) as Mayo points out " observations…deviate from testable predictions, even when the scientific claims from which they are derived are approximately true descriptions of some aspect of a phenomenon…. [This is]…due to the incompleteness of observations, inaccuracies of measurements, and the general effects of environmental perturbations and 'noise'" (1983: 27)."[56] So how do we get around these disturbing factors? To successfully circumvent these challenges we need some principle of evidence to assess when data constitute evidence for a hypothesis (see Mayo & Cox (2006)). As Mayo has often suggested (Mayo (1996), Mayo & Spanos (2006), Mayo & Cox (2006; 2010)) we can start off by acknowledging minimally that if our data could never indicate that our hypothesis is false, even when it is, then that data clearly would not provide evidence for its correctness.

Naturalists suggest we reject *a priori* reasoning and look to the sciences or to case studies of 'good' science to discover how scientists actually warrant scientific claims using empirical data. Error Statisticians agree with both these suggestions but differ radically from other naturalist approaches about how best to proceed meta-methodologically, as well as, where to look to do this. A commonly held view both among laypersons and scientists is that it is their methods for testing inferences that underwrite the objectivity of accepted scientific claims. So let us look where scientists generally claim to look (i.e., how they claim they justify their inferences)—let us look at scientific methods. But how are we to do this?

Deborah Mayo (1996) first introduced the idea that methods could and should be evaluated and assessed based on their empirical properties or abilities to detect, to eradicate or to

[56]For clarity and stylistic reasons, I've switched the clauses in this statement. The original begins "due to the incompleteness of observations… ., cause observations to deviate from testable…"

otherwise control errors in empirical inquiries.[57] In section 1, keeping in mind the needs of "an activist learner in the midst of inquiry," I further develop and argue that this is the best road into naturalism for philosophers. In section 2, I distinguish the ES account from Faust and Meehl's use of statistics for meta-analyses of science. In section 3, taking some cues from Popper, I develop explicitly the ES approach to the use of case studies, current or historical, for achieving philosophical goals. Though following Lakatos' dictum above, the end results are decidedly un-Lakatosian.

4.1 Appraising Methods on the Error Statistical (ES) Account: Why do we even look at methods to begin with? What work do we expect a good or reliable method to do? What makes a method unreliable? *In short, why do we bother with methods at all?* The answer to the above questions for error statisticians is that methods are important because errors are important. For Error Statisticians to determine whether or not a specific method is causally efficacious in a particular context requires one to determine the empirical properties inherent in that method, which enables its use to detect errors, subtract them out, or control for them with the goal of achieving reliable inferences in spite of errors.

Although whether or not any one method will work in a situation is context specific there is a general goal that methods are called upon to promote; namely, to provide reliable procedures for arriving at true claims (or correct models).[58] Mayo has captured this error probing role of methods in the following principle of evidence that she has proposed in several places (1996; 2008):

> ***Severity as a principle of evidence***: One has satisfactory evidence for a hypothesis or claim only to the extent that the ways the claim could be incorrect have been well probed by procedures that with very high probability would have

[57]This approach was further developed in Mayo & Miller (2008).
[58] By true claim, I mean not necessarily literal truth, as scientists could well be interested in testing how well a model, though known to be false, accurately captures some feature of a phenomenon of interest or not.

> signaled the presence of the errors, if they existed, and yet no errors are found; instead our observations are of the sort that are to be expected given the *absence* of error. In such cases we can say that the error has been reliably or severely probed.

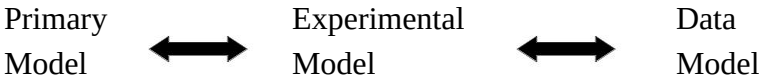
This principle comes directly out of her concept of "arguing from error." Indeed it is simply her "argument from error" rewritten to emphasize its evidential import.

> ***Argument from Error:*** It is learned that an error is absent when (and only to the extent that) a procedure of inquiry (which may include several tests) having a high probability of detecting the error if (and only if) it exists nevertheless fails to do so, but instead produces results that accord well with the *absence* of error (Mayo 1996: 64, italics in original, see also p. 445).[59]

Rather than looking at error as only a block to reliable inference, Error Statisticians, following Mayo's lead, focus on and take advantage of errors to learn about the world. Thus in chapter 1, we saw that a concern with biased assignments and natural variations between samples (e.g., human subjects) lead to the use of randomized clinical trials. Cueing in non-human and human animals could give rise to a variety of blinding methods for use in different situations as the occasion demanded (e.g., blinding and double-blinding set-ups in medical trials as well as the use of placebos, etc.) which in turn can lead to more knowledge about human subjects. Below, I'll illustrate how having the ability/inability to rule out particular errors allows us to justify the reliability/ unreliability of our inferences using a much discussed and common method: experimental replication. But first, *real* science experiments are extremely complicated. Even in only slightly complicated investigations, and especially when dealing with complex phenomena, scientists need a systematic and organized way to track potential errors, so let me take a brief detour into strategies for breaking up experimental inquiry into manageable chunks.

[59] It is important to read this definition carefully. The 'the' in the second sentence, "detecting *the* error" is referring to 'an error' in the first sentence of the definition. It should not be read as a universal and does not imply all errors but a specific error or type of error that is being ruled out.

4.1.1 A Piecemeal Approach to testing. Mayo advocates taking a 'piecemeal' approach to testing in order to grapple with and learn from errors. There are three basic models composing her approach to experiment which feed into one another:



The role of the primary model is to break a substantial inquiry or theory into a hypothesis (one or more local questions as Mayo (1996: 130) phrases it) that is amenable to severe testing, i.e., that can be reliably probed for potential errors. For example, an investigation into the effects of acid rain on seepage lakes in the Northern USA was split off into many different questions, including one about the effect of pH on the survival rates in two species of zoo plankton.[60]

The experimental model plays two roles as fitting to its linking role above: (1) It translates the primary model into a particular experimental set-up often by re-phrasing or capturing the primary hypothesis using one or more canonical models (e.g., coin-flipping mechanism or bulls' eye target to indicate chance), which allows (2) relating the data to these experimental questions (e.g., a series of heads and tails). In the lake experiment, survivorship was determined by relative abundance of each species, which was measured using relative biomass from samples taken through time. Finally, rarely is raw data used directly as evidence in an inquiry. Instead, the data or observations are modeled in order to more aptly bear on or answer the experimental question at hand. For example, a series of observations may be averaged to get a more accurate measurement of temperature, weight or melting point. (See also Bogen and Woodward who make a similar point in their distinction between 'data' and 'phenomena.')

[60] See Miller and Frost for details. I will discuss this case further below and in the next chapter.

Data models, the third component of Mayo's piecemeal approach to testing, includes both pretrial and post-trial questions (i.e., pretrial, how should we model the data to answer experimental questions, and post-trial, were the assumptions of the data generating process, e.g., randomization, etc., met). See Mayo 1996 Chapter 5 for a detailed discussion of her piecemeal approach to experiment. In the lake experiments discussed below, Monte Carlo experiments were run to try to determine the effect that violating replication had on calculating p-values; see Miller & Frost for details.

This piecemeal, divide and conquer approach to errors and testing, is especially useful in making inferences about large-scale phenomena, e.g. about the effects of acid rain, global warming, evolutionary theories, etc. Let me try to build up this picture of how normative naturalists of the error statistical stripe would go about their business (and prescribe for others as well). There are numerous debates about replication both in object level sciences (e.g., surrounding gravitational detectors, large-scale ecosystem experiments in ecology) and in meta-level studies about replication as a method for reliable inferences (e.g., debates over breaking out of so called "experimenters' regresses" between Franklin and Collins), so let us examine replication as a methodological rule from our new error statistical perspective.[61]

4.1.2 Replicating the phenomenon. First, let's be clear that by replication we do not mean duplication. While most people can figure out that exact duplication would be *practically impossible*, it takes a little more thought to realize that it is not only impossible but is *highly undesirable* as well. Why? If you could precisely and exactly repeat my experiment, you would also duplicate any and all mistakes or errors that I had made. Thus your ability to duplicate my results is not a reliable or severe test of their correctness or of the reliability of my procedures. For a simple example when testing an owner's claim that his horse Hans could add, skeptics had

[61] See also Miller 1997.

the owner repeat the test but with the difference that the horse was wearing blinders and so could not see his owner. This replication of the experiment differed crucially from the original in that the blinders made unconscious cueing between owner and horse almost impossible. Perhaps not unsurprisingly, when the blinders appears, the phenomenon, Hans the horse's arithmetic ability, disappeared. This simple example illustrates that it is replication with the specific aim of error detection that makes replication "work" for objectivity.

So one of the most relevant features is that the set-up should ensure that the phenomena should be reproducible but that specific (potential) sources of error have been eliminated or their influence controlled. That is, the thing about reproducibility is that it suggests that one has identified and captured a real effect, and not merely hooked onto a chance effect or artifact of one's experimental/observational set-up or situation. Attempts to replicate phenomena in this stringent way provide rich case studies and arenas of debate that with careful scrutiny can illuminate the specific errors that are of concern in an area of study. Such discussions are critical, especially as many experiments are too costly or rare to be able to be replicated physically, but one can discuss what a replication *would* require. This could in turn allow one to perhaps find other, easier or cheaper ways to see if the error of concern can be detected and controlled or subtracted out, etc. There is another sense of replication that I discuss in section 3.1.4 below that illustrates this point and is still a center of controversy where scientists (ecologists in this case) are looking to philosophers and others (e.g., statisticians) for help.

4.1.3 Replication as a Methodological Rule. Replication is a standard and perhaps one of the most well-known of scientific methods. We could simply claim that replication is a valuable standard method because all the scientific authorities concur, as Giere suggests. However, this doesn't explain to us why or how it works, even ignoring the problems with the sheer circularity

of the reasoning to begin with! Also, such an appeal to authority would get the causal relationship backwards, for we think that the reason the authorities have reached a consensus on using this method is because the method has some property or properties that makes it valuable for testing their inferences in a reliable and objective manner, not that their consensus makes the method desirable. Plus, consensus on when a replication has occurred or what it would take for an experiment to be considered a replication is not clear cut and indeed is often a source of heated dispute and controversy among the experts in a field (see Collins (2006)).

Even though methodologically replication is considered an important method for supporting scientific claims, explaining exactly why it is, is a bit trickier. One reason is because when it comes to replication debates, we need to be aware that there are many types and/or levels of replication sought. Whether or not replication is achieved and why and what type is desirable, will depend on the hypothesis under scrutiny, the testing methods used, as well as the data actually generated or gathered. Let us turn now to the second sense of replication as a method and some of the controversies that arise when it is violated in large-scale ecosystem experiments to help shed light on why it is considered such a valuable experimental component or rule.

4.1.3.0 Replicate units: the case of pseudo-replication: The other sense of replication as a method refers not to repeating entire experiments but instead to an experimental design that incorporates multiple experimental (e.g., manipulated) and control (un-manipulated reference) units. A unit here is understood as anything or group of things or persons that may be regarded as a discrete entity for the purposes of the experiment. What constitutes a unit of interest for any one investigation is determined in reference to three scales—spatial, temporal and biological aggregation—for manipulation and analysis (Frost, et al. (1988)).

The idea behind having multiple units to run a manipulation on is to act as a check on errors arising from natural variation and from stochastic events that could act as potential confounding factors. Having multiple units is a requirement for randomization. Together, replicate units that are randomized work together to ensure that differences between units are distributed between manipulated and control units in such a way to wash out the effects of unknown differences e.g., due to natural variation (Fisher (1922)). This use of replicate units is a methodological requirement for ensuring the validity of certain statistical tests, e.g., significance testing.

It is the second sense of replication above that pervades experimental discussions of Before-After-Control Intervention (BACI) designs used in ecological studies of large-scale phenomena in complex systems e.g., acidification of lakes (Hurlbert (1984), Stewart-Oaten, et al. (1986, 1992); Miller & Frost). Here controversy revolves around the validity of applying statistical techniques where the underlying assumption of replicate units (and thus also randomization) is violated.

Manipulations at small scales can be replicated in both senses of the term: repetitions of the experiment, each with several replicates at the unit of analyses. Replications in either sense of the term can become prohibitive at larger scales for many reasons including sheer impossibility (e.g., the unit of interest cannot be replicated, the earth for example) or for ethical or financial reasons (e.g., in ecological experiments with pollutants we do not want to destroy multiple lakes, etc.) For example, if we look at studies of large-scale phenomena, replicated experiments, though doable, may not be suitable. In order to discover and quantify the effects of acid rain on lake ecologies in the mid-West, a large-scale Before-After Control-Intervention

(BACI) experiment was designed and implemented at Little Rock Lake (LRL) in northern Wisconsin.

4.1.3.1 BACI Design: BACI studies compare two locations, a control and an intervention site, over an extended period both before and after an experimental manipulation at the intervention site. In such a design, there is only one experimental unit and one control unit, and samples are taken over time in both units (Stewart-Oaten, et al. (1986)). The BACI design is the most "common type of 'controlled' experiment in field ecology [that] involves a single 'replicate' per treatment" (Hurlbert (1984, 1999)). To argue for the sensitivity of this type of experiment, experimenters need to show that the two units (manipulated and control) are very similar. (For a simple example of this type of qualitative argument, see Fisher (1947: 22-25).

The LRL acidification experiment conducted in northern Wisconsin is an example of this type of design (Brezonik et al. (1993); Frost et al. (1999)). The lake was divided into two separate basins by a flexible, inert barrier. Both sides were monitored for one year prior to the barrier being placed and for another half year after the barrier had been inserted. The utility of the BACI deign in inferring causation is augmented by the use of a series of interventions. The treatment side was acidified in steps that approximated 0.5 pH units in three two-year periods from a starting pH of 6.1 to a final pH of 4.7, close to the average pH of rain in the region (Watras and Frost (1989)). These three stages represent separate interventions, each of which can be evaluated separately. While not an "active intervention or treatment," the recovery period can be seen as yet another separate intervention in that it provides a changing pH environment and ecosystem responses can be studied during this period (see Frost et al. (1998).

The experiment was conducted to gain insight into the ecological effects of advancing acidification in seepage lake systems. Many different types of ecological data (chemical and

biological responses, etc.) were collected and monitored from this one long-term whole-ecosystem experiment using samples taken every two weeks when the lake was ice free and every five weeks when ice covered. The problem in using such samples (e.g., measurements taken on the same experimental unit through time) as a substitute for replication of experimental units is that events that happened in the unit previously can have a substantial influence on future samples. The samples are not independent. This lack of independence renders many standard statistical procedures inappropriate (cases of pseudo-replication, according to Hurlbert (1984)) if mechanically applied.

4.1.3.2 Pseudoreplication defined: Hurlbert coined the term "pseudoreplication" to refer "not to a problem in experimental design (or sampling) per se but rather to a particular combination of experimental design (or sampling) and statistical analysis which is inappropriate for testing the hypothesis of interest" (Hurlbert (1984:190)). This is much too broad a definition. There are many ways to misapply statistical analysis to a host of experimental designs, and many ways to utilize sampling techniques that are inappropriate for testing particular hypotheses.

Here, I will restrict the use of the term "pseudoreplication" to refer specifically to cases where the 'inappropriateness' comes from violating the assumption of independent replication underlying the use of statistical tests for inferring that differences between a manipulated and control basin are due to the manipulation. This restriction includes using samples from the same unit through time as independent samples. The errors from violating replication for inference about "treatment effects" as discussed by Hurlbert (1984) and the solutions suggested by Stewart-Oaten, et al (1986, 1992) have important implications for interpreting evidence from BACI experiments above and beyond the use or misuse of statistical tests. Violating replication

can be equally or even more problematic for non-statistical data modeling techniques (e.g., graphing) that Hurlbert suggests should be used to avoid pseudoreplication (Miller and Frost).

4.1.3.3 Replication as a check on natural variation: According to Hurlbert (1984) "Replication *controls* for the stochastic factor, i.e., among-replicates variability inherent in the experimental material" (191; italics in original). Control is italicized to show that he is using the term in the broader sense he discusses (*ibid.,*191) as one of the "features of an experimental design that reduce or eliminate confusion" (e.g., experimental error). The first error of concern in BACI designs that Hurlbert expresses is that the two units are not identical. According to him, the validity of comparisons between the manipulated and control unit for treatment effects is justified only if the two units are *identical* not only at the time of manipulation but throughout the experiment with the exception of the treatment effects. He states: "the supposed 'identicalness' of the two plots almost certainly does not exist, and the experiment is not controlled for the possibility that the seemingly small initial dissimilarities between the two plots will have an influence on the …[effect under test]" (Hurlbert (1984:193)). Natural variation assures us of that point. Of course, depending on the size of the treatment effect, this may not be a problem. To make this point in an admittedly exaggerated way, if one is studying the effect of radiation, even though other cities are not identical to Hiroshima, the effect of a nuclear bomb is pretty unambiguous. Hurlbert agrees that the size of the expected treatment effect can compensate for a lack of replication. "When gross effects of a treatment are anticipated or when only a rough estimate of effect is required, or when the cost of replication is very great, experiments involving unreplicated treatments may also be the only or best option" (*ibid*. 199-200).

Hurlbert agrees that significance tests can be validly used to determine whether there is a significant difference between two locations; nonetheless, he claims: "the lack of significant differences prior to manipulation cannot be interpreted as evidence of such identicalness. This lack of significance is, in fact, only a consequence of the small number of samples taken from each unit" (*ibid*: 200). This is because:

> We know on first principles, that two experimental units are different in probably every measurable property. That is, if we increase the number of samples taken from each unit… our chances of finding a significant premanipulation difference will increase with increasing number of samples per unit. These changes will approach 1.0 as the samples form an experimental unit come to represent the totality of that unit… . (*ibid*: 200)

The problem above is what is referred to as the large *n* problem and was discussed in the last chapter. The problem is that as the number of samples increases, the sampling procedure becomes increasingly sensitive to even minor differences between populations. This would be a case in which the sampling for finding differences is too sensitive, so the test based on it actually would be less severe compared to one with fewer samples for picking up relevant (versus irrelevant) difference between populations. However, this argument would apply equally well to small-scale experiments, as no set of replicates is ever absolutely identical.

While BACI designs cannot use multiple replicates to cancel out noise from random natural variation, this does not mean that other reliable arguments for similarity cannot be forwarded to defend the robustness of the results against this particular error. Thus, to place the unobtainable demand for identicalness on BACI designs seems unwarranted. The real question is not whether the two units are identical—they are not—but whether they are similar enough in the relevant aspects that we can infer that subsequent changes between them are likely to be caused by the manipulation. Adequate sampling regimes before and after start-up can resolve this

issue. (For details on how such arguments were made at LRL, see Stewart-Oaten, et al (1986), Frost, et al (1988), Brezonik, et al (1993), Sampson, et al (1995), and Watras and Frost (1989).)

4.1.3.4 Replication as a check on stochastic events: The second problem Hurlbert raises for cases of pseudo-replication is that given the two units are sufficiently similar in the relevant aspects to start with; will they remain that way over time? There could be an uncontrolled extraneous event that affects one but not the other plot, thereby increasing the dissimilarity between them that may even, in a worst-case scenario, be falsely attributed to the treatment under test (see Hurlbert 1984:191-3). Neither the similarity between the two units nor the size or intensity of the treatment can ameliorate this second situation. We can agree with Hurlbert that replication and interspersion of treatments provide the best insurance against chance events producing such spurious treatment effects, but that does not mean it is the only insurance (method) for detecting and mitigating this particular error. Moreover, Hurlbert's turn to graphical and other non-statistical techniques as a solution to pseudoreplication only sweeps the problem of violating replication under the carpet. All of the errors Hurlbert raises are errors in the procedure (BACI design) used for *generating* data and so would also pose a threat to non-statistical representations of the data as well.

Frost et al (1988:3) "[e]mphasized simple, primarily graphical, presentations of their data. And have avoided any strictly statistical assessments of the significance of treatment effects…due to the fact that whole ecosystem experiment is a contentious area in terms of statistics." They graph biomass (us/L) for both the control and manipulated LRL basins against year (1984-1990). A bar graph is used to show the interannual species turnover rates for each basin (1985-1990). The graph for K. taurocephala (Frost, et al (1988:5)) shows massive spikes and increases in the acidified basin. It is visually quite dramatic. However, without replication,

these spikes could simply be the result of a stochastic event that occurred only in the manipulated basin at the time of startup. When replication is violated, the errors it controls for are potentially present in the data regardless of how it is presented unless the BACI design is supplemented with other procedures for detecting and/or arguments for ruling them out. Turning to graphical presentations to avoid pseudoreplication merely makes potential errors less obvious. This is not to deny the fruitfulness of such approaches for modeling data.[62] I am only denying that doing so removes the errors for which proper replication controls.

Stewart-Oaten, et al. (1986: 935-6) partition the types of stochastic events Hurlbert refers to based on a combination of the spatial and temporal characteristics of their possible effects thus:

1. Short-term effects spread over a large area;
2. Short-term effects localized to one experimental unit;
3. Long-term effects spread over a large area;
4. Long-term effects localized to one experimental unit.

Knowledge about possible "artifacts" or confounders (i.e., events that are mistakenly attributed to but not caused by the manipulation) is especially critical if one is using samples through time on the two units in lieu of physically independent replicates. This is because an event that happens in a unit may continue to affect subsequent samples from that same unit as the samples are not independent. This is the reason that Hurlbert condemns using samples through time as if they were independent samples as a clear case of pseudoreplication (Hurlbert (1984:193)).

Stewart-Oaten, et al. (1986) argue that chance events that have effects of short duration are harmless provided we sample over a sufficiently long time. The time between samples must be long enough for the process to "forget" its remote past—i.e., "correlation between deviations that are far apart in time is close to zero" (*ibid*: 932). Sufficiently long sampling periods are

[6262] See Spanos (1986, 1999) for the many uses of graphical representations for finding violations of assumptions (e.g., normality, independence, etc.) in misspecification testing.

needed to check that past events will not build up and influence the next sample so that subsequent samples may be treated as independent. Further, if such short-lived effects occur over a large area, they will also be distinguished from the manipulation effects, as they will also be manifested in the control unit. For the same reason, events with long-lasting effects that occur over a wide area are also with careful sampling and experimental design, not an insurmountable obstacle to the robustness of results from BACI design.

But, as Stewart-Oaten et al point out "the independence assumption should also be checked against the data. There are, in fact, powerful tests for doing so (*ibid*)." In short, we can turn to sampling procedures and other tests to supplement the BACI design in order to mitigate these three potential errors. No one is claiming that this is easy to do, only that it is possible in theory to construct such experimental arguments from error as Mayo calls them.

Stewart-Oaten et al (1986: 937) still share Hurlbert's concern that there "remains the possibility of a large, long-lasting but unpredictable effect occurring at about the same time as the start-up and affecting one location much more than the other." Unlike the other three possibilities, the possibility of a long-duration event with localized effects could not be ruled out (e.g., could go undetected) even with a long sampling procedure in place. However, they also feel that replicated and randomized studies can also suffer from similar errors, as "their results may be due not to the treatment but to the way it is administered", e.g., placebo effects, cage effects, contaminates, etc. I will return to these criticisms in the next chapter, when I look at Longino's account of critical discursive interactions in order to better show how, I hope, applying an ES approach to them can work with and strengthen her social account.

4.1.3.5 BACI versus Bottle experiments: Given all the problems and work entailed in running a BACI experiment, why would one want to use them? A case in point is in determining

effects of acidification on various species of zooplankton survival. Replicated and reproducible bottle experiments yield one answer: K. taurocephala will be adversely affected, have much lower survival rates and even do worse than other zooplankton species. But, although this answer about survival is reliable for bottles, it does not translate or extend into natural environments. Indeed, the reverse is the case: in natural environments such as lakes and bogs, K. taurocephala flourishes and quickly outstrips other species of zooplankton. (We can see this as a portability issue in a Latourian (1986) sense or as an issue of 'external validity' in an experimental economics sense (see Guala; Ross). That the bottle results do not hold in 'natural environments, e.g., lakes' was demonstrated in the Little Rock Lake (LRL) BACI experiments.

Notice in the above discussion on replication, the reliability or unreliability of a method for a specific inference was assessed based on its error probing, detecting or eliminating capacities or lack of them. *Methods matter, because errors matter* and methods are designed and implemented to probe for errors, to detect, eradicate or in some way help us control for them so that they do not lead us astray when we test our inferences. Debate over these issues and the lack of consensus it reveals does not show that evidence, method and reason are failing in science and that other social factors are determining experimental knowledge, as Collins would have it. Rather, debating these errors is the motor which drives knowledge and (most importantly) learning.[63] So we can argue that while no one method will do, nonetheless in general we assess each method based on its error probing capacities in a particular situation under scrutiny.

4.1.4 Localization: The "local" nature of method assessment. It is important to note the "local" or "topical" nature of the above reliability assessment. A method (M) is assessed as unreliable not in isolation but only within the context of a specific inference and based on its ability/inability to detect and/or control errors in that hypothesis. Thus bottle experiments were

[63] I discuss this more thoroughly in chapter 6.

reliable for bottle phenomena but not for extensions to larger environments. The BACI was robust against 3 types of causes of error, only one cause gave rise to un-reliability. Knowing this, however, doesn't mean we give up, but rather it means that we need to do more work.

The key point is that reliability or severity is *always* assessed relative to (1) the particular hypothesis (e.g., the occurrence or non-occurrence of differential death rates of *K. taurocephala*), (2) the method used to generate or gather data (e.g., Replicated and randomized bottle experiments or BACI experiments) and (3) the actual outcome/result[64] (e.g., *K. taurocephala* is booming). [65] Even though the assessment is local, the principle guiding it (severity) provides a universal principle for evaluating evidence. And it is with this claim that Longino and I part company, for she does not think there are any general epistemological principles we can draw upon in assessing and justifying methods. All such justifications are, in the end, local for her. I will pursue this point in the next chapter.

4.1.5 Methodological Plurality So in studying one phenomenon, for example, the effects of acidification on lake-zooplankton, we have two different methods. The replicated laboratory studies in bottles are questionable in regards to their external validity. Do the bottle experiments mimic real environmental conditions successfully enough? The BACI experiment and other observational studies provide good reasons to suggest that they do not. Still there is an error in the BACI, and that hole, too, needs to be plugged. Notice, while I am discussing multiple methods here, the underlying reasoning behind the use of all of them is the same—to rule out errors that would allow the data generated to "fit" the hypothesis under test even if it were false.

[64]Of course, in most inquiries, the outcome is modeled (e.g., measurements averaged, noise subtracted out, etc. These data modeling procedures are part and parcel of (2) the test(s) (methods) used to argue that the data constitute evidence for the inference.

[65] This relativity of errors to a specific hypothesis also unlocks many of the difficulties in determining when alternative perspectives may or may not help in an investigation, something Longino and others have been unable to determine and which I discuss at length in Chapter 6.

This is the key to reliable or severe testing: not only must the data "fit" the hypothesis, but it must be generated in such a way that such a close fit would be highly improbable if the hypothesis being tested were false (see Mayo (1996:180) for this version of her severity criterion). This additional requirement is crucial for data to actually provide evidence for a hypothesis: data must not only fit the hypothesis, but must also act as probes for errors in it. And this requirement works as a general criterion or principle of evidence.

Thus, while I endorse the plurality of methods in science and meta-science studies, I reject the brand of epistemological pluralism (local epistemologies) that Helen Longino draws from similar situations (2006: 178-9). Longino takes the lack of methodological unanimity between field studies versus laboratory experiments, for example, as indicative of the necessity for philosophers and other humanists to take a "pluralist-epistemologies" approach to science. Briefly, her view is that the scientists in each approach will support their own methods (and the findings from using them) against other competing findings and methods. This is because each group has, to varying degrees, different aims and work with different methodological and metaphysical assumptions to which they are often blind.[66] As methods serve different functions and have different strengths and weaknesses, which each community values differently based on their various goals and, to them, transparent background assumptions, then the justification of a method is ultimately local—as is the justification for its use in any particular situation.

Longino; however, does not want her epistemological pluralism to wind up leading us into either circular justification of methods or Kuhnian incommensurability. Instead, she promotes the view that objectivity is garnered through critical discussions between these competing communities. This is because as they do not share the same methodological or

[66] This is pure, hardcore Kuhn. Kuhn's view of the nature of scientific paradigms directly leads to this circularity thesis for justifying evidence. The difference between Longino and Kuhn is she sees incommensurability as less of a problem than Kuhn did.

metaphysical assumptions, members of one community can better expose the faulty assumptions held by members of the other community. These discursive interactions, if properly absorbed by each community,[67] can improve their knowledge. Longino's account of pluralism and localization, however, denies that there are any general shared epistemological principles to underwrite or facilitate these discussions for objectivity, leaving it instead to work itself out, however it will, by sheer dint of plurality of viewpoints constrained by her four norms, which I will discuss in the next chapter.[68] This I think is a problem but one that can be resolved, and her account strengthened by incorporating Mayo's severity principle (or at least something very similar to it) into it. Another benefit of taking the error probative approach is that it provides suggestions for how different types of methods and experimental scales can be linked together to probe higher level phenomena or theories (e.g., about acid-rain in the mid-west) as suggested above.

4.1.6 Models of error (systematization, partitioning.) Errors pose an obstacle to objectivity. So we need to be able to detect and control them. But, a critic may ask, how can we ever even begin to exhaust all the various sources of error? There are two answers that work together here to allow us to get a handle on this error elimination task. One is the acknowledgement that though there are probably an infinite number of specific errors, they run to a few canonical types. These include but are not limited to:

1. Mistaking experimental artifacts for real effect;
2. Mistaking chance effects for genuine correlations or regularities;
3. Mistakes about a quantity or value of a parameter;
4. Mistakes about a causal factor;
5. Mistakes about the assumptions of experimental data;

[67] Her norm of "uptake" requires absorption of some, though not all, criticisms, which I will discuss in the next chapter.

[68] This plurality includes a hodge-podge of methods and criteria that fall within basic empiricism and logic, but without taking a stand on any of the disputes within and between various schools of thought within that context, e.g., Bayesian versus standard statistical practices have equal footing within that plurality.

6. Mistakes in linking statistical inferences to substantive scientific hypotheses. (A combination of Mayo (1996:18) and Mayo & Spanos (2010: 19))

As errors run to type, we can partition the space of potential errors into a manageable framework. Methods are chosen based on the standard types of errors that an inference is liable to as well as the way in which we are generating data that could be in error. That is, based on the inference we want to make, the test of the inference and the data actually generated, along with methods used, all work together to determine the reliability of the inference or alternatively the severity of the testing procedures. It is this ability to localize sources of error based on the intersection between test method, the specific hypothesis and localization that allows us to partition the space of potential errors into a manageable sequence.

Second, we can often research separately and learn enough about general types of mistakes to see how they would affect results and take this into account when assessing our data. Thus, we do not need to require the complete elimination of error, only methods that allow us to measure or otherwise take account of its impact. Fortunately, this is an empirical question and open to empirical inquiry and testing. We can test, often by inducing the error on purpose, in order to see how it would affect an experiment and thus get a handle on it. It is this focus on error that makes "the conglomeration of tools from statistics and modeling, data analysis and experimental design" (Mayo & Spanos 2010: 34) so attractive to naturalists of the error statistical stripe, (see also Mayo 1996). And even though in some, perhaps even many, cases, we will not be able to get a hold of every error, the extent to which we can will improve the objectivity and reliability of the inferences we wish to sustain.

4.1.7 Statistical methods as error exemplars. Even when statistical tools cannot be directly applied in an inferential or experimental set-up, they can nonetheless provide methodological guidance both to the sources of errors and what would be needed to control or

correct or detect sources of error. I disagree with Hurlbert's 1984 claim that graphical methods could replace statistical significance tests, thus solving the pseudoreplication problem (because replication and randomization are not assumptions underlying the validity of using these techniques) for BACI-like experiments. I disagree because the error of concern—the possibility that the occurrence of an unknown event with long lasting effects in one basin but not the other could confound the result—remains unresolved. While using graphical techniques is not technically invalid in the same way applying statistical techniques is, they still do not pose a severe test of the hypothesis as they also would not detect such an error. Thus, this way of modeling the data is not a reliable method for sustaining the type of causal inference scientists were after. Here again, we can see the fruitfulness of statistical models and methods even where they are not being used, but as exemplars of errors that they highlight (or that attempting to meet their assumptions highlights) for making reliable inferences. Attempting to meet their stringent demands provides guidance for non-statistical methods as well. It is the focus on assessing errors that provides the link between qualitative and quantitative reasoning.

4.2 Statistics as Toolkit. Mayo initially suggested, and other error statisticians concurred, that meta-analysts should *look at the tools in statistics and allied disciplines such as experimental modeling and design*, which have been designed to overcome specific problems in inference in the face of error and incomplete data. I discussed this direct appeal to statistics in the last chapter. What I want to do next is to distinguish the ES recommendation for using statistics as a form of and an exemplar for inductive reasoning given in chapter 3 from Faust and Meehl's (2002) appeal to statistical techniques for doing meta-scientific inductions on the history of science. This then sets up the ES approach to the use of historical case studies in philosophy of science.

4.2.1 Meta-statistical induction on the historical record: the Faust-Meehl Thesis. David Faust and Paul Meehl advocate using statistical techniques such as linear regression methods to analyze the historical record in order to better meet many of the descriptive and prescriptive goals that historians and philosophers of science hold. Such studies would "supplement and extend more commonly used case study methods," which have several weaknesses (Faust & Meehl 2002: S185). They have identified four major problems with the case study method currently in use:

1. The data base is massive and growing rapidly. …[it is] nearly impossible for anyone to master and continuously track more than a relatively small proportion of this data base S186
2. [R]elations between the methods that scientists apply and the outcome of their efforts are largely probabilistic, not deterministic. [For example] [g]ood or even excellent methods do not guarantee success nor do bad or poor methods always lead to failure. S 186
3. Lacking representative sampling, one lacks the data base needed to best answer or resolve these types of inherently statistical questions [e.g., of the means/ends type in #2 above] S188.
4. The case study method is directed toward identifying or accruing instances that illustrate or support a position, and therefore is likely to produce skewed or grossly skewed samples. This allows any proponent to find many supportive instances. S187

Faust & Meehl remark on many of the same problems in using historical data to study the efficacy of methods that Mayo (1996) and Mayo & Miller (2008) also have found. For example, correlating methods to ends is not a one to one relationship. Several different methods may produce the same desired outcome, the "many to one" problem as they call it (S187). On the other hand, a procedure or method "can produce inconsistent or varying levels of success. The relationship here is the one to many (ibid.)."

The stochastic nature of the correlations between methods and goals, on their view, requires implementing statistical methods, which by their very nature are designed to deal with both the stochastic nature of the data set as well as the complexity of the factors

involved in drawing inferences about the relationship between methods and goals.[69] These two characteristics combined with the sheer size of the historical record, allow proponents to find supportive instances for any program, and I would add for opponents to find counter-instances as well. Thus, their proposal, captured in what they call the Faust Meehl thesis (FMT) is that rather than choose singular paradigmatic instances (e.g., as exemplified by the case study approach) to provide evidence for a thesis; instead, we "need to incorporate some form of representative sampling of scientific episodes. Obtaining representativeness will generally require random sampling of a sufficient number of episodes (although this number may not need to be nearly as large as one might suppose)" (*ibid* S188). Once we have representative samples, we can run regressions on a whole matrix of variables and see what features end up singly or in combination (and in what degrees or orderings) to promote various aims and goals scientists hold. Let us note that with the advent of computerization, this is much easier to do as much of this information is electronic.

Faust and Meehl provide several reasons why statistical methods would be more objective and advantageous for answering many of the meta-analytic questions philosophers and historians have regarding science. These include the fact that the case study method is often "directed toward identifying or accruing instances that illustrate or support a position, and therefore is likely to produce skewed, or grossly skewed, samples" (*ibid.* 188). Their approach could certainly help to mitigate against the cherry picking of particular episodes to support or topple one or another meta-philosophical program. The fact that a 'legion' (to use their term) of psychological and other research studies have shown that human beings are simply not very good at complicated weighing of multiple factors, which are probabilistically tied to outcomes, also

[69] Reaching a goal may require the combination, and correct weighting, of several factors.

suggests human minds are not well-equipped to disentangle or weave together the multiple factors needed and to what degree to achieve various aims or goals (see their discussion on S189). For example, in trying to determine explanatory success, how does one weight various factors such as fruitfulness, simplicity, accuracy and a host of other characteristics for providing enduring or successful explanation, especially as these may very well vary across different types and styles of explanation and theory construction?

Their goal is to provide more rigorous methods, ones that have "greater precision and accuracy, especially around matters that require complex data integration (e.g., which factors in which combination best predict the long-term fate of theories should provide improved guidance (S189-90)." What the FMT is proposing is a more rigorous and quantitative means/ends approach to the meta-study of scientific methods—Laudan on statistical steroids. While their methods will produce better and more accurate correlations for the reasons given above, it still will produce only correlations! While their approach is more cognizant of the biased nature of the record and the need for selection principles (randomization etc.), still it suffers many of the defects of Laudan's approach.

4.2.2 Problems with means/ends correlations. The FMT approach, though statistically sophisticated, still relies on means/ends correlations as pioneered by Laudan. This can be seen in their discussion on page S190—justifying why past performance is indicative of future success, they apply the straight rule (Laudan's rule 1) to the use of control groups, and then conclude that by using statistical tools to make these types of means/ends correlation then "greater precision and accuracy, especially around matters that require complex data integration, …should provide improved guidance" (*ibid*.). But, and for the many difficulties they raised about the nature of methods and aims, the "one to many" and "many to one" problems, this sort of straight line

reasoning of "it has worked and so will continue to work," only tells us that control groups work in specific circumstances but not why they worked. That is, what property does the method of control groups have that allow for inferences sustained based on the data from using them so reliable?

So, while their method may uncover which method was used in a particular situation, the "why", the explanation of success (or failure) remains missing. Ideally, what we really want to know is what it is about using method M—(or a complex combination of Methods, M1, M2..Mn)—that *caused* success or failure in achieving goal G? If we could unearth the reason why the method worked in those circumstances, it seems to me we would be able to provide much better guidance, as we would be able to tell a causal story about its past success, rather than have to rely on correlations with successful applications. It is this type of causal reasoning that the ES approach supplies using Mayo's concepts of severity and arguing from error, which tell us to look not at correlations but at the properties of methods for detecting and controlling errors in testing claims. In this case, one reason control groups are useful is because they provide the foil for what would be the case if a treatment were not given, thus isolating treatment effects from factors that could be occurring during the experiment as those factors would affect both treatment and control groups.

Means/ends correlations will not explain *why* certain tools, including theirs, work when they do, and conversely why they don't work in other cases.[70] Notice that in asking these types of causal questions, we are asking about tools, not history. Answering them will take us further afield and into how error statisticians appeal to statistics and allied fields, which I covered in the

[70] Their psychometric tools do seem promising for specific types of human judgment questions like measuring reliability of expert judgments, predictive or useful/redundant criteria used in grant evaluations, etc., but are not applicable for the meta-methodological questions that I am addressing in this thesis.

last chapter or into a different interpretation of how best to use historical (and contemporary) case studies, to which I now turn.

4.3. Case studies. We can agree with Kuhn and others that the place to look for insights into actual practices is at the frontiers of science, where science is in the making. Indeed going beyond them, we claim this is also where normative advice can best be developed. And with a lot of diligence and a little luck, we can also find this type of information in the historical record, sometimes openly, other times buried. The ES approach to current or historical case studies is to use them as heuristics to provide the type of information on evidence and inference in the face of uncertainty and error.

Kuhn's meta-methodological legacy, accepted by proponents and opponents alike, can be summed up by his final question in his introduction to SSR. "How could history of science fail to be a source of phenomena to which theories about knowledge may legitimately be asked to apply?" *While we can agree that the history of science can provide a source of phenomena to be explained, this does not mean that the history of science itself will provide us with an explanation,* especially as we know that the history of science itself is not composed of full and unbiased descriptions of all aspects of practice but similar to the fossil record is incomplete, biased and opportunistic. Just as many parts of the geological record have not been preserved by fossilization, so too much of history is not preserved. And just as certain types of fossils, those with hard parts and advantageous burial, (e.g., marine organisms) are opportunistically preserved, so too are the successes of science preserved while the "failures" and paths not taken (e.g., the soft bodies) are erased by neglect or the winds of change. Cataclysmic events both personal and social, death, wars, and other tragedies as well as indifference can erase records just as the weather and climate erases the geological record.

However, given all this frailty and incompleteness, just as we can harvest reliable information from the fossil record about prehistoric life and evolution, the nature of the historical record does not preclude our harvesting reliable information and insights into scientific practice if we are clever inquisitors. Furthermore, we ourselves are biased insofar as we must acknowledge that there is no "view from nowhere or from everywhere" as one learns in any decent introductory historiography course (Collingwood, Jenkins, Novick, Popper). Every history is situated in some frame of reference. There are surprises of course which serve to push at us and make us question our frameworks or assumptions.

4.3.1 Situated Histories. Minimally, the histories we produce are situated and contextualized by the very questions we ask; however, it is this very contextualization, if done well, which allows historical research to shed light on philosophical questions (e.g., about methodology, networks of communication, etc.). In fact, we couldn't even begin to do history without some assumptions about what we want to know and what type of information would be relevant—at most we would have an undifferentiated, homogenous list of whatever information was available. [71] And as histories are always incomplete, what may be available may also be very distorting—relevancy criteria may help us identify such distortions in the record.

Nonetheless, given that we have learned when looking to the history of science and looking to the various human sciences that there are no neutral descriptions, "pure" observations, etc., this does not mean there are not better and worse descriptions or that we cannot have more or less "objective" descriptions if appropriately circumscribed. Clearly, the very questions we

[71] It has been argued that our very concept of time (e.g., linear versus cyclical) fundamentally has influenced the type of history we do. The concept of linear time and linear history (that is one with a definite beginning, middle and end) "discovered" by the early Hebrew tribes, is quite different to ancient Greek concepts and this metaphysical difference of assumptions is reflected in the types of histories written by each and the lessons they drew from them. The former focus on change and differences between past and present, often to find progress while the latter focus is on sameness, similarity and repetition An excellent example of a cyclically based history and historiography is Ibn Khaldun's The Muqaddimah, while Vico is famous for introducing the concept of a spiral view of history, which is seen as an amalgam of the two views of the unfolding of history through time.

ask of the historical record help circumscribe and determine which features of it are relevant and which irrelevant for answering specific questions. Our particular line of questioning may also hinder us or lead us astray and as Popper exhorts us, we should be prepared to revise our framework in light of our research. Thus, our very questions force us to make preliminary evaluative assumptions, but in this our situation is no different from any other empirical researchers, such as scientists. This is the position Popper takes in his discussion of situational logic as an approach to history (Popper (1994)).

4.3.2 Popperian Historiography. For Popper, the starting point of empirical inquiry (scientific, historical, etc.) is always rooted in a problem we want to solve. Moreover, in searching for empirical information, the problem at hand determines/provides the necessary relevancy criteria—at least to get us started. He endorses the view that "… the only point which determines the choice [e.g., of information, documents, etc.] is the historian's purpose, the questions he is asking" (*ibid.* 145).[72] He clearly felt that as our research progressed, we would need to revise our attempted solutions and indeed the problem itself (or how it was posed) in light of new findings and evidence. Not just any data or historical information will do, instead we need to find historical data that is telling in regards to conjectured solutions to our historical problem. "You cannot start from an observation; you have to know first *what* to *observe*. That is, you have to start from a problem" (*ibid.*). Moreover, data that would provide evidence for one or another solution in one context could be absolutely meaningless in another. Popper illustrates this last point by stating that even though a train ticket may prove his innocence by providing an alibi and hence become an important historical document, in general he would not recommend people save train tickets as a method for doing good history.

[72] Popper is quoting Elton, p. 128 here.

Popper admonishes us to turn to the historical record with the goal of falsifying our historical conjecture/problem-solution, to detect errors in our thinking and historical inferences. For Popper, this "problem-oriented" style of historiography solves the problem of 'historical relativism':

> Admittedly, our conjectures are relative to our problems, and our problems are relative to the state of our knowledge. And admittedly, there may be much in the momentary state of our knowledge that is erroneous. Yet this does not mean that truth is relative. It means only that the elimination of errors and the approach towards the truth are hard work. There is no criterion of truth. But there is something like a criterion of error: clashes arising within our knowledge or between our knowledge and the facts indicate that something is wrong. In this way knowledge can grow through the critical elimination of error (*ibid.* 142-3).

Unfortunately Popper could never cash out the advantages of a "criterion of error" in a positive light. His dogmatic denial of induction never allows him to say anything beyond "our conjectured solution is not yet falsified."

4.3.3 Popper's crucial insight. Nonetheless, the crucial insight from Popper is that both our choice of case studies and which aspects of them are important to study are determined by the problem (hypothesis) at hand. Taking Popper's error criterion a step further, in order to test our conjectured solution (meta-theories) we need information that can probe them for errors. That is, we need to be able to argue that not only does our hypothesis fit some aspect of the record, but also because confirmations are cheap, that such a good fit would be highly unlikely if that hypothesis were incorrect. This is how Mayo has interpreted Popper's call for "sincere attempts to falsify" our theories: she requires severe tests, i.e., tests that meet her severity criterion.

We can in turn apply a similar approach to case studies by delimiting the case study in light of our questions/problems and probing our conjectured solutions for errors. But how are we to proceed? The complexity of science and scientific practice as a rich human practice requires a

certain splitting apart by our own questions. Such division of questions not only makes practice manageable for scrutiny and analysis but also helps determine which features of the many that could be studied are relevant for the question at hand.[73] Such a piecemeal approach should not be confused with an algorithmic approach however. Our first attempts at division may lead us to a dead end or we could discover that we need a different division or to look elsewhere, perhaps to social or economic factors, to understand why a particular group chose to follow the path they did.

4.3.4 Case studies are not the objects of inquiry. Case studies are used as an aid to inquiry. But it is the properties of methods that are the objects of empirical scrutiny in the ES approach. To this end, case studies are useful to the extent that they provide information about methods, evidence and inference. Error Statistical claims about reliability and scientific methodologies require that for case studies to be useful, the type of empirical evidence needed for methodological scrutiny requires access to the actual data, data generating techniques and other modeling and interpretation techniques used. If the methods used do not have the capacities claimed for them for detecting experimental errors or if the data is such that it does not provide a reliable test of the inference in question, or is simply not available, then we need to look elsewhere to determine how a case was closed.

Nor, given human nature, should we be surprised that other elements were involved in either closing or extending disputes; however, similar to the ES approach to methods, these other factors simply because they are concurrent do not provide evidence that they are the factors responsible for various ends (e.g., reliability) unless we can determine the mechanisms, e.g., their inherent properties that promote reliability. So the supposed fit of case studies to a method is not

[73] Philosophers are beginning to develop a more nuanced and sophisticated approach to case studies as can be seen in recent philosophical works utilizing extended case studies by Chang and Weber.

being used to support or deny the epistemic virtues of that method. Instead, case studies are useful for what they reveal about methods, the information they contain about how various methods function in particular contexts. It is by focusing on mechanisms, that is, the properties of methods, which allows us to break away from the standard method of using case studies to find correlations, and instead seek a causal story for the success/failure of achieving certain ends using various means.

The same is true for technological artifacts. For example, historical archeology and the annals school engage in reconstructing instruments as they were originally made to test more reliably or severely what could and could not be detected using those instruments. The recreation of historical instruments and experiments has moved from the physics lab, where modern equipment was often used, into the historical lab to assess the viability of the observations and inferences that could be made and sustained using similar instruments and techniques of the time being studied.

Remember, the key feature of the error statistical concept of evidence is the requirement that for data to be construed as evidence for an inference, it must not only fit the hypothesis but must in some way provide a severe test or a severe or reliable error probe of the inference in question. That is, not only must the hypothesis fit the evidence but it must also fit so closely that it would be very improbable if the hypothesis were false. It is this requirement that data go beyond merely fitting a hypothesis that distinguishes the error statistical account of evidence and approach to methodology from other empirical (i.e., naturalistic) approaches.[74]

Thus the mere ability to find cases that fit one's account does not provide evidence for that account, unless one can argue that such a close fit would be highly improbable if the account

[74] While Popper also required that a test must be severe, he was never able to adequately explain or define his notion of severity. Hence his account, though a testing rather than an evidential relationship account, remains a primarily logical affair, rather than an empirical account of testing.

were in fact false. If it turns out that a case study did not fit one's method, for example, if Eddington's eclipse experiments did not provide a severe test of the deflection of light, one of the cases Mayo uses to illustrate her approach, this would not falsify her account. This is because merely fitting a case is not a severe test of a hypothesis (in this case a methodological hypothesis) and so does not provide evidence either way. On an ES approach, remember, we are looking to the properties of the method to warrant its use, not whether it fits a case or not.

4.4. Conclusion: Error Statistics--a Third Way into Naturalism. The above discussion about how the ES concept of evidence and approach to testing is empirical, i.e. is naturalistic, is very different from standard philosophical concepts of naturalism in both philosophy of science and its allied sister subjects in STS. The naturalist trend has not been to focus on the empirical properties of methods and tests for underwriting evidential concepts but (1) to engage in one or another type of "meta-induction" on the history of science (as do Laudan briefly described in chapter 1 and Faust and Meehl as described in this chapter) or other use of case studies, or (2) to use actual sciences (e.g., biology, cognitive science, psychology, sociology) or specific scientific theories to describe and prescribe methods for achieving epistemic goals such as truth, reliability, predicitive success, etc., as discussed in the previous chapter. Error Statistics bucks this trend and in so doing offers a viable "third" way for being naturalistic that is to look at the empirical properties of methods for detecting, controlling, or eradicating error to underwrite a strong sense of objectivity. But, I can well imagine Worrall asking, what assumptions does this account rest on?

4.4.1 The Error Statistical Assumption: The ES account of testing and evidence rests on the following assumption:

> ***ES assumption***: Warranted experimental arguments demand severe or reliable probes of error.

If we understand what error means, we can then reason out why using methods to detect and control for errors will result in more reliable inferences, hence, to warrant an experimental or empirical argument requires severe or reliable probes of error. The ES principle of evidence based on both Mayo's formal concept of severity as well as its informal counterpart, her argument from error, are equally justifiable or plausible as Worrall puts it, in light of the terms used, and hence could be seen to constitute a minimal *a priorism* on his account. I accept his claim that an account of reasoning, such as I have been spelling out here, must be plausible and if plausibility makes such an assumption *a priori*, then yes, if one sets the bar that low, the ES approach constitutes a case of Worrall's minimal *a priori* but to the extent that physics or biology are *a priori* because they use mathematics. But, if conceptual clarity is all that he means by *a priori*, then it is a trivial claim poses no real hurdle for claiming that both the ES account and its application are thoroughly naturalistic, that is, empirical.

4.4.2 Naturalistic teeth. What underwrites the usefulness of the severity rationale, what puts the teeth into, broadly speaking, error statistical methods, is whether or not they work in the world (e.g., a bachelor is a bachelor even if none exist), but the ability to detect a specific type of error is irrelevant if such errors did not exist or if we were unable to detect them or if we were chasing down the wrong type of error.

But some will complain or secretly hope (e.g., Worrall), that this conceptual clarity makes the ES approach at heart an *a priori* one albeit perhaps more general and fundamental than other approaches. Not at all. The fact that there is conceptual clarity is trivial. This would be similar to claiming that physics is *a priori* because it uses mathematics or perhaps more clearly that determining the number of cords of wood needed to get one through the winter (to heat one's house) is *a priori* because it relies on making calculations! But this is absurd, the amount

of wood required is an empirical matter based on the properties of the wood; the features of the furnace used; the climate in one's region as well as the desired temperature and insulation in one's house. The appropriateness of using mathematical operations in this situation and the answers arrived at are empirical matters. Determining severity and constructing arguments from error to determine when one has evidence for an inference and to overcome errors is equally empirical.

The real bite in the ES approach is that assessing and evaluating severity depends on the way the world is and how, that is the way in which, various methods really "work" in it to facilitate reliable inferences. Indeed, one of the crucial aspects of this account is that if the world were to change our methods are such that they would indicate this, which means we have here an account of induction that does not rely on a principle of the uniformity of nature and hence avoids Hume's problem of induction (see Mayo (2003, 2005)). For our purposes here, the naturalistic point to be made is that if the world changes, then our methods--having the empirical properties they do (tied into the world as they are)--will detect and indicate any such changes. This is the real self-correcting thesis that Kitcher so desired, but failed to provide as seen in the last chapter.

4.4.3 Error Statistics & Case Studies. It is the methods that are the objects of empirical scrutiny in the ES approach—case studies are used as an aid to inquiry. To this end, Error Statistical claims about reliability and scientific methodologies require that in choosing and assessing case studies, the type of empirical evidence needed requires access to the actual data, data generating techniques and other modeling and interpretation techniques used to gather and turn data into evidence for an inference.

To illustrate more clearly the distinction I am drawing here, I will use a hammer example. We can look in carpenter's tool box to see what tools she uses to build a house. We can also look to instruction guides and books (e.g., statistics, experimental design manuals, etc.) or read description of past house building feats (e.g., historical case studies) or look to current ones (Home and Garden shows, magazines like *This Old House*, etc.) to figure out how these tools work—for example a hammer. By understanding the properties of a hammer (focused weight onto a small surface like a nail head), we can also see and understand why some shoes (a clog for example) would be potential substitutes in an emergency. But it is the understanding of how the hammer works that explains how to drive a nail home or analogously, by understanding how methods work to drive the reliability of an inference home.

It is the understanding of how methods work to detect errors and control them that underwrites the objectivity of inferences. Because these procedures to generate, interpret and model data can be assessed based on their features or properties for probing specific inferences for how they could be false, that is for errors in making such an inference. If the method used has the characteristics, which would likely have pinpointed the error if it existed, but not otherwise, then if it does not detect the error, that is good reason—provides good evidence—for declaring the error is absent. Understanding how methods function in inquiries; that is for understanding the rationale behind their use and their properties for error detection and control—is equally as important in assessing and justifying objectivity in the use of formal quantitative methods as it is for assessing qualitative ones.

Chapter 5: A Methodological Conundrum in Longino's Social Epistemology & Suggestions for How to Resolve It

> I argue for a normative social element as part of the meaning of "knowledge," i.e., that epistemic acceptability of content (or epistemically justified acceptance of content) requires the satisfactory performance of certain kinds of social interactions. (Longino (2002: 574))

5.0 Introduction. The recognition that social factors invariably enter in obtaining knowledge is often thought to preclude objectivity; objectivity is generally thought to demand freedom from the biases of personality, social, and cultural values because these are seen to open the door to allowing claims to be the result of prejudices and wishful thinking, perspectives and interests, rather than what is the case. A classic ploy to save objectivity, however, is to redefine objectivity or rationality in terms of social factors. Kuhn famously declared that what is objective and rational is just what a group of scientists working within a shared paradigm (having been educated in certain ways) come up with—there is no deeper overarching notion of objectivity they have to live up to. "As in political revolutions, so in paradigm choice—there is no standard higher than the assent of the relevant community (Kuhn 1996: 94)." This is problematic because a group can give highly unreliable guidance, and the result is the exact opposite of what the goal of objectivity is supposed to have provided.

The well-known philosopher of science Helen Longino offers a sociological conception of objectivity. A question arises as to whether it too, like Kuhn's, is just defining objectivity as whatever a certain kind of group, appropriately composed, arrives at. This question arises because, in line with many sociologists of knowledge, she denies there are any overarching principles or logics of objectively rational method or inquiry above and beyond the specific standards adopted by local communities.

> [T]he only non-question begging response to challenge must be: '"We are open to criticism, we do change in response to it, and while we may not have included all

> possible perspectives in the discursive interactions that underwrite our methodological procedures, we've included as many as we have encountered (or more than others have).".…The point is that there is nothing further, that appeal to standards or methodological norms beyond those ratified by the discursive interactions of an inquiring community is an appeal to transcendent principles that inevitably turn out to be local. (Longino (2002: 174))

Her conception of social objectivity above requires identifying, understanding and augmenting the plurality of perspectives between and within groups and their interactions for achieving consensus through critical discursive practices, or more simply, criticism.

One reading of Longino is to see her contribution as specifying and extending the composition of Kuhn's "quite special groups"[75] in science in order to secure a more objective appraisal of theories. At the same time, Longino adamantly wants to preserve a stronger epistemological sense of objectivity that is more appealing to traditional goals held by philosophers of science. These include denying Kuhn's incommensurability thesis and its attendant claim that theories are always necessarily circularly justified in science. In particular, she sees paradigms or "local epistemologies" as having considerable overlap, at least in aims and goals, which allows them a better ability to talk to rather than through one another.[76] This is seen in her call above for multiple perspectives to be brought to bear in underwriting methodological procedures.

5.0.1 A dilemma The dilemma I pose is this: either—despite her denial—she must implicitly appeal to overarching principles for critiquing the results of the various social groups, or else Longino's view simply will take us back to a Kuhnian-style social relativism about knowledge. I think the good news is that she does, implicitly, accept a principle of method, one that can be

[75] "To discover how scientific revolutions are effected, we shall therefore have to examine not only the impact of nature and of logic, but also the techniques of persuasive argumentation effective within *the quite special groups* that constitute the community of scientists. To discover why this issue of paradigm choice can never be unequivocally settled by logic and experiment alone…" .(Kuhn: 1996: 94, italics added)

[76] Her norms of public standards and uptake attest to this view.

seen to underwrite both her numerous case studies as well as her own method of multiple perspectives (MMP), as I call it; and, further by making this implicit principle in her method explicit, we can resolve some of the conundrums that have been raised against her social epistemology (e.g., Kitcher 2002) while strengthening some of its perceived weaknesses (e.g., as admitted to in Longino 2002). So, another reading, one which I think is more in line with her stated goals and numerous case study examples, is to see her as providing a method for instantiating what Deborah Mayo has identified as the severity principle to a specific subset of cases, e.g., in which social (group) assumptions play a significant and detrimental role. To get to this point; however, will require several passes at understanding in order to pull apart the many tangled strands of thought woven into Longino's social epistemology.

5.0.2 Chapter overview I begin with a quick sketch of Longino's social approach to objectivity in science and raise some potential concerns about it. Next, I discuss Popper's social epistemology and the contrast between it and Longino's, as this sets out the sometimes subtle but very real differences between their two approaches to criticism. I then shift our focus onto the roles of pluralism and localization—two new experimentalist themes—as providing key characteristics of Longino's MMP and Mayo's ES in a way that argues for the compatibility of the two approaches to work hand in hand for objectivity rather than in apparent contradiction. Mayo's Severity Principle (SEV) is the foundation of her error statistical approach to philosophy of science and statistics. In the previous quote, Longino denies any such overarching principles exist, hence, the presumed incompatibility between the two accounts. I will argue that the extent to which Longino's MMP is successful for securing objectivity rests on the extent to which it facilitates Mayo's more general concept of severity.

To sum up this chapter, I want to argue that Longino can resolve the dilemma I posed above by appealing to overarching aims similar to those Deborah Mayo endorses and extending them into her social context, and at the same time I want to augment Mayo's error statistical program by using Longino's work on social methods to suggest how ES can be extended to qualitative social methods.

5.1 Barebones Sketch of Longino's Social Epistemology. This section provides a background summary of her argument for the social nature of scientific objectivity[77] and of the several components comprising her epistemology.

5.1.0 The social nature of objective inquiry. Longino (1990), after reviewing common ideas about objectivity, argues that "the objectivity of science is secured by the social character of inquiry" (1990: 62). She builds up to this claim by drawing several distinctions based on commonly accepted views about the features of scientific practices. Science is commonly understood to be objective in two senses: in its content in so far as it makes correct claims about the world and the objects in it—scientific realism; and second, in its mode of inquiry insofar as scientific views appeal to "nonarbitrary and nonsubjective criteria for developing, accepting, and rejecting the hypothesis and theories that make up the view" (*ibid*). Together these two claims compose the common view of scientific objectivity. This is very close to the strong sense of objectivity that I have speaking about as a cornerstone of Mayo's error statistics—though realism is too strong a term, as it will be for Longino as well, and both of us, along with Popper, substitute intersubjective for non-subjective.[78]

[77] I closely follow her argument for the deeply embedded social nature of scientific knowledge given in Longino (1990), chapter 4.

[78] Both Mayo's and Longino's accounts are modeling accounts of scientific epistemology and equally compatible with scientific realist and anti-realist positions.

Longino continues by stating that the objectivity of methods is also divided into two. First, the objectivity of data, which is based upon data being gathered using accepted procedures without cheating or tampering with it in ways that would invalidate it. Second is the objective development of hypotheses and theories. Drawing on the well-known philosophical distinction between the context of discovery, which is seen as subjective, and the context of justification, which is seen as objective, Longino explains that while the initial development of a theory may be entirely subjective, e.g., discovered in a dream or by accident, these non-empirical aspects are quickly discarded and replaced by empirical factors of observation, experimentation, deduction and other commonly accepted forms of theoretical justification.

Importantly, because here is where Longino begins to depart from her predecessor Popper's social epistemology, science is a fully social endeavor, not social in the simple sense of being conducted jointly by and between several individual scientists. Longino points out, that in order for someone to follow a method, there has to be a commonly accepted method held (at least abstractly) by the community of scientists. There must be some common context for interaction. To understand this sociality, two shifts need to be made. First, science must be understood as a practice or activity rather than as a passive body of knowledge. Second, this practice or activity must be understood primarily as a social practice.

Longino (1990: 67) pulls out three reasons or aspects which characterize the deeply social nature of scientific practice from the work of Marjorie Grene (1985): (1) Scientists need the tools, previous theories, data, etc. from previous scientists. (2) Science is a learned activity and so scientists must be taught how to make observations, measurements, and techniques from those who have already learned them, and so on. (3) Science only has value as part of the social structure and so also affects the society in which it is practiced. From the reasons paraphrased

above, Longino concludes that scientific practice is irredeemably a social practice and cannot be done by an individual. The knowledge that results from this collaborative effort is not just the sum of individual knowledge but instead the process of scientific socialization integrates individual knowledge into a social whole that is much greater than the sum of its parts. Longino agrees with Popper that criticism is the vehicle for the growth (and integration) of this knowledge. At the same time she rightfully holds his feet to the coals for his conventionalism about the status of 'observation statements' in both corroboration and falsification, as we will see.

For Longino, criticism is not just directed at theories but also at the very methods of science.[79] This is accomplished by attacking the data from experiments and, key to Longino's approach, the background assumptions underlying experimentation and the data derived from it. In Longino's view: "Criticism is thereby transformative…As long as background beliefs can be articulated and subjected to criticism from the scientific community, they can be defended, modified, or abandoned in response to such criticism" (2002: 178-9).

The trick for Longino, in securing objectivity, is determining how we can direct criticism at the various background assumptions, especially those which justify method and data as 'objective,' the determination of which Popper writes off to convention. In this respect we can see camaraderie between Longino and Mayo in that both have dispensed with a white glove logical analysis between data and hypothesis to really grapple with the actual processes and assumptions for generating data in the first place.

[79]To be fair, for Popper, observations, data, were conventionally accepted but as he pointed out, an observation or background hypothesis though accepted for present purposes, could itself one day be on trial and attempts made to falsify it (1962: 38, fn 3.) However, his was more of a white glove logical approach, compared to the gloves-off , dirt under the fingernails approaches to data that both Longino and Mayo take.

Unlike claims that rest on observation or method that can fall in the path of criticism, according to Longino, criticism itself can maintain objective equilibrium in light of criticism. This is because criticism is a social endeavor, and because it "is a characteristic of a community's practice of science rather than of an individual's that makes it possible to criticize *every* aspect of it" (*ibid*., italics added)." Longino provides four norms or criteria that communities must meet for securing claims of objectivity both for their scientific claims and for their own critical discursive practices upon which much of the objectivity of their scientific claims rest.

5.1.1 Longino's four norms of objectivity in social epistemology. Longino develops norms that emphasize the structure and role of scientific institutions for fostering criticism and hence objectivity. We can look at her four norms for science as far more realistic and practical replacements for Merton's more idealized norms of universalism, communality, disinterestedness and organized skepticism for achieving the goal of communal objectivity. Furthermore, she argues that institutions and institutional practices and norms affect the actual content of knowledge in significant ways (e.g., such as determining which types of entities can and cannot be considered candidates for causal agents.)

It is important to note at the outset that for Longino, objectivity is not a binary (on or off) concept but instead admits of degrees and the degree to which an enterprise is objective is ascertained by how well a group instantiates her four norms. These norms are designed to secure objective inference for the group:

1. **Venues:** There must be publicly recognized forums for the criticism of evidence, of methods, and of assumptions and reasoning.
2. **Uptake:** There must be uptake of criticism. The community must not merely tolerate dissent, but its beliefs and theories must change over time in response to the critical discourse taking place within it. This standard does not require that individuals or

research groups capitulate to criticism but that community members pay attention to and participate in the critical discussion taking place…

3. **Public Standards:** There must be publicly recognized standards by reference to which theories, hypotheses, and observational practices are evaluated and by appeal to which criticism is made relevant to the goals of the inquiring community.
4. **Tempered Equality:** [C]ommunities must be characterized by equality of intellectual authority….A diversity of perspectives is necessary for vigorous and epistemologically effective critical discourse. [Equality is tempered or qualified as follows:] …The social *position or economic power of an individual or group in a community ought not to* determine who or what perspectives are taken seriously in that community.[Note, unlike political equality which only requires an equal chance of being included, here all (relevant) perspectives must be included, see her footnote 15 in 2002: 131.] The exclusion of women and…racial minorities …constitutes ….a cognitive failing…even if the absence …was self-chosen. Longino 2002: 129-131.

These are norms for effective criticism, or what Longino calls the "critical discursive practices" of a community. Her first norm, venues, is designed to ensure that criticism is possible and also that it is valued. Indeed, Longino suggests that science ought to value criticism of existing theories, data and methods *as much as* it does the discovery of new theories, facts and methods.[80] Hence, journals of criticism and public recognition of good criticism must become a major component in the social structure of science. This makes sense—if criticism carries the objective warrant for data and hypotheses in science, then surely it must be accorded equal standing and reward, especially if we want our best scientific minds to work on it.

Equally obvious, if criticism is offered (and even heard) but not followed, that is, if it has no impact on the knowledge practices of a community, then it is impotent. That is, if criticism is most often ignored or consigned to the dust bin then it will not be a very effective method for ensuring objectivity. Therefore, Longino requires that criticism must not only be acknowledged but further uptake must occur, that is, the theories and practices of a community must change and conform to relevant criticism. Her specification for how, why and when this uptake must occur and guidelines for relevancy are vague, which I will discuss later. But for now, the take-home

[80] "In addition, critical activities should be given the same weight or nearly the same weight as is given to 'original research'…" (*ibid*. 129).

lesson from Longino is that mere lip service to criticism is not enough for ensuring objectivity—a real change in intellectual practice must occur in light of relevant criticism of it.

Longino's third norm for achieving social objectivity, which requires there be shared public standards, does not imply unanimity about a single set of necessary or sufficient standards either within or across scientific communities (*ibid*.: 130-1). Instead, for Longino public standards simply means that a community needs a common understanding of what is required for criticism to be made relevant to the community's goals. Under this umbrella of common understanding would be shared referring terms, principles of inference and values/aims, all of which makes criticism possible and allows scientists to communicate and be accountable to one another in the scientific community.

This common understanding, including community wide goals, may be implicit, which Longino thinks is fine (ibid. 130). The point of having *public* standards is that they require "individuals and communities [to] adopt criteria of adequacy by which they may be nonarbitarily evaluated…with respect to shared values and standards" (ibid. 130-1). These standards themselves are not set in stone[81] but are also open to critical evaluation by the community and are "themselves subordinated to its [a community's] overall cognitive aims which will be implicit in its practices even if not fully explicit" (*ibid.*130)."

Longino's fourth and final norm is tempered equality. This norm requires that, with certain qualifications, intellectual authority is granted equally to all members of the community. Some qualifications are included, such as the fact that different humans have different intellectual capabilities, whether these are due to innate abilities or having had different

[81] Clearly adopting a Laudanesque approach to standards, Longino states: "…standards are not a static set but may themselves be criticized and transformed, in reference to other standards, goals, or values held temporarily constant." (2002: 131) Her approach to the social criticism of standards and methods, even aims, is very reminiscent of Laudan's (1977) rectilinear model of scientific objectivity and progress.

opportunities that would improve their abilities and so their authority, such as better or more advanced educational opportunities, etc.

She also distinguishes intellectual authority from cognitive authority, with the latter referring to those specially trained in the topic under discussion. Determining how to temper equality, therefore, must be decided based on how it works to achieve the goals for which this norm was devised. The reasoning behind this fourth norm is the view that "a diversity of perspectives is required for …epistemically effective critical discourse,…[one] in which all relevant perspectives are represented" (ibid. 131). This is important because *consensus* must be "the result of critical dialogue in which all relevant perspectives are represented (ibid)." In particular, Longino stresses that consensus cannot be the result of economic or social power if objectivity is the goal. Consensus must be achieved by "the persuasive effects of reasoning and argument…" and the result of unforced, un-coerced assent to "the substantive and logical principles used in them" (*ibid.* 131-2).

Because she rejects the existence of any general principle of inference or evidence, which is why she claims focusing on reasoning and argument alone will not do, then a question arises as to how to weed out relevant from irrelevant viewpoints. For consideration of exclusion, she offers the above rejection of two criteria, economic and social power, as reasons to reject consensus. Shortly thereafter, she states that female and minority viewpoints not only cannot be excluded, but more strongly, she claims they must be represented or else the scientific community loses epistemic warrant for its claims (ibid.: 132). We will need to unpack the slew of background assumptions and case studies that underlie this demand because, superficially, I for one would see this as engaging in reverse discrimination. This is because if we inquire as to what indicates a minority (it is not simply a question of population numbers), then it seems that

we have to refer to economic and social power, or the lack thereof, and this was seen as one of the things that we were not to bring to bear on consensus formation.[82] I will return to this later.

5.1.2 One norm to rule them all—why the fourth norm is the key. Lloyd Ericson notes that "[b]ecause the criteria all hinge on the fourth, it is by that fourth criteria that the objectivity of science is decided…." Let us figure out what would lead him to make such a claim. Looking at Longino's first norm, having venues only ensures a public forum for airing criticisms, not that the criticisms are a result of or come from diverse viewpoints. Public standards also do not ensure diversity of viewpoints, they only provide for common points of engagement, and further, these points are determined (validated) by the community (communities) involved. Also, if the community is not diversified, or its members granted equal intellectual authority, then determining uptake, which criticisms must be absorbed and which can be rejected, is again left to the community, which if the fourth norm is not in effect would probably reflect local power structures, not a search for diversity.

The success or failure of Longino's first three norms as platforms for facilitating criticism and its absorption all depend to a great extent on how well her fourth norm is implemented. It is the fourth norm that injects multiple points of view into the community mix and it is the fact of multiple points of view being brought to bear on questions of evidence and inference that is the source of objectivity on her account. Hence, the first three norms all hang on the fourth, and so too the fate of objectivity as well.

Ericson goes on to assert: "While it is certainly hopeful that an equality of intellectual authority exists in the scientific community, it does not exist; nor can its possible existence ever

[82] From the perspective of standpoint epistemology, minorities, in certain situations are accorded great epistemic knowledge as they have insight both of their own minority situation and the dominant position (e.g., the people they work for in the upper classes), while the reverse is not true. See S. Harding (1991) for an application to science. Longino, however, seems not to propound either perspective (e.g., men versus women) as superior but simply as bringing different perspectives to bear.

be verified." Let us leave aside the empirical question of whether his assertion is true or false,[83] the upshot is that as it stands, if the fourth norm is not met, then his claim is that her account of multiple perspectives will simply reflect the local power structures in place. In which case, according to Longino's own perspective, criticism need not (and probably will not) lead to objectivity, or at least not to a sufficiently high degree of objectivity, for underwriting scientific claims.

This concern arises because there is no *general* principle, other than an injunction for plurality, to assess criticism or the constitution of the group, that is, whether or not the points of view brought to bear on a case, are relevant or not. There may be local constraints, and especially constraints through commonly accepted methods and observations, but those are not enough to reign in subjectivity and bias according to Longino, which was why we needed the social account to begin with. This focus is established in the second quote from her that I gave at the beginning of this chapter (i.e., the one in which methodological procedures were justified because they had undergone criticism from "as many as possible [viewpoints], or at least more than others.") Now Longino sees her social epistemology as supplementing, not entirely replacing standard principles of empiricism, reasoning and logic. So Longino does have a role, a large role, for principles of evidence and reasoning to play, though by themselves they are not decisive for objectively warranting hypotheses based on data. But let us see how far they can rein in the problems above for objectivity on her social account.

5.1.3 Principles of reasoning & evidence on her account: Longino asserts that the four norms above, especially her fourth one, are necessary but not sufficient conditions for objectivity (1990: 80). To achieve sufficiency, her account

[83] Most likely it varies from community to community and like the tides ebbs and flows, which of course would be one reason why Longino sees objectivity as coming in degrees.

> ...does not dismiss the work done by philosophers of science concerned with elaborating principles of reasoning or of evidence. Indeed, it presupposes basic empiricism and logical norms, while remaining neutral with respect to ongoing philosophical debates (e.g., the status of Bayesianism, multi-valued logics). Instead, it proposes to add to those norms traditionally studied by philosophers (and about whose precise nature there is ongoing disagreement) those that (ought to) govern the social interactions also partially constitutive of scientific knowledge.

By basic empiricism, she is referring to 'empirical adequacy,' the requirement that theories fit the phenomena, for some notion of fit, which can be quite loose).[84]

The primary logical norm that she endorses for reasoning is the principle of non-contradiction. Her statement above presupposes quite a plethora of (sometimes conflicting) methods and principles of reasoning. Why is this problematic? Accepting such a slew of vague and broad generalities makes it possible to support/reject almost any position or criticism, even contradictory ones. This is especially true as many of them are in direct contradiction to one another, for example, Bayesian versus Frequentist statistics. Her neutrality to these debates makes it virtually impossible to pinpoint success or failure in bringing any of these tools to bear in compelling or forcing consensus formation. In short, she has so many, often contradictory, tools to turn to that she (or scientists adopting her account) can explain any success or failure as fitting or "conforming" to her account, whether the explanation from her account is correct or not (i.e., the real cause of the success or failure in a particular case).[85]

Let us turn now to her notion of conformation as a better substitution for traditional philosophical accounts of confirmation and see if this last piece of the puzzle can help tie together the social and methodological strands in her account. How these strands are linked is important because methods and principles of reasoning are supposed to constrain consensus, but

[84] Such a fit can be quite mercurial on her "conformational" account, which allows for partial representations that can quite comfortably conflict. See the next section 5.1.4,

[85] We can see this as exemplifying the one-to-many and many-to-one problems that Faust and Meehl discussed as an objection to the standard case study method in use by philosophers of science, which I discussed in the last chapter.

at the same time, consensus determines which methods and principles of reasoning are warranted in particular scientific inquiries. Her account seems to be engaged in what Harry Collins (1985; 2004) calls an "experimenters' regress," which results because there is no independent principle to break the justificatory circle that locks consensus (which justifies methods) and methods/principles (which justify consensus) on her account.[86]

5.1.4. Conformation Longino sees maps as providing a good model for a variety of epistemological concepts. Maps are used to represent and, in order to be a good map, this representation is necessarily partial, "otherwise it fails to be a map. The best map is the one that best enables it users to accomplish their goals…"(2006: 116-7). Just as "maps fit or conform to their objects to a certain degree and in certain respects…" she is "proposing to treat conformation as a general term for a family of epistemological success concepts including truth but also isomorphism, homomorphism, similarity, fit, alignment, and other such notions. …This approach avoids the crudity of a binary evaluation" (*ibid*.: 177). Her concept of conformation is a very loose notion of fit, similar to that used by Ron Giere in his modeling account, as she points out (*ibid*.: 116 fn 29). However, there is an important aspect about maps that, like Giere, she overlooks. And that is, a key factor in using maps successfully to meet your goals requires knowing where the potential sources of errors in a map lurk—that is, those areas where the representation as presented would not allow one to achieve their goals.

A clear example of this point can be drawn from the practical science of navigation. When laying out courses for a very long voyage, one would not want to plot the route on a standard Mercator projection chart (map). This is because the earth is a sphere, and the Mercator projection does not take that into account. For short distances, the errors are negligible using the

[86] I discuss Collins experimenters' regress in Chapter 6 and argue that Mayo's severity concept can break it. I think the same is true for Longino's account as I will explain later in this chapter.

flattened projection, but over long distances, they are huge. Using such a flat plot to go from California to Japan, the shortest distance is actually to arc up over Hawaii. (You have probably already seen similar great circle routes in the back of in-flight magazines.) Similarly, flying from NYC to London, one would follow a curved path up over Greenland. In fact, the appearance of Greenland on the Mercator projection illustrates well the error that arises from representing the earth as flat rather than curved, for it looks larger than the continent of Australia!

So how does one plot out a long distance route? The common solution is to use a great circle chart, one in which the curvature is maintained in plotting lines, for laying out the entire trip. So why not just use these charts, and skip the Mercator projections all together? There are two main reasons. First, they show too much area and so are unusable during the trip to plot the actual course over ground on, which especially with the effect of currents, wind, and wave action is quite different to the plotted course. It would be like using the map of the USA in the front of an Atlas to draw out a sightseeing trip around Blacksburg and Virginia Tech for a visiting student—there is not near enough the resolution required at that scale for that purpose. Second, actually steering a great circle path would require that the helmsperson keep incrementally changing course (compass) headings, an impossible task for a human.[87] So the way to plot a long voyage is first to draw it out on a great circle chart, then break down the great circle plot into legs of the trip (rhumb lines) that are drawn on the smaller scale Mercator charts and that the helm will steer.

Notice, a key consideration above was what *errors* there would be in each choice of projection for which task. This emphasis on the failure of maps is also seen in the mandatory use of publications for correcting charts, such as *The Notice to Mariners* and *Sailing Directions* that navigators both use and also are expected to submit corrections to, if they find errors on their

[87] The exception is if one is going due north or South, or following the equator, for those three are great circles.

charts. Thus, in keeping with Longino's analogy, it is equally important not only that one's representation "fit" the goal one is using it for but also the ways in which it would not fit has equally well been ruled out. Longino's analogy of mapping, which she uses to underwrite and justify her account of conformation, if the analogy is to hold, implicitly relies on meeting the second part of Mayo's severity criterion, That is, for Longino's map analogy to hold, she must appeal to some criterion of "not fitting" in assessing conformation claims. I think this is implicit and by making it explicit, we see the potential compatibility between her approach and Mayo's. Mayo, too, allows for a wide latitude of notions of "fit," (what Longino calls "epistemological concepts of success") to be drawn upon in her account of severity.

In discussing some of the elements of Longino's account above, I have raised some initial concerns about it. But to really understand her position and her meta-methodological approach, and see if and how the dilemma I posed at the beginning of this chapter can be resolved, we need to scrutinize this approach more closely. In particular, let us look at criticism and its role in promoting objectivity in the face of Duhemian concerns about the underdetermination of hypotheses by evidence and the role Longino sees her norms, particularly the crucial fourth norm, playing in resolving that philosophical problem. The underdetermination problem is the impetus for Longino's rejection of Popper and her turn to her MMP as the source of scientific objectivity.

5.2 Contrasting Popper's & Longino's accounts of criticism. Popper was the first social epistemologist, and he promoted criticism as the vehicle for objectivity. I begin with a brief recap of Popper's view on the social workings and value of criticism and then turn to Longino's criticisms of him. This sets up many of the problems she sees herself as solving (or attempting to solve). Further, in drawing the contrast between them, we can more clearly distinguish the

novel features she claims for her approach. Longino (2002) also draws a similar contrast to start her project.

Objective inquiry for Popper is based on his method of conjectures and refutations. If an anomaly is observed, a scientific way of dealing with it is to modify the theory or hypothesis—not save it at all costs. That would not be objective. For Popper, objective criticism is effective just to the extent it is capable of falsifying a position, and a claim or hypothesis is supported only to the extent that it has withstood serious criticism and yet no flaws are found. In that case, as Mayo (1996, 2008) emphasizes, we may say the claim is corroborated or that it has passed a severe test. *We objectively orient our theories to the world by trying to falsify them.* The trouble is that even if we verify the anomalous observation we are confronted with the classic dilemma as to which of the various background hypotheses or if the theory itself is to blame. Duhem's problem seems unsolvable with only the machinery of Popperian falsification at hand. We will return to this shortly.

On the positive side, Popper would insist on something that it seems to me we all would: namely *if a hypothesis passes a test or agrees with data but the test had no chance of finding evidence against the hypothesis, then there is not good grounds for inferring it.* So for example, if a group of observers holds implicit assumptions so that they would never entertain or look for observations other than those supporting hypotheses about the dominance of males in leadership roles in primates, then the fact that they report observations in sync with this hypothesis is actually very poor evidence for it.[88] This is because they had no chance of rejecting the claim even if it were false.

[88] Here, the humans are the "detecting" apparatus, and in this case, the "machine" is flawed and so, too, are the outputs of those machines. It could easily be the case that their observations are "artifacts" of their biases or that they were literally unable to "see" and hence could not detect relevant data for the phenomenon (e.g.., non-males in leadership roles) under observation.

The italicized statement above is a quite general principle of evidence and inference. Mayo calls it the weak severity principle. In methodological terms, we can state it so:

> ***Methodological version of the Weak Severity Principle:*** If a method has no chance of detecting an error, then the fact that the error is not found is not evidence that the error is absent.

If Longino will adhere to this much as a general principle of evidence and inference both for empirical evidence and for social critical discursive practices (e.g., for evaluating group judgments and/or group composition), then this may be all she needs to tie the two together without winding up in an experimenter's regress and also fend off Erickson's charge that power is the final arbitrator in group judgments.

5.2.1 Moving beyond Popper. Both Longino's and Mayo's projects start with Popper's work and then seek to improve upon it. Mayo's entire philosophy can be summarized as trying to provide ways to implement each of the pieces for falsifying and corroborating hypotheses. We do have methods that are highly reliable at detecting errors, say instrumentation, statistical analyses, etc., and Mayo thinks that is what philosophers of science and epistemologists ought to be developing. She gives a full account of the different strategies for determining the capacities of tests to unearth errors (Mayo 1996). That is, to determine error probabilities. However, the follow-up work would be to instantiate them for various contexts. The worries that arise in say investigating Hormone Replacement Therapy (HRT) will differ from those that arise in studying gravitational theory.

Longino can be read as trying to fill out a strategy for very specific contexts—mainly those where there are likely to be deep seated and hence unarticulated group biases. The most prominent candidates would be where social and cultural biases loom, as they are absorbed at an early age, rather than learned, say, in graduate school, where students are indoctrinated into the

scientific groups' methods and metaphysics.[89] Longino's method is to interject multiple perspectives into critical discussions in order to detect these, to the group who holds them, transparent background assumptions, which are a source of error in their inferences. Clearly these background assumptions must be causing some inferential error or mistake to be made; otherwise, if they weren't, no-one would care. This sort of common sense reasoning, about when biases would matter for inference, already cracks open the door to seeing her account as providing a strategy to unearth errors.

But does that reading short change her account? That is, is she giving us new ways to implement a general error probing strategy in the scrutiny of hypotheses, in which case it falls under the error statistical project? Or—as she surely thinks given her rejection of any general epistemological principle—is she doing something entirely different?

> I argue for a normative social element as part of the meaning of "knowledge," i.e., that epistemic acceptability of content (or epistemically justified acceptance of content) requires the satisfactory performance of certain kinds of social interactions (Longino 2002: 574).

And, if so, how is that so very different from:

> [I]t would be a mistake to think that scientists are more 'objective' than other people. It is not the objectivity or detachment of the individual scientists but of science itself (what may be called 'the friendly-hostile cooperation of scientists'—that is, their readiness for mutual criticism) which makes for objectivity (Popper 1994: 93-94).

5.2.2 Longino's Criticisms of Popper: Longino charges Popper with holding a view she calls the rational-social dichotomy. This is the view that social factors only introduce irrational elements into inferences and any component identified as rational, e.g., scientific methods, must be non-social.

[89] An excellent autobiographical story detailing how graduate students are taught and apprenticed into a group's "paradigm" problem solving methods and thinking can be found in Pepper White's *The Idea Factory: Learning to Think at M.I.T.*

> *Popper vacillated in his conception of criticism. At times he wrote as though criticisms is wholly a matter of logical relations*...That a theory is or is not a solution to a problem situation, that it does or does not contradict another theory, that it does or does not have a particular empirical consequence, are all matters of determinate logical relations, independently of their being thought. Criticism is just correctly identifying the consequences of a theory and comparing them to the empirical basis of science. ...*At other times, he writes as though criticism is a social matter—an affair of competing scientists trying to demonstrate the inadequacies of one another's theories by means of alleged observational discrepancies or conceptual and metaphysical shortcomings* (Longino 2002: 7).

I think her charge above results from a simple misunderstanding of how to disentangle having a general logic or principle of criticism (e.g., Popperian falsification) from implementing that logic or principle in specific local contexts (e.g., a social matter of critical discussions). It is important to raise this criticism and what I perceive as her confusion above, because, as I will argue below, I think she makes a similar mistake in regards to her own brand of social epistemology.

Popper throughout his work argued that it was important to bring different perspectives to bear on a topic in critical discussions. However, he also argued that relevant criticism was always an attempt to falsify a position. Therefore, different perspectives were fruitful just to the extent that they brought potential falsifiers to the discussion. Falsification provided an overarching principle of reasoning for assessing the various perspectives. Longino denies there are any *overarching* aims or principles of evidence for assessing hypotheses.[90] I suspect that she does not realize that she implicitly holds fast to what I have identified above as the "minimal severity principle for evidence" because implementing it is a local affair, and context specific.[91] That is, this background assumption in her account is invisible to her. That she must hold at least some such evaluative principle of evidence becomes clear in her second criticism of Popper.

---

[90]"…there is nothing further, that appeal to standards or methodological norms [e.g., severity criterion] beyond those ratified by the discursive interactions of an inquiring community is an appeal to transcendent principles that inevitably turn out to be local" (Longino (2002: 174)).

[91]I just want to stress her that implementation is always context specific and local as laid out in the severity function (SEV), even though the principle itself is quite general. See my chapters 3 & 4 here for a fuller discussion of SEV.

Longino's second criticism of Popper cannot be so easily dismissed (2002: 6-7). There are, as both she and Mayo discuss, many familiar problems that make getting an account of Popperian learning off the ground difficult. (1) Observations are theory-laden. (2) Determining whether a falsification can be ascribed to the falsity of the theory under test or one of the auxiliary assumptions used in testing it cannot be resolved with a simple appeal to the logic of *modus tollens* (i.e., Duhem's problem[92]). And, (3) Popper's rejection of any form of induction really limits the ability of scientists to learn anything positive from the errors they do detect and uncover (Mayo 1996, 2010: chap 1). His logic of falsification is simply too thin to deal with the complexity and nitty-gritty details of actual testing. As Mayo (1996: 2) explains, in order to get the additional information needed to get a falsification off the ground, Popper with his strict allegiance to making scientific testing a deductive affair backed himself into a corner and so had to write off determining what would count as an observation, an acceptable assumption or auxiliary hypothesis, etc., as conventional decisions rather than as resting on experimental evidence.

For Popper, the specter of criticism forced researchers into letting go of personal idiosyncrasies and biases, because they knew if they did not, that future scientists would find them out. Longino thinks his method is too weak to detect and control biases in the background assumptions used in testing and criticizing theories. Let me be clear here— for Longino, the really troubling assumptions are not about which problem to work on, (e.g., interests like studying Rheumatoid Arthritis because you have it), but are instead what she calls "substantive and methodological" assumptions. Substantive assumptions are metaphysical in nature, often implicit, which determine what kinds of things (e.g., hormones versus chi lines) can be causes and so also determine when a correlation may be taken as evidence of a causal phenomenon

[92] MT: H → e, ~e, therefore ~H but in real testing: (H & A_1…& A_n) → e, ~e, therefore ~H v ~A_1…~A_n.

rather than dismissed as an epiphenomenon. Methodological assumptions concern questions about which are the best methods to use in any particular case (e.g., laboratory studies or field studies to answer specific ecological questions).

Let us step back for a minute and notice that *she criticizes Popper's account based on its inability to detect, much less assuage, these types of assumptions (2002: 5-7). But in making this criticism, she is implicitly accepting and explicitly appealing to what Mayo has identified as a minimal requirement for evidence (words in parentheses are mine:*

> **Minimal severity principle of (lack of) evidence:** "If a method has no chance of finding fault in a hypothesis, (including hypotheses about background assumptions), then the fact that no fault was found is very poor evidence for the hypothesis (or assumption) in question."

The fact of her criticisms of Popper—that she doesn't think he succeeds in detecting and controlling bias, prejudice, etc.; that his account has minimal severity and hence must be rejected or revised in light of its errors and our ability to correct them; and that this is a legitimate criticism of his approach, all indicate that she must accept Mayo's minimal severity criterion for she is adhering to it at least at the meta-level. This brings us right back to where we started—principles of evidence versus consensus for justifying objectivity.

5.2.3.Transparent Background Assumptions As pointed out at the beginning of this chapter, for Longino "[a]s long as background beliefs can be articulated and subjected to criticism from the scientific community, they can be defended, modified, or abandoned in response to such criticism" (1998: 179). But she proposes that often these assumptions are invisible to those who use them and therefore to make them visible, different people, who hold different (and to them perhaps equally transparent) assumptions, need to be brought in.

> Both ascertaining the evidential relevance of data to a hypothesis and accepting a hypothesis on the basis of evidence require reliance on substantive and methodological background assumptions….In general, the assumptions on which

> it is permissible to rely are a function of consensus among the scientific community, are learned as part of one's apprenticeship as a scientist, and are largely invisible to practitioners within the community. Although invisible, or transparent, to members of the community, these assumptions are articulable and hence, in principle public. (Longino 2002:104)

But perhaps the deepest assumptions are those unconsciously learned from growing up and living within a particular social-economic milieu and cultural context.

> Feminist scholars have demonstrated how assumptions about sex and gender structure a number of research programs in biological, behavioral, and other sciences. Historians and sociologists of racist practices and ideologies have documented the role of racial assumptions in the sciences. The long-standing devaluation of women's voices and of those of members of racial minorities means that such assumptions were for a long time protected from critical scrutiny. (Longino 2002: 132)

She supports her claims above and her recommendation that detecting and fixing false assumptions requires multiple points of view be brought to bear in an inquiry based on several case studies, notably ones where social frameworks were appealed to in generating and assessing data. Given the transparency of these assumptions, how do we determine whose perspective to bring in?

5.2.4. Three main sources of alternative perspectives There are three main sources that can (and ought) to be mined for alternative perspectives, which can be found in Longino's account: (1) diversity in the socio-economic and cultural backgrounds of the actual members who make up the community of scientists; (2) diversity of scientific disciplines and sub-disciplines brought to bear on an inquiry; and lastly (3) concerns and criticisms from the larger social community in which a scientific community is embedded, e.g., the public.

To start, in her fourth norm, she gives an across-the-board recommendation that women and minorities should be brought into the sciences for considerations not only of social justice but cognitive success:

> The exclusion of women and members of certain racial minorities from scientific education and the scientific professions constitutes not only a social injustice but a cognitive failing. Similarly, the automatic devaluation in Europe and North America of science from elsewhere constitutes a cognitive failing. (*ibid.*)

Even if women and minorities did not or do not want to participate in the sciences, they must be encouraged to do so as their absence carries grave epistemological consequences.

5.2.5 Reverse Discrimination? In a nutshell, her normative proposal for this segment of multiple perspectives is to institute reverse discrimination based on race, gender, and economic background. This is done in the hopes that diversity of intellectual viewpoints will be found among those forming the complement of the social and cultural features of the traditional majority of scientists. This is because the only way to detect these really deep seated and unconscious social values would be to introduce different types of people into the membership of the scientific community, people of different social and cultural backgrounds captured partially at least by race, gender, or ethnicity. "Effective critical interactions transform the subjective into the objective, not by canonizing one subjectivity over others, but by assuring that what is ratified as knowledge has survived criticism from multiple points of view" (2002: 129). Now another source for this type of diversity can come from (and should be sought from) the larger surrounding community, which is in general more diversified than the scientific community.

An illustrative case of this would be seen in the criticism of and pressure from powerful women (e.g., congresswoman Patricia Schroeder) to replace anecdotal evidence for the success of HRT in everything from slowing hot flashes to rejuvenating appearance to curing breast cancer and heart disease with more rigorous randomized clinical trials and observational studies (see Women's Health Initiative (WHI) study).

Let us grant that clashes between the different groups will detect and discover assumptions, but how do we know we have the right mixture to detect biases? What sort of relevancy criteria will ascertain, for example, which is the right group composition for the task at hand? Will the racial and ethnic characteristics, sexual and gender orientation, etc., assure that what is ratified by those who "have" those features is objective? While I agree that there are very real cases of blind bias in science and testing, I think there are many deep problems with Longino's proposed method for solving this problem.

Now, Longino does state that these differences are to work only at detecting and bringing deep seated biases into the light. Arguing that they have affected the promulgation and evidential assessment of hypotheses requires recourse to scientific methods and principles. But now two problems arise. First, as this source of alternative perspective will often come from the larger public, how do we determine which perspectives will be relevant—men, women, children; Indian, Native Americans, Caucasian, Chinese; Stakeholders or non-stakeholders? Well, I can imagine readers here saying, that would depend on the type of inference or study being entertained. Women, as the intended end users of HRT, and thus the ones that would be affected adversely by side effects, etc., seem an obvious choice for inclusion as they are the primary stakeholders. We can well imagine that in many cases those to be affected or helped by a drug may not be the best choice, as depending on their condition, they may be willing to adopt and support any half-baked idea or accept the most minimal amount of evidence in hopes of a cure. And also, how can they be heard, or who will act as their advocates to provide the scientific methods, principles, and arguments to present their criticism? One could make the case that women's points of view were heard in the HRT case due to money and power, which afforded the new trials.

Second, we need to remind ourselves that it is the very principles and methods that are themselves often under attack (substantive and methodological assumptions) and to which we are to bring alternative perspective to bear, so the standard methods/principles may not apply, in which case, where do we go to fulfill this part of Longino's program?

5.2.6 Multiple scientific perspectives Many will say, and have said, that the above focus on the general social make-up of members misrepresents Longino's position. In the majority of her case studies, she looks not to the non-scientific social backgrounds of members of the scientific community but to their "scientific backgrounds/training" to mine as a source of alternative perspectives to inject into scientific discourse. Here, the perspectives being drawn upon are more in alignment with a Popperian sense of theoretical perspectives or of Kuhnian paradigms. Examples of this sort of multiplicity of perspectives would include, for example, laboratory versus field experiments in ecological controversies, Bayesian versus Frequentist statistics for calibrating radiocarbon dates, and so on.

Longino identifies four different theoretical approaches—behavior genetics, neurobiology, developmental systems, and social-environmental approaches—in the scientific study of behavior, and so on. She argues that:

> [i]n the critical interactions of proponents of these different approaches with each other, several crucial issues emerge in addition to problems inherent to specific methodologies; the characterization of the casual milieu, the nature of causal action/interaction, and the questions held to be important (Longino 2006: 107).

Longino again has found an example where bringing different perspectives to bear in a case will aid objectivity. But if her MMP is to carry more weight than say, the tongue in cheek injunction from Chalmers "to take evidence seriously," then a the real problem, which still remains to be adduced, is how do we determine which perspectives to bring to bear, and further, within the context of critical discursive interactions, how do we adjudicate among the various criticisms?

This is especially critical if we want to have a normative account, not merely engage in hindsight reconstructions.

5.2.7 Relevancy, or what is really doing the work? Whenever it comes to discussing which views need to be taken into account or to which parts of criticism attention needs to be paid, or which parts taken up and in light of which the community will modify its views, the word that Longino keeps plugging in is "relevant." Relevant is doing a lot of the work for her, in what seems to me to be at times a question begging way. We need to ask: what is relevant? And if the answer is whatever finds the error is relevant, then relevant to what? If the assumptions are transparent, then which views would be relevant? Remember, it is unknown what the assumptions are (or even if there are any) that may be or are guiding/biasing the work in question. That was the reason for bringing in alternative perspectives to begin with. But we need a principled way to weed out irrelevant from relevant view points, to winnow out the chaff from the wheat.

Simply having more and more views brought to bear on a topic is neither necessary (many errors can and have been found without the introduction of such alternative perspectives) nor sufficient (we may have a lot of perspectives brought to bear and still not uncover an error) to get objectivity, and as Longino states this strategy of multiplicity may often lead to cacophony. If the point is just to bring in as many views as possible on the off chance that one of them will find an error, that seems to me to be a pretty risky approach. Nor do her other norms really help here—uptake alone may simply be gratuitous. As she suggests though rather ambiguously, uptake must reflect only the good stuff. And we also assume journals will only present the relevant criticism, but both of these already presume that we know what the relevant criticism or perspective is on the topic.

Further, her qualification that what counts both for presenting criticism, and importantly for accepting or uptaking and assessing criticism is divorced from the "perspective," and instead is to be based on the merits of principles of reasoning, and on evidence presented in the argument. But isn't this precisely what is being contested— methods, principles of reasoning, e.g. about casual factors, etc.—and further, given her account of conformation which is so broad and at the same time to contradictory methods and principles behind them, it is really difficult to see how any of these ways of winnowing out irrelevant from relevant viewpoints is achieved except in hindsight, if at all.

Paying attention to arguments of other camps, their principles of reasoning and methods, etc., won't really help within the context of Longino's approach because her neutrality on these very issues makes anything pretty much acceptable. Instead, what I suggest is that it is only by seeing "relevancy" as falling under Mayo's more general concept of severity that will allow us to get a handle on when and why criticisms and proposals are and ought to be brought to bear in particular investigations. And, not surprisingly, these will vary from investigation to investigation as a severe test in one inquiry may well be an insevere test in another (see Mayo 1996). This move towards severity as underwriting relevancy in her approach also strengthens Longino's account by buttressing it against justified complains of vagueness. Further, by localizing (the relevancy of) perspectives and criticisms within the severity triad,[93] we keep intact Longino's emphasis on both the plurality and the localness of methods, including her own MMP.

As I have already covered a case demonstrating criticism from a variety of perspectives at work in the LRL BACI experiment in the previous chapter, let us use that as an example for MMP from different "scientific group perspectives." (See Miller and Frost for more details.)

[93] See my chapter 3 within, severity is a three place function: hypothesis, method, data.

5.2.8 Ecological perspectives: Acid Rain Experiments First, this experiment shows how interests or the perspective of the larger community in which science is embedded affect the study. The Little Rock Lake (LRL) Before-After-Control-Intervention (BACI) experiment was devised to test the effect of acidification on lake ecologies and the final pH level chosen was chosen to reflect local acid rain composition. We can see that in entertaining this question about this form of pollution is in direct response to the concerns of the larger community. So of all the possible sorts of manipulations that could be run on LRL, we can understand why this pH level was considered relevant.

Criticism from a laboratory perspective was directed at LRL BACI because it violated a standard rule of experiment, that of having multiple replicates of both the control and experimental units, which also should be randomized in order to claim statistical validity, for example, running significance tests using the data from the experiment. The hallmark of a BACI experiment is that there are only two units involved—the manipulated (intervention) and un-manipulated (control) unit, albeit each unit is of large scale (1/2 a lake basin each). The *relevancy* for why such a violation was problematic could be (and was) *cashed out as an error of concern* that is uncontrolled for in the design BACI experiments. The problem could be seen arising from an inability to rule out a potential source of error—that of an event occurring in one basin but not the other and that had long term effects that would confound the results of the experiment. [94]

On the other hand, from the point of view of field ecologists, *criticism* can be and was directed at laboratory studies of the effect of acidification on target species, which were done in bottles, as not being extendable to real world environments. This problem is known as the problem of "external validity" in economics (see Guala 2008). The laboratory experimenters'

[94] Other types of events could be accounted for with proper sampling regimes, see chapter 4 here and Miller & Frost.

simplifying assumptions were overly simple. This *error of extension* is what made the field ecologists' criticism relevant to the laboratory experimenters. *Relevancy in both cases of criticism and its 'uptake' is cashed out in specific errors regarding the ability of the data to pose a severe test of the inferences being entertained in each case.* It is the ability or inability to make an argument from error that provides both a necessary and sufficient criterion for claiming relevancy in assessing and acting on criticism and hence onto the point of view being brought to bear on the subject.

What about other suggestions, such as Hurlbert's (1984) to use graphical techniques for the presentation of results, which would avoid the charge of technical invalidity in using statistical tests? It avoids charges of invalidity because replication and randomization are not explicitly technical assumptions underwriting the validity of graphical presentations. Or Bayesian perspectives that claim their perspective should be adopted as replication does not pose a problem for them to calculate quantitative evidential measures in BACI experiments? Well, if we understand that the error is one of design and remains in spite of these multiple ways of presenting the data, then these alternative methods seem not to be relevant, at least not in the ways that aim at sustaining objective inference.

We should note here that on Longino's account, if we look at what she says and ignore her examples and not try to extract similar principles from them, which often seem to be picking up on severity or similar type error criticisms to work, then it seems the Bayesians are actually being the most objective, at least insofar as their social interactions with the rest of the scientific community go. Why do I make this claim? Well, they are willing to include many points of view into their calculations (expert opinions within the field of ecology generally but not solely). Also they have done work to uptake input from other methods— for example they are at pains to show that their calculated outcomes can simulate outcomes that would be expected from Frequentist methods, like significance testing. In point of fact, they are today the ones trying to develop a mixed approach (i.e.,

one that draws on a variety of statistical perspectives).[95] Further, they are working to meet the goals ecologists hold: providing a quantitative measure to outcomes, etc. By the lights of MMP, these Objective Bayesians are in their critical discursive practices the most "objective" game in town. Their selling point is that they will allow ecologists far more inferential freedom and yet their practices is also quantitative.

Now, Frequentist ecologists have also developed methods for simulating and testing what types of error could be expected to result in the ensuing p-value associated with the LRL experiment. This has been done both by ARIMAs and by a new technique, called RIA, randomized intervention analysis, which itself was developed and tested as part of the LRL experiment. The upshot of these methods was to put bounds on how far off the measured p-value could be from the actual (but unobtainable) p-value under various circumstances, e.g., types of errors (see Miller & Frost).

This type of reasoning does not allow for the carte blanche inferences offered by the Bayesians. Further, this modeling perspective is itself firmly entrenched within the Frequentist camp. As far as meeting diversity of perspectives requirements then, Bayesians are clearly more objective on Longino's account, though in Ecology they are still a minority view. From a Frequentist point of view, as they are seen as sweeping errors under the carpet rather than dealing with them openly, the Bayesian perspective is judged to be less objective. Nor, as Longino suggests, can we simply turn to the principles of evidence at work in the two accounts—they are in conflict and so cannot resolve this situation. So should we pull in another view here to adjudicate? Or, as I have been suggesting, do we need now to turn to an overarching principle for determining relevancy in assessing and gauging the multiple perspectives in hand and what they are bringing to the table? (And also, ideally, for indicating what type of perspective would be needed to find the errors, or mistakes, being made in the current ones being held.)

[95] See for example *Statistical Science* **18**(1) (2003) Berger 2003 and commentaries.

5.3 Two roads in the wood. I think the best reading of Longino is that she doesn't want to avoid principles or aims. We want severity, but in order to implement it, we need the social mixture. So the unique feature of her account is:

> **Longino's definition of objectivity:** H is adequately objectively ascertained by consensus *iff* that consensus was achieved because H has survived scrutiny from multiple relevant perspectives—by a socially and culturally diversified group of scientists.[96] (To be consistent and reflexive, this would also apply to STS claims.)

Objectivity here is primarily a social feature but now we have two different routes open to justify it, one following the path of Kuhn, the other following in the footsteps of Popper.

5.3.1 Kuhnian path: For Kuhn, objectivity is defined as whatever the group accepts (for Kuhn, the group was a scientific research group). Longino adds her requirement that the members of the group be culturally and socially diversified. And there is nothing more to it, nothing deeper to discover as relevancy, which is doing a lot of work on her account, is determined by the group.

> **Method K:** Accept whatever the group says as objective fact, as correct once consensus is achieved—however it is achieved.

But this is an unreliable method for accepting hypotheses. Even if we add constraints from methods and principles of reason as Longino wants:

> **Method K_L:** Accept whatever the group says as objective fact, as correct once consensus is achieved—however it is achieved provided this consensus is achieved based on using whatever methods and principles of reason the group has deemed relevant.

The problem here is that because Longino provides no overarching principles for determining relevancy, whether of method or of data, or for criticism, or for perspectives, etc., then relevancy, which is doing a lot of the winnowing work for her, is too broad and vague a concept to really constrain Method K above. Her minimal empiricism and logical rules do not act as

[96]The cultures being referred to here can be scientific or originate from the larger surrounding society.

much, if any, constraint within her account of conformation. If a group decides to determine species survival rates in an acidic environment based on replicated, randomized laboratory bottle experiments Method K_L above provides no principled way of denying or affirming objectivity to that consensus from a consensus based on running a BACI experiment. This is why it is so unreliable a method for securing/justifying objectivity even though from outside, it may be descriptively accurate. Surely for a consensus to be reliable, then the reasons and evidence it is based upon cannot themselves be entirely group specific or justified only by group consensus. And clearly Longino doesn't want to shield off critical scrutiny of consensus practices in the scientific community, else why would she demand diversity? I repeat…*Longino definitely does not want to go down this road, for her work is all about "breaking out of what the group accepts," busting through those community held biases and prejudices.* Social criticism for her is much closer to Popper's conception of objectivity.

5.3.2 Popperian path. For Popper, objectivity is the result of critical evaluation based on a principle(s) or logic of evidence. For Popper, it was his falsifiability criterion.

> **Method P:** a method or principle of evidence needs to be critically evaluated based on its ability to detect the falsity of hypotheses. That is, on how it works/fails to work in specific contexts for detecting, eradicating or otherwise controlling error.[97]

I have already stated that for Longino to get her criticism of Popper off the ground, she must accept the negative version or weak version of Mayo's severity principle (or some principle akin to it)—in particular, she must accept that: "If bias has made it so that a scrutiny had no chance of finding fault in a hypothesis, including hypotheses about background assumptions, then the fact that no fault was found is very poor evidence for the hypothesis or assumption in question."

[97] See Mayo & Miller for a discussion on how to go about making these types evaluations.

However, I think it is fair to say that she also feels we do make progress in detecting biases, else why propound a method for doing so? And hence, she also embraces the strong or positive version of Mayo's severity principle:

> ***(Full) SP:*** If a scrutiny (e.g., on meeting the standards set by Longino's critical discursive interactions) had a good chance of finding a specific error (e.g., bias, prejudice, etc.) and yet, no error is found, then we have good evidence that the error is absent.

Elsewhere, Longino explicitly accepts severity for assessing inferences based on evidence. (She can hardly do otherwise for she accepts all principles of logic, traditional empiricism and philosophies of science regardless of contradictions that position leads her to.)

5.3.3 Principles for in practice but not across practices? While Longino sees the value in this (severity based) sort of principled criticism, she thinks it only works in cases where there is shared content and practices, but will not work across practices (2002a: 165). She claims there is an entire class of substantive and methodological assumptions that are closed off to a similar type of severity assessment because these assessments are "often blind to the deepest assumptions. Awareness of values and presuppositions is imposed on inquiries through interactions with those who do not share them (*ibid.*)." Now this is an empirical claim, open to empirical scrutiny. However, to test this claim, one must also rely on having some general principle of evidence to assess whether or not the ways in which it could be false have been well probed. This in turn requires examining and understanding why her method works, when it does, and why it fails to work when it should. That is, we need to understand the empirical properties of a method in order to determine when and where best to apply it and also to see if it can and should be supplemented, modified or otherwise improved.

To be clear here: Longino claims that Error Statistical reasoning and principles can only work within groups but not across groups because:

> …the processes that are available to minimize the influence of values, such as intersubjective criticism are only partially effective barriers. While they can make visible and available for consideration and adoption or rejection) some value-laden assumptions, those shared by all members of the scientific community will remain hidden. Such assumptions build commonly held values into the accepted background in the context of which data are evaluated and inferences are made and thus hide those values from scrutiny. (1990:223)

Unfortunately this only shows she doesn't understand the approach, which provides an account of how to spell out the numerous assumptions and tests and methods and checks demanded to generate, model, and use data to reliably learn from errors and biases. What she overlooks, and what is the real strength of ES account is that determining the severity of tests is not localized relative to the community's aims, but more closely, it is relativized to the claims of the specific hypothesis under test and what one wants to (and can) learn. Mayo-style severity requires that for data to be evidence for a hypothesis not only must the data fit the hypothesis (for some specified notion of fit) but that such a close fit would be highly unlikely if the hypothesis is false. And it is in cashing out what it would mean for the hypothesis to be false that we can delineate errors of concern, including biases, etc.

One of the key innovations of ES is that hypothesis testing is a local affair, which Mayo has captured in the "hierarchy of models" picture of testing she proposed (based on earlier work by Patrick Suppe). This requires one to take a "piecemeal" approach to testing. That is, an inquiry must be broken down into manageable pieces. By manageable is meant broken down in such a way that error at all levels of testing (in setting out the hypothesis, in the experimental models, the data models and the links between them) can be partitioned off in ways that allow for them to be reliably detected, eradicated or otherwise controlled. Thus to even begin an empirical

inquiry requires that the hypothesis that is tested and finally inferred or rejected has been narrowed down and rewritten/revised into a testable form.

The role for experimental models is to set up situations such that empirical evidence can be generated that will provide information about some aspect of the phenomena in question, and the data models allow us to render the data so generated into a form that is telling about that experimental aspect (see Mayo 1996: chapter 5). In the vernacular, the experimental "trick," or far more aptly, skill, is to realize that one is learning in an environment full of errors and uncertainty and to narrow down an inquiry so as to be able to learn something of interest despite them. This view is quite in sync with Longino's (2007) view of testing as inherently local and context dependent.

Longino's contribution can be seen to fill out a narrow case where the subject matter tends to open the door to a very specific kind of bias—only part of the much fuller set of biases that Mayo discusses for objective inquiry. *So I think Longino is best understood as not rejecting this principle of evidence but as instead trying to work out social ways of implementing it.* The novel feature of Longino's brand of criticism is her requirement that the makeup of the membership in all (objective) scientific communities adhere to her norm of "tempered equality." She derived this norm from examples where after women were introduced into a field in significant numbers, accepted observational evidence was challenged and in many cases overthrown or re-interpreted.

For example, the detection of the (mis)use of human social organization to understand primate societies specially leadership choices, the "disappearing women" problem in archeology—i.e., even on the traditional man the hunter, women the gatherer paradigm, why is it as soon as the gathering activity becomes the basis of society that the women "disappear"

leaving men as the inventors of "agriculture?" However, this view of "disappearing women," while suggestive, did not stand or fall on women promoting it but required specific lines of evidence to show that women could have and did in some instances make this transformation, which is in keeping with Longino's view.

Alison Wylie points out in archeology, while women brought many of the biased assumptions to light and challenged traditional theories, the success of those challenges was not due to their having a special feminine point of view but because they could provide evidence to back up their claims and also appeals to logic (e.g., inconsistencies like in the disappearing women example).[98] But not any old evidence or appeal to logic will work. What ties all of these challenges together and makes for "relevancy," which is left open on Longino's account, is that each criticism or point of view was relevant only in so far as it pointed out an error in the specific inference in question such that it could be shown that the claim in question did not pass a severe test or error probe. Without this methodological principle for determining relevancy of criticism, her critical discursive practices would go nowhere.

More recently, when women achieved positions of power and money, and more say in science, that the evidence for Hormone Replacement Therapies (HRT) was suddenly cast into doubt, quickly found to be biased, and was subsequently overthrown. What do these case studies prove? For one, it certainly shows that where biases are in play which carry risks for a specific group or in other ways go against their interests, then it seems in those cases, at least some of them, that the stakeholders as those members affected by the research results are known may be more likely to unearth such biases. But it is not enough to show that bias exists, what needs to be shown is that those biases affect the severity of the test or otherwise negatively affect the results.

[98] E.g., Scientists look for evidence in current 'primitive' hunter gather societies to see if women are in charge of agriculture, and agricultural innovations, and in recent change-over societies, etc.

For example, sexist attitudes allowed the medical community to accept a much lower standard of proof—relying on anecdotes and hearsay—when dealing with women's health issues and drugs rather than holding to higher standards, such as randomized controlled trials. But as this case suggests, though women's input detected the use of lower standards, the rationale for requiring randomized controlled trials was based on standards of evidential rigor for severe tests. Randomized controlled trials allow one to argue that even though various confounding factors (age, money, health, etc.) are operative in the control and treatment group, they are equally operative in the two groups and hence will cancel out. Notice here, we can appeal to an overarching principle of evidential warrant, severity, which is quite general and can be appealed to in elucidating and arbitrating cases where different methodological assumptions and points of view come into conflict.

Introducing a different perspective, however, does not seem to have the same ability to cancel out confounding factors. Revisiting the primate example, we have men making observations of alpha males as the way for understanding troop leadership dynamics and choices, and we can introduce women, who instead see females coaxing and manipulating to get their way—how do we determine which is correct? How could we tell if both views are incorrect or biased? Longino also comments that perhaps neither side has the correct view. Okay, but now what should we do? Should we add yet another perspective? No, instead we need to look at and make the claims under test exact, and hence testable, so that we could determine what observations would be needed or could not occur if the H were false.

And here again, it may turn out that setting up a randomized observation protocol could work better for controlling biased observations than introducing observers with different biases. Moreover, knowing which biases we are concerned with can help determine which types of

observers would be in the best position to identify errors. But once we can do that, it may turn out that we don't need different observers but more aware observers. My point here is that by focusing on error, we can get a handle on assessing and criticizing methods that come from a variety of viewpoints and diversity of scientific practices. Longino's account, as it relies on viewpoints to criticize methods and unspecified methods to determine the relevancy of viewpoints, uptake, criticisms, etc., winds up in a *cul de sac* for adjudicating the very type of Duhemian disputes that she set out to resolve. But by switching out her reliance on the term relevancy and instead substituting Mayo's concept of severity, Longino can go much, much further in showing how, when, where and why particular viewpoints ought to be brought to bear in specific inquiries, as error probes, for ascertaining and warranting objectivity.

The point I want to stress here is not that diversifying the body of scientists will never work but that it works only in specific situations, in particular those where observers and researchers draw (implicitly or explicitly) on human social experiences to design their observational & experimental protocols and apply in their observations when such use is not only unwarranted (almost all of the time) but will lead to mistaken inferences. Not to minimize her contribution, but we also need to understand that her method works by introducing alternatives but those alternatives (whether in hypotheses or experimental designs or data modeling) alone ca not determine which one is correct. Longino herself always ends up calling on traditional scientific intuitions and methods (even though she is unable to articulate them clearly and they often conflict and contradict with one another) to warrant which perspective is correct at the end of the day. This makes it even more puzzling why she wants to reject similar principles at work in choosing perspectives that would be helpful in specific cases.

5.3.4 Standard cases where gender, etc. really do count in experimental design. Does this mean that the characteristics of the person conducting inquiry never matter? No. For example, in certain psychological studies, the gender of the interviewer can affect the outcome of responses, which introduces errors into the study and hence must be accounted for. But note, this is not a blanket recommendation that fits all inquiries—only those in which it can be shown to detect or correct specific errors in specific hypotheses. Moreover, we can do experiments to aid us in the detection of these types of errors as seen in studies of the effects of the gender, age and demeanor of surveyors on responses to surveys which is one area of error research. In making medical claims about women or for testing air bag safety and efficiency, women test subjects will be needed; we can no longer assume that the male is the "prototypical" or "generic" human. No error statistician will dispute this point! The key criterion here for including a viewpoint is that the view itself must be (conceivably) relevant for probing an inference for specific errors.

In sum, the nature of the hypothesis will (in some, if not all cases) suggest potential errors that may be implicit in it, while in other cases the method for gathering data may also suggest specific types or needed partitions of errors. While inserting people with different viewpoints or scientific backgrounds *may in some cases* be the most *efficient way* to detect and overcome biases it is neither necessary nor sufficient for doing so. But these considerations are all made based on basic principles of evidence—that to have good evidence requires inferences be severely tested, that a point of view only carries weight if and to the extent that it can be used to find errors in a particular inference. Judging the relevancy of criticism on the characteristics of the person giving it, except in very special cases, is a very unreliable method of criticism.[99] Notice this same line of reasoning also holds when the perspectives being considered are

[99] This is not to say the famous scientists won't have an easier time airing his views, but this intrusion of power is viewed as detrimental to objectivity and hence calls for methods to weed it out.

scientific, whether they are molecular biology or nuclear physics, based in field ecology or experimental ecology, etc.

5.4 Conclusion: Re-evaluating Longino's Method of Multiple Perspectives. It seems to me that the failure of Longino's MMP is a failure of appropriate localization in two ways. First, rather than taking a cue from the new experimentalism and "going" local, advocates of this method see it as a universal algorithm insofar criteria could be set out for which perspectives would constitute the relevant ones for science, then those perspectives would be called upon in each and every testing situation regardless of the inference of interest or the availability of other, perhaps better methods for detecting or controlling biases. For example, in the case of Hormone Replacement Therapies (HRT) this would lead to undesirable consequences, for we have much more reliable methods like randomized controlled studies for probing errors, including those due to sexist stereotypes and chauvinism. Second, Longino's method of localization to a view point whether it originates in broader social culture or a specific scientific culture, ignores how we can use the inference of interest to help delineate 'transparent' hidden assumptions as the ES triad forces us to do.

What Longino has shown is that we need to avoid biases if they prevent severe or reliable evidential scrutiny. She has failed to show that there are person or group specific criteria that will accomplish this. So either Longino must deny that she and others have grounds other than their own prejudices/biases for choosing which mix of perspectives are objective or else, she must admit that she is appealing to some minimal principle of evidence like Mayo's SP in choosing whose views to include and whose to exclude.

Longino's account has no mechanism for localizing or relativizing 'viewpoints' to specific inferences similar to the ES severity triad. But if she did, if she would admit that she and

others are holding something like Mayo's severity principle, then the conundrum would disappear. Moreover, having a principle of evidence, even a weak one, to refer to, would allow her to better target particular types of studies, methods, specific theories, etc., for detrimental biases and try to work out solutions to those problems.

In short, while others have taken the fact that social factors invariably enter into obtaining knowledge as grounds for rejecting principles of evidence and inference and incorporating purely social factors in their stead, I want to make the bolder proposal that this fact instead shows the viability of social methods for pursuing objectivity to the extent that their properties for detecting errors can be scrutinized and justified based on principles of evidence and inference in the same way that other methods are assessed, evaluated and justified. That is, rather than redefining and weakening objectivity in terms of social factors, instead epistemologists should look more deeply at and develop stronger, more accurate social methods with properties that help secure objectivity.

Chapter 6: Sociological and Anthropological Approaches to Experiment

[O]ne can never fully capture the world in formula or description, because the meanings of formula descriptions cannot be separated from what is done with them. And people learn what is to be done with them by being immersed in social groups as well as being taught explicitly. (Collins 2007:746)

Methodologically, ANT [Actor Network Theory] has two major approaches. One is to "follow the actor," via interviews and ethnographic research. The other is to examine inscriptions. (Nancy Van House).

6.0 Introduction. Philosophers, like Helen Longino, are not alone in championing a social epistemology for understanding how scientific knowledge is produced and warranted;[100] nor are philosophers alone in pursuing naturalist and new experimentalist accounts of science. Many sociologists rebelling against logical empiricism in philosophy and the Mertonian school in sociology,[101] claim that sociological analyses should not be confined to the study of the institutions and social practices of science but are pertinent for understanding and justifying the actual contents of scientific knowledge claims. Their claim is that the acceptance and rejection of scientific theories is better understood by looking at the dynamics of social interactions between different groups of scientists. More strongly, several schools insist that epistemological factors ought to be ignored entirely as explanatory vehicles for closing controversy as they either play no role (e.g., Bruno Latour's call for a moratorium on epistemology) or not a decisive role (e.g., Harry Collins' Empirical Program of Relativism, EPOR).

Their claims, however, are based on a misunderstanding. In rejecting epistemology, what they are *really* rejecting, and rightly so, is a particular approach to scientific inference that was championed by logical empiricists and positivists. This is the view that an adequate epistemology of science will be an algorithm for relating evidence to hypotheses (and evidential relationship).

[100] In other words, social epistemologies cover both the context of discovery and the context of justification.
[101] See Zuckermann (1988) for an in-depth discussion of the split and suggestions for rapprochement between Mertonian scholars and micro-sociologists.

Much like a mathematical function, such an evidential relationship was to be a purely logical construct, independent of what type of hypothesis was under scrutiny, of the means used to obtain the evidence, and so on. Let me stress here that the hope was to develop a scientifically accepted universal logic or method that would relate any individual piece of evidence to any hypothesis. Of course, if science operated based on such a universal method, then in applying it (i.e., being scientific) there could not be any disagreement, certainly not prolonged, unless one of course departed from the rules, but to do so, of course, would be to act irrationally.

6.0.1 A false dilemma. The dilemma these sociologists offer us is briefly the following: theory choice is either determined by following some algorithm relating given evidence to given competing hypotheses, or else it is determined by non-epistemic, but rational, social factors. Clearly, when we look at the disagreements, especially those where the methods themselves and the status of data as evidence are in dispute, then given the above characterization, epistemology is not the determining factor. Therefore, if epistemic factors are not decisive, it must be that social factors are decisive.

The dilemma is valid but it is not sound because the first premise is false. This assumption is false because there are several non-algorithmic epistemologies of science available (chapters 2-5), and further, I have argued that at least one of them, the error statistical approach (chapters 3 & 4), is quite viable for weighing in on these very types of long running methodological and evidential disputes. Because the dilemma they set up is false,[102] any conclusions drawn from it are questionable. I illustrate the failure of this very common

[102]Outline of the basic argument:
Premise 1: Either theory choice is determined by epistemological factors (understood as some type of an evidential relationship algorithm) or sociological factors. (E v S)
Premise 2: There is no epistemological algorithm for the choice. (~E)
Conclusion: Therefore theory choice is the result of a social resolution of conflict. :. (S)

sociological argument by looking at the work of Harry Collins and Bruno Latour in this chapter. (I also point out what I find to be strengths in their approaches.)

6.0.2 Collins' meta-methodological linchpin. Collins' EPOR is very much in the spirit of Kuhn both in embracing a sociological explanation of scientific evidence, method and theory choice and in championing the view that the best case studies for learning how scientists produce and certify knowledge claims are to be found at the frontiers of science. For Collins, a key step for claiming that social rather than epistemological factors are the determining factors in closing scientific debates is achieved by looking at cases that have not yet been closed. I focus on this meta-methodological claim in this chapter.

6.0.3 Latour's meta-methodological linchpin. Latour denies both social and epistemological explanations of technology and science, or technoscience as he calls it. He does not want to distinguish the two. Latour claims both society and nature are consequences of technoscience, and thus as a meta-methodological rule, neither can be assigned as a cause of scientific knowledge. Instead, he wants to view all things (animate and inanimate, human and nonhuman) as being enmeshed in and constructed by webs of power relations, enrolling and being enrolled by one another along contingent historical trajectories. For Latour, knowledge exists to the extent that it is "picked-up" by others and passed onto future generations. He is specifically concerned with what economists call "external validity"—how results, techniques and knowledge created within a controlled laboratory environment are reliably (or in his words, stably) extended out from the lab into the wider world.

6.0.4 Stories from the frontier. This move away from rational reconstructions, which rely on already knowing which "side" won before they even get off the ground, to look at how closure was achieved regardless of final outcome, is similar to the error statistical approach to

normative naturalism. Both micro-sociologists and error statisticians argue that the most instructive case studies are to be found at the frontiers of scientific research where learning is still going on. As Mayo emphasizes throughout her work, this frontier where the data are inexact and noisy and the arguments are still being formed and the data's status as evidence is under debate is precisely where researchers interested in epistemological questions about evidence and inference should turn their gaze. But where the ES school sees a (rather messy) conglomeration of tools and arguments from statistics, experimental design, indeed a hodgepodge of strategies used by humans to detect and correct errors, micro-sociologists see a variety of social interactions and tactics and other activities engaging both humans (agents) and non-humans (actants). Notable among these activities are "negotiation" (Collins) and "ally enrollment" (Latour) to achieve closure and consensus and to bridge tacit knowledge between individuals and research groups.

Although they grant that "reasoned arguments" are presented by scientists during debates, they claim that such arguments are either not themselves decisive (Collins, et al.) or can be better understood as small arms weapons in a larger arsenal of enrollment tactics (Latour, et al.). These sociologists purposely put aside epistemological rationales in their attempts to pursue social science accounts of science. They assert that their accounts describe how scientific knowledge, in fact, is obtained.[103] One of the most interesting facets of these accounts is that they embrace the "reflexive view," i.e., that an empirical account of science must itself toe the line set for empirical accounts.[104] Thus, it seems a useful exercise to assess the empirical properties of their

[103] This turn from studying social institutions to the contents of scientific knowledge is pace David Bloor's command to the troops. An aside: Popper, too, saw the method of science as primarily social—though to be sure *not at all* in the way they are going about it, see previous chapter.
[104] E.g., Bloor, Collins, Latour, Harding, Haraway

methods just as we would for any other scientific method. The extent to which their methods are reliable and their positions can be reliably maintained is the focus of this chapter.

6.1. Methodological Relativism: Collins (1985, 2004 and 2007) argues for methodological relativism as the correct approach for sociologists studying scientific controversies. He claims it is a thoroughly scientific method as well:

> Methodological relativism is little more than the scientific prescription to investigate one cause at a time by holding everything else constant. In this case, where the science is contentious, we hold the science constant (treated as a not causally contributing variable), and concentrate on the social variables. This does not mean that scientific arguments do not have to be explained, and G*ravity's Shadow* should be unobjectionable to the most adamant realists unless they think the story of the science can be told without discussion, and a great deal of discussion, of the social (Collins 2004).

He claims his method approximates ones used in controlled experiments where potential causes are literally held constant or physically screened off in such a way as to isolate the results of a trial from potentially being affected by them. This physical isolation is why scientists can "treat" them as not causally contributing variables. Physical isolation ensures that the only material cause acting is the one of interest, which is the one being manipulated.

Now, one does not always need literal manipulation and control to sustain this type of reasoning, which Mayo calls an argument from error. The error here would be mistakenly identifying one factor as the cause of an effect one is interested in, when in fact, the effect is due to some other factor operating. Randomized controlled studies achieve this same isolation by distributing confounding factors between the control and experimental group in such a way as to permit the argument that even if such factors were operative, which they probably are, they would affect both groups equally. Provided the assumptions of the randomization are met, then any difference in results between the two groups may reliably be ascribed to the presence of the manipulation in the one group and its absence in the control group. Does Collins' EPOR method

have the empirical characteristics to physically isolate confounding factors, or allow one to make an argument that the method enables one to argue *as if* they had isolated confounding factors? A natural question here of course, is how does he actually hold the science constant?

6.1.1 Holding the science constant: 3 assumptions. Collins attempts to show that his EPOR program is a "controlled" experiment used to test a causal claim, an experiment in which he holds the science constant and manipulates social conditions to see if they affect how closure of a controversy is achieved. His stance is based on three assumptions: 1. each side in a debate is acting rationally; 2. the methods or techniques are themselves under dispute and so cannot compel closure; and 3. the correct result or answer is unknown, hence truth cannot explain closure except in hindsight. It is these three assumptions, I think, that underlie Collins' belief that he has screened off all but social factors in the debates at which he looks. The contentious part of his system is his rejection of epistemological factors as the deciding force in resolving scientific controversies. This raises another question, however. Are there epistemological criteria independent of the truth to guide decisions? I return to this when I discuss his experimenters' regresses.[105]

6.1.2 EPOR as a technique for double blinding. What properties does Collins' method have to control which errors? As I will explain, I do *not* think he as adequately controlled for confounding factors (in this case the epistemological one, i.e., holding the science still). Instead, what his method is good for is controlling bias both on the part of the investigators (i.e., the sociologists) and their experimental subjects, here scientists involved in a controversy—that is we can see it as ensuring double-blinding of all concerned. The analogy here is to see the

[105] More literally, Collins argues that provided these assumptions hold, then epistemological factors are effectively randomly distributed across all sides of a controversy and hence not causally effective for closure. So he will need only to look at differing social actions during a controversy, not actually manipulate them, to determine what caused closure.)

sociologists as doctors and the scientists involved as patients assigned to different treatment groups. The (double) blinding techniques in his and his students' studies are achieved to a great extent insofar as the "correct" theory (i.e., outcome) has not yet been achieved, i.e., neither side in the dispute nor the sociologists looking at the dispute know which side has "the truth" (e.g., a real effect or the correct theory) or a falsehood (e.g., an artifact or a plausible but incorrect theory). Similar to a double blind trial, neither side knows who got the truth or the placebo, so to speak. Thus his procedure, at least in choice of case studies, does seem to have the property of not allowing hindsight to bias the analysis (unlike rational reconstructions wherein the end result influences and constrains reconstruction), meeting assumption three of his model.

But the hypothesis that EPOR was (originally) designed to probe or test was not which social structures or factors facilitate closure but whether or not social factors rather than methodological or epistemological factors determined closure. Collins claims his method will screen off epistemological factors (hold the science steady), but while blinding techniques will stop bias, it does not isolate the system so that only social factors would be operative, unless one views epistemological factors as operating within some type of shared algorithm. This brings us back to the three assumptions I identified in the last section. In particular, in order to screen off the science, assumption 1, everyone is rational, must be interpreted to mean that everyone follows the same algorithm for assessing evidence and inferences, so that is not different between various sides in a controversy. In short, all accept the same recipe and hence rationality excludes prolonged disagreement and debate. (Only tidy logics and Bayesians need apply!)

6.2 Experimenters' Regresses. For Collins, it is at the frontiers of science, where methods and techniques are themselves being developed and justified, that epistemology gets jettisoned. If there is no algorithm for success, then suddenly scientists appear to be

helpless in learning about phenomena or replicating experiments based on epistemological factors alone. Thus while "ordinarily, the discussion of methodology provides a warrant for study's findings; it explains why the findings are to be believed" (Collins 2007:749), this will no longer be the case when working with novel phenomena. But this only works if there is one method, one algorithm to turn to, rather than, as the ES approach argues, a general type of reasoning that can be appealed to in developing multiple methods and tests.

Controversy and the lack of an algorithm seem to argue at most that replicating and learning about novel phenomena is not an algorithmic process but truly difficult and will take ingenuity, not that they must suddenly turn to social avenues for closure or that their scientific culture must provide the answer. Indeed, it seems, at the frontiers of science, scientists *are being forced out of their normal cultural ways of thinking* and doing things—now they are engaged in exploratory work and can no longer appeal to "black boxed" technologies, hence it is here that the real work of learning must go on. This is also an area fraught with error and uncertainly, and thus if novel phenomena are to be learned about and/or stabilized, then scientists must be able to learn from their mistakes, for surely here is where they will make them. At most, Collins has provided evidence to sustain the weaker claim that the process of scientific inferences based on evidence, including inferences about the evidence itself (e.g., the status of the "data" as "evidence") is not ruled by an algorithm in novel research areas.

6.2.1 Testing the Extent of Tacit Knowledge. Collins claims that knowledge is in good part "tacit" or learned implicitly because:

> [O]ne can never fully capture the world in formula or description, because the meanings of formula descriptions cannot be separated from what is done with

> them. And people learn what is to be done with them by being immersed in social groups as well as being taught explicitly. (*ibid.* 746)

The claim above provides the basis for his experimenters' regresses. Can we reliably draw this conclusion that a good part of knowledge is inherently tacit? No, and Collins actually provides an interesting method for delimiting or determining the extent of tacit knowledge. He attempts to mark off tacit knowledge learned by "cultural immersion" from that which can be passed on in written form by trying to teach non-social creatures knowledge that social creatures like humans learn socially.

One area of investigation is machine learning, for machines are examples of "non-social" creatures. A common example Collins provides are bread machines (non-social learners) versus human bakers. The bread machine basically learns all the bare essentials (what we call necessary and sufficient conditions) but without any of the frills or special touches that individual bakers utilize in making leavened breads. But such differences should not be surprising, remember our earlier discussion, in a replication we are concerned with replicating a phenomenon not duplicating an experiment. Moreover, the point of replicability in this sense is to detect any errors in the original design so as to be able to argue that a real effect has been detected and the phenomenon is not a mere artifact of the experimental set-up (which could include technology, the lab, the persons, etc.). Note also, the problem/differences here can be due less to programming tacit knowledge than to cost efficiency. Many cooks use the dough cycle on the bread machine and then remove the dough ball and spread a layer of olive oil on the outside of it and bake it in the oven for the "homemade" crispy artisan crust. One could design a machine or series of machines to do this but the cost would be much higher than the simple machine now on the market.

The point for philosophers and scientists is that the ability to produce a real effect requires one to know the necessary and sufficient conditions behind its production—not every variation or refinement possible.[106] One key point in replicating novel phenomena and problems in new detecting technologies arises from the fact that what assumptions are background and which are integral to the phenomenon as well as what can be written off as ceteris paribus and what cannot (i.e., all other things normal is still being explored in the non-normal process) are unknown. Collins makes two claims. One is that actually going on site facilitates learning how to build a new technology or replicate a process. This is a claim about expediency in the learning process and is true and fairly un-contentious. The second is that this immersion is not only expedient but necessary and is contentious.

Indeed in his footnotes, where he puts examples of tacit knowledge, one sees that this knowledge can be "brought out into the light" as it were. In the one case the problem of replicating an experiment was due to a wire being too short and it turned out on investigation that there was a critical length needed, and in the other case the color of a wire (red) caused problems in replicating a laser, as the red dye caused interference. Until an experiment can be replicated and scientists *know* why (e.g., color, length) then the scientists have failed to capture, or more accurately to understand well a real effect. They may 'have' it without knowing which facets of the experimental set-up can be written off to ceteris paribus clauses and which actually are critical for producing the effect of interest.

What Collins has shown is that replicating experiments and getting novel technologies to work before knowing exactly how and whether they are working requires great attention to detail, that "unimportant" things may turn out to carry great consequences for success and that

[106] For example, in making leavened bread, one needs a sugar source for the yeast, but honey, maple syrup, sugar of any kind (e.g., brown or white, cubed or loose) will work. We need not fully but only adequately indicate the need for food for the yeast to eat, to make the bread rise.

experimenters must depend on a lot of trial and error type of thinking to get things working and replicating, if it is possible. There is no algorithm for what will be important in any one case, but that does not mean it cannot be discovered and articulated. So, social interactions and access to machines are important for reproducing experiments. But they are neither necessary nor sufficient, because while an experimenter may fail to figure out how to make the machines work because of a lack of social interactions, social interactions are not necessary for getting them to work. This is because machines do not fail because of a lack of social interactions—the wires or something of that sort make them fail. Further, having social interactions will not guarantee that the experimenter will be able to figure out what is going wrong with their machines.

But as we are interested in experimental learning, let us instead ask the reverse question—had the first group not tried to make the machine prior to "immersion" would either camp have discovered the importance of the color red for the wire? No, that is why it is important in claiming that one has detected a real effect (i.e. that one's detector is working) that one does not merely "copy" a machine, but tries to build it from scratch or to detect the phenomenon using another type of detector. This is the way that otherwise encultured and 'tacit knowledge' becomes articulated.

The same is true for replication.

> Experimenters' regress, it will be recalled, draws on the fact that the results of the second experiment cannot bear on the results of the first experiment, unless it is agreed that both experiments were competently done; where there is deep conflict about what the conclusion should be, such agreement is unlikely. The normal criterion for confidence—that the experiment produces a correct outcome—is not available, where what counts is correct outcome is a very subject of the controversy. (Collins 2007: 788)

However, the fact that scientists are exploring what is needed as a criterion of success, just shows how much more important making assumptions explicit is at the frontiers of science. Indeed, it is where there are no clear criteria of success that methodology becomes "un-black boxed" and

scientists can no longer rely on "tacit" knowledge of how to do things, for they have entered now into new territory where assumptions and methodological concerns must be justified.

From the ES perspective, to claim one has evidence for a hypothesis requires showing that one has done a good job of ruling out the ways in which it would be a mistake to infer that particular hypothesis. This hypothesis is often couched in terms of the absence or presence of a specific error, e.g., the effect is real and not an artifact of one's detection machine, etc. While there are an infinite number of potential errors, in many actual testing situations the number is manageable for two reasons. First, there are a handful of general error types into which individual errors may be classified or partitioned that Mayo has identified and I have discussed in previous chapters. To remind us of them, I re-iterate them here:

> (a) mistaking chance effects or spurious correlations for genuine correlations or regularities
> (b) mistakes about a quantity or value of a parameter
> (c) mistakes about a causal factor
> (d) mistakes about the assumptions of the data or experiments
> (Mayo 2003: 102).[107]

Second, testing itself is done by taking a piecemeal approach, which again has the aim of making error detection and management tractable.

To argue that one's data is evidence requires arguing that the procedures or methods used to generate and interpret the data had a good chance of discovering errors in the hypothesis under scrutiny, if present, but not otherwise. This does not require that we need to know the truth beforehand but that we need the ability to recognize or control errors. This holds for methods in general, including our own methods for scrutinizing scientific episodes. If our methodology is such that we would make mistakes about, for example, a causal factor, then any inferences made

[107] Another error can be found in Mayo & Spanos (2010: 19) and that is: Mistakes in linking statistical inferences to substantive scientific hypotheses. There are several versions of these canonical errors, with Mayo 1996 leaning more towards the informal, experimental articulation and subsequent versions such as Mayo & Spanos 2010 learning toward the more formal statistical articulation of them.

(or cases analyzed using that method) are liable to be very unreliable. When looking at some of the methodological assumptions being made by STS researchers, we find good motivations for methodological assumptions; however, the assumptions and hence the methodology founded on them are extremely problematic.

6.3 Conflating realism and epistemology. Collins shares a common new experimentalist/naturalist mistake of conflating realism with epistemology. To see this, first, we need a closer look at the scale and concepts of cause being appealed to in Collins' EPOR. Collins claims that, like any scientist, he is exploring the world using the explanatory level of his particular science. So while physicists look at quarks and atoms and electrical forces, chemists occupy themselves with molecular bonds, biologists with cells and genes, anatomists look to organ systems, and astronomers planets, each takes his causal entities as "real"—Collins, as a practicing sociologist, will study social groups and social forces and take those as real.

On one hand, this position is quite innocuous and unobjectionable for studying phenomena produced by social groups and forces. But it is also disingenuous as seen in his footnote 13, reproduced here:

> I endorse realism is an attitude with scientists of their work and for sociologists at theirs. Thus, I write this book in an attitude of "social realism"—that is to say, I treat the social as real. The difficulty is that for me to treat the social facts of science as real, I need to treat the natural facts in a different way. I believe all of us should be capable of stepping outside ourselves from time to time so as to understand that there are other ways talking about are taken for granted worlds. (Collins 2007: 14)

We need not deny that such causes are real or even empirically adequate in order to raise the question that is the crux of the disagreement between philosophers and sociologists: are these causes are appropriate for the epistemological task they have been set to. That is, are they

necessary and sufficient conditions for producing the phenomena in question?[108] Another question to ask is: why is there this conflict between social and natural worlds? Why must one be real, while the other then must be held as non-real somehow or other? This is merely asserted and perhaps may seem true to Collins, but he has not given the rest of us a reason to accept that natural and social worlds cannot be "real" side by side. Surely in external reality both natural and social factors are causally active, which is why they could be confounders and hence the need for isolation?

6.3.1 Relativism. Moreover, his relativism seems, as Collins hints, to be somewhat arbitrary:

> On the whole, the exact choice about where to relativize and where not to relativize is not very carefully worked out; most of the time it does not need to be. Most of the time, some things are treated as scientific facts and some as facts in the making depending on the dynamics of the story. (Collins 2004:758)

As Collins points out in the passage below, his justification for regarding sociological variables as causal and scientific methods, arguments, and so on as "constant" and hence non-contributing variables in a case, is to avoid "circularity."

> Most of the time, the principle of methodological relativism, when it is applied to facts-in-the-making, needs be seen as no more than a version of the methodological guidelines in every science: concentrate on the explanatory variables. In this case, implies that the science be "held constant," as it were. For facts-in-the-making, the science was not to be taken to explain itself on pain of circularity and\or the dimming of the sociological case. (Collins 2004:758)

The term 'science' is pretty vague here but surely we would want to "dim" the sociological case if it were not the case. But in holding the science constant and only appealing to social factors, he assumes the very claim he wants to justify, i.e., that social variables are the decisive causal factors. At the same time, he dismisses any other explanation for knowledge as circular;

[108] Think back to our discussion in chapter 2. Even if we could reduce cognitive functions to a bio-neural chemical story that would not provide a framework or tools for answering epistemological concerns regarding which of two theories is the better theory.

remember for his brand of social realism, there can be no other explanatory variable than the social.

But given his meta-methodology, we have no way to tell whether the social explanation is incorrect, for it is the only contender—there is no chance of finding a different explanatory variable for closing a case including truth, nature, argument, etc., as Collins himself points out.[109] Nor am I overstating the case here:

> The analyst had to ignore the scientific facts of the matter on pain of producing a circular argument: "this truth came to be established because it was true. Scientific truth had to drop out of the explanatory equation as the new way was to make sense." (*ibid.* 792)

We can certainly agree with Collins' statement above—truth will not provide the warrant for accepting how a fact came to be accepted as true except in hindsight and manifestly this is not the epistemological task. And we can also agree that in many of the debates he has looked at that all parities are acting rationally.

Does this mean then that non-epistemological factors are at work that is, is this sufficient to sustain his claim that epistemological and methodological considerations have been "shielded" off in these debates? No! They would only be shielded if one presupposes that scientific rationality is reducible to having one shared 'method' or 'rule' and hence acting rational would imply unanimity. But of course, this is exactly what the disputes he is analyzing are about, i.e., they are all disputes about the validity of the methods themselves, and thus whether or not one's data can provide reliable, or indeed any, evidence for a claim generated using a disputed method.[110]

---

[109] We can not even broadly say "non-social" in this case, because whatever social explanation is provided must pass, thus this is a very unreliable method.

[110] I emphasize again that having data is not the same thing as having evidence. My experiment could produce all sorts of data, but if the mechanism generating the data was an artifact of my detecting machine, instrument or biases, then this data would not provide reliable evidence for or a severe test of my claim. And while substantiating such

The key point I want to make here is that for his method to be successful in isolating the epistemological factors in these types of debates depends crucially on holding a very specific view of the nature of these factors; in particular one must embrace an algorithmic view of scientific epistemology. But if one does not hold this view, which I, and most philosophers, do not, then his argument that he has isolated these factors fails.

But as I have argued in earlier chapters, one can hold that there is a general style or pattern of reasoning which guides and provides shared epistemological criteria, but given the multitude of strategies and methods and the plethora of phenomena under investigation, sharing even this much of an epistemological standard does not imply unanimity of method. Such tasks (i.e., implementing and instantiating this type of reasoning in various contexts) are manifestly local and context dependent.[111] However, it does set a standard of evidence that if not met, indicates objective grounds for disagreement as well as for identifying points where agreement, if it were achieved, was obtained based on non-epistemological factors. Indeed, if we were trying to be "scientific" as he indicated in the quote at the beginning of this section, we would want to manipulate several variables or causes *and* require that those items we claim are causal actually have the properties that would allow them to act as causes in the case under consideration, whether social or epistemological.

Collins persuasively claims in his online article "Sociology of Science for Non-sociologists" that he is interested in "distal" rather than "proximate" causes. Which means, as a

---

claims is inherently a local task, and an arduous one, these local methods and strategies can be seen as local instantiations of a more general pattern of reasoning, arguing from error, or formally, severe testing, which then provides general shared criteria for assessing their success and failure.

[111] The context I am referring to here, as explained in Chapters 1 and 3, refers to the specific hypothesis under test, data generating procedure used, and the actual data produced as encapsulated in the severity function SEV(Test T, outcome **X**, inference H).

sociologist, he is trying to look at the effect of "social collectivities" on science without denying that other factors (including scientific arguments, etc.) are proximate causes in closing controversy. Most epistemologists could accept this; indeed, EPOR could then be recast as doing Mertonian sociology but in a more sophisticated experimental manner.[112] But of course, it is his experimenters' regresses and the mechanisms that he postulates for the closure of controversies that belie this claim and clarify the differences between him and Mertonian-inclined philosophers.

For Collins, social distal causes close controversies because the usual methodological and epistemological factors cannot be appealed to, and this is the case for him not only during short lived controversies but is also the case in long term, literally decades long, controversies. His views here are very similar to both Lakatos' and Duhems' views that any group of scientists with enough funding and ingenuity can 'legitimately' save their theory by appealing to auxiliary assumptions. This is another point where error statisticians will often part company with sociologists. Error statisticians will claim that one must have (and sometimes can have) good epistemological grounds for blaming auxiliary assumptions, but to continue to do so without such grounds is to follow a very unreliable procedure and to continue to do so is to be unscientific. But how, if the methods themselves are under contention, is assessing blame and burden of proof to be done? Collins' work is meticulous and provides a great deal of information for how such arguments may be built and sustained.

From an ES perspective, the burden of proof falls on the person claiming to have detected a phenomenon to provide evidence for that claim. In the case of Joseph Weber,[113] while his claim

[112] This is the same general line of inquiry that Kitcher (2000) cajoled Collins and others to pursue (Philosophy of Science Association biennial meeting 1998) and that he also follows (e.g., Kitcher 1993, 2001).
[113] Joseph Weber's claims to have detected gravity waves and the reception and ultimate denial of them by the scientific community are the primary focus of Collins (2004).

that he had detected gravitational waves with his apparatus seemed plausible at first, to be accepted it would need to be severely tested. And as time went on, and others attempted to test his claim by replicating his experiments, his claim was rejected by the larger scientific community. It was rejected not only because no one could replicate it but also and indeed more so because serious errors in his work were brought to light as Collins details in his 2004. These included but were by no means limited to: incorrect statistical analyses, that evidence based on simultaneous readings between labs were actually not simultaneous due to being in different time zones (so detecting a gravitational wave at 1100 at two different sites, was not simultaneous but instead occurred an hour apart; and further, these errors, though known, were hidden and later an attempt was made to justify the discrepancy in an *ad hoc* manner. However, these arguments are epistemological or methodologically based insofar as they are arguments about the empirical properties of specific methods for detecting or controlling specific errors in order to claim that the data produced are evidence for the claim being made. And so they are objective not only in being inter-subjectively accessible but in warranting that one's tests or methods tests one claims against the world. Nor in making this claim do I have to say that scientific practice is free from the influence of personal idiosyncrasies or that subjective elements were not at play as well. Indeed one of the main reasons for requiring severe testing, for probing for and arguing from error, is to detect those very factors, among a slew of other errors. So, the point is not to deny subjectivity but to try to neutralize its role in the long run, and to flip Collins, the long run is often a lot shorter than sociologists claim.

At some point, and this will vary with field, subject, tests, phenomena etc., there comes a time when one can, with good reason, deny a claim (or accept a claim) as being reliable. And further, one needs to be very clear about what claim or portion thereof has reliably or severely

passed a particular test, or error probe. For example, in his case, what has been ruled out was that Weber had reliable evidence for gravitational waves, that his equipment was functioning in the way he claimed (due to the irregularities in his method and data analyses, for one thing). But this is not to say that gravitational waves don't exist or even to say that he hadn't by accident at one time or another detected them or something (unknown) associated with them. The claim is that Weber does not have evidence for having done so. And, the objectivity of science hangs on the requirement that one must provide reliable evidence for claims.

6.3.2. Necessary and sufficient conditions. The point is learning is hard work, lots of trial and error, and while it can be facilitated by on site visits, the actual arguments, the trying to figure out what conditions can be written off as *ceteris paribus* and which ones are actually necessary or sufficient conditions for producing a phenomenon is why replication is so difficult but also why it is so valuable. For example, in the one case, the color of the wire actually did make a difference and hence color could not be written off as part of "all other things being normal" or cached under the background that color shouldn't make a difference. When dealing with novel phenomena basic assumptions are questioned—this is why replication is valuable, because when everything else seems to be the same, our focus is turned to previously unquestioned assumptions when things don't work out.

6.4 Actor-Network Ethnographic Theory: While Bruno Latour is often seen as the odd man out in science studies because he does not seem to fit neatly fit into any category or discipline, e.g., philosophy, sociology or history; nonetheless his work has been and continues to be highly influential especially his "actor network" methodology. Let us briefly look at actor-network theory as a method for understanding "technoscience" by (1) unraveling his methodological

rules, (2) his input about external validity and (3) the "meta-physical twist" he brings to the discussion.

Latour and Callon in *Laboratory Life* and Latour in *Science in Action lay* out their methodological program underlying actor-network theory. Latour based his methodological approach to laboratory studies on ethnographic methods of observer/participant studies. However, while bringing no "scientific methodological presuppositions" to bear, he also does not attempt to uncover any "reasoning" principles behind activities but only looks at physical outputs as purely material objects rather than material products embedding ideas.

What do I mean? Latour focuses on the production of inscriptions (e.g., papers, charts, computer outputs, etc. as the goal of scientific lab production, purposely not appealing to any concept of knowledge production. That is, knowledge is itself not a category of description available for him—part of his moratorium on epistemology. However, when he states that the aim of all the activities in a lab, including dissecting rats, etc., are written inscriptions, he does not answer the question as to why did scientists need or feel the need to inject and then dissect rats to begin with in order to produce these particular types of outputs. Moreover, as a philosopher/ethnologist, he does seem to be prejudiced into interpreting scientific activities in light of his own activities, which consist primarily of written outputs (descriptions or representations) rather than manipulations of organisms, machinery, or more generally *pace* Hacking, "interventions." Nonetheless, despite the above reservations, his actor-network methodology has been widely embraced by STS researchers and the outside world, so let us examine it in a bit more detail. It is also well worth the effort to see how his metaphysical twist can sit well within the ES tradition.

6.4.1 Latour's seven rules of method. Latour (1987: 258-259) provides seven rules of method:

> *Rule one* We study science *in action* and not ready-made science or technology; to do so, we either arrive before the facts and machines are blackboxed or we follow the controversies that reopen them.
> *Rule two* To determine the objectivity or subjectivity of the claim, the efficiency of perfection of the mechanism, we do not look for their *intrinsic* qualities, but at all the transformations they undergo *later* in the hands of others.
> *Rule three* Since the settlement of a controversy is the *cause* of Nature's representation, not its consequence, we can never use this consequence, nature, to explain how and why a controversy has been settled.
> *Rule four* Since the settlement of a controversy is the *cause* of Society's stability, we cannot use society to explain how and why a controversy has been settled. We should consider symmetrically the efforts to enroll human and nonhuman resources.
> *Rule five* We have to be as *undecided* as various actors we follow as to what techno science is made of; every time an inside/outside divide is built, we should study the two sides simultaneously and make the list, no matter how long or heterogeneous, of those who do the work.
> *Rule six C*onfronted with the accusation of irrationality, we look neither at what rule of logic has been broken, nor at what structure of society could explain the distortion, but to the angle and direction of the observer's *displacement* and to the *length* of the network thus being built.
> *Rule seven* Before attributing any special quality to the mind or to the method of people, let us examine first the so many ways through which inscriptions are gathered, combined, tied together and sent back. Only if there is something unexplained once the networks have been studied shall we start to speak of cognitive factors.

I note right off the bat that his second rule of method is in direct contradiction to the error statistical approach to studying methods. Where the error statistician urges us to look at the empirical properties of methods to explain their use—how they function in science or to put a Latourian twist on it " to watch the rules in action," Latour explicitly denies that any intrinsic properties of machines, or facts, or methods can explain their use, success or failure. He refers to six principles he holds to justify his methods above, which I will discuss as they arise below.

For example, he justifies the seemingly odd view that the properties of machines, methods, etc., cannot be drawn upon to explain their success with his first principle that "[t]he fate of facts and machines is in later users' hands; their qualities are thus a consequence, not a cause of collective action" (259). On one hand, this first principle is true, indeed it is almost trivially true. If an idea, technology, or method does not get "picked up," used, or passed on, then it dies becoming unknown. But while true, that is quite different than his principle that the qualities of a machine are a consequence of collective action. *A key question is why did it get picked up? What about it made it attractive (or in his words) allowed it to "enroll" others, or in the case of humans, what attracted them about it?*

Most of us want to argue that the qualities of a machine provide one of the causes, albeit probably not the only one, for its fate. The mere fact that the idea has been used does not explain "why" it has been adopted. If we are not to look to society or to "nature" as the cause of its success, then it seems the only place to look is to the thing itself and how it knits the two together (or enrolls the other two). I think Latour has taken the old adage that "the proof of the pudding is in the eating" a bit too literally!

6.4.2 Hormone Replacement Therapy Network. For example, the "fate" of Hormone Replacement Therapy (HRT)—its use—has fluctuated over time. Using Latour's network theory, we could trace out how in the early 70's the HRT network grew—more women were using it, more doctors were prescribing it and magazines were advertising it. It was the miracle cure. However, in the nineties, after several of the Women's Health Initiative (WHI) trials were stopped prematurely due to findings of increased cancer risks, we saw the pro-HRT networks beginning to break up—go into decline. Now on one hand, Latour's network theory seems to offer an explanation—the Networks for HRT were breaking up and others, challengers, were

being built (e.g., many women quit taking it, Congress got involved, there were more medical studies, hence doctors doing stuff with it, etc.) but this alone— while providing insight into how change occurred—does not explain *why the networks were changing nor whether this change was/is desirable (and why.)* To illustrate this problem, ANT would indicate that the results of the randomized controlled WHI trials were decisive in this change, but it does not explain why they were decisive. Ultimately, Latour at some point needs to be able to explain, if not epistemologically then ontologically, why, for example, randomization counts.

One way we could approach this on a metaphysical account would be to include errors as part of the network—certainly they should be able to fit neatly into his ontology, for mistakes are "real" and can be captured not only in inscriptions but in other "physical" ways, that is, they can be detected, measured and identified. We would also need to add another "measure" to Latour's networks, one that rewards the elimination of certain actants (errors). That is, we can look at Latour's networking as a game of Monopoly—besides getting more property as a way to win, if you enroll an error, your network may be penalized by having to "skip go and go directly to jail", but if you enroll a method that can detect or control or eradicate an error, it is like having a "get out of jail free" card. My point here is not to make fun of Latour but to explore how the ES approach can be given a metaphysical spin without too much of a stretch.

6.5 Latour's phenomenological approach. The reason that Latour's account can only touch the surface here is captured by his rules 2-4 which in essence capture the acausal nature of his network methodology. In short and as becomes clear in the last 3 rules, Latour is trying to offer a metaphysical account, a phenomenological account of epistemology. Hence there are no reasons, much less causes, either in nature or society or cognitive factors—there are only

"things" (some are living, i.e., human beings, rats, microbes and others are non-living, e.g., needles, facts, rocks, test tubes, etc.). Further, these things interact in the *same* way—enrollment.

Latour has offered an accumulative, magnetic account of how things (beings and things) group and regroup. He has limited his description of science to a study of such groupings. Success, then, on this view, is based on the size (i.e., length) of the grouping. Moreover, outside of physical presence there are no other characteristics either of things, societies or cognition that exist, much less that are allowed to be appealed to in order to explain the shifting groupings. We can see in Latour's work perhaps the best attempt at a totally neutral description. But even here, we are reminded that in choosing which things to observe, (and how do we know we have observed everything in the network?), some theory is required to guide us. Am I being too harsh on Latour?

No, his principles make this phenomenological approach explicit. Thus his third principle clearly tells us that only "a gamut of weaker and stronger *associations* exist; thus understanding *what* facts and machines are is the same task as understanding *who* the people are" (259). His method has no power to distinguish correlations from causes for in the end there are only associations, the networks. This view that as far as explanatory power, there exist only 'associations' and 'things' is re-iterated in his sixth principle that "[h]istory of technoscience is in a large part the history of the resources scattered along networks to accelerate the mobility, faithfulness, combination and cohesion of traces…" (ibid). This is a very metaphysical approach to epistemology, indeed, it seems as if Latour would replace epistemology with ontology. Such a view is supported by his call for a "10 year moratorium on epistemology."

The rules above seem to be an attempt to provide an acausal description or explanation for "technoscience." But I do not see how his first principle can really underwrite this approach

if what we want to know is why certain practices were accepted by later users, etc. When I say Latour's method seems superficial in many ways, I do not mean to disparage its usefulness entirely. It is, however, in my eyes, inadequate for epistemological tasks. For other tasks, however, it could be a very strong method. Thus, if one is curious as to what scientists do, how they spend their time, what sort of tasks being a scientist (successful or otherwise) entails or what new/old skills would be beneficial to teach future scientists, then this is a great method for it actually enjoins us to "follow the scientists around" and see the wide variety of activities they engage in.

6.5.1 The daily life of scientists. Similar to John Ziman's work, so too has Latour's work been an excellent resource, a method for pointing out that scientists do not just spend their time in lab coats tinkering with experiments, etc. Instead, these studies have shown how much time is spent getting funding, financing and managing labs, and in administrative duties, etc. Thus, they are excellent resources for inquiry into the day-to-day functioning of a lab or a scientist's work life. Indeed, this is one reason businesses want to hire Latourians by not pre-selecting what tasks to include/exclude but where success can only be determined in hindsight by users adopting methods, etc. It seems to me that when all is said and done what Latour has really perfected for the study of laboratory science is to extend Galbraith-type time and motion studies to more 'theoretical' work areas—from studies of factory efficiencies into laboratory efficiencies. But if we want to be prescriptive, to give advice, it seems we need to look at the properties a method has that would make later users *want* to use it.

Latour's acausal or phenomenological approach is problematic on several other fronts. Not only does it not provide epistemological insights, but it seems to be useful only in hindsight, indeed the entire causal order seems to be backwards. Thus, rule two and even three seem rather

odd at first glance, because they appear to get the time order wrong. Causally speaking, users accept certain kinds of facts, methods and machines, or refuse them because they're somehow efficient for the work at hand or they have some intrinsic property which makes them useful. But to say something is useful because people use it is circular. By simply following the things and people around indiscriminately, without trying to weight objects or take into account purposes, if one only cumulates based on number, volume and association, then sheer numbers supply the only criteria for success. But even in common life, we would not want to make such a claim for methods or knowledge—this is but an *ad populum* argument carried to ontological extremes.

By looking at the surface of things, of practices, of activities we do gain insight into what people do, and the various roles scientists and equipment and theories may assume in a variety of situations, but doing this cannot explain why they assume or why it is thought they ought to assume such roles in the first place. Latour's idea, encapsulated in rule three and four, is that the settlement of controversy and claims are the cause of nature's representation and society's stability. Thus we can use neither nature nor society to explain our views about them, which is really to deny that we can explain anything at all. We can only follow and record associations and group things, but not explain why the groupings occur other than through the kinematics of the grouping itself. This is a truly physical theory of the growth of scientific knowledge but it also suffers a similar weakness as cognitive approaches and other reductionist approaches (e.g., to brains or psychology) insofar as it will never provide normative insight into which theories, or facts or methods should be embraced, but only an accountant's book listing those that have been. Science and technology are successful just because they were successful.

In summary, Latour's actor-network theory seems to provide a great method for discovering how people (and things?) spend their time—that is, how various objects associate in

space and through time. It can also indicate how various objects become introduced to one another. What it does not tell us is *why* these objects associate and that is the paramount question most methodologists are interested in, especially those who want to be normative and figure out how we can do better in empirical investigations.

6.6. Conclusions. The methods Collins and Latour promote have both strengths and weaknesses as discussed in this chapter. The main problem is that neither promotes the ends their use is claimed to serve specifically as either partial or full replacements for epistemology. However, they do promote other important ends. The method offered by Collins, for example, does not isolate epistemological factors out of a case study. It does, however, serve to mitigate biases that occur from post hoc analyses. Latour's method serves to highlight the multitude of tasks required for the conduct of modern scientific investigation, including the many administrative and financial considerations. Both authors' methods also serve as useful methods for investigating and addressing questions about how knowledge and technology are transmitted and changed from one environment to another.

Chapter 7: Taking Evidence Seriously: Minimal Severity, Objectivity and Naturalist Meta-methodologies.

While philosophers and sociologists of science may debate some of the precise qualities that define science, they all agree that research conducted with a predetermined outcome is not science. (MacGarity & Wagner 2008: 11)

We intuitively deny that data $\boldsymbol{x_0}$ *are evidence for* H *if the inferential procedure had very little chance of providing evidence against* H, *even if* H *is false. We call this the weak severity principle. (Mayo & Spanos 2010: 21)*

7.0 Taking evidence seriously: a concept of minimal severity: The basic idea of where objectivity would enter for the error statistician is that in order to take evidence seriously it cannot be that one has already predetermined what inference is going to result from the evidence (see Mayo 1996, 2010). Pretty clearly it would be regarded as unscientific (or bad science or the like) if you said you had 'good evidence' for a hypothesis when, no matter what the data were, you were going to either interpret the data so that it supported the hypothesis or find some other way to support it. The trouble here is that you are going to be able to claim to have evidence in favor of a hypothesis even if that hypothesis is false. In other words, in such an admittedly extreme case, there is no chance of the procedure (or test) detecting a flaw at all. The strategy or method above exemplifies the most extreme case of not taking evidence seriously and in the error statistical approach it would be said that such a method (or test) has no probative power whatsoever; it is the most 'insevere' test one can come up with.

The basic idea of error statistics (ES) then would be to evaluate methodological rules and principles by considering the ways in which they can promote the goal of using evidence seriously. In order to come up with an account that is not overly simple and yet can do this work of evaluating methods, one needs to be much more sophisticated about the multiple steps involved in the collection, modeling, and interpretation of data, using the results to achieve some sort of statistical inference from that data, and subsequently to take that inference and use it to

evaluate substantive theories or reach decisions.[114] In order to do this, Mayo (1996) has proposed an iterative 3-tier hierarchy of models—models of the primary hypothesis, experimental models, and data models—so that at each step of the way, one could evaluate the procedures used in collecting, modeling, and interpreting the data by considering whether each step is promoting this goal of unearthing biases and errors.

The idea motivating this ES approach is really quite simple. It is to take very literally and very seriously that we learn from our mistakes. This view leads us to have a kind of minimal principle for taking evidence seriously, which is that the data can be regarded as providing "genuine evidence" for an inference *only if* the method that it results from would have at least some probability of having uncovered the falsity or flaws in reaching that inference if it is false. This is the principle , we have been calling it Mayo's minimal severity principle, upon which this dissertation has been built upon for assessing the various types of methods and meta-methods on offer. Again, it is can be stated formally thus:

> **Minimal severity principle of (lack of) evidence:** If a method has no chance of finding fault in a hypothesis, (including hypotheses about background assumptions), then the fact that no fault was found is very poor evidence for the hypothesis (or assumption) in question.

Now sometimes the ES account appeals to probability in the formal mode (discussed in chapter 3) but more importantly it emphasizes the use of probability in order to capture more informal, intuitive reasoning (chapter 4).[115] Furthermore, the principle of evidence above, while seemingly very weak, is actually much stronger than mere falsificationist requirements. For while a hypothesis may be logically falsifiable it could very well be that the procedure(s) in use make it virtually impossible for any such falsifying evidence

[114] For more on the ES approach to substantive theories, see Mayo 2010: 28-57.
[115] See Spanos 2009 for how the formal mode can be used to investigate informal, albeit highly sophisticated, inductive reasoning to better understand and assess historical episodes, such as the discovery of Argon, which he analyzes, without falling into anachronistic traps.

to result. The procedure in use may be so biased—perhaps because we refuse to consider anything that goes against our hypothesis, (e.g., assigning leadership roles to males, the pharmaceutical drug being tested is ineffective, or that floor mat entrapment of the accelerator pedal is due to a design flaw and not flawed drivers, etc.)—that it had no chance to produce conflicting data. So, the error statistical idea is to take seriously potential flaws in order to achieve and promote objectivity by trying to avoid them.

7.0.1 Error probabilities. When error statisticians talk about trying to evaluate the probative capacity or severity of a test, they are talking about the ability of the test to uncover errors, and this is the idea of an error probability. The probability refers to the capacity of the test to unearth various errors and mistakes. This is why the approach is called, very generally, the error statistical account. Probability is never assigned as a degree of belief in or a measure of the truth of a hypothesis. Now, in this dissertation, I was primarily interested in seeing how much mileage I could get from this principle in applying it to assess naturalist meta-methodologies commonly found in the philosophy of science and Science and Technology Studies (STS), areas to which it has not been previously applied.

I looked at some well-known naturalist accounts from each of the two main routes into naturalism commonly taken by philosophers of science and researchers in STS. Following in the footsteps of Kuhn, we have the historical approach, which looks to the history of science to provide empirical evidence to warrant philosophers' and STS researchers' meta-claims about science. My interest was solely in meta-claims about scientific methods dealing with evidence and inference. The second route into naturalism we can call the science of science approaches, as here naturalists appeal to the various sciences to underwrite their meta-claims about how science is done or why scientific knowledge is so reliable, objective, etc. Rather than following

either of these approaches to naturalism, I have followed up and extended Mayo's 1996 proposal for a third way into naturalism, which is to take a causal approach to methods for understanding and justifying (as well as rejecting) claims of objectivity and reliability in scientific inference. Why did I follow this route and what does it mean for meta-methodology?

7.1 Uncovering the mechanisms. Meta-methodologists of the historicist schools, whether naturalists like Laudan or apriorists like Worrall, provide meta-methodologies that are themselves highly unreliable. This is because their accounts/meta-methodologies rely on questionable correlations. What is really needed is an account which can uncover the mechanisms, e.g., provide a causal explanation for the success of particular methods in specific contexts for generating data and justifying that that data indeed provides evidence for or against some particular hypothesis of interest. Such an account could check the "tendency to take a historical episode, view it as exemplifying [one's] preferred approach, and then read our intuitive endorsement of that episode as an endorsement of that approach" (Mayo 2010: 80, see also Mayo 1988 for how to apply severe testing principles on this meta-level.) Without a causal explanation, then we really are left with making ad hoc use of constructed principles of inference based on questionable correlations (see chapters 1 and 4; Mayo & Miller).

So the real problem with the way naturalists use history lies in their searching for correlations when what we are really interested in is a causal explanation behind why particular methods enable us to achieve specific aims in specific contexts—in short we need to uncover the mechanisms. By mechanism, I mean those the properties inherent in a method that would make its application help us to achieve a goal or that its violation would cause us not to achieve that goal. *This is an empirical question.* We must look to the methods themselves and their properties in order to explain how their application (violation) allowed us to achieve (fail to achieve)

specific goals (e.g., rule out particular errors, make reliable inferences, etc.) in concrete specified situations.

Before the advent of the ES methodology, the most promising source for such an account was what we can call the "New Experimentalism"[116] movement in philosophy of science (e.g., Hacking, Cartwright, Galison, Franklin, etc.). Members of this loosely defined school focus on local practices at the level of individual experiments to find how "evidence *qua* evidence" achieves epistemic warrant and acceptance. One problem seen in their overall approach, though, is that their focus is often so local that sometimes it *seems as if nothing general can be said about scientific practice.* To date, Mayo's error statistical approach to experimental learning is the only philosophy of science that takes the insights of the new experimentalists and uses them to erect an account of and justify general principles of reasoning for assessing and evaluating the evidential well-foundedness of some theories (i.e., for assessing and evaluating scientific methodologies and experimental inferences).

7.1.1 Assessing Severity. The error statistical account of normative naturalism recommends evaluating methods based on their possessing characteristics for detecting; circumventing or otherwise controlling specific errors (see chapters 3 & 4; Mayo & Miller 2008). These properties are evaluated relative to (1) a specific hypothesis, (2) data set and (3) data generating mechanism used. This localization is captured in the notion of a Severity function abbreviated: SEV(hypothesis *H*, Test T, and data ***x***). That is assessing severity will also be a function of (relative or localized to) a particular inference under test, hypothesis *H*; by a specific data set ***x*** that was generated by a particular test T (Mayo 1996, Mayo & Spanos 2006; Mayo & Cox, 2006; 2010).

[116] *Pace* Ackermann (1989); Mayo (1996)

I have illustrated this approach using several examples; the most extended one revolved around replication as a method both at the level of entire experiments as well as at the level of experimental units. Replication at the latter level is required for performing valid statistical tests (e.g., significance tests), but, I argued, as the same errors pose a potential threat when non-statistical (e.g., graphical) data analysis is substituted in place of these statistical tests then those methods, though not strictly invalid, are still unreliable. Avoidance of error, not validity, is the real issue that needs to be addressed in pseudo-replication debates in ecology. Throughout, the main thrust of this thesis has been my attempt to show that this same strategy can and should be used when evaluating meta-methods, that is, when philosophers and STS researchers employ empirical methods to study scientific methods. This is simply the reflexivity thesis in action.

7.2 Benefits of the ES approach. I have argued that there are several benefits to assessing methods in this manner. Perhaps the greatest is that such a meta-methodology, judging methods based on their intrinsic empirical properties for detecting, avoiding or otherwise controlling errors, does not wind up in either "pernicious relativism" or "vicious circularity" as feared by John Worrall. This is because the severity assessment made in any particular experimental or observational situation provides the benchmark for measuring methods.

Admittedly there may be several methods that can meet that benchmark, and this is all to the good for an error statistician, whose concern is with eliminating error not with promoting one particular method for doing so in all circumstances. Nonetheless, while multiple methods may be sufficient to avoid a particular error, those that are not sufficient will be eliminated. While no one method may be necessary for reliable inference; identifying, eradicating or controlling error that a particular inference is liable to *is* necessary for objectivity. Eliminating or measuring and somehow controlling the error just is the concern for error statisticians in discussing objectivity.

Circularity is avoided because though the overarching principle of reasoning (severity) is quite general, specific implementations are localized as captured in the severity triad discussed in the last section.[117]

Moreover, as the focus is on the reduction of error, not any one method *per se*, this ES account explains why there is such a diversity of methods. Just as there are many ways to get to top of a mountain, there are many ways to detect and control an error. Indeed almost any complicated investigation will require the use of multiple methods to reliably or severely probe the ways it would be a mistake to infer a hypothesis of interest.

7.2.1 Is ES Worrall a priori? But, some may ask, isn't the injunction to "avoid error" an *a priori* imperative for the ES method? That is, does this account constitute a case of Worrall's (1999) minimal *a priorism*? I do not believe this is the case. Why? While the injunction to avoid error can be justified in light of its terms, or as Worrall would put it, is plausible in light of the terms used, this fact of plausibility on its own is meaningless without specific empirical input. The interpretation and application of the SEV principle will always be a local empirical matter as is the determination as to whether it has or has not been satisfied in a specific case. Thus assessing severity, or more informally, an argument from error is always based on and justified by the way the world is, not from the definition of the terms used.

On Worrall's account, *a priori* status is assigned to the justification of actual methods employed to which a case is referred; hence the warrant given to an episode is *a priori* justified. For example, on his account, double blind trials are justified only in reference to being an instantiation of some other, more general *a priori* justified principle, in this case the principle that all plausible alternatives must be ruled out. We should note here that this justification fails the ascertain-ability requirement I gave in chapter 1 (see also Mayo 2010, chapter 1). For how

[117] SEV(hypothesis *H*, Test T, and data **x**).

can we know if all plausible alternatives have even been thought of much less ruled out? This is not a problem on Worrall's account as he sees no need for any empirical warrant to be given to this principle. Its acceptance, according to him, is purely *a priori*. In contrast, on the ES approach, the justification, e.g., that an error was detected, that the method succeeded in ruling out an error—is always an empirical claim. The ruling out of errors, the warrant for claiming success or failure in ruling out a specific error is empirical all the way down—it is always contingent on the way the world is and on the empirical properties that a method possesses (or lacks) for dealing with errors that arise in actual experimental and observational contexts. There is nothing *a priori* in that assessment.

The only *a priori* aspect to the account that I think can be held up to it, is in the definition or understanding of errors as impediments to reliable inferences, to objectivity. But even there it could be the case that we lived in a world where errors did not exist or we could not make them, or even if we did, we could not know this, in which case, the ES approach would be a curious footnote to epistemology but have no real world bite to it. That none of those situations are the case—that is, errors exist, we can and do make them, and we know that we make them or are liable to make them—are once again, all empirical claims. And, unlike Worrall's minimal *a priorism*, it is in reference to these empirical claims that error statistical reasoning is justified. So all in all, I contend ES is a robustly naturalistic approach to evidence and inference—one that depends on and reflects the way the world is and one that would change accordingly i.e., if (and as) the world changes our methods are such that they would detect and indicate such changes. Indeed, our most successful methods have and will continue to evolve to account for such changes in either the natural or social environments in which they are employed (e.g., as seen in the development of new methods to deal with placebo effects, etc.).

As I discussed in chapter 2, Worrall also claims that one cannot be a full-fledged naturalist without embracing either a vicious circularity or a pernicious relativism. This is because in choosing which scientific episodes to study, or which sciences are relevant for epistemology, one must already presuppose a concept of "good" scientific practice or scientist. Minimally one must at least possess a concept of what makes a practice "scientific." He argues that if naturalists made their assumptions explicit, they would see their accounts rest on an *a priori* foundation/rule for determining good science or else they would detect the circularity or relativism inherent in their views.

While both these challenges pose tremendous problems for naturalist meta-methodologies that they must address, nonetheless in this dissertation, I have taken a different approach to evaluating meta-methodologies. I have tried to see if the meta-methods being proposed have the properties necessary or sufficient to the task(s) set them by their authors, especially to uncover any major errors, which as Worrall pointed out are often rooted in their starting assumptions. In short, I've attempted to see if those methods have the empirical properties to produce data that could provide a severe or reliable test of the claims made using them (see chapters 2, 5, and 6).

But can error statistics give naturalists what we need in STS? That is, will following its prescriptions help underwrite the reliability of our empirical claims about science? This is an empirical question and can only be determined by looking at the properties of the methods available and how they function in specific contexts. Only then, on a case by case basis, can such judgments be made, and that is a thoroughly empirical assessment. Most of the naturalist methods I looked at failed to have all the properties claimed for them, properties that would detect or control errors in the specific theses they wished to sustain. However, my analysis, did

not simply dismiss such methods, but showed how several of them could be modified, strengthened or else applied to other types of meta-claims and thus achieve at least some, if not all, of their goals. For example, in chapter 5, I suggested that Longino could incorporate severity into her method of multiple perspectives to try to delimit in advance which points of view could be relevant in an investigation, which has been an outstanding problem for her.

7.3 Locating Objectivity & Progress. In chapter 1, I pointed out that one way to parse the various naturalist attempts was based on where each looked to secure the objectivity (or stability) of scientific inferences, that is, what components involved in experimental or observational processes were credited with turning 'raw' data into 'evidence' for inferences on their accounts. Naturalists who look to the biological and human sciences can often be seen as locating the source of objectivity in the individual human being, for example Giere; while others (e.g., Longino, Collins, Latour) look to social institutions and practices as the principle source of objectivity. In the later cases, objectivity may be embedded either in social interactions of one or another type (critical discussions) or in the use of a scientific method, (e.g., sociological attempts at controlled experiment, various anthropological research tools.) In these last two, consensus in the one case determines objectivity while in the other case stability is built between outsiders (observer participants) and insiders (the various tribes of scientists), at least in the approaches under scrutiny here (chapters 5 & 6).

However, unless we allow them to simply define the problem of objectivity away or otherwise sweep it under the carpet, these approaches—that rely on the correctness of the human, biological or social sciences they are drawing upon—need to justify that their methods have the ability to detect and control errors in producing data that is in choosing which science, scientific theory or case studies to look to—and thus provide evidence for the specific claims they are

drawing. This requirement above comes because, for error statisticians, objectivity is secured only to the extent that severity is met in any particular case whether at the methodological or meta-methodological level. I have argued in chapters 2, 5 and 6 that these approaches have significant problems in doing this, some of which I will review below.

7.3.1 Philosophical Naturalism Under philosophical naturalism I considered Ron Giere. He is pretty content to say that a true naturalism is more descriptive than normative, which is not an uncommon view. The trouble there is that Giere will appeal to what scientists are doing while completely omitting the importance of being able to critically evaluate their doings. For him and others like him, such evaluations at most come from other scientists and perhaps common sense but not philosophy *per se*. So what is the point then of us as philosophers if our job is simply to do a kind of journalistic recording of what various scientists say and do?

Giere does make some appeals to cognitive science as a way to assure ourselves that we are not getting it too wrong. But these very naturalistic appeals to cognitive science, psychology, and the like already come with the built-in assumption that the methods and theories on which these cognitive science results are based are themselves unproblematic. Nothing could be further from the truth. And as philosophers it really behooves us to be the ones critically evaluating these tools, if we are to use them, however indirectly, especially as these are some of the most interesting, questionable and problematic sciences out there. We should not be coming to do our philosophical work building upon their tools, particularly given just how shaky these tools are. Much of the current methodological literature in these fields is debating the (mis)use of statistics, disparagingly referred to as "voodoo statistics". Some of their other apparatus (e.g., fMRI) has actually produced evidence that fish think, though, in this case, the cognizing fish was not a red herring but literally a dead salmon (Madrigal 2009). (I thank Emrah Aktunc for these

examples.) It is utterly crucial that we have the tools to critically evaluate these results and methods.

The tools referred to above because of the uncertainties and limited data, generally involve statistical methods. The statistical methods on which they rely have themselves been open to a variety of philosophical and foundational criticisms. These include the more familiar formal ones such as between objective frequentist accounts and Bayesian accounts (see chapter 3). But further, the whole idea of connecting the results of statistical inferences to substantive notions, unless the philosopher of science can critically evaluate the methods on which these cognitive and similar naturalist accounts are built, is really just going to beg the whole question and in effect just take us back to a kind of journalistic reporting on what scientists are doing. This is why I spent some time articulating as clearly as I can how the ES account has been able to deal with these methods (chapter 3) and I considered specially some criticisms that have arisen in areas with which I am most familiar such as with debates about replication (Chapter 6) especially as they arise in Before-After-Control-Intervention (BACI) experiments in ecology (Chapter 4; see also Miller and Frost).

The beauty of appealing to these statistical methods is that it gives us the kind of iconic canonical examples showing how error probabilities can be obtained and how one can critically evaluate the assumptions of statistical models and tests, in order once again to evaluate the extent to which they can prevent, or get in the way of, arriving at a test with good type 1 and type 2 errors. These are the more formal counterparts of the kinds of errors that we consider informally not only in science but in everyday life. So it is extremely important to understand them both to raise the level of discussion that is so often based on the results of those methods, (e.g., from cognitive science and so on), and to begin to allow us to critically evaluate them. That is what a

naturalistic philosophy of science should be. If we are going to take seriously the naturalistic mantra that we philosophers of science should use the methods of science to do philosophy, then what we ought to do is consider the methods scientists actually use in order to avoid biases, obtain objectivity, and critically evaluate another scientist's tools—not require that everyone already agrees on methods, for we would never make progress that way.

Now, it is true that in debates it may very well be that scientists disagree as to what are the best methods to use, for example to ascertain when a replication has been done (Collins 1994; 2004), or to determine the risks associated with hormone replacement therapy (HRT) (see NIH WHI study) and so on. But because they have recourse to this overarching principle of severity, scientists can critically evaluate the evidence for various claims about the above questions. One of the great things that the more formal levels of critique does for us, is provide well understood and worked out exemplars of when and where various procedures, such as hunting for the data, violating replication, etc., prevent the data being taken seriously as evidence and where it does not. One wants as much as possible to have recourse to these canonical tools—where the assumptions, methods and errors are well understood—as they can also provide a kind of exemplar for more informal cases.

Yet, another naturalistic philosopher to appeal to cognitive science, psychology, and notably decision theory is Philip Kitcher. What I found interesting in his work was the way in which we are going to get the constraints needed for objectivity. This is done by way of a concept he has that we will demand explanatory unification across various phenomena. Kitcher uses explanatory unification in two ways—one as an exploratory device for coming up with and articulating theories, the other as a warrant for objectivity. I dealt only with the latter use.

In chapter 2, I showed that where his account correlates with error statistical criteria, that is, where it actually has the ability to probe for errors in regards to a specific phenomenon or inference, then it can be empirically justified and we can provide a rationale for how pursuing explanatory unification promotes objectivity. Not any unification will do. Similar to Aristotle's use of cross referencing the senses to get ensure reliable inference (e.g., the sense of touch corrects the visual impression that a stick in the water is straight, not crooked), so, too, explanatory unifications where phenomena in one experiment work to cross check the assumptions or possible errors in another experiment are the types of unifications that work for objectivity. They work because in this setting, the diverse phenomena act as error probes (function as arguments from error) of different parts or potential errors in an explanation. In short, unifications that work as error probes are the ones that act as objective constraints on explanations—but it is the probative nature, the cross checking of errors that is doing the work for objectivity, not the unification *per se*.

The existing literature fails to address the weaknesses of these naturalists' accounts on their own terms much less to strengthen or replace them with more satisfactory ones. My goal has been dual—not just to show how they fail but to try to set the stage for new avenues for satisfactory approaches, which can incorporate the good insights that are within them.

7.3.2 Social epistemologies Next, I looked at accounts that come broadly speaking under social epistemology. In chapter 5, I looked at Helen Longino as one who sees herself as providing us with an account of objectivity. The trouble is that her definition of objectivity fails to identify requirements to achieve taking the data seriously. Insofar as one accepts even that most minimal requirement for what it would mean to have objective evidence for something, it will not do. What is her account? In a nutshell, it will be based on bringing multiple perspectives

(points of view)—based on describing certain features of the audience, which can range from their scientific background/specialties (e.g., experimental versus field ecologists) to considerations of their gender or ethnicity—to bear on a case. Although there will be many certain special cases where that will correlate with being able to bring to bear a perspective that might indeed be more likely to uncover biases, what is really doing the work is the ability to uncover those biases—so not any perspective will do, nor will any one scientists even with the "desired" features necessarily be able to uncover bias. In many cases, the bias will be uncovered by other means.

So her method of multiple perspectives (mmp) as a universal criterion is neither necessary nor sufficient. What is necessary and sufficient, and which her method may, in special circumstances, be able to facilitate, I argued was the ability to promote either the critical evaluation of lack of severity or achieve a more severe, more probative, less biased result. What is important to understand is that without an account of severity, we have no principled way of choosing potential perspectives to include and implementing her method will lead to cacophony as she herself fears it could. The question an error statistician wants answered when it is suggested that a variety of diverse perspectives be brought to bear on a case, is how can each perspectives work in this case to successfully probe for errors? The answer would vary depending on the hypothesis under scrutiny, the methods used to generate data as well as the data itself. This is what I think Longino also really desires and so I suggested in Chapter 5 that by seeing her method as one of many possible strategies for implementing severity (in special cases), it would succeed. But even in those special cases where it happens that the bias of the individuals may indeed be inclined to prevent them from seeing things objectively—even there we need to say more about what these pluralistic perspectives are actually supposed to do in

order to discern errors and biases reliably. Without severity as a guide, then it is too easy in every one of her cases, if we apply her method solo, that we can easily come up with perspectives that would fail utterly to discern errors and that increase rather than decrease objective assessments of evidence and inference. This is similar to the case I argued for Kitcher's meta-methodology. So that was my overarching goal and I illuminated what is going on by considering the presuppositions of these accounts that lead them into the problems they end up with and suggested way to fix or strengthen them using severity criteria.

A similar presupposition comes up when we consider Harry Collins, a sociologist, who also seems to grow out of the sort of Kuhnian framework, which posits that a scientist works and tests always from within a paradigm. And according to this view that paradigm gives the scientist the methodological tools, which he cannot question and had to use, while scientists in another paradigm have their tools, and so it is no wonder circularity results when they go to test the paradigms theories. This is the very type of approach that Mayo's entire philosophy of ES grew out of by opposing that and showing explicitly by means of forward looking methods how we can debate rules even when there is complete disagreement. There is one thing that the scientists do have in common which really does explain why they can make progress even if they do not settle a controversy, and that is this overarching meta-level principle about the minimal requirement for what counts as taking evidence seriously. Although scientists may take those principles for granted, our job, as philosophers, is to make them explicit. This same principle should apply to our own accounts if they are to be based on empirical evidence. The account that I am situating these other approaches in, I believe, gives us a handle on these problems and further can be used to improve current naturalistic accounts like those of Giere, Kitcher, Longino, Collins and Latour.

Some of these accounts, such as that of Collin's, first need to resolve violations of the underlying assumptions of their methods before they should claim they have evidence for their claims (i.e., the assumption that everyone is rational in a controversy does not suffice to hold off or isolate scientific considerations from influencing learning and the closure of controversy). Still others, such as Longino's account, could actually end up promoting methods that have worse error properties than those they are to replace in some cases. But what allows for progress, I have tried to both argue and show, is that it is possible to assess these methods on their own grounds—the problems they've set out to solve, their assumptions, and the claims made for them—and their shortcomings are based on their properties or lack thereof in the face of specific errors. [118] By pointing out these errors, we now have grounds for improving them or modifying their application rather than simply dismissing them out of hand.

7.4 Conclusion. I have tried to show and argue that ES provides for a general and systematic naturalistic way to scrutinize both object level (e.g., scientific) methods and meta-methods (meta-science). This is because the empirical focus for ES is on the properties methods have for detecting, eradicating or otherwise controlling specific errors in making specific inferences based on particular data sets. It will be understanding the strengths and weaknesses of methods for detecting errors, etc., that will allow for inter-subjective assessments of methods based on their properties rather than the continuous infighting that has been the case in the comparative approaches (e.g., as seen in Kitcher's critique of Collins; the exchanges between Collins & Franklin; and Longino and Kitcher). This is because in our critique of methods, we

[118] Their extensive use of case studies to illustrate and support their methodological claims often disguise the weaknesses in their own methods, for the case studies they have chosen often embody better, more reliable methods than the one they have identified (or claimed to have identified) based on them. However, a case study does not count in one's favor if it has not been correctly analyzed though the presence of the case study, which is often very well written up, is important for it provides empirical data for various meta-methodologists to re-analyze and discuss, so that we can become clear on our own errors and interpretations.

will assess methods based on their properties for achieving their goals while avoiding specific errors, not on how they compare with other methods. Thus, rather than telling Collins to leave epistemology and scientific methodology alone and go and be a good Mertonian, as Kitcher suggested several years ago, my critique of Collins' assumptions was based on whether they were met and hence whether or not his studies achieved the isolation (or holding steady) of "scientific factors" as he claimed. All the criticisms I offered in this dissertation were "self-contained" in this way, in that the errors I pointed out resided in the properties of the methods being proffered, or in the assumptions being relied upon for producing data to assess (i.e., to be evidence for) their meta-level claims about scientific methodology. I did not rely on and in fact feel that I have shown the way to break out of what Longino has called the *a priori* rational/social dichotomy that has characterized much of the meta-methodological disputes during the "science wars" of the last two decades.

Bibliography

Achinstein, P. (1998). Why Philosophical Theories of Evidence Are (and Ought to Be) Ignored by Scientists. *PSA 98 Part II: Symposia Papers:* S180-S192.

Achinstein, P (2010). Mills' Sins in Mayo, D. & A. Spanos (eds.). *Error and Inference: Recent Exchanges on the Philosophy of Science, Inductive-Statistical Inference, and Reliable Evidence*. Cambridge: Cambridge University Press: 170-88.

Ackermann, R. (1989). The new experimentalism. *British Journal for the Philosophy of Science* **40**: 185-90.

Actor network theory home page: http://carbon.cudenver.edu/~mryder/itc/ant_dff.html

Aktunc, E. (unpublished manuscript). 'Voodoo Correlations', Salmon thoughts, and the promised science of fMRI.

Backhouse, R. E. (1994). *New Directions in Economic Methodology*. New York, London: Routledge.

Bauer, H. (1994). *Scientific Literacy and The Myth of the Scientific Method*. University of Illinois Press.

Berger, J. Could Fisher, Jeffreys and Neyman Have Agreed on Testing? *Statistical Science*, **18**(1): 1-12

Berger, J. (2006). The case for objective Bayesian analysis. *Bayesian Analysis* **1**(3): 385-402.

Bloor, D. (1991). *Knowledge and Social Imagery* 2nd Edition; Chicago: University of Chicago Press (1st Edition (1976): London & Boston: Routledge & K. Paul).

Box, G. E. P. (1976). Science and Statistics. *J. Am. Stat. Assn* **71**: 791-799.

Box, G. E. P., and G. Tiao. 1975. Intervention Analysis with Applications to Economic and Environmental Problems. *J. Am. Stat. Assn* **70**: 70-79.

Brezonik, P. L., J. G. Eaton, T. M. Frost, P. J. Garrison, T. K. Kratz, C. E. Mach, J. H. McCormick, J. A. Perry, W. A. Rose, C. J. Sampson, B. C. L. Shelley, W. A. Swenson, and K. E. Webster. (1993). Experimental Acidification of Little Rock Lake, Wisconsin: Chemical and Biological Changes over the pH Range 6.1 to 4.7. *Can. J. Fish. Aquat. Sci.* **50**: 1101-1121.

Callebut, W. (1993). *Taking the naturalistic turn, or, How real philosophy of science is done.* Chicago: University of Chicago Press.

Campbell, D. T., and J. Stanley. (1981). *Experimental and Quasi-experimental Designs for Research*. Boston: Houghton Mifflin.

Carpenter, S. R. (1999). Microcosm Experiments Have Limited Relevance for Community and Ecosystem Ecology: Reply. *Ecology* **80**: 1085-1088.

Carpenter, S. R., S. W. Chisholm, C. J. Krebs, D. W. Schindler, and R. F. Wright. (1995). Ecosystem Experiments. *Science* **269**: 324-327.

Carpenter, S. R., J. J. Cole, T. E. Essington, J. R. Hodgson, J. N. Houser, J. F. Kitchell, and M. L. Pace. (1998). Evaluating Alternative Explanations in Ecosystem Experiments. *Ecosystems* **1**: 335-344.

Carpenter, S. R., T. M. Frost, D. Heisey, and T. K. Kratz. (1989). Randomized Intervention Analysis and the Interpretation of Whole-Ecosystem Experiments. *Ecology* **70**: 1142-1152.

Carpenter, S. R., and D. W. Schindler. (1998). Workshop on Ecosystem Manipulation. *Ecosystems* **1**: 321-322.

Cartwright, N. (1983). *How the Laws of Physics Lie* Oxford: Oxford University Press.

Cartwright, N. (1999). *The Dappled World: A Study of the Boundaries of Science*, Cambridge: Cambridge University Press.
Chalmers, A. F. (1999). *What Is This Thing Called Science?* 3rd ed., University of Queensland Press, Australia.
Chalmers, A. F. (2010). Can scientific Theories be warranted? in D. Mayo & A. Spanos (eds). *Error and Inference: Recent Exchanges on the Philosophy of Science, Inductive-Statistical Inference, and Reliable Evidence*. Cambridge: Cambridge University Press: 58-72.
Chang, H. (2004). *Inventing Temperature*, Oxford and New York: Oxford University Press.
Churchland, P. (2001). Eliminative materialism in *Voices of Wisdom: A Multicultural Philosophy Reader* 4th Ed. (Ed) Gary Kessler, Belmont, California: Wadsworth: 467-475.
Collins, H. (1985). *Changing Order: Replication and Induction in Scientific Practice.* London: Sage.
Collins, H. (1994). A Strong Confirmation of the Experimenters' Regress, *Studies in the History and Philosophy of Science* **25**(3): 493-503.
Collins, H. (2004). *Gravity's Shadow*. Chicago: University of Chicago Press.
Collins, H. (2007.) Webpage entrance for LIGO and other papers: http://www.cf.ac.uk/socsi/contactsandpeople/harrycollins/grav-wave-1.html
Collingwood, R. G. (1956). *The Idea of History* New York: Oxford University Press.
Cunningham, A. and P. Williams (1993). De-Centering the 'big picture': The Origins of Modern Science and the Modern Origins of Science. *British Journal for the History of Science* **26**: 407-432.
Dawid, A. P. (2000). Causal Inference without Counterfactuals. *J. Am. Stat. Assn* **95**: 407-424.
Dennis, B. (1996). Discussion: Should Ecologists Become Bayesians? *Ecol. Appl.* **6**:1095-1103.
Dennis, B. (2004). Statistics and the Scientific Method in Ecology. Chapter 11 in Taper, M. L., and S. R. Lele, eds., *The Nature of Scientific Evidence: Statistical, Philosophical, and Empirical Considerations.* Chicago: University of Chicago Press.
Doppelt, G. (1990). The Naturalist Conception of Methodological Standards in Science: A Critique. *Philosophy of Science* **57**(1):1-19.
Drenner, R. W., and A. Mazumder. (1999). Microcosm Experiments Have Limited Relevance for Community and Ecosystem Ecology: Comment. *Ecology* **80**: 1081-1085.
Edwards, D. (1998). Issues and Themes for Natural Resources Trend and Change Detection. *Ecol. Appl.* **8** :323-325.
Efron, B. (1979). Bootstrap Methods: Another Look at the Jackknife. *Ann. Stat*. **7**: 1-26.
Efron, B., and G. Gong. (1983). A Leisurely Look at the Bootstrap, the Jackknife, and Cross-Validation. *Am. Stat.* **37**: 36-48.
Elton, G. (1967). T*he Practice of History* London: Fontana Press.
Ericson, L. (2005) Longino's Community of Science and the Power Structures of Community, http://www.ericsonhome.net/loyd/pdfs/community%20of%20science.pdf
Faust, D. and P. Meehl (2002). Using MetaScientific Studies to Clarify or Resolve Questions in the Philosophy and History of Science. *Philosophy of Science* **69**(3) Supplement: Part II Symposia Papers.
Fisher, R. A. (1947). *The Design of Experiments.* 4th ed. London: Oliver and Boyd.
Fisher, R. A. (1971). *The Design of Experiments.* 9th ed. New York: Hafner.
Forster, M (1988) Unification, Explanation, and the Composition of Causes in Newtonian Mechanics, *Studies in History and Philosophy of Science***19**: 55-101.

Forster, M (2004) Chapter **3**: Simplicity and Unification in Model Selection (Last updated March 6, 2004) in *Occam's Razor and the Relational Nature of Evidence*, on line Manuscript: http://philosophy.wisc.edu/forster/.

Franklin, A. (1986). *The Neglect of Experiment*. Cambridge: Cambridge University Press.

Franklin, A. (1994). How to Avoid the Experimenters' Regress. *Studies in the History and Philosophy of Science* **25**(3): 463-491.

Freedman, D., R. Piscani, R. Purves, and A. Adhikari. (1991). *Statistics.* 2nd ed. New York: Norton.

Frost, T. M., D. L. DeAngelis, S. M. Bartell, D. J. Hall, and S. H. Hurlbert. (1988). Scale in the Design and Interpretation of Aquatic Community Research. In Carpenter, S. R., ed., *Complex Interactions in Lake Communities.* New York: Springer.

Frost, T. M., S. R. Carpenter, A. R. Ives, and T. K. Kratz. (1995). Species Compensation and Complementarity in Ecosystem Function. In Jones, C. G., and J. H. Lawton, eds., *Linking Species and Ecosystems.* New York: Chapman and Hall.

Frost, T. M., P. K. Montz, and T. K. Kratz. (1998). Zooplankton Community Responses during Recovery from Acidification: Limited Persistence by Acid-Favored Species in Little Rock Lake, Wisconsin. *Restor. Ecol*. **6**:336-342.

Frost, T. M., P. K. Montz, T. K. Kratz, T. Badillo, P. L. Brezonik, M. J. Gonzalez, R. G. Rada, C. J. Watras, K. E. Webster, J. G. Wiener, C. E. Williamson, and D. P. Morris. (1999). Multiple Stresses from a Single Agent: Diverse Responses to the Experimental Acidification of Little Rock Lake, Wisconsin. *Limnol. Oceanogr.* **44**:784-794.

Galison, P. (1987). *How Experiments End*. Chicago: University of Chicago Press.

Giere, R. (1999). *Science without Laws.* Chicago: University of Chicago Press.

Giere, R. (2006a). *Scientific Perspectivism*. Chicago: University of Chicago Press.

Giere, R.:(2006b). Perspectival Pluralism. *Minnesota Studies in Philosophy of Science* **19**: 26-41.

Giere, R. (2001). Critical Hypothetical Evolutionary Naturalism. In *Selection Theory and Social Construction: The Evolutionary Naturalistic Epistemology of Donald T. Campbell*, ed. Cecilia Heyes and David L. Hull, SUNY Press: 53-70.

Giere, R. (2003). The Role of Computation in Scientific Cognition. *Journal of Experimental & Theoretical Artificial Intelligence*, **15**: 195-202.

Giere, R (2007) Interview on naturalism: http://www.youtube.com/watch?v=eyWkGl6v5Sc

Gilinsky, N. L. (1991). Bootstrapping and the Fossil Record. In Gilinsky, N. L., and P. W. Signor, eds., *Analytical Paleobiology*. Knoxville, TN: Paleontological Society.

Glymour, C. (1992). Invasion of the Mind Snatchers in R. Giere (ed) *Cognitive Models of Science*. Minneapolis: University of Minnesota Press: 465-475.

Glymour, C. (2010). Explanation and Truth in D. Mayo & A. Spanos (eds). *Error and Inference: Recent Exchanges on the Philosophy of Science, Inductive-Statistical Inference, and Reliable Evidence*. Cambridge: Cambridge University Press: 331-350.

Gonzalez, M. J., and T. M. Frost. (1994). Comparisons of Laboratory Bioassays and a Whole-Lake Experiment: Rotifer Responses to Experimental Acidification. *Ecol. Appl*. **4**: 69-80.

Gooding D., Pinch, T., and Schaffer, S. (1989). Preface & Introduction in *The Uses of Experiment: Studies in the Natural Sciences* D. Gooding, T. Pinch and S. Schagger (eds). Cambridge: Cambridge University Press: xiii-30.

Gottelli, J. J. and G. R. Graves. (1996). Null Models in Ecology. Washington: Smithsonian Institution.

Greenland, S., J. M. Robins, and J. Pearl. (1999). Confounding and Collapsibility in Causal Inference. *Stat. Sci*. **14:** 29-46.
Grene, M. (1985). Perception, Interpretation and the Science, in *Evolution at a Crossroads*, (eds.) David Depew and Bruce Weber Cambridge MA: MIT press: 1-20.
Guala, F. (2008) *The Methodology of Experimental Economics*, Cambridge: Cambridge University Press.
Hacking I. (1983). *Representing and Intervening: Introductory Topics in the Philosophy of Science*. Cambridge: Cambridge University Press.
Hans the Horse: http://www.bbc.co.uk/dna/h2g2/A2390104
Harding, S. (1991). *Whose Science? Whose Knowledge?: Thinking from Women's Lives*. Cornell University Press.
Hurlbert, S. H. (1984). Pseudoreplication and the Design of Ecological Field Experiments. *Ecol. Monographs* **54**:187-211.
Huston, M. A. (1999). Microcosm Experiments Have Limited Relevance for Community and Ecosystem Ecology: Synthesis and Comments. *Ecology* **80**:1088-1089.
Jenkins, K. (1991). *Rethinking History*, New York, London: Routlege.
Kellert, S., H. Longino, and C.K. Waters (eds) (2006). *Scientific Pluralism* Minneapolis MN: University of Minnesota Press.
Kitcher P. (1992). The Naturalists Return. *The Philosophical Review* **101**(1): 53-114.
Kitcher P. (1993.) *The Advancement of Science: Science without Legend, Objectivity without Illusions*. Oxford: Oxford University Press.
Kitcher P. (2000). Reviving the Sociology of Knowledge *Philosophy of Science* **57**: 44–59
Kitcher P. (2001). *Science, Truth, and Democracy*. Oxford; New York: Oxford University Press.
Kitcher P. (2002). The Third Way: Reflections on Helen Longino's The Fate of Knowledge *Philosophy of Science* **69**(4): 549-559.
Kosso, P. (1989). Science and Objectivity. *Journal of Philosophy* **86** (5):245-257.
Kuhn, T. (1962). *The Structure of Scientific Revolutions*, Chicago: University of Chicago Press.
Kuhn, T. (1970). Logic of Discovery or Psychology of Reaseach? in I. Lakatos & A. Musgrave (eds). *Criticism and the Growth of Knowledge*, Cambridge: Cambridge University Press: 1-24.
Kuhn, T. (1977) Chapter 13: Objectivity, Value Judgment, and Theory Choice in *The Essential Tension*. Chicago: University of Chicago Press: 330-339.
Lakatos, I., & A. Musgrave (eds). (1970). *Criticism and the Growth of Knowledge*, Cambridge: Cambridge University Press.
Latour, B. (1987). *Science in Action: How to Follow Scientists and Engineers through Society* Boston: Harvard University Press.
Latour B. & S. Woolgar (1979). *Laboratory Life: The Social Construction of Scientific Facts* Beverly Hills: Sage Publications.
Laudan, L. (1977). *Progress and Its Problems: Towards a Theory of Scientific Growth*, Berkeley and Los Angeles: University of California Press.
Laudan, L. (1990). Normative Naturalism *Philosophy of Science* **57** (1):44-59.
Leplin, J. (1990). Renormalizing Epistemology, *Philosophy of Science* **57** (1): 20-33.
Longino, H (1990). *Science as Social Knowledge: Values and Objectivity in Scientific Inquiry.* Princeton NJ: Princeton University Press.
Longino, H. (2002a). *The Fate of Knowledge* Princeton NJ: Princeton University Press.

Longino, H. (2002b). Science and the Common Good: Thoughts on Philip Kitcher's Science, Truth and Democracy. *Philosophy of Science* **69**(4): 560-568

Longino, H. (2002c). Reply to Kitcher. *Philosophy of Science* **69**(4): 573-577.

Longino, H. (2006). Theoretical Pluralism and the Scientific Study of Behavior in *Scientific Pluralism*, Kellert, S., H. Longino, and C.K. Waters (eds) pp.102-131.

MacGaritiy, T and W. W. Wagner (20008) *Bending Science: How Special Interests Corrupt Public Health Reseach*, Cambridge MA: Harvard University Press.

Madrigal, A. (2009) Scanning Dead Salmon in fMRI Machine Highlights Risk of Red Herrings, *Wired Science* September 18, 2009. http://www.wired.com/wiredscience/2009/09/fmrisalmon/

Mayo, D. (1983) An Objective Theory of Statistical Testing *Synthese* **57**: 297-340.

Mayo, D. (1988) Brownian Motion and the Appraisal of Theories in A. Donovan, L. Laudan, and R. Laudan (eds.), *Scrutinizing* Science, Kluwer, Dordrecht: 219-243. (reprinted by Johns Hopkins University Press, 1992.)

Mayo , D. (1991) Novel evidence and severe tests. *Philosophy of Science* **58**: 523–552.

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

Mayo, D. (2003a). Severe Testing as a Guide for Inductive Learning, in H. Kyburg (ed.), *Probability Is the Very Guide in Life*. Chicago: Open Court. 89-117.

Mayo, D. (2003b). Could Fisher, Jeffreys and Neyman Have *Agreed*? Commentary on J. Berger's Fisher Address". *Statistical Science* **18**: 19-24.

Mayo, D. G. (2004a). An Error-Statistical Philosophy of Evidence. Chapter 4 in Taper, M. L., and S. R. Lele, eds., *The Nature of Scientific Evidence: Statistical, Philosophical, and Empirical Considerations.* Chicago: University of Chicago Press.

Mayo, D. G. (2004b.) Rejoinder. Chapter 4.3 in Taper, M. L., and S. R. Lele, eds., *The Nature of Scientific Evidence: Statistical, Philosophical, and Empirical Considerations.* Chicago: University of Chicago Press.

Mayo, D. (2005a). Peircean Induction and the Error-Correcting Thesis, in R. Mayorga (guest ed.), *Peirce-spectives on Metaphysics and the Sciences* Transactions of the Charles S. Peirce Society. 299-319.

Mayo, D. (2005b). Evidence as Passing Severe Tests: Highly Probed vs. Highly Proved in P. Achinstein (ed.), *Scientific Evidence*, Johns Hopkins University Press, Baltimore: 95-127.

Mayo, D. (2005c) Philosophy of Statistics, in S. Sarkar and J. Pfeifer (eds.) *Philosophy of Science: An Encyclopedia*, Routledge, London: 802-815.

Mayo, D. (2006). Critical Rationalism and Its Failure to Withstand Critical Scrutiny, in C. Cheyne and J. Worrall (eds.) *Rationality and Reality: Conversations with Alan Musgrave*, Kluwer series Studies in the History and Philosophy of Science, Springer, The Netherlands: 63-99.

Mayo, D. (2010). Severe Testing, Error Statistics, and the Growth of Theoretical Knowledge in Mayo, D. & A. Spanos (eds) *Error and Inference: Recent Exchanges on the Philosophy of Science, Inductive-Statistical Inference, and Reliable Evidence*. Cambridge: Cambridge University Press.

Mayo, D. & D. Cox (2006). Frequentist Statistics as a Theory of Inductive Inference, *Optimality: The Second Erich L. Lehmann Symposium*, (ed. J. Rojo), Lecture Notes-Monograph Series, Institute of Mathematical Statistics (IMS), **49**: 77-97.

Mayo, D. & D. Cox (2010) Frequentist Statistics as a Theory of Inductive Inference in *Evidence and Inference* in Mayo, D. & A. Spanos (eds) *Error and Inference: Recent Exchanges on the Philosophy of Science, Inductive-Statistical Inference, and Reliable Evidence*. Cambridge: Cambridge University Press.

Mayo D. & J. Miller (2008). The Error Statistical Philosopher as Normative Naturalist. *Synthese* **163**(3): 305-314.

Mayo, D. & A. Spanos (2004). Methodology in Practice: Statistical Misspecification Testing *Philosophy of Science* (Symposia), **71**: 1007-1025.

Mayo, D. & A. Spanos (2006) Severe Testing as a Basic Concept in a Neyman-Pearson Philosophy of Induction, *British Journal of Philosophy of Science*, **57**(2): 323-357

Mayo D. & A. Spanos (2007). Philosophical Scrutiny of Evidence of Risks: From Bioethics to Bioevidence, *Philosophy of Science*, 73(**5**): 803–816, 2006.

Mayo D. & A. Spanos (2008). Risks to Health and Risks to Science: The Need for a Responsible "Bioevidential" Scrutiny, *Biological Effects of Low Level Exposures, Newsletter* **14** (3), 18-22.

Mayo, D. & A. Spanos (eds) (2010) *Error and Inference: Recent Exchanges on the Philosophy of Science, Inductive-Statistical Inference, and Reliable Evidence*. Cambridge: Cambridge University Press.

Mayo, D & A. Spanos (forthcoming) "Error Statistics' in Prasanta S. Bandyopadhyay and Malcolm Forster (eds.) the *Handbook of Philosophy of Science* (7). Elsevier: Amsterdam, The Netherlands.

McGarity, T. & W. Wagner (2008). *Bending Science: How Special Interests Corrupt Public Health Research* Cambridge: Harvard University Press.

MeGill, A., (ed.) (1994). *Rethinking Objectivity*. Durham & Longs: Duke University Press.

Michaels, David (2008). *Doubt is Their Produce: How Industry's Assault on Science Threatens Your Health*. Oxford: Oxford University Press.

Miller, J. (1997). *Enlightenment Error and Experiment: Henry Cavendish's Electrical Researches*. Master's thesis, Virginia Tech.

Miller, J. (2004). Rejoinder in Taper, M. L., and S. R. Lele, eds., *The Nature of Scientific Evidence: Statistical, Philosophical, and Empirical Considerations*. Chicago: University of Chicago Press: 258-274.

Miller, J. and Frost, T. (2004). Whole-Ecosystem Experiments: Replication and Arguing from Error in Taper, M. L., and S. R. Lele, eds., *The Nature of Scientific Evidence: Statistical, Philosophical, and Empirical Considerations*. Chicago: University of Chicago Press 221-248.

Musgrave, A. (1974). Logical versus Historical Theories of Confirmation, *The British Journal for the Philosophy of Science* **25**(1): 1- 23.

National Institutes of Health (NIH). Women's Health Initiative (WHI) study. http://www.nhlbi.nih.gov/whi/

Neel, J. (2002) The Marketing of Menopause: Historically, Hormone Therapy Heavy on Promotion, Light on Science at http://www.npr.org/news/specials/hrt/.

Novick, P. (1988). *That noble dream: the "objectivity question" and the American historical profession* Cambridge: Cambridge University Press.

O'Keefe and Nadel (1978). *The Hippocampus as a Cognitive Map*. Oxford University Press.

Parker, W. (2008) Computer simulation through an error-statistical lens, *Synthese* (2008) **163**:371–384.

Pearl, J. (2000). Comment on Dawid. *J. Am. Stat. Assn* **95**:428-431.

Pickering, A. (1995). *The Mangle of Practice: Time, Agency & Science* Chicago: University of Chicago Press.
Popper, K. (1962). *Conjectures and Refutations: The Growth of Scientific Knowledge*, New York: Basic Books.
Popper, K. (1970). Normal Science and Its Dangers in Lakatos, I., & A. Musgrave (eds). *Criticism and the Growth of Knowledge*, Cambridge: Cambridge University Press: 51-58.
Popper, K. (1994). *The Myth of the Framework: In Defence of Science and Rationality* (edited by N.A. Notturno). London: Routledge.
Quine, W.V.O. (1951). Two Dogmas of Empiricism Originally published in *The Philosophical Review* **60** (1951): 20-43. Reprinted in W.V.O. Quine, *From a Logical Point of View* (Harvard University Press, 1953; second, revised, edition 1961).
Quine, W.V.O. & J.S. Ullian (1978). *The Web of Belief*. New York: Random House.
Rasmussen, P. W., D. M. Heisey, E. B. Nordheim, and T. M. Frost. (1993). Time-Series Intervention Analysis: Unreplicated Large-Scale Experiments. In Scheiner, S. M., and J. Gurevitch, eds., *Design and Analysis of Ecological Experiments*. New York: Chapman and Hall.
Reckhow, K. H. (1990). Bayesian Inference in Non-replicated Ecological Studies. *Ecology* **71**:2053-2059.
Richardson, A. (2006). The Many Unities of Science: Politic, Semantics, and Ontology in *Scientific Pluralism* Kellert, S., H. Longino, and C.K. Waters (eds) pp. 1-25.
Robins, J. M., and S. Greenland. (2000). Comments on Dawid. *J. Am. Stat. Assn* **95**: 431-435.
Rosenberg, A. (1990). Normative Naturalism and the Role of Philosophy, *Philosophy of Science* **57**(1): 34-43.
Ross D. (2005). Economic Theory and Cognitive Science: Microexplanation. A Bradford Book: MIT Press: Cambridge, MA.Roth, P.
Roth P. (1999) The Epistemology of "Epistemology Naturalized," *Dialectica* **53**(2)
Rubin, D. B. (2000). Comments on Dawid. *J. Am. Stat. Assn* **95**:435-438.
Salmon, W. (1989). *Four Decades of Scientific Explanation*, Minneapolis: University of Minnesota Press.
Salsburg, D. (2001). *The Lady Tasting Tea: How Statistics Revolutionized Science in the Twentieth Century*. New York: W. H. Freeman and Company
Sampson, C. J., P. L. Brezonik, T. M. Frost, K. E. Webster, and T. D. Simonson. (1995). Experimental Acidification of Little Rock Lake, Wisconsin: The First Four Years of Chemical and Biological Recovery. *Water Air Soil Poll*. **85**:1713-1719.
Scheffler, I. (1982). *Science and Objectivity*, 2nd Ed. Indianapolis, Indiana: Hackett Publishing Company.
Schindler, D. W. (1990). Experimental Perturbations of Whole Lakes as Tests of Hypotheses Concerning Ecosystem Structure and Function. *Oikos* **57**:25-41.
Schindler, D. W. (1998). Replication versus Realism: The Need for Ecosystem-Scale Experiments. *Ecosystems* **1**:3231-3334.
Schindler, D. W., K. H. Mills, D. F. Malley, D. L. Findlay, J. A. Shearer, I. J. Davies, M. A. Turner, G. A. Linsey, and D. R. Cruikshank. (1985). Long-Term Ecosystem Stress: The Effects of Years of Experimental Acidification on a Small Lake. *Science* **228**:1395-1401.
Shapere, D. (1982). The concept of observation in science and philosophy. *Philosophy of Science* **49**(9): 231-67.

Signor, P., and N. Gilinsky. (1991). Introduction in Gilinsky, N. L., and P. W. Signor, eds., *Analytical Paleobiology*. Knoxville, TN: Paleontological Society.
Sinks, T. & W. Wagner & D. Farquhar (2007) "The Science and the Law of Toxics," *Journal of Law, Medicine & Ethics: Special supplement*: *Public Health and the Law*. **35**(4):63-68.
Solomon, M. (2001). *Social Empiricism*, Cambridge MA: MIT Press.
Spanos, A. (1986). *Statistical Foundations of Econometric Modeling*. Cambridge: Cambridge University Press.
Spanos, A. (1999). *Probability Theory and Statistical Inference*. Cambridge: Cambridge University Press.
Staley, K., J. Miller and D. Mayo (eds) (2008). *Synthese* **163**(3) Error and Methodology in Practice: Selected Papers from ERROR 2006.
Statistical Science (2003) **18**(1): 1-32 Berger article Could Fisher, Jeffreys and Neyman Have Agreed on Testing? and Commentaries.
Steel, D. (2001). Bayesian Statistics in Radiocarbon Calibration**.** *Philosophy of Science* **68**(3), Supplement: Proceedings of the 2000 Biennial Meeting of the Philosophy of Science Association. Part I: Contributed Papers (Sep. 2001), pp. S153-S164
Stewart-Oaten, A., J. R. Bence, and C. W. Osenberg. (1992). Assessing Effects of Unreplicated Perturbations: No Simple Solutions. *Ecology* **73**:1396-1404.
Stewart-Oaten, A., W. W. Murdoch, and K. R. Parker. (1986). Environmental Impact Assessment: "Pseudoreplication" in Time? *Ecology* **67**:929-940.
Urban Dictionary: http://www.urbandictionary.com/define.php?term=truthiness
Vasudevan, Anubav (2004) Animating the EPR-Experiment: Reasoning from Error in the Search for Bell Violations, VT electronic dissertation: http://scholar.lib.vt.edu/theses/available/etd-01102005-145952/
Watras, C. J., and T. M. Frost. (1989). Little Rock Lake (Wisconsin): Perspectives on an Experimental Ecosystem Approach to Seepage Lake Acidification. *Arch. Environ. Contam. Toxicol.* **18**:157-165.
Weber, M. (2005). *Philosophy of Experimental Biology* Cambridge: Cambridge University Press.
White, P. (1991) *The Idea Factory: Learning to think at M.I.T.* New York: Dutton.
Wilson, R. (M.D) (1968) Feminine Forever (At any age, you can be...). New York: Pocket Book.
Woodward, J. (1989). Data and Phenomena. *Synthese* **79**:393-472.
Woodward, J. (1997). Explanation, Invariance, and Intervention. *Philosophy of Sciences* **64** (proceedings): S26-S41.
Woodward, J., and J. Bogen. (1988). Saving the Phenomena. *Phil. Rev*. **97**(3):303<n>352.
Worrall, J. (1985). Scientific discovery and theory-confirmation. In Pitt, J.C. (Ed.), *Change and progress in modern science* (pp. 301–332). Dordrecht: Reidel.
Worrall, J. (1989). Fresnel, Poisson and the White Spot. In Gooding, D., Pinch, T., & Schaffer, S. (Eds.), *The uses of experiment* (pp. 135–157). Cambridge: University of Cambridge Press.
Worrall, J. (1999). Two Cheers for Naturalised Philosophy of Science—or: Why naturalized Philosophy of Science is Not the Cat's Whiskers. *Science & Education* **8**: 339-361.
Worrall, J. (2010). Error, Tests and Theory Confirmation in D. Mayo & A. Spanos (eds). *Error and Inference: Recent Exchanges on the Philosophy of Science, Inductive-Statistical Inference, and Reliable Evidence*. Cambridge: Cambridge University Press: 125-144
Wylie, A. (2002). *Thinking from Things: essays in the philosophy of archaeology*. Berkeley: University of California Press.

Ziman, J. (1994). *Prometheus Bound.* New York: Cambridge University Press.
Zuckerman, H. (1988). Chapter 16: The Sociology of Science. In Neil J. Smelser (ed.) *The Handbook of Sociology*. Newbery Park CA: Sage Publications. 511-574.