The New Experimentalism, Topical Hypotheses, and Learning from Error[1]

Deborah G. Mayo

Virginia Polytechnic Institute and State University

1. Introduction: The New Experimentalists

Following a period during which philosophers of science focused on theory to the near exclusion of experiment, a number of philosophers, historians and sociologists of science have, in one way or another, turned their attention to experimentation, instrumentation, and laboratory practices.[2] Considerable work in philosophy of science of the last decade reflects this surge of interest in experiment, as promoted by Ackermann, Cartwright, Franklin, Galison, Giere, Hacking and others. Where has this movement taken us and where do we still have to go?

In asking this question, my focus is on that subset of the experimentalist movement whose members, following Ackermann (1989), I dub the "New Experimentalists". Although their agendas differ, members of this group share the core thesis that aspects of experiment might offer an important, though largely untapped, resource for addressing key problems in philosophy of science. In particular, their hope is to find ways to steer a path between the old logical empiricism, where observations were deemed relatively unproblematic, and the more pessimistic post-Kuhnians, who take the failure of logical empiricist models of appraisal as leading to underdetermination and holistic theory change, if not to denying outright the role of evidence in constraining appraisal. To steer this path it is suggested that we clear away the obstacles created by old-style accounts of how observation provides a basis for appraisal (via confirmation theory or inductive logic) and repave the way with an account rooted in the actual procedures for arriving at experimental data and experimental knowledge.

Why is it thought that turning to these experimental practices will offer up new pathways for grappling with philosophical problems about evidence and inference? The answer, as I see it, can be summed up with Ian Hacking's apt slogan: "experiment may have a life of its own" (1983, 160).

There are three main senses in which the life of experiment may be independent of theories and theorizing, and each corresponds to an important theme brought out by the New Experimentalist work. First, the claim of an independent life for experiment, the one initially emphasized by Hacking (1983), asserts that the *aims* of experimental

PSA 1994, Volume 1, pp. 270-279

inquiry may be quite independent of testing, confirming or filling out some theory. Instead, actual experimental inquiries focus on manifold local tasks: checking instruments, ruling out extraneous factors, getting accuracy estimates, distinguishing real effect from artifact, and estimating the effects of background factors.

The second reading of the slogan asserts that experimental data may be *justified* independently of theory, that experimental evidence need not be theory-laden in any way that invalidates its role in grounding experimental arguments. "A philosophy of experimental science", insists Hacking, "cannot allow theory-dominated philosophy to make the very concept of observation become suspect." (1983, 185)

A third reading of the slogan asserts that experimental knowledge may be retained despite changes of theory. Says Galison, "experimental conclusions have a stubbornness not easily canceled by theory change." (1987, 259) This suggests that experimental knowledge may serve not only in adjudicating between rival theories, but also as a basis for progress in science.[3]

In exploring these three themes the New Experimentalists have opened up a new and promising avenue for grappling with key challenges currently facing philosophers of science. Less clear is whether the new attention being accorded experiment has paid off in advancing solutions to these problems. Nor is it clear that they have demarcated a program for working out a philosophy or epistemology of experiment. For sure, they have given us an important start: their experimental narratives offer a rich source of illustrations of how experiment lives its own life apart from high-level theories and theorizing. But something more general and more systematic seems to be needed to show how this independence is achieved and how it gets us around the problems of evidence and of inference in so-called theory dominated philosophies. My aim in this paper is to suggest why the New Experimentalism has come up short and propose a way to remedy this. I will illustrate a portion of my proposal utilizing Galison's (1987) interesting experimental narrative on neutral currents. All references to Galison will be to this work.

2. Getting Small: Topical Hypotheses and the Local Discrimination of Error

To begin, I suggest we pursue seriously the first reading of the slogan, "experiment has a life of its own". Galison states it clearly:

> [E]xperimentalists' real concern is not with global changes of world view. In the laboratory the scientist wants to find local methods to eliminate or at least quantify backgrounds, to understand where the signal is being lost, and to correct systematic errors. (245)

For Galison, the question "How do experiments end?" (as in the title of his book) asks "When do experimentalists stake their claim on the reality of an effect? When do they assert that [it]...is more than an artifact of the apparatus or environment?" (4) The answer, in a nutshell, is only after having sufficiently well ruled out or subtracted out various backgrounds. Accordingly, a central experimental task is investigating and debating claims about backgrounds.

More recently, Hacking refers to the kind of claims experiment investigates as "topical hypotheses"—like topical creams—in contrast to deeply penetrating theories. Hacking claims:

> It is a virtue of recent philosophy of science that it has increasingly come to acknowledge that most of the intellectual work of the theoretical sciences is con-

> ducted at [the level of *topical* hypotheses] rather than in the rarefied gas of systematic theory. (Hacking 1992, 45)

To their credit, the New Experimentalists have been the leaders in this recognition. At the same time I think this points to the reason the New Experimentalists have come up short. The reason, as I see it, is that the experimental practices that have the most to offer in understanding these local tasks are still largely untapped. These are the activities involved in experimental design, experimental modeling, and data analysis—activities which, in practice, receive structure from statistical methods and arguments.

This is not to say the experimental narratives do not include the use of statistical methods. In fact, their narratives are chock full of specific applications of statistical techniques, e.g., techniques of data analysis, significance tests, confidence interval estimates, and other methods from what I propose to call *standard error statistics*.[4] What has not been done is explain how these methods are used to accomplish reliably the local tasks of arriving at data, learning about backgrounds, and so on.

In rejecting old-style accounts of confirmation as the wrong way to go, the New Experimentalists seem dubious about the value of utilizing statistical ideas to construct a general account of experimental inference. Theories of confirmation, inductive inference, and testing, were born in a theory-dominated philosophy of science, and this is what they wish to move away from. The complexities and context dependencies of actual experimental practice just seem recalcitrant to the kind of uniform treatment dreamt of by philosophers of induction. And since it is felt that overlooking these complexities is precisely what led to many of the problems that the New Experimentalists hope to resolve, it is natural to find them skeptical of the value of general inference accounts. Ironically, where there is an attempt to employ formal statistical ideas to give an overarching structure to experiment, some New Experimentalists revert back to the theory-dominated philosophies of confirmation, testing, and decision, particularly Bayesian philosophies (e.g., Franklin 1986, 1990).

The central position of what may be called "theory-dominated" philosophies of confirmation or testing is that the task of a theory of statistics begins with data or evidence already in hand, and seeks to provide some uniform rule (akin to deductive logic) to relate evidence (or evidence statements) to any theory, hypothesis, or decision of interest. Most commonly, the rule is to operate by providing some quantitative measure of support, confirmation, credibility or probability to hypotheses. Examples are the inductive logics of Carnap and of subjective Bayesians.

Galison is right to doubt that it is productive to search for "an after-the-fact reconstruction based on an inductive logic" (3). Such accounts, at their best, serve to reconstruct scientific inferences after-the-fact, rather than capture the methods actually used, though I will not argue this here. Where the New Experimentalists shortchange themselves is in playing down the use of local statistical methods at the experimental level—the very level they exhort us to focus on.

Those philosophers of statistics who have entered the experimentalist discussions (e.g., Howson and Urbach 1989) have encouraged this downplaying of the methods from standard error statistics. Embracing the theory-dominated philosophy of subjective Bayesian confirmation theory, Howson and Urbach reject standard error statistics as inappropriate, and regard its widespread use in experimental practice as unwarranted. Now it is true that the conglomeration of local tools comprising standard error statistics looks inadequate from the perspective of the aims of theory-dominated confirmation theory, because they do not provide a uniform quantitative measure of the bearing of evidence on hypotheses. But when it comes to the New Experimentalist aims, exactly the reverse is the case. Standard error statistics provide just the tools needed for investigating the topical hypotheses in experimental learning.

After all, if what we want are tools for discriminating signals from noise, ruling out artifacts, distinguishing backgrounds, and so on, then we really need tools for doing that. And these tools must be applicable with the kind of information scientists actually tend to have.[5] The conglomeration of methods and models from standard error statistics is the place to look for forward-looking procedures to obtain data in the first place, and which are apt even with only vague preliminary questions in hand. As such, these tools can provide the needed structure to the practices given a central place by the New Experimentalists.

3. Arguing From Error

Rather than approach the statistical tools in their formal setting, I shall begin right off with how I think they are used in experimental learning. Their aim, as I see them, is to direct experimental activities so as to allow us to give experimental arguments. The arguments follow a pattern of what might be called *an argument from error* or *learning from error.* The overarching structure of the argument is guided by the following thesis:

> It is learned that an error is absent when (and only to the extent that) a procedure of inquiry (which may include several tests) with a high probability of detecting the error if it existed, nevertheless failed to do so.

Such a procedure of inquiry, we can say, is one with a high capability of severely probing for errors—we may call it a *reliable (or highly severe) error probe.* According to the above thesis, we can argue that an error is absent if it fails to be detected by a highly reliable error probe.

Alternatively, the argument from error can be described in terms of a test of a hypothesis, H, that a given error is absent. The evidence indicates the correctness of hypothesis H, when H passes a severe test—one with a high probability of failing H, if H is false. An analogous argument can also be given to infer the presence of an error.[6]

Standard error statistics provides tools for reliable error probes that are robust across different scientific domains, with very minimal assumptions. The New Experimentalist offerings reveal (whether intended or not) the function and rationale of these statistical tools from the perspective of actual experimental practice—the very understanding missing from theory-dominated perspectives on scientific inference. Standard statistical tools, thus understood, can return the favor to the New Experimentalist program. Its already well-worked-out models and methods, I believe, provide the needed general framework for pursuing the different ways in which experiment lives a life of its own.

Here I shall focus on a first step, corresponding to the first construal of our slogan. This first step is to utilize the New Experimentalist narratives, together with this thesis about arguing from error, to understand the role of error statistics in distinguishing genuine effects from artifacts.

4. Distinguishing Effects From Artifacts: Galison and Neutral Currents

Galison's (1987) work is especially congenial. I shall follow a portion of his discussion of the discovery of neutral currents. Although by the end of the 1960s, Galison tells us, the "collective wisdom" was that there were no neutral currents (164 , 174), soon after (from 1971-1974) "photographs...that at first appeared to be mere curiosities came to be seen as powerful evidence for" their existence. (135) I am just going to focus on one particular analysis for which Galison provides detailed data. Abstracted from the whole story, this part will obviously not give an understanding of either the theory at stake or the sociological context. But it is sufficient to bring out the answer to Galison's key question: "[H]ow did the experimentalists themselves come to believe that neutral currents existed? What persuaded them that they were looking at a real effect and not at an artifact of the machine or the environment?" (136)

Here are the bare bones of the experimental analysis: Neutral currents are described as those neutrino events without muons. Experimental outcomes are described as muonless or muonful events, and the recorded result is the ratio of the number of muonless and muonful events. (This ratio is an example of what is meant by a statistic—a function of the outcome.) The main thing is that the more muonless events recorded, the more the result favors neutral currents. The worry is that recorded muonless events are due, not to neutral currents, but to inadequacies of the detection apparatus.

Experiments were conducted by a collaboration of researchers from Harvard, Wisconsin, Pennsylvania, and Fermilab, the HWPF group. They recorded 54 muonless events and 56 muonful events giving a ratio of 54/56. The question is: Does this provide evidence of the existence of neutral currents?

> For Rubbia [from Harvard] there was no question about the statistical significance of the effect. ...Rubbia emphasized that 'the important question in my opinion is whether neutral currents exist or not... The evidence we have is a 6-standard-deviation-effect.' (Galison, 220)

The "important question" revolved around the question of the statistical significance of the effect. I will refer to it as the *significant question*. Galison puts it this way:

> Given the assumption that the pre-Glashow-Weinberg-Salam theory of weak interactions is valid (no neutral currents), then what is the probability that HWPF would have an experiment with as many recorded muonless events as they did? (220)

Three points need to be addressed: How might the probability in the significant question be interpreted? Why would one want to know it? and, How might one get it? While the answers to these questions are found to be problematic from the point of view of theory-dominated accounts of inference, this is not the case were one to adopt the point of view of the New Experimentalism. I will consider each in turn.

(i) *Interpreting the significant question*

What is being asked when one asks for the probability that HWPF would have an experiment with as many recorded muonless events as they did, given no neutral currents? The question, in statistical language, is: How (statistically) significant is the number of recorded excess muonless events? Here I want to explain the significant question informally.

The experimental result, recall, was the recorded ratio of muonless to muonful events, namely, 54/56. The significant question, then, is: What is the probability that HWPF would get as many as (or more than) 54 muonless events, given there are no neutral currents? One way to cash out what is wanted is this: How often, in a series of experiments such as the one done by HWPF, would as many muonless events be expected to occur, given there are no neutral currents?

But there is only this one experimental result, not a series of experiments. True, the series of experiments here is a kind of hypothetical construct. What we need to get at is why it is perceived as so useful to introduce this hypothetical construct into the data analysis.

(ii) What is the value of answering the significant question?

The quick answer is that it is an effective way to distinguish real effect from artifacts. Were the experiment so well controlled that the only reason for failing to detect a muon is that the event is a genuine muonless one, then artifacts would not be a problem and this statistical construct would not be needed. But artifacts are a problem. From the start a good deal of attention focused on the backgrounds that might fake neutral currents. (Galison,177) A major problem was escaping muons. "From the beginning of the HWPF neutral-current search, the principal worry was that a muon could escape detection in the muon spectrometer by exiting at a wide angle. The event would therefore look like a neutral-current event in which no muon was ever produced." (Galison, 217)

The problem, then, is to rule out a certain error: construing as a genuine muonless event one where the muon simply never made it to the spectrometer, and thus went undetected. To relate this problem to the significant question, let us introduce some abbreviations. If we let hypothesis H be

H: neutral currents are responsible for (at least some of) the results

then, within this piece of data analysis, the falsity of H is the artifact explanation:

H is false (the artifact explanation): recorded muonless events are due, not to neutral currents, but to wide-angle muons escaping detection.

Our significant question becomes:

What is the probability of a ratio (of muonless to muonful events) as great as 54/56, given that H is false?

The answer is the *significance probability* or *significance level* of the result.

Returning to the relevance of knowing this probability (the significance level), suppose it were found to be high. That is, suppose as many or even more muonless events would occur frequently, say more often than not, even if H is false (and it is simply an artifact). What is being supposed is that a result, as or even more favorable to H than the HWPF result, is fairly common due, not to neutral currents, but to wide angle muons escaping detection. Were that so, the HWPF result clearly does *not* provide grounds to rule out wide-angle muons as the source (the artifact explanation). Were one to proceed by taking such a result as grounds for ruling out the artifact explanation, one would be wrong more often than not. That is, the probability of correctly detecting the artifact explanation (not-H) would be less than .5. The procedure would be an unreliable error probe. Since high significance level means low reliability, results are not taken to indicate H unless the significance probability is low.

Suppose now that the significance probability is very low, say 0.01 or 0.001. This means that it is extremely improbable for so many muonless events to result, if H were false and the HWPF researchers were really only observing the result of muons escaping. Since escaping muons could practically never be responsible for so many muonless events, their occurrence in the experiment is taken as good grounds for rejecting the artifact explanation. That is because, following an argument from error, the procedure is a highly reliable probe of the artifact explanation. This was the case in the HPWF experiment, although the probability in that case was considerably smaller. But how do you get the significance probability?

(iii) How is the significant question answered?

The reasoning I just described does not require a precise value of the significance probability. It is enough to know that it is or is not very low—that the procedure is or is not fairly reliable. But how does one arrive at even a ballpark figure? The answer comes from the use of various standard statistical analyses, but to apply them (even qualitatively) requires information about how the artifact in question could be responsible for certain experimental results. Statistical analyses are rather magical, but they do not come from thin air. They send the researcher back for domain-specific information. Let us see what the HWPF did.

The data used in the HWPF paper is as follows: (Galison, 220)

Visible muon events	56
No visible muon events	54
Calculated muonless events	24
Excess	30
Statistical significant deviation	5.1

The first two entries just record the HWPF result. What about the third entry, the calculated number of muonless events? This refers to the number calculated or expected to occur because of escaping muons. This calculation comes from separate work deliberately carried out to find out how an event can wind up being recorded "muonless", not because no muon was produced (as would be the case in neutral currents), but because the muon never made it to the detection instrument.

The group from Harvard, for example, created a computer simulation to model statistically how muons could escape detection by the spectrometer by exiting at a wide angle. This is an example of what is called a "Monte Carlo" program.

> By comparing the number of muons expected not to reach the muon spectrometer with the number of measured muonless events, they could determine if there was a statistically significant excess of neutral candidates. (Galison, 217)

In short, the Monte Carlo simulation afforded a way (not the only way) to answer the significant question.

The reason probability arises in this part of the analysis is not because the hypothesis about neutral currents is a statistical one, much less because it quantifies credibility in H or in not-H. Probabilistic considerations are deliberately *introduced* into the data analysis because they offer a way to model the expected effect of the artifact (escaping muons). Statistical considerations, we might call them "manipulations on paper" (or on computer), afford a way to subtract out background factors that cannot literally be controlled for. In several places, Galison brings out what I have in mind:

> In a sense the computer simulation allows the experimentalist to see, at least through the eye of the central processor, what would happen if a larger spark chamber were on the floor, if a shield were thicker, or if the multiton concrete walls were removed.

> The Monte Carlo program can do even more. It can simulate situations that *could never exist in nature*. ... One part of the Gargamelle demonstration functioned this way: suppose the world had only charged-current neutrino interactions. How many neutral-current *candidates* would there be? (265)

It was calculated that 24 muonless events would be expected in the HWPF experiment due to escaping muons. Next, Galison explains, "they wanted to know how likely it was that the observed ratio of muonless to muon-ful events (54/56) would fall within the statistical spread of the calculated ratio (24/56), due entirely to wide-angle muons." (220) The difference between the ratio observed and the ratio expected (due to the artifact) is 54/56 - 24/56 = 0.536. How improbable is such a difference even if the HWPF experiment *were* being done on a process where the artifact explanation is true (i.e., where recorded muonless events were due to escaping muons)? This is "the significant question" again, and finally we can answer it.

The simulation lets us model the relevant features of what it would be like were the HWPF study actually experimenting on a process where the artifact explanation is true. It tells us it would be like experimenting on a process that generates ratios (of m events to m-less events) where the average (and the most likely) ratio is 24/56. (This corresponds to the hypothetical sequence of experiments we spoke of.) The statistical model tells us how probable different observed ratios are, given the average ratio is 24/56. In other words, the statistical model tells us what it would be like to experiment on a process where the artifact explanation is true; namely, certain outcomes (observed ratios) would occur with certain probabilities. (Most experiments would yield ratios close to the average (24/56); the vast majority would be within two standard deviations of it.)

Putting an observed difference between recorded and expected ratios in standard deviation units allows one to use a chart to read off the corresponding probability. The standard deviation (generally estimated) gives just that—a standard unit of deviation that allows the same standard scale to be used with lots of different problems in different scientific domains. Any difference exceeding two or more standard deviation units corresponds to one that is improbably large (occurring less than 2% of the time).

Approximating the standard deviation of the observed ratio shows the observed difference to be 5.1 standard deviations.[7] This is so improbable as to be off the charts; so, clearly, by significance test reasoning, the observed difference indicates that the artifact explanation is untenable. It is practically impossible for so many muonless events to have been recorded, were they due to the artifact of wide angle muons. The procedure is a highly reliable artifact probe.[8]

This is just one small part of a series of experimental arguments that took years to build up. Each involved this kind of statistical data analysis to distinguish real effects or signals from artifacts, to estimate the maximum effect of different backgrounds, and to rule out key errors *piece-meal*. They are put together to form the experimental arguments that showed the experiment could end.

5. Conclusion

The New Experimentalists are right to insist on the centrality of the tasks of distinguishing and subtracting out backgrounds, quite apart from the aim of testing high-level theories. They are also right to suppose that experimental practices offer especially powerful tools for these local tasks. While their experimental narratives offer a rich source of illustrations, something more general is needed to understand how experimental practices accomplish these tasks. In this paper I have showed how a standard error statistical tool (significance tests) together with an experimental narrative, can serve to articulate the procedure for distinguishing artifacts in an important class of cases. The next step or set of steps would be to explore how a handful of standard (or canonical) statistical models permit analogous arguments from error to be substantiated across a wide spectrum of experimental inquiries. These, still mostly untapped, tools, I believe, are the key to advancing solutions to the problems about evidence and inference that the New Experimentalist movement set for itself.

Notes

[1] This research was supported by an NSF award in Studies in Science, Technology and Society. I gratefully acknowledge that support.

[2] A collection of this work may be found in Achinstein and Hannaway (1985). For a good selection of interdisciplinary contributions, see Gooding, Pinch, and Schaffer (1989).

[3] Giere (1988) and Hacking (1983) have especially stressed how this sort of progress is indicated when an entity or process becomes so well understood that it can be used to investigate other objects and processes.

[4] I use this label rather than the labels often given to specific components of this methodology, e.g., Fisherian tests, Neyman-Pearson or Orthodox statistics, because the latter are associated with certain inference philosophies that do not necessarily reflect the uses of these methods in experimental practice.

[5] In contrast, to get a Bayesian inference going, an agent requires a prior probability assignment to an exhaustive set of hypotheses, among other things.

[6] I discuss severe tests in Mayo (1991). A full discussion of arguing from error, and a development of the corresponding error statistics approach occurs in Mayo (forthcoming).

[7] The standard deviation is estimated using the recorded result and a standard statistical model. It equals

$$\left(\frac{24}{56}\right)\sqrt{\frac{1}{24}+\frac{1}{56}} = 0.105. \text{ (Galison 1987, 220-221)}$$

[8] Galison points out that a different analysis of the HWPF data resulted in a different level of significance—still highly significant. The error statistics approach does not mandate one best analysis—several are used to check and supplement one another.

References

Achinstein, P. and Hannaway, O. (eds.) (1985), *Observation, Experiment and Hypothesis in Modern Physical Science*. Cambridge, MA: MIT Press.

Ackermann, R. (1985), *Data, Instruments, and Theory.* Princeton: Princeton University Press.

_ _ _ _ _ _ _ _ _ . (1989), "The New Experimentalism", *The British Journal for the Philosophy of Science* 40: 185-190.

Cartwright, N. (1983), *How the Laws of Physics Lie.* Oxford: Clarendon Press.

Franklin, A. (1986), *The Neglect of Experiment.* Cambridge: Cambridge University Press.

_ _ _ _ _ _ _ . (1990), *Experiment, Right or Wrong.* Cambridge: Cambridge University Press.

Galison, P. (1987), *How Experiments End.* Chicago: University of Chicago Press.

Giere, R. (1988), *Explaining Science.* Chicago: University of Chicago Press.

Gooding, D., Pinch, T., and Schaffer, S. (eds.) (1989), *The Uses of Experiment: Studies in the Natural Sciences.* Cambridge: Cambridge University Press.

Hacking, I. (1983), *Representing and Intervening.* Cambridge: Cambridge University Press.

_ _ _ _ _ _ _. (1992), "The Self-vindication of the Laboratory Sciences", in A. Pickering (ed.), *Science as Practice and Culture.* Chicago: The University of Chicago Press, pp. 29-64.

Howson, C. and Urbach, P. (1989), *Scientific Reasoning: The Bayesian Approach.* La Salle: Open Court.

Mayo, D. (1991), "Novel Evidence and Severe Tests", *Philosophy of Science* 58: 523-552.

_ _ _ _ _ _ . (forthcoming), *Error and the Growth of Experimental Knowledge.* Chicago: The University of Chicago Press.