SCRUTINIZING

SCIENCE

Empirical Studies of
Scientific Change

Edited by

ARTHUR DONOVAN
U.S. Merchant Marine Academy, New York

LARRY LAUDAN and RACHEL LAUDAN
University of Hawaii

KLUWER ACADEMIC PUBLISHERS
DORDRECHT / BOSTON / LONDON
Brownian movement (discovered by the botanist Brown in 1827) refers to the irregular motion of small suspended particles in various fluids which keeps them from steadily sinking due to gravitation. Attempts to explain this phenomenon are linked with the atomic debates from the late 19th century to the early 20th. In particular, the testing of the Einstein–Smoluchowski theory of Brownian motion by Perrin is considered to have provided the long sought-after evidence in favor of the molecular-kinetic theory of gases against classical thermodynamics. While this episode has often served as a case study for philosophers, such case studies, as remarked generally by Laudan et al. (1986, p. 159) "are not 'tests' of the theory in question at all, but applications of the theory to a particular case". The aim of my discussion will be to employ this episode to carry out genuine tests, focusing on the following set of interrelated theses on theory appraisal: (T2.3, T2.4, T2.5, T2.6, T2.8 and T2.10). As many of these pertain to evaluating guiding assumptions, I shall suggest briefly how my results bear on a thesis on the appraisal of guiding assumptions (GA1.1), and a thesis on scientific revolutions (GA4.1). The atomic debate as a whole is too broad to be taken up here. I restrict my focus to theories of Brownian motion, those put forward both prior and subsequent to those explicitly linked with the kinetic theory in the 1870s.

But what justifies taking my analysis as a genuine test of the theses, in contrast to the mere 'applications' mentioned earlier? Does not our lack of a theory of science open any alleged test of rival theories of science to the charge of circularity? The answer, I claim, is no. So long as the theses aim to accord with actual scientific inquiry – as the theorists under consideration all do – non-circular criteria for testing them exist. It needs only to be assumed that the criticism raised against 'mere applications' is indeed a criticism. The complaint, as put by Laudan et al. (1986, p. 159) is that "such applications, by treating the model in question as unproblematic, fail to be probative ...". If the model is assumed,
clearly the study has very little, if any, power to reveal its inadequacies. To deem this a criticism is tantamount to requiring that tests of theses have sufficient power to reveal correctly where the theses fail to describe science, as well as where they succeed. Otherwise the theses will fail to be constrained by how science actually functions.

This plausible demand alone provides the basis for two (meta-level) criteria for testing these theses. I shall frame them in the (attempted neutral) language of my test results, that is, in terms of assertions about whether a test case indicates (or is a good or poor indication of) a thesis $H$. They may be stated as:

Two (Meta-level) Criteria: A good test of a thesis $H$ must have little chance of incorrectly indicating thesis $H$, as well as a good chance of correctly indicating $H$.

(A quantitative construal of this occurs in Note 4.) My intention in explicitly setting out the criteria I am using is not only to facilitate carrying out my tests, but to facilitate their critical scrutiny by others.

A shared thrust of many of these theses is a downplaying of the importance of empirical data methods in appraising theories and guiding assumptions. In addition to testing these theses, my study serves to test, or at least uncover, the presuppositions underlying this post-positivist tendency. In the final wrap-up I shall remark on the bias that these presuppositions introduce into the testing program and suggest how it might be avoided. I begin by testing (T2.3):

*The appraisal of a theory is based on the success of the guiding assumptions with which the theory is associated.*

Sets of guiding assumptions, which I shall abbreviate throughout as GAs, are supposed to include not only specific theories, but ontological claims about what exists and methodological claims about how they should be studied. The debates I am considering as to the cause of Brownian motion correlate with a number of opposing views which may be seen as sets of GAs (e.g., molecular vs. energeticist ontologies; mechanical vs. phenomenological explanation, atomic vs. continuous metaphysics, statistical vs. non-statistical models, the method of hypotheses vs. positivistic methodologies, realism vs. instrumentalism, and others). Thus, if there are genuine causal connections between appraising theories and their associated guiding assumptions, as (T2.3) alleges,
there is a good chance they would be operating in this case. But that just means it is easy for an analysis of this episode to have a high chance of indicating such interconnections, as a number of discussions of this episode show. Such an analysis satisfies only one of the two criteria I am assuming are required for a good meta-test. A second requirement is that a meta-test does not construe the episode as indicating (T2.3) too readily. That is, the test must have a reasonably good chance of counter-indicating (T2.3) if the theory appraisal actually was not based on the success of GAs. This requires getting clear on what might be intended in denying (T2.3).

Thesis (T2.3) appeals to many contemporary theorists of science because its denial is typically assumed to be an endorsement of positivist views of theory appraisal, in which theories were thought to be evaluated by some sort of logical comparison with the evidence, in isolation. Granted, this episode counts against that sort of positivistic theory appraisal. However, my test of this episode shows the supposition that either a theory is tested in isolation (along positivistic lines), or else (T2.3) is correct (and its appraisal is based on the success of its associated set of GAs), is a false dilemma. For, neither need be the case. As in the rest of my study, I shall focus on the early theories positing causes of Brownian motion (around 1827–1877) as well as the evaluation of theories explicitly based on the kinetic theory of gases (e.g., the theory of Einstein and Smoluchowski).

Thesis (T2.3) is counter-indicated by the early theories about the cause of Brownian motion, because they were vigorously debated and appraised from its initial discovery by Brown in 1827 – even before it was known whose problem it was. (It could have been chemistry, biology, physics or something else.) Brown tested and rejected his own initial view that the particles were self-acted (by finding that the motion occurred in inorganic as well as organic substances). Brown’s experiments were recognized by Faraday (in an 1829 lecture) and others to have ruled out all of the causes of the motion suggested up until that time (e.g., unequal temperatures in the water, evaporation, air currents, heat flow, capillarity, motions of the observer’s hands, and others). (See, for example, Jones (1970, p. 403).) Thus, theory appraisal took place, but without knowing to which set of GAs (or even which science) it belonged.

It might be countered that (T2.3) refers only to appraising more
full-blown theories, such as those explicitly based on the kinetic theory of gases, e.g., by Delsaux in 1877. Even with this restriction, (T2.3), as I understand it, misdescribes the episode. For starters, it is not clear that there is a set of ‘GAs with which the theory is associated’. Some critics of the atomic-kinetic accounts (Ostwald, Duhem,) endorsed an energeticist ontology; others, like Mach and Planck, did not. A number of scientists (e.g., Clausius, Kelvin, and Born) worked within the school of the phenomenological (‘pure’) theory of thermodynamics, whose validity did not depend on any theory about the nature of matter, while they also contributed to atomic theory in different publications. Secondly, conflicting appraisals of the kinetic account were urged by scientists who held the same assessment of the success of associated GAs. This was so in what was considered at the time to be the most important challenge to the kinetic explanation, provided by Nageli, a German cytologist. Nageli shared with Delsaux the belief in the success of the atomic-kinetic viewpoint, admitting that “the idea that the molecules of a gas travel past each other with large velocities has entered physics and, because of its irrefutable proof, has found general agreement . . .”. 6 Ironically, from the point of view of (T2.3) Nageli’s (and his followers’) denial of Delsaux’s theory (attributing Brownian motion to molecular collisions) was based on accepting the molecular-kinetic model of gases. For they use the molecular magnitudes given by the kinetic theory (e.g., Avogadro’s number N) to argue against molecular collisions as their cause. (Nageli calculated that a million water molecules would have to hit Brownian-particles at the same moment from the same direction to account for a single jerk of the particle. But since the collisions would come from all different directions, he argued, the particle should not appear to jerk at all. He was in error.) The main outlines of the statistical argument by which Smoluchowski answers Nageli’s objection will arise later.

Thesis (T2.3) comes closest to describing portions of the dispute in which certain anti-atomists (e.g., Ostwald, Mach (at times) and the mathematician Zermelo), insisting on the absoluteness of the Second Law of Thermodynamics, base at least part of their opposition to the molecular-kinetic theory on the fact that it would allow exceptions to this principle. But even here there is evidence that the disputants on both sides did not regard such criticisms as an adequate basis for appraising the kinetic theory. (That this is so for anti-atomists like
Ostwald will be indicated in testing, and affirming, (T2.10) on crucial experiments.) Perrin makes it clear that without experimental arguments, the disputants were not yet at the level of scientific theory testing. While "the beautiful discoveries that we owe to molecular hypotheses" make Perrin hesitant "to support this radical opinion" of the anti-atomists, Perrin continues:

But since one has no direct proof of the existence of molecules, he can only go by esthetic reasons if he has not succeeded in proving by experimental arguments that the second law does not have the character of absolute rigor, in the name of which one would sacrifice the molecular theories (Perrin, 1950, p. 57 emphasis added).

It is not that our episode indicates that theory appraisal takes place without reference to facts from associated sets of GAs. It indicates only that theory-testing need not appeal to the overall success of the GAs. And where there is such an appeal to GAs it does not seem to count as the basis for an adequate theory appraisal. Theory appraisal, in this episode, seems to be based neither on appraising GAs nor on positivistic-style algorithms (relating statements of evidence and theories). Instead, it is based on statistical and experimental reasoning. Examining this reasoning shows this episode is seriously at odds with thesis (T2.6):

The appraisal of a theory is usually based on only a very few experiments, even when those experiments become the grounds for abandoning the theory.

Lakatos' endorsement of (T2.6) is expressed as a contrast between his methodological falsificationism and claims espoused by Popper. Lakatos (1978, p. 35) claims his own view "denies that 'in the case of a scientific theory, our decision depends upon the results of experiments' . . . It denies that 'what ultimately decides the fate of a theory is the result of a test, i.e., an agreement about basic statements'." It is true that actual scientific practice shows the inadequacy of this version of a Popperian test, as of other positivistic models of testing. But in so doing, these episodes show that theory appraisal requires many more, not less, experiments than positivist models suppose. For, the fact that theory appraisal depends on intermediate theories of data and experiment, and that data is theory-laden, inexact, and 'noisy', only underscores the necessity for numerous experiments, shrewdly interconnected, if one is to learn from data. From the initial discovery of Brownian motion by Brown in 1827, each inquiry into its cause is a story of hundreds of
experiments. They may be grouped into two main classes: (1) experiments to (arrive at and) test theories attributing Brownian motion either to the nature of the particle studied, or to factors external to the liquid medium in which the Brownian particles were suspended (e.g., temperature differences in the liquid observed, vibrations of the object glass); (2) experiments to test the quantitative theory of Brownian motion put forward by Einstein and (independently) by Smoluchowski. I shall abbreviate the Einstein–Smoluchowski Theory of Brownian motion as the ES Theory.

Each molecular-kinetic explanation of Brownian motion (first qualitatively proposed by Wiener in 1863) spurred a flurry of experiments by biologists and physicists aimed at refuting it. Each non-atomic theory tried, say by devising a new way to view Brownian motion as due to temperature differences, triggered a new set of experiments to refute the challenge. The enormous variety of organic and inorganic particles studied is too numerous to list in full, but it includes sulfur, cinnabar, coal, urea, india ink, and gamboge. Equally numerous were the treatments to which such particles were subjected in the hope of uncovering the cause of Brownian motion – light, dark, country, city, red and blue light, magnetism, electricity, heat and cold, even freshly drawn human milk.

Counter-indicating (T2.6), our test case shows experiments to be the main basis for checking predictions, as well as for learning how better to run experiments. The major scientists working on this problem (e.g., Brown, Wiener, Ramsay, Gouy, Perrin and Smoluchowski) would begin by carrying out experiments to exclude all exterior causes – checking and rechecking even those suspected factors that had already been fairly well ruled out. (Even after Perrin and the general acceptance of molecular theory, experiments continued, using ever-improving methods to observe hundreds of thousands of microscopic grains.) By the end of the 19th century the most favored explanations attributed the cause to heat in some way (e.g., theories of Exner, Dancer, Quincke). Ironically, finding that the same Brownian particles could be used over and over again, sometimes conserved on slides for 20 years, compelled experimenters who had been searching for non-kinetic explanations to admit this as strong evidence for the kinetic explanation. It indicated the motion was eternal and spontaneous, in accordance with the kinetic account. But most required many more quantitative experiments before abandoning non-kinetic explanations.
BROWNIAN MOTION

It is only by keeping in mind that a great many causal factors had been ruled out by experiments prior to Perrin's tests (around 1910) of the ES theory of Brownian motion, that the work his experiments performed can be understood. Because his experiments are so central in testing all of the other theses I plan to consider, I shall discuss a few points in some detail. The central problem with appraising the ES theory was that for many years experimenters were simply measuring the wrong thing. What they thought had to be checked was whether the molecular effects on the velocity of Brownian particles accorded with that hypothesized by the kinetic theory. But this mean velocity had been ascertained by trying to follow the path of a Brownian particle, inevitably yielding a measured path much simpler and much shorter than the actual path, which changes too fast. An important advantage of the ES theory was that it provided a testable prediction that made no reference to this unmeasurable velocity (and at the same time explained why attempts to measure it had failed). Instead, it was put in terms of the expected displacement of particles. The displacement of a Brownian particle is the total distance it travels in any direction (say along the x-axis of a graph) as it weaves its zig-zagged path, and this could be measured using the microscopes of the day. As is usual — even in testing theories not themselves statistical — the testable prediction was of a statistical distribution law. Here the distribution law specifies how frequently Brownian particles would be expected to be displaced along the x-axis by certain amounts over a given time \( t \). If molecular agitation (as described by the kinetic theory of gases) causes Brownian movement, then the displacement of a Brownian particle is Gaussian distributed about its mean (which by symmetry is 0) with variance equal to \( 2Dt \) (where \( D \) is the coefficient of diffusion and \( t \) is the time). (Simply put, this tells us that displacements near 0 are most likely to occur, while those further from 0 are increasingly less likely.) Since Avogadro's number \( N \) is a function of \( D \), the prediction of the kinetic theory (for a given type of particle) can be stated as a predicted standard deviation. Once \( D \) is estimated, Avogadro's number \( N \) can be calculated and compared to values hypothesized by the kinetic theory. So the crux of Perrin's appraisal of the ES theory is evaluating the statistical hypothesis:

\( H \): The experimental displacement distribution is from a population distributed according to the Gaussian distribution \( M \) with parameter value a function of \( N^* \),
where \( N^* \) is the (probable) value for \( N \) hypothesized by the kinetic theory. However, the sample data can be used to this end only if it can be seen (i.e., modelled) as the results of observing displacements from the hypothesized Gaussian process. Being a good experimenter, Perrin realized that appraising \( H \) requires two broad steps which I shall call \textit{Step (I)} and \textit{Step (II)}. \textit{Step (I)} consisted in checking whether the results of the experiment actually performed follow the given statistical distribution \( M \), and \textit{Step (II)} involved using estimates of \( D \) to estimate (or test) values of \( N \). While statistical theory was not fully developed at the time, the methods (e.g., \textit{chi-square} tests) employed by Perrin were fairly routine. \textit{Step (I)} involved a \textit{(chi-square)} test of the claim

\[(j) \quad \text{The data approximates a (random) sample from the (hypothesized) Gaussian process } M.\]

In other words, the \textit{denial} of \( j \) is \( j^\dagger \):

\[(j^\dagger) \quad \text{The assumptions of the experimental model } M \text{ are violated; the sample displacements are characteristic of systematic effects.}\]

Being able to rule out \( j^\dagger \) indicates that violations are sufficiently negligible for the purpose of estimating the parameter \( D \). The necessity of doing so introduces the need for a large number of experiments. The aim is to learn enough about potential violations to either generate data that avoids them regularly, or to compensate them in the analysis. \textit{Step (II)} calls for numerous experiments for a different reason. Essentially it is because the statistical prediction is an assertion of what would be expected in a large number of experiments. As well as indicating the importance of large numbers of experiments, the statistical nature of the theory appraisal indicates well (T2.5):

\textit{The appraisal of a theory is based on phenomena which can be detected and measured without using assumptions drawn from the theory under evaluation.}

This is the case because the statistical data analysis concerns only low level, independently checkable theories. The availability of the data
analyses (thanks to Perrin) and the clarity of the derivation of the Gaussian model by Einstein and others\textsuperscript{10} is such that there is a very good chance of uncovering such question-begging assumptions, were Perrin's theory appraisal to require them. Thus, this is a good test case of (T2.5).

Without actually getting down to the nitty-gritty of the data analysis in Step (I), however, Perrin's avoidance of such assumptions is likely to be overlooked, and the reason why the results of Perrin's Step (II) were so convincing overlooked as well. (Feyerabend (1978, e.g., p. 39) is a case in point.) What made Perrin's results (at Step (II)) so telling was his ability to ensure the particles observed had approximately the same size, that they could be counted, weighed, and a host of extraneous factors controlled or 'subtracted out' – even Einstein expressed surprise. The general argument in ruling out possible external factors was that were Brownian motion the effect of such a factor, then neighboring particles would be expected to move in approximately the same direction; the movement of a particle's neighbors would not be independent of its own. Thus, the object of Perrin's inquiry was to determine whether the movement of Brownian particles exemplified a type of random phenomenon, known from simple games of chance. As Perrin (p. 112) shows, Einstein's derivation of the displacement distribution depends on "making the single supposition that the Brownian movement is completely irregular \ldots." Checking this single supposition was a matter of testing if the Brownian movement exemplifies the type of random phenomenon known from simple games of chance. Thus, this supposition could be checked 'without using assumptions drawn from the theory under evaluation' as (T2.5) requires. These checks made use of a completely independent statistical theory of the random walk phenomena. The displacement of a particle may be seen as the result of \( k \) steps, where at each step the particle has an equal chance of being displaced by a given amount in either a positive or negative direction, as described in Note 9. (This is called a simple random walk.) Were it incorrect to assume \( j \), there would be an unequal chance of being displaced by an amount and these dependencies would be detectable.\textsuperscript{11} This single supposition is precisely what is checked in Step (I), and doing so avoids question-begging assumptions.

Consider one of Perrin's checks, based on observed displacements of 500 grains of gamboge (a type of microscopic vegetable particle). The positions of each are recorded every 30 sec on paper with grids of
squares, as shown in Figure 1. The values are then shifted to a common origin in order to count the number found various distances from this point. As Perrin reasons,

Fig. 1. Tracing of horizontal projections of lines joining consecutive positions of a single grain every 30 sec (Perrin, p. 115).

The extremities of the vectors obtained in this way should distribute themselves about that origin as the shots fired at a target distribute themselves about the bull’s-eye. (p. 118)

Fig. 2. Observed distribution of displacements of 500 gamboge grains.

Here again we have a quantitative check upon the theory; the laws of chance enable us to calculate how many points should occur in each successive ring (p. 118, emphasis added).

The number observed in each ring is close to the hypothesized number. That is, the observed displacements do not differ in a statistically significant way from what ‘should occur’ according to hypothesis $j$. (They are not improbable under $j$.) This indicates that $j$ holds, only because the experiments were deliberately designed so that were the model inadequate, it would almost always yield differences that were
statistically significantly different from what is expected according to \( j \). On such grounds, \( j' \) is ruled out by Perrin and others, and \( j \) is upheld:

In short, the irregular nature of the movement is quantitatively rigorous. Incidentally we have in this one of the most striking applications of the laws of chance (Perrin, p. 119, emphasis added).

The multiple experiments for which Perrin stresses the need (e.g., Perrin, p. 96) are those deliberately designed so that if one misses a bias, another is likely to find it. The bias of concern here is that some regularity of molecular motion has been concealed. Such a regularity was the main alternative to Perrin's causal hypothesis \( H \).\(^{12}\) He rules it out by finding that the Brownian motion is totally irregular (as described in \( M \)). But does the theory appraisal in this episode seem correctly described as relative "to rival theories in the field", as asserted in (T2.8)?

*The appraisal of a theory is relative to prevailing doctrines of theory assessment and to rival theories in the field.*

Focusing on the latter part of this thesis, as I understand it, the answer seems to be no. A test claim \( H \) always has a rival but it may be a vague 'something other than \( H \)'. Since Brown's tests in the 1820s, scientists talked of testing theories as to the cause of Brownian motion (as traced in testing (T2.6)) where the only 'rivals' were assertions denying the hypothesized causal factor. As in checking whether experimental assumptions are met or not (\( j \) vs. \( j' \)), the only rivalry is between a claim that a given model adequately captures certain aspects of the Brownian movement, and a claim that the motion is so far from what the phenomenon described in the model is likely to produce, as to indicate it is not so produced. Here the testing follows the logic of Fisherian statistical significance testing, in use at the time. Discrepancies either are or are not detected from a single hypothesis. (It is more like a Popperian refutation.\(^{13}\))

But it might be countered that these low level theories do not provide a good test of what is asserted in (T2.8). After all, testing Einstein's theory was regarded as a test telling against the rival Classical Thermodynamics. My response is twofold. First, to the extent that Perrin's tests tell against Classical Thermodynamics — by showing that Brownian motion violates a non-statistical version of the Second Law — that counter-evidence is already provided by the simple statistical tests at
Step (I). (This is discussed in my test of (T2.10).) Second, even if Perrin's tests of (the parameters of) Einstein's theory of Step (II), the testing strategy is by local statistical hypotheses. Having checked that the assumptions of the experimental model have been met, by ruling out \( j^1 \) in Step (I), Perrin remarks

To verify Einstein's diffusion equation, it only remains to see whether the number [obtained for \( N \) by substituting the estimate of \( D \) into the equation \( D = RT/\pi \eta s \)] is near \( 70 \times 10^{22} \) (p. 132).\(^4\)

The 'alternatives' at this stage of appraisal (i.e., Step (II)) are the possible values of the parameter space of \( N \) or \( D \). This is still hardly more of a rival assessment than in testing \( j \) and \( j^1 \) in Step (I).\(^5\) More than this is typically meant in asserting (T2.8).

How does the test of the predicted value for Avogadro's number (Step (II)) proceed? Nearly all the estimates of \( N \) from several experiments were (statistically) insignificantly far from that predicted by the kinetic theory, \( N^* \) (i.e., \( 70 \times 10^{22} \)). Perrin declares:

It cannot be supposed that, . . . values so near to the predicted number have been obtained by chance for every emulsion and under the most varied experimental conditions (p. 105).

The many instances this case affords of such (so-called) 'arguments from coincidence'\(^6\) recommend it as a good test of theses on the role of 'surprise' and 'novelty' in science, such as (T2.4):

*Theory appraisal is based entirely on those phenomena gathered for the express purpose of testing the theory and which would be unrecognized but for that theory.*

This statement is interestingly ambiguous, and whether it is indicated by this test case depends on its interpretation. It is counter-indicated if understood as first intended by Lakatos (e.g., Lakatos, 1978, p. 38). On this *temporal view* of novelty, as Lakatos asserts, (p. 184) "The best novel facts were those which might never have been observed if not for the theory which anticipated it." Brownian motion, of course, had been recognized since 1827. But Zahar's (1973, p. 103) definition of novel prediction, which Lakatos subsequently endorses (e.g., Lakatos, 1978, p. 192), would not be forced to deny that Brownian motion was novel. For it presumably did not belong to the problem situation which gener-
ated the molecular-kinetic theory of gases. The gas theory, according to Einstein (1949, pp. 46–49), arose from a different aim, also of interest in this connection:

My major aim in this was to find facts which would guarantee as much as possible the existence of atoms of definite finite size. In the midst of this I discovered that, according to atomistic theory, there would have to be a movement of suspended microscopic particles open to observation, without knowing that observations concerning the Brownian motion were already long familiar (emphasis added).

But what about Einstein’s theory of Brownian motion that Perrin appraised? In one sense it certainly seems that Brownian motion was part of the problem situation that yielded the specific theory that became testable. As mentioned in testing (T2.6), it was to avoid the anomalies for the kinetic theory due to attempting to measure the mean velocity of Brownian particles that Einstein formulated his theory of Brownian motion without any reference to this velocity. So even if construed using Zahar’s redefinition of novel fact, it is not clear that (T2.4) would count experiments on Brownian motion as grounds for appraising the ES theory. Moreover, as Musgrave (1974) argues, Zahar’s account suffers from being too ‘psychologistic’. Worrall’s (1985) reformulation of novelty attempts to avoid this charge by making novelty, not a matter of whether the theory was devised to explain the fact, but whether the fact was used to construct the theory. So long as evidence e was not used to construct theory H, and H entails e, e satisfies Worrall’s criterion of use novelty for testing H. But for Worrall, this is not only a necessary but also a sufficient criterion for e to provide support for H. (See for example, Worrall, 1985, p. 318.) As a result, even if Worrall is right, and determining if this criterion is satisfied is an objective historical matter, the resulting account seems counterindicated by the episode under consideration.

For, if it is granted that estimates of kinetic parameters, such as Avogadro’s number, constituted novel facts in Worrall’s sense (which it must be for Perrin’s experiments to be taken to support the ES theory), successful predictions of these parameters would also seem to have warranted the kinetic theory, long before Perrin’s tests. As Brush (1968, p. 380) notes:

Several independent methods of determining these parameters had been known since 1870 or before, to say nothing of the many successes of kinetic theory in predicting the properties of gases.
Worrall's account would fail to explain why scientists accorded a great deal more weight and importance to Perrin's experiments than to these earlier appraisals. Similarly, since any fact satisfying Worrall's criterion for novelty is on par— all providing 'severe tests' for $H$— his account would render inexplicable both the scientific criticisms of the experiments prior to Perrin's and the pains scientists took in constructing and interpreting their experiments so as to avoid such criticisms. So all of these senses of novelty appear to render (T2.4) a poor description of this episode. However, this episode itself, I think, indicates a more appropriate construal of the requirements for a good test under which (T2.4) accords well with this episode. What it indicates is that what mattered for an adequate appraisal was not an issue of when, why, or how theories were constructed, but rather, of the manner in which a theory was tested. The aim was to deliberately construct or design a test that had the ability to generate evidence $e$ that would be highly surprising (extremely improbable) if (and only if) a given theory were an incorrect account of the origin of $e$. Then the occurrence of such evidence is a good indication that the theory was not an incorrect account of $e$'s origin. This is our (meta-level) notion of a good test now operating on the object-level. Successful fits with kinetic predictions are highly probable given the correctness of the kinetic explanations; but, contrary to the criterion of 'use novelty', this was not a sufficient warrant for their acceptance. Such successful predictions were disparaged (by pro- and anti-atomists) when such good fits (between $H$ and $e$) were seen to be probable even if the kinetic theory did not give the correct causal account (of the origin of phenomena such as Brownian motion). Perrin's experiments avoided just such disparagement.

Consider again Perrin's Step (II), where the predicted value for Avogadro's number was tested against observed estimates. It was not the concordance of estimates itself that mattered; it was that experiments were deliberately designed (using what was learned at Step (I)) so that such concordance would be overwhelmingly surprising (so inexplicable) if there were a real discord between the hypothesized value, $N^*$, and the 'true' value for $N$ (where the 'true' value is the mean of $N$ in a population of experiments). It was this assurance that was lacking in experiments prior to Perrin's.

It should be noted that the rationale for deeming novel predictions as better evidence than non-novel ones, in the existing senses of novelty, runs counter to the major theories of statistical testing favored by
philosophers (Bayesian and Likelihood schools).\textsuperscript{19} For these theories are based on the principle that given the same evidence $e$ and hypothesis $h$, the same evidential weight (that $e$ affords $h$) should be assigned.\textsuperscript{20} If this principle is strictly adhered to, when or how the data or theory is generated cannot matter in assessing the evidential support provided. To the extent that the historical record indicates the importance of novelty (in any of the senses mentioned here), it simultaneously counter-indicates these statistical philosophies.

The same notion of novelty that renders (T2.4) a good description of this episode underlies the importance often attributed to crucial experiments. It provides a good test case for (T2.10):

The appraisal of a theory depends on certain tests regarded as 'crucial' because their outcome permits a clear choice between contending theories.

With regard to causal theories appraised before and after the appraisal of the ES theory, it has already been noted that fully-fledged rival theories may be absent. Nevertheless, (T2.10) seems well indicated, if understood in a statistical sense suggested by the notion of 'novelty' arrived at above. A test is crucial between theories $H$ and $H^1$ to the extent that the test has a high probability of yielding one sort of answer (i.e., one set of experimental outcomes) if $H$ is the case and another set if alternative $H^1$ is. Letting the former set of experimental outcomes be just the ones taken to indicate $H$ (as against $H^1$) gives a test with a very low chance of indicating $H$ erroneously. At the same time the test will have a good chance of detecting $H^1$ if $H$ is not the case. Again, these are just the criteria for a good meta-test defined at the outset.\textsuperscript{21} In the extreme case, the test correctly discriminates between $H$ and $H^1$ with probability one. This corresponds to the strongest type of crucial test: one whose result discriminates unequivocally between $H$ and $H^1$. The constant experimentation to appraise proposed causal factors reflected the importance attached to evidence that would fairly unambiguously discriminate between causal factors, in this sense. A test result may not logically compel rejecting a theory (as Duhem and others point out), but once the only available explanations that could be appealed to are fairly well ruled out, there are excellent grounds for doing so. Granted, Perrin's work does not provide a crucial experiment between kinetic theory and classical thermodynamics taken as whole sets of guiding assumptions in the sense often attributed to Popper (as is
consistent with its counter-indicating (T2.3)). But that does not mean it is not an instance of a crucial experiment in the sense meant by Einstein, Perrin and others.

Mach and Ostwald attacked the molecular-kinetic theory alleging that the theory was not needed, that a phenomenological description such as thermodynamics contains sufficient information, and avoids various problems that plagued atomic theory (e.g., the use of entities deemed hypothetical). Responding to this allegation, Einstein (1926) begins by stressing that the two theories give conflicting predictions about Brownian notion:

If the movement discussed here can actually be observed (together with the laws relating to it that one would expect to find), then classical thermodynamics can no longer be looked upon as applicable with precision to bodies even of dimensions distinguishable in a microscope: an exact determination of actual atomic dimensions is then possible. On the other hand, had the prediction of this movement proved to be incorrect, a weighty argument would be provided against the molecular-kinetic conception of heat (Einstein, 1926, pp. 1-2).

Perrin viewed the aim of his experiments as providing such a crucial test:

I have sought in this direction for crucial experiments that should provide a solid experimental basis from which to attack or defend the Kinetic Theory (Perrin, p. 89).

The kinetic theory, in contrast to the classical theory, views dissolved molecules as differing from suspended particles only in their size; their motion would be the same. If Brownian motion could be explained as caused by something outside the liquid medium or something within the particles themselves, then it would not be in conflict with the classical theory. If, on the other hand, it could be shown that Brownian motion was caused by a molecular motion in the liquid medium, as given in the kinetic theory, then it would be in conflict. Moreover, it would show a statistical process was responsible (and that the Second Law requires a statistical rendering). So Perrin's appraisal does test the ES theory against a non-statistical version of the Second Law of Thermodynamics, while its result is still consistent with the Second Law statistically rendered. Perrin states:

'Thus the practical importance of Carnot's principle for magnitudes and lengths of time on our usual dimensional scale is not affected (p. 107).'

That the kinetic theory indicated a non-statistical version of the Second Law was argued by Boltzmann around 1870. That Brownian motion,
indeed spontaneous, would be an exception to the (non-statistical version of the) Second Law was also recognized before the ES theory was formulated. For, if it is true that without temperature differences in the system, a Brownian particle denser than water rises spontaneously, then it constitutes a case in which part of the heat of the medium is transformed into work. This recognition is explicitly discussed by Gouy around 1890. Poincaré, persuaded by Gouy’s arguments, declares:

... but we see under our eyes now motion transformed into heat by friction, now heat changed inversely into motion and that without loss since the movement lasts forever. This is the contrary of the principle of Carnot (Poincaré, 1904, p. 610).

In what I referred to as Step (I) (see test (T2.6)), Perrin showed that one can generate at will an observable process due to an agitation not attributable to the particles or to external energy sources. Thus, in carrying out Step (I), Perrin demonstrates the existence of violations to the (non-statistical version of) the Second Law. Perrin often describes his experiments as methods for generating such violations:

Briefly, we are going to show that sufficiently careful observation reveals that at every instant, in a mass of fluid, there is an irregular spontaneous agitation which cannot be reconciled with Carnot’s principle except just on the condition of admitting that his principle has the probabilistic character suggested to us by molecular hypotheses (Perrin, 1950, p. 57, emphasis added).

Even ardent anti-atomists (probably excepting Mach) construed Perrin’s experiments as providing this crucial evidence. On the basis of such experimental evidence, even Ostwald came to reverse himself on the atomic-kinetic energy theory in 1909:

I have convinced myself that we have recently come into possession of experimental proof of the discrete or granular nature of matter... this evidence now justifies even the most cautious scientist in speaking of the experimental proof of the atomistic nature of space-filling matter.23

3. SUMMARY OF RESULTS AND IMPLICATIONS FOR THESES ON APPRAISING AND CHANGING GA’s

My test results have indicated (T2.4, T2.5 and T2.10), and counter-indicated (T2.3, T2.6 and T2.8). These results point to the important role of experimentation (as against T2.6) in which novel facts (T2.4) may be deliberately used to construct crucial tests (T2.10) without question-begging assumptions (T2.5). The overall thrust, then, points to the importance of empirical methods. To the extent that accepting the
ES theory is linked to a change in GAs, the results of my testing the above theses on theory appraisal are relevant for testing theses on the importance of empirical methods in appraising and changing GAs. In particular, my results appear to indicate thesis (GA1.1):

_The acceptability of a set of GAs is judged largely on the basis of empirical accuracy._

In asserting (GA1.1), Kuhn correctly notes that the discrepancies between theories and data and between rival theories are typically quite small. I suggest that _that_ is why statistical considerations are important; it is not because scientists in this episode are interested in which theory is most probable or best supported (according to some probabilistic measure) that statistics enters. It is because in order to make what is a small difference, on our measuring scale, big enough to matter, they needed to consider what would result in a large number of experimental outcomes (contrary to (T2.6)). As with the examples Kuhn (1977) gives as illustrations, I believe examples from this episode show how difficult it is to explain away established quantitative anomalies, and to show how much more effective these are than qualitative anomalies in establishing unenforceable scientific crises (Kuhn, 1977, pp. 210–211).

Their being ‘unenforceable’ indicates the concern scientists had with getting the cause of Brownian motion correct. There is no doubt that the dispute involved disagreement over whether mechanical or other (e.g., energeticist) explanations are preferable in science and in corresponding differences in world-views (GAs). But because the final evaluation turned on quantitative experimental results, it counter-indicates Kuhn’s incommensurability claim in (GA4.1):

_During a change in GAs (i.e., a scientific revolution) scientists associated with rival GAs fail to communicate._

As this case shows, through shared methods of experimental design, quantitative analysis was robust across rival GAs. Yet Kuhn rejects probabilistic testing across GAs:

If, as I have already urged, there can be no scientifically or empirically neutral system of language or concepts, then the proposed construction of alternate tests and theories must proceed from within one or another paradigm-based tradition (Kuhn, 1962, p. 146).

So Kuhn’s endorsement of (GA4.1) is based on a positivist assumption that this episode shows to be in error. Members of any of the rival GAs
one cares to identify can be seen to have agreed that the argument first raised by Nageli (based on the popular assumption that the molecules would have to move in a coordinated fashion to cause Brownian motion) posed a real challenge to the kinetic account. Nor did their allegiances prevent them from being convinced that the statistical argument Smoluchowski elucidated answered Nageli’s criticism. Representatives of rival GAs could similarly agree that Perrin’s experiments, if successful, would show Smoluchowski’s statistical argument was instantiated. The arch anti-atomist Ostwald tried out kinetic ideas for his own purposes and, though he was disappointed, what matters for our purpose is that even he had no difficulty in using notions from incompatible GAs.

The appraisal and acceptance of the ES theory of Brownian motion corresponded to a change in beliefs about fundamental entities, about the existence of molecules and in the particulate nature of matter. It also led to a change in scientific methodology; a new limit to experimental accuracy due to Brownian fluctuations and ‘noise’ in measuring systems was introduced. The entry of statistical validity into physics also conflicted with the cherished philosophical conception that physics discovers wholly exact laws. This episode is a good illustration of how empirical tests constrain beliefs in each of these GA components, of how a reasoned appraisal of GAs themselves is possible. For example, in appraising the ES theory, methods of theory assessment were used intersubjectively (i.e., across sets of GAs), although these very methods led to a change in testing methods (by leading to that theory’s acceptance)! Perhaps this illustrates how something like what is suggested in Laudan’s (1984) reticulated model takes place.

4. A POINT ON SYSTEMATIC BIAS IN META-METHODOLOGY

As Laudan et al., (1986, p. 155) note, one of the main points that the theorists here considered share is the rejection of positivistically inspired logics of induction and confirmation. At the same time, however, my tests consistently revealed the theses on theory testing to be rooted in positivist testing models. These models sought to provide quantitative measures of support (confirmation, truth-likeness, etc.) that evidence affords hypotheses. Any complete theory of testing must involve statistical ideas of data generation, modelling, and analysis. But there are
statistical theories of testing very different from those of the positivist models that seem more adequate (e.g., those developed by Giere and other philosophers including the present author, based on current statistical methods). Nevertheless, even if I am correct to allege that the theses on theory testing are based on positivist models, this does not create any bias for the present meta-level testing program. At worst, the theses would misdescribe scientific episodes, and good tests should detect this, as I believe some of my tests did. What is worrisome for the program is positivistic assumptions about experimental methods in carrying out tests of these theses. For they are likely to create systematic bias that will prevent such meta-level tests from being good tests.

Suppose a tester fails to find positivistic views of theory testing playing an important role in a given historical case – which seems likely, given it is generally agreed that such views inadequately capture actual science. A tester who operates with a positivistic conception of experimental testing must take this failure to indicate that what is important is something other than experimental testing. Then, when the historical study does show correlations between the holding of rival theories and various extra-scientific differences (e.g., in subjective beliefs, pragmatic interests, cultural background, etc.), these are assumed to be a major basis of theory appraisal. The danger, then, is that starting out with a positivistic view of theory testing is likely to lead to severe bias in testing theses against historical cases – one which will not be detectable afterwards. Such tests have little if any chance of not pointing to the importance of non-empirical or non-scientific factors. They thereby violate the meta-criteria for a good test I set out at the start.

Acknowledgment: I would like to thank Alan Musgrave for useful suggestion on an earlier draft.

Department of Philosophy,
Patton Hall,
Virginia Polytechnic Institute,
Blacksburg, VA 24061,
U.S.A.
NOTES

1All page number references to Perrin, except where noted, refer to Perrin (1923).
2In addition to being discussed by the theorists considered here, and in discussions
concerning these models, e.g., by Clark, Gardner, and Worrall, it has served as a key
illustration for views held by Cartwright, Hacking, Harman, and Salmon.
3My aim here is to avoid the assumptions that go along with viewing evidence or
instances as providing degrees of support or confirmation for or against assertions.
4This can be expressed more formally by considering two possible results of a test of a
thesis as follows:
Results:
(1) Test T indicates thesis H holds in the test case e [i.e., T indicates H(e)]
(2) Test T counterindicates a thesis H holds in the test case e [i.e., T counterindicates
H(e)].

Then T is a good test of H to the extent that both
(i) Prob (Result (1) given not H(e)) is very low
and
(ii) Prob (Result (2) given H(e)) is very low.

Such probabilistic considerations enable one to talk of degrees of goodness as well.
5In this respect our episode accords with a main thesis of Giere's model of science.

6This quote is from a passage of Nägel's cited and translated by Brush (1968), p. 336.
The source is Nägel (1879), "Über die Bewegungen kleinster Körperchen", Sitzungsber-
ichte der mathematisch-physikalischen Classe der K. Bayerischen Akademie der Wissen-
schaften zu München 9, pp. 389-435.

7An excellent sourcebook detailing these modern experiments is Wex (1954).

8Values obtained for the mean velocity of agitation by attempting following as nearly as
possible the path of a grain, gave the grains a kinetic energy 100,000 times too small.
According to Einstein's theory, mean velocity in an interval t is inversely proportional to
root t; it increases without limit as the time gets smaller. Thus, the meaningless results
were just what would be expected. Brush (1968, p. 369) claims "One can hardly find a
better example in the history of science of the complete failure of experiment and
observation, unguided (until 1905) by theory, to unearth the simple laws governing a
phenomenon." While Einstein's theory did in fact serve that role, it did not require that
theory's acceptance. Moreover, the statistical point that was needed here was already
made in 1854 by Thomson (Lord Kelvin) with respect to the problem of laying the Atlantic
Cable. Thomson defends his theoretical prediction against apparent experimental refu-
tation, by explaining the source of the diversity of measured values of the velocity of
electricity to the time required in making the measurements. Brush (1968, pp. 369-370)
remarks:

Apparently the scientists who attempted to measure the velocity of particles in
Brownian movement later in the nineteenth century had not followed the dispute
about Thomson's law of squares in the electric telegraph problem, and they obtained
a similar collection of wildly varying results (none of them in agreement with [what
would be expected according to the kinetic theory]).
The standard deviation (square root of the variance) is the displacement in the direction of the $x$-axis which a particle experiences on average (root mean square of displacement). The importance of this statement of variance, for the experimental determination of $D$, is that it states that the mean square displacement of a Brownian particle is proportional to the time $t$. This suggests, for example, that a model for Brownian motion is provided by viewing a particle as taking a random walk. Since it has the same chance of being displaced a given amount in the positive or negative direction, on average, after $k$ steps, the displacement would be 0. Occasionally, more steps will be to one direction than to the other, yielding a non-zero total displacement. That the variance is proportional to the time (the number of steps) corresponds to the fact that the more steps taken, the larger the value that this non-zero total displacement can have.

In addition to Einstein, see Parzen (1960) and Wax (1954) for such derivations.

In addition to deriving the statistical distribution in the ES Theory from random walk models, it was analogously derived from the recognition that the displacement of Brownian particles is distributed like the winnings of a gambler who stands to win or lose a fixed amount $x$ with equal probability on each trial. Clear derivations of this distribution occur in Einstein (1926, Chapter V), and Parzen (1960, pp. 374-376).

Perrin had to run a number of separate experiments to justify his use of Stokes's formula, whose applicability to the fluid at hand had been open to question. See, for example, Perrin (1923), p. 129.

Indeed, Popper's approach is seen by some statisticians, such as Oscar Kempthorne, as an embodiment of the Fisherian logic.

$\gamma$ is the viscosity of the fluid, $T$, its absolute temperature, $R$, the gas constant, $a$, the radius of the particles.

Admittedly, because Step (II) tests a predicted value of a parameter such as $N$ against rival values (in a parameter space), it is more like (what later came to be called) a Neyman–Pearson test than the Fisherian test at Step (I). But this still does not involve the sort of fully-fledged rival presumably meant in (T2.8).

Examples of philosophers who identify it as such are Cartwright and Salmon. I discuss Cartwright's and my own treatment of Perrin's experiments as arguments from coincidence in Mayo (1986).

This can often be made quantitatively rigorous: One constructs a test to generate a type of outcome $e$ that would be highly surprising (improbable) if a theory differed from a correct account of some parameter by more than a specifiable amount $\delta$. Then the occurrence of $e$ is a good indication that the theoretical value of the parameter does not differ by more than $\delta$ from the correct value. This was indeed the case for Perrin's test of $N$ at Step (II). See Note 18.

That is, $N$ is itself a parameter (a mean) of a population distribution from which observed estimates of $N$ are random samples. While actual experiments do not give the precise true value of this parameter, they tell us if estimates are discrepant from the true value by certain amounts. They do this by ruling out discrepancies which, with overwhelming probability, could not have produced observed estimates.

Popper stresses this, calling such approaches 'justificationist'. Neyman–Pearson theory contrasts with these other schools on this point, for error probabilities are influenced by the manner of data and hypotheses generation. Interestingly, this is one of the main reasons Neyman–Pearson statistics is criticized by opponents.
BROWNIAN MOTION

20This is expressed formally in the Likelihood Principle.

21This notion of a statistically crucial test seems to be congenial to Laudan’s idea of solving empirical problems approximately. A statistical test might be seen as specifying a rule whereby a given test outcome either deems the theory a solution or not a solution (or a fairly good, or poor one). Then what is provided by a ‘good’ statistical test (i.e., one with sufficiently low error probabilities) are assurances that a theory will not often be deemed a solution to the problem when it is far from being one, and not often be deemed a poor solution when it is close to solving the problem. This also suggests that we can consider a test (statistically) crucial between any two claims for which the test is a good one, according to my statistical criteria.

22Perrin calculates, for example, that to have a better than even chance of seeing a brick weighing a kilogramme suspended by a rope rise to a level by virtue of its Brownian motion, one would have to wait more than $10^{20}$ billion years. So the statistical version involves a classical frequentist interpretation of probability.

23This translated quotation is given in Brush (1968), p. 381. The source is Wilhelm Ostwald (1909) Grundriss der allgemeinen Chemie. Leipzig: Verlag von Wilhelm Engelmann, p. 4. Quotation is from the “Vorbericht”.

24The error in Nageli’s challenge was not experimental but mathematical-statistical. Gouy was first to come close to answering this objection by citing the ‘law of large numbers’. Smoluchowski gave a more rigorous explanation based on a statistical argument of which Nageli and others had been unaware. The argument essentially shows how a gambler can lose a great deal of money, even with an even chance of winning or losing a fixed amount on each trial, provided he plays long enough (See Note 11). Analogously, unlike what Nageli supposed, the jerks of Brownian particles do not need coordinated motion to explain them; with enough hits, the average displacement can be large, even when each hit has an equal chance of moving the particle to the right or left a given amount.

REFERENCES


Gouy, L. (1888), 'Note sur le mouvement brownien', Journal de Physique 7: 561–564.
Harman, G. H. (1965), 'Inference to the Best Explanation', Philosophical Review 74, 88–95.