ARCHIMEDES
New Studies in the History and Philosophy of Science and Technology

Wrong for the Right Reasons

Edited by
Jed Z. Buchwald and
Allan Franklin

Springer

Archimedes

Volume 11

Archimedes

NEW STUDIES IN THE HISTORY AND PHILOSOPHY OF SCIENCE AND TECHNOLOGY

VOLUME 11

Archimedes has three fundamental goals; to further the integration of the histories of science and technology with one another: to investigate the technical, social and practical histories of specific developments in science and technology; and finally, where possible and desirable, to bring the histories of science and technology into closer contact with the philosophy of science. To these ends, each volume will have its own theme and title and will be planned by one or more members of the Advisory Board in consultation with the editor. Although the volumes have specific themes, the series itself will not be limited to one or even to a few particular areas. Its subjects include any of the sciences, ranging from biology through physics, all aspects of technology, broadly construed, as well as historically-engaged philosophy of science or technology. Taken as a whole, *Archimedes* will be of interest to historians, philosophers, and scientists, as well as to those in business and industry who seek to understand how science and industry have come to be so strongly linked.

Wrong for the Right Reasons

Edited by

JED Z. BUCHWALD

California Institute of Technology,
Pasadena, USA

and

ALLAN FRANKLIN

University of Colorado,
Boulder, USA

Springer

A C.I.P. Catalogue record for this book is available from the Library of Congress.

ISBN-10 1-4020-3047-9 (HB) Springer Dordrecht, Berlin, Heidelberg, New York
ISBN-10 1-4020-3048-7 (e-book) Springer Dordrecht, Berlin, Heidelberg, New York
ISBN-13 978-1-4020-3047-5 (HB) Springer Dordrecht, Berlin, Heidelberg, New York
ISBN-13 978-1-4020-3048-2 (e-book) Springer Dordrecht, Berlin, Heidelberg, New York

---

Published by Springer,
P.O. Box 17, 3300 AA Dordrecht, The Netherlands.

Printed on acid-free paper

Printed in the Netherlands.

TABLE OF CONTENTS

LIST OF CONTRIBUTORS

Jed Buchwald
Division of Humanities and Social Sciences
California Institute of Technology
Buchwald@its.caltech.edu

Xiang Chen
Department of Philosophy
California Lutheran University
chenxi@clunet.edu

Allan Franklin
Department of Physics
University of Colorado
Allan.Franklin@Colorado.EDU

Marjorie Grene
Department of Philosophy
Virginia Tech
mgrene@vt.edu

Alexander Jones
Department of Classics
University of Toronto
alexander. jones@utoronto.ca

William Newman
Department of History and Philosophy of Science
Indiana University
wnewman@indiana.edu

Lawrence M. Principe
Department of the History of Science and Technology
and Department of Chemistry
The Johns Hopkins University
Imafp@jhu.edu

Alan Shapiro
School of Physics and Astronomy
University of Minnesota
ashapiro@physics.spa.umn.edu

George Smith
Philosophy Department
Tufts University
gesmith@mit.edu

N. M. Swerdlow
Department of Astronomy and Astrophysics
University of Chicago
nms@oddjob.uchicago.edu

JED Z. BUCHWALD AND ALLAN FRANKLIN

INTRODUCTION: BEYOND DISUNITY AND HISTORICISM

One of the recent trends in the history of science, or "science studies" in general, has been the tendency to regard science as purely local and contextual. Perhaps this was a needed remedy for the universal claims prevalent in much of the late 20th century. Nevertheless we believe that the pendulum has swung too far. To say that everything in science is local and contextual implicitly downplays the importance of understanding the arguments that scientists use, and have used, to support their hypotheses or experimental results in any manner that breaches locality. It also suggests that we should not look for similarities in method between scientific disciplines or within a single discipline at different times. Many contemporary historians of science think that such attempts are doomed to inevitable failure. There would on this account be no point in investigating, for example, the similarities and differences between the experiments that demonstrated the nonconservation of parity in weak interactions and the experiment that decided between three different proposed mechanisms of DNA replication. Although they were performed at the same time, they were in different scientific fields, which on the recent view have nothing worth remarking in common.

The authors of the essays in this collection disagree with the current trend. It seems to us that there is considerable value in critically examining the evidence cited, how it was acquired, and what arguments were given for the correctness of both the data and the conclusions reached. These are, after all, crucial elements in the means by which scientists or natural philosophers have sought to persuade themselves and one another to this or that view. In a series of studies that range in topic from ancient astronomy through alchemy to 20th century theories of radioactive decay, the essays presented here argue that wrong hypotheses and results can be detected and judged, and that doing so has historical virtue. To be sure, the articles examine those episodes in the light of then-contemporary standards, but they also investigate both the evidence and the arguments offered. They show that there is more to a scientific error than merely noting that x, or even a large number of xs, working in a particular locale, just believed that y was wrong. Of course, one might argue that it is hardly of any compelling interest to know that x was right and y wrong, at least for matters long gone. So what? We would answer that it's not a question of handing out report cards on the past, but rather of thoroughly understanding what was done. This, we contend, can be achieved only by means of a mastery of contemporary technique that is sensitive to issues of argument and evidence – and that therefore uncovers apparent *lacunae* and problems. In our view the history of science is not just about causes, whether these be social, cultural or a complex mix of the two. It is also about understanding.

J.Z. Buchwald and A. Franklin (eds.), Wrong for the Right Reasons, 1–16.

To place these studies in their appropriate context, let us look briefly at what some earlier commentators on science had to say about scientific errors. In 1840 the Reverend William Whewell wrote:

> The discovery of general truths from special facts is performed, commonly at least, and more commonly than at first supposed, by the use of a series of suppositions, or *hypotheses*, which are looked at in quick succession, and of which the one which really leads to truth is rapidly detected, and when caught sight of, firmly held, verified, and followed to its consequences . . . it has very often happened in the history of science, that the erroneous hypotheses which preceded the discovery of the truth have been made, not by the discoverer himself, but by his precursors; to whom he owed the service, often an important one in such cases, of exhausting the most tempting forms of error. [Whewell (1840), vol. 2, pp. 207–208]

For Whewell past science was overwhelmingly wrong science, its mistakes having been swept aside by the discovery of truth. Pierre Duhem in 1906 offered a rather different image of mistaken science:

> Since logic does not determine with strict precision the time when an inadequate hypothesis should give way to a more fruitful assumption, and since recognizing this moment belongs to good sense, physicists may hasten this judgment and increase the rapidity of scientific progress by trying consciously to make good sense within themselves more lucid and more vigilant. [Duhem (1974), p. 218]

More than a half century of extensive scientific change separates Whewell's truth from Duhem's good sense. Though Duhem attributed scientific progress to judgment rather than to Whewell's seemingly absolute detection of truth, both agreed that the scientifically less good must eventually give way to the better.

Today, in the early years of the 21st century, Duhem's and Whewell's views on progress would not command wide assent among historians of science; they would, in fact, be thought altogether irrelevant, at best quaint. The rapid spread of historicist conviction, together with the near-universal focus on purely local matters, has coupled to the conviction that nothing of any interest is to be gained by examining what common links might exist between widely-different forms of scientific work. Statements such as Whewell's and Duhem's sound unbearably naïve to many contemporary historians, the product of minds too thoroughly captured by the beliefs of their times.

One inevitable consequence of these views has been to make questions concerning scientific mistakes and errors both utterly boring and entirely contingent on the standards held at a given time. Whether these standards might transcend the moment in an interesting way, whether we might for example ask whether Ptolemy worked with nature in ways not altogether different from, say, Newton would today not be thought an interesting or even a meaningful question. Different times, different beliefs, different sciences. No more need be said.

None of the historians writing for this book doubts that standards change in myriad ways, that judgment and skill are irreducible aspects of scientific work, or even that

'truth', whether in science or anywhere else, is at least awfully hard to come by. Yet every one of the accounts that follow is informed by its author's conviction that meaningful statements can be made about mistakes and errors in science, and that these statements reach beyond the momentary and the local. But what might count as a mistake that transcends the purely local? Is it possible to make assertions about scientific error that will not inevitably fall to the twin bulldozers of historicism and disunity? Can we compare Ptolemy with Newton without being silly?

To do so we have first to draw as clear a distinction as we can between deliberate attempts to mislead – to commit *fraud* – and the situational complexity that makes difficult choices inevitable. Consider, for example, Heinrich Hertz's production in the late 1880s of a printed trilogy that explicitly claimed to represent the actual sequence of experiments and thoughts that eventuated in his discovery of electric waves late in 1887. Careful historical research, as well as the discovery of Hertz's laboratory notebook, show unequivocally that this cannot be correct. That, in fact, Hertz had considerably altered the true course of events in ways that made his path seem to be much more logical and linear than it was [Doncel (1991) and Buchwald (1994), p. 300]. Did Hertz commit fraud? By today's standards it is entirely possible that Hertz would be rather severely criticized for having misled his readers.

But were his actions fraudulent in any truly meaningful sense? Hardly. Hertz did not intend to mislead his readers in order to create in them a false sense of his experimental and logical acumen. He knew that his results were difficult and novel. In writing his trilogy Hertz deliberately chose to lead his readers step by step through unfamiliar territory, to guide them by the hand. This is not fraud; it is good pedagogy. No doubt Hertz might have remarked in a note that he was not *literally* recounting the paths he had followed, but that would in fact have only raised questions about his results, deflecting attention from them to their production. Attention of a kind that would inevitably follow in any case, as others attempted to replicate Hertz's results.

Fraud has no doubt existed in every area of human endeavor that involves persuasion. Science is certainly no exception, since claims made by its practitioners are crafted to convince others. They are in the best sense of the term rhetorical, but they are not *merely* rhetorical. That is, scientific claims do not aim to convince just because their promoters want others to think as they do. The scientific practitioner certainly does want others to hold certain beliefs, and he or she will try to build an appropriately persuasive rhetoric to further belief. Those who successfully persuade "born-again" Americans, apparently in the millions, that they will soon by transported rapturously to heaven in their Chevrolets and Toyotas at the Second Coming certainly do no less. But there is a difference. The scientist knows that any attempt to persuade may come to shipwreck on the shoals of a future observation, experiment or calculation, and that this does not lie altogether under his or her powers of persuasion. Whereas if the Rapture doesn't happen tomorrow, then believers will just wait for another day.

Many contemporary historians and sociologists of science would no doubt argue that scientific persuasion differs only in degree, and not in kind, from its other forms. After all it's always possible to quarrel *ad infinitum* with this or that result. Even a child caught with chocolate on his face and his hand in the cookie jar can always just deny

having eaten one. And sometimes it might even be true. If this were normally the case, then there would be little point to no-cookie rules or to the articles in this collection. But attentive consideration to precisely what arguments were concerned with, and to how they were settled or just disappeared, does not show that scientists for the most part behave either like liars or like mere believers. They behave instead like skillful practitioners of complex systems that are linked to a natural world which is beyond their complete control. Things do not inevitably go the way the scientist expects them to, no matter how much he or she tries to manipulate the outcome. Practitioners who think otherwise, who therefore believe in the possibility of permanent scientific fraud, do not last long – at least, there is no sign of a long-lasting, fraudulent system in the history of science.

Most past science must of course be thought of as retrospectively mistaken, though hardly fraudulent. After all, a great deal of what scientists in the past believed is no longer held to, and even much of the past's experimental practice has long since vanished. The contemporary physicist does not think that an electrified metal sphere and a non-electrified one push one another away after they kiss because balloons of electric matter press outwards against one another by contact. Many 18th and even early 19th century natural philosophers did think just that, and several of them were quite good at providing supporting experimental evidence. Moreover, some of these same experiments involved skills that physicists have long since lost, making their close recreation today a very difficult matter indeed – and an utterly irrelevant one for the modern physicist. Pushy electric atmospheres were *mistakes*, and experiments that supported these mistakes were either badly done (and therefore involved *errors in practice*) or badly understood (which involved *errors in understanding*). Most contemporary physicists would likely say something of the sort when told the historical details.

But what sorts of errors have we to do with here? For clarity we should distinguish at least between those that could not possibly have been known to be mistakes at the time, and those that could have been. And we should distinguish as well between errors in practice and understanding that could be corrected only in later years and those that might have been corrected at the time. As an example, consider a calculation carried out by the English physicist J.J. Thomson in 1881. For reasons having to do with issues within contemporary British field theory, Thomson wished to determine how a moving, charged, conducting sphere interacts with a magnetic field in its environment. During the course of his analysis, Thomson noted that the moving sphere would alter the magnetization in the surrounding medium. That alteration had itself to be taken into account in computing the resultant effects on the sphere – but Thomson forgot to include it. As a result, he found that the force on the sphere includes a factor of half the medium's magnetic permeability which should not be there, had he done the calculation correctly [Buchwald (1985), p. 271]. Here we clearly have to do with poor practice, because Thomson had erred even according to his own terms (and corrected the result years later without comment). This can reasonably be thought of as an *absolute error*, one that transcends the time and locality of its production. A sufficient condition for an absolute error is its having been produced as a result of a failure in

skilled practice. Had Thomson been more skillful in looking after all the factors that he had deployed, then he would not have erred: his mistake was certainly corrigible, and we will call a case like this one of *intrinsic corrigibility*.

Failure in skill is not however a necessary condition, since absolute errors may be produced by people who work correctly and skillfully according to their own lights, but who could reasonably have known that they were in error had they paid closer attention to their contemporaries or been somewhat more careful in their experiments. Here one too easily enters slippery historical ground, because what might appear to be corrigible error may on closer inspection not be so corrigible after all. Moreover, there seem to be comparatively few historical instances of this sort of *extrinsic corrigibility*, at least for moments close in time to the initial work. We can however find examples in which subsequent work by others produces countervailing results that the initial investigators do recognize as showing that something was wrong with their own productions. In fact, this situation is altogether common and even constitutes a hallmark of scientific work. Though the process usually involves a great deal of point and counterpoint, the disputes are often resolved, with one side granting the cogency of the other's arguments. Of course, resolved disputes do not generally concern matters that reach deep into the hearts of competing scientific systems, where consensus is certainly difficult and even rare. But when scientific disputes are settled, we can ask whether the resolution was just the product of local happenstance, or whether it has a more universal character. Can we understand the arguments used by those scientists to resolve the dispute in a historically reasonable way? If so, then we may have license to move beyond the historicist, localizing imperative.

In the kinds of situations that the philocontingentist has in mind, we cannot, and should not, engage in judgments since to do so will inevitably distort historical circumstance. And, certainly, no good historian of science would today argue, e.g., that Maxwellian field theory was in any reasonable sense compellingly superior to Wilhelm Weber's electrically-active particles *circa* 1865, or that Count Rumford's kinetic conception of heat was superior to the French predilection for caloric *circa* 1810. But just as hard cases make bad law, so do extreme situations of this sort make for bad science history. We need a finer touch than broad comparison permits. That, for example, the deeper reaches of Augustin Fresnel's understanding of light *circa* 1820 was not persuasive to an opponent like Jean-Baptiste Biot does not entail that no aspects of Fresnel's scheme were unavoidably compelling even to Biot. As a matter of historical fact, Fresnel was able to produce a quantitative account of diffraction that even determined opponents of his system of wave optics (including Biot himself, Poisson and Laplace) accepted in public. While their acceptance did entail rejection of the belief that rays of light do not influence one another, it did not also entail acceptance of the deeper reaches of wave optics. And anyone who looks at Fresnel's results must, we would argue, come to the very same conclusion that these skeptical judges of his work did. Fresnel engaged in corrigible work on diffraction that critical contemporaries sought ways to reject; having failed to do so, they accepted the results. Here, we can say, an attempt at extrinsic correction failed, with acceptance following rapidly.

The proper subject of this book concerns the consequence of an impoverishment that leaves the historian blind to differences that offer much help in understanding the paths that scientists have taken. If we refuse to consider with due subtlety and care the quality of work done by our subjects then we may very well miss much about that work that seemed to be compelling at the time. Conversely, considerations of quality may lead the historian to wonder just why work that seems to be so rich was not picked up at the time, and this may in turn open an entire field of historical investigation that would be forever closed if we were simply to assume that *x* ignored *y* because of anything *but x*'s justified conviction that *y*'s work was as a matter of fact not good. Of course, a conviction is a state of mind, and mental states are difficult and occasionally impenetrable subjects of inquiry. In which case it might be utterly pointless to ask whether *x* truly thought that *y* was mistaken even though *y*'s work seems to be quite rich, and just what reasoning supported *x*'s belief. If we cannot look into beliefs, then we can hardly look into cognitive reasons for beliefs. We must rather take them as being bound to something entirely different in kind, such as social or cultural circumstance. But there are ways to get behind belief, ways that take the historian into the nitty-gritty work of his or her subjects and that can open questions concerning the justification of belief to answers that go beyond the exclusively social or cultural.

If we do not understand that J.J. Thomson was mistaken in his calculation of the force on a charged sphere moving in a magnetic field then we will not likely understand why it was that subsequent corrections of his calculations, which were usually silent in respect to Thomson's original, passed without comment even by Thomson himself. We will not understand, that is, why belief concerning the inadequacy of Thomson's original calculation was thoroughly justified. However, in order to do this the historian must master the methods deployed by his or her subjects, must enter as fully as possible into their practice, must be able to do very much what they did. This isn't easy, and the payoff might be small; but then again it might be quite significant as well. We must for example work through Fresnel's calculations, and understand what he did in the laboratory, in order also to grasp just how persuasive many of his claims seemed to people like Laplace and Poisson, who were no friends to his overall program. In both of these instances, an overly historicist attitude, or at least one that simply takes words of acceptance or rejection as they stand, might lead one to invent chimerical explanations for silence or agreement, when they are entirely unnecessary. Nor are these at all unique examples; they could easily be multiplied, and have been by historians of physics, astronomy, and mathematics who seek to master the long-gone practices of their subjects. Every one of the studies in this collection attempts to get at scientific practice in essentially this way.

Alexander Jones discusses Ptolemy's use of observational data. He points out that models in the *Almagest* are prior to the particular deductions that lead to them from observations, and moreover that at least part of the observational evidence itself "appears to have been tampered with, or even fabricated, in order to make the deductions work". Jones provides three instances of what he terms Ptolemy's "sharp practice in deriving ostensibly definitive results from refractory data", two from the

Almagest and one from the *Geography*. These examples show Ptolemy either (or both) in equivocal engagement with his predecessors or else skewing a calculation to arrive at a desired outcome. However, Jones argues, specific problems that Ptolemy faced, rather than simple issues of misrepresentation or outright fraud, go far to explain what he was doing. For example, in one instance involving the length and (especially) the constancy of the solar year, Ptolemy appears to have fabricated a putative observation of his own by extrapolating from ones made by his predecessor Hipparchus, in an apparent effort actually to undermine Hipparchus' authority. But things are not quite what they seem. Ptolemy, Jones notes, lacked contemporary observations that could settle these questions. He had however to provide a solar theory that could in principle be derived from observations, even if the latter were fictitious. That, together with computational simplicity, dictated Ptolemy's method here. The issue seems to be one of appropriate pedagogy, much as in the late 19th century case of Hertz, rather than sly misrepresentation.

The second case discussed by Jones involves Mercury's apsidal line. Here Jones detects a much more serious matter, because Ptolemy produces a "small, slow effect that could never be detected from the unmanipulated data". Why did he do so? Again, things are more complicated than they seem. The problem, Jones shows, derived from the necessity that a complete model for each of the planets required deciding what to do with the apsidal line over the long run. Ptolemy knew that procedures for locating the line were not at all robust, and to decide the matter he chose to use a "traditional assumption", namely that the lines must be fixed sidereally, though he in effect disguised the assumption by using a tropical instead of a sidereal reference frame. He was, Jones argues, deploying a "conservative rule of thumb" in order to reach a solution, and this was in fact treated as an "explicit methodological principle by other Greek scientific writers, in particular Hipparchus".

Overall, Jones finds that Ptolemy's work in the *Almagest* and the *Geography* shows that the power of his mathematical technique was limited by deficiencies in observational data or by the complexity of the phenomena itself, that Ptolemy was unwilling to leave loose ends untied, and that he had a tendency to accept his predecessors' parameters. Every one of these characteristics can be found in subsequent centuries as well.

N. M. Swerdlow also discusses Ptolemy, specifically his theory of planetary latitudes. Here we have to do with difficulties raised by Ptolemy's very close reliance on observation for deriving both models and parameters. This resulted in an intricate theory that, Swerdlow shows, was "wrong for the right reasons". Indeed, the models for latitude are the most complex in the *Almagest*, so complex that Ptolemy felt compelled to remark that simplicity in nature may not concur with our preconceptions, that, as Swerdlow remarks, "nature is not necessarily simple according to our way of thinking".

The problem reflected the strength of Ptolemy's empiricism. His derivations for the superior planets required planetary observations at opposition and as near as possible to conjunction, with the center of the epicycle located at each of the latitudinal limits. These conditions occur together infrequently, e.g., once in 30 years for Saturn.

There are great observational problems involved, the observations that Ptolemy used, probably just conventional estimates, were not especially accurate, and his method of derivation was itself highly sensitive to inaccuracies. The inferior planets have entirely different motions in latitude and provided even greater difficulties and models of greater complexity. In his *Planetary Hypotheses* Ptolemy "finally got nearly everything right", Swerdlow notes, by altering the assumptions that he made concerning the constraints on orbital inclination. Ptolemy did not indicate how he had come to the corrected structure, except to say that the process somehow involved "continuous observations".

In their paper on analytical techniques in alchemy, Newman and Principe show how alchemists, concerned with issues of probation and accuracy, devised and marshaled an increasingly large armory of analytical practices over the course of a millennium. In some cases these techniques were used to test the success of specific transmutational trials, but, later, also to discover more about the composition of mixed substances in general. Their point is two-fold. First, to note how even those engaged in a practice so widely seen today as "misguided" were keenly interested in verifying the outcomes of their processes – no matter how futile we might think them – and amending their theoretical conceptions as a result. Even if such techniques did not in and of themselves rule out the possibility of metallic transmutation (which would have amounted to the assertion of a universal negative) the techniques proper increased in sophistication, application, accuracy, and importance over time. Second, to show that the analytical tools recognized as crucial for chemistry from Lavoisier onwards are linked to their steady development through alchemical searches for the secrets of nature. This link argues for a continuity of interests and practices over time – a continuous tradition in the best sense of the term – and also for a set of evaluative criteria (and the desire to implement them) that partly transcend local differences of a social or cultural nature. While each step in these developments must be understood and explored in its proper and complete context, an appreciation of the grander sweep of historical events is possible only with a synthetic approach that allows local contexts, at least temporarily, to recede from the foreground – allowing a glimpse of the forest rather than just the trees.

Marjorie Grene extracts from the controversies surrounding Harvey's discovery of the blood's circulation an account of innovative natural philosophy caught, as it were, in a conservative trap. Whereas many of Harvey's contemporaries were prepared to accept his claim for circulation, many fewer would follow his conceptions of the heart's specific functions in systole and diastole, and his further assertion that blood does not change color in the heart but in the lungs. In fact Harvey's contemporaries tended to see him in these respects as much too radical, whereas on the whole he was thought much too conservative – precisely because Harvey was not a mechanist, retaining as he did an Aristotelian causal structure and elements of a vitalist understanding – unlike Descartes, who retained much traditional medical understanding of the heart's behavior, while nevertheless treating animals as "ingenious machines". Here, then, we have a situation in which the form of natural understanding that, in several respects at least, eventually predominated, namely mechanism, was nevertheless

a clear drag on physiological claims that were also eventually to predominate. Both sides in this story have much to recommend them from the standpoint of later science: Cartesian philosophy for its rejection of final causes; Harvey's physiology for its particular claims, nicely backed by careful observation. One might say that Harvey's theories are retrospectively correct, his philosophy wrong, and the other way around for his contemporaries, but the point is that in both cases there were excellent reasons for holding contrary views. In which case it's perhaps possible to conclude that good philosophy may produce bad science, whereas good science may employ bad philosophy.

Alan Shapiro examines the background to Newton's infamous Experiment 8 in the *Opticks*, which, through its dispersion law, forbid the construction of achromatic lenses. "In analyzing Experiment 8", Shapiro writes, "undue, but quite understandable, attention has been given to the role of Newton's belief in the impossibility of constructing an achromatic lens. Historians have never been able to explain why Newton became convinced of the impossibility . . . ". Shapiro shows that Newton, in his discussion of Experiment 8, was certainly "guilty of falsification" in claiming that he had from observation found no colors when emergent and incident rays are mutually parallel since he had much earlier claimed otherwise. Here we have the interesting case of a false claim that was actually unnecessary since, Shapiro argues, all that Newton needed to establish the point was to show that a compound water–glass prism should be nearly achromatic when the total deviation vanishes. And this followed from the observation, also in the *Opticks*, that a glass prism and one filled with water have the same (or nearly the same) dispersions. There are other issues as well. First of all, it's clear that, though Newton doesn't say so, his experimental claim and the dispersion law that he provides as a conclusion from it are in fact compatible with one another only for small-angled prisms, as Samuel Klingenstierna noted in 1754 (which, however, Newton likely knew, as Shapiro also shows). Second, Newton did not attempt to test his law for anything other than the water–glass pair.

Shapiro sets the entire issue in a different manner from the way Newton did in the *Opticks*. "If", he remarks, "we look at the problem the other way, by moving from the theorem to the experiment, then the problems of non-experiment and vague deduction become tractable". The dispersion law was originally produced without any thought to its implications for achromatism, just because Newton was aiming early on at a mathematical science of colors. What would such a science look like, or, better put, what would having it enable one to do? To have a mathematical science of colors meant, in Shapiro's words, that "if the refraction of a single ray at one angle of incidence is known in any medium, then the refractions of all other rays in that medium may be determined for any angle of incidence without any additional measurements" (provided that the refractions for a given incidence are known for all the rays in any one medium). That was ever Newton's goal, as were attempts to avoid giving any purchase to Hooke's claims concerning color. The latter motive accounts for the fact that in a late and unsent addition to the draft of a letter to Hooke of 1672 Newton had described an Experiment 9 that he had not performed, but that he had (as it turns out, incorrectly, calculated), whose results are diametrically opposed to those

presented in his *Opticks* account of Experiment 8. For Experiment 9 was inimical to Hooke's theory of color, whereas Experiment 8 was not. Of course, Hooke and his theory were very much live issues in 1672, whereas by 1704 neither was.

Newton's work on refraction was driven and governed primarily by theoretical considerations rather than experimental results. Shapiro remarks of Newton's *Optical Lectures* that though he "did of course make some measurements, . . . these (or at least the ones he chose to make public or preserve in his papers) were just sufficient to derive the fundamental parameters for his dispersion model". Nevertheless, the measurements of refraction and dispersion that Newton did make were governed by an acute sense on his part of how best to minimize unavoidable inaccuracies. This he did, first, by setting his prism at minimum deviation, and, second, by then directly measuring the entire length of the spectrum – thereby avoiding the introduction of the systematic errors that could have occurred if Newton had independently measured the indexes at the spectral extremes. "Newton", Shapiro notes, "has grasped the essential virtue of his method of minimum deviation: the errors too are at a minimum." Moreover, Newton would vary each parameter "slightly, while keeping the others fixed, to see how it affects the observed result can be applied experimentally, and this was undoubtedly one of the principal means of estimating errors in the era before formal error analysis had developed."

Here, then, we find a rich stew in which the young Newton was first faced with the need to contravene Hooke's theory, and led thereby through incorrect calculations to claims that would do so, only eventually to completely contradict these very claims in the interests of a mathematical science of colors. And when Newton did measure he did so with an acute understanding of how to minimize what we would today call random and systematic error – this over a century and a quarter before the proper beginnings of error analysis. In recent years several historians have argued that a thorough concern with the vagaries of measurement appeared only with formal error analysis. One colorful and engaging book on the origins of the metric system claims for example that the two main astronomers involved, Méchain and Delambre, could not distinguish, except in a vague and 'intuitive' way, between errors that make instruments inaccurate (and so are systematic) and errors that make them imprecise (and so are random) – that Méchain in fact lacked "a concept of error to help him identify the source of" a conflict between results taken at two different winter locations [Alder (2002), p. 299]. While it was certainly the case that neither Méchain nor anyone else at the time uncovered the specific problem with the device that led to the discrepancy, and that it was only found many years later by an astronomer who was sensitive to instrumental imperfections, this hardly implies that scientists before the 19th century were limited to 'intuition' (whatever that might mean) in their understanding of such matters. To establish a claim of this kind requires more than a singular example of failure; it requires a careful, comparative study of a range of measuring experiments. And in the case of Newton at least it is precisely this very distinction between the random and the systematic that he was not only aware of but careful to control.[1]

George Smith examines Newton's account of motion in resisting media in Book II of the *Principia*. In the *Principia*'s first edition, Smith notes, Newton provides an

erroneous solution to the problem of fluid efflux from a vessel. Using modern analysis, Smith shows that this solution implies that the resistance force due to inertia alone on the front of a moving sphere yields a drag coefficient of 2. On the other hand, Newton's experiments with pendulums in water, air and mercury give a total drag coefficient of 0.7, and, further, that the inertial part of this would also be 0.7. Consequently the combination of erroneous solution with experiment seems to imply that, somehow, 1.3 of the theoretical coefficient is somehow cancelled, with the remainder due entirely to inertia. This cancellation, which has an analog in Newton's own computation, was attributed by him to the action of the fluid on the rear of the sphere, and it enabled him to claim that the inertia of the Cartesian plenum would necessarily affect the paths of comets.

Once Newton was shown that his solution to the efflux problem was flawed, he provided a new one along with experiments involving vertical fall in water, instead of pendulums. These experiments now gave a total drag coefficient of 0.5, as did Newton's new solution for the coefficient due to inertial drag on the front of a moving sphere. On this basis Newton could now attribute all resistance to fluid inertia without requiring any sort of compensation at the rear of the moving sphere. Throughout these calculations Newton had assumed that all resistance effects can be represented by the sum of three discrete terms, each one corresponding to a specific action (*viz.* inertia, viscosity, etc.).

The question that Smith addresses concerns a possible conflict between Newton's well-known aversion to hypotheses and what appears to be his having tailored theory "to match a known experimental result". That is, Newton's three-term sum for resistance is not derived from laws of motion so much as justified by plausibility arguments; even worse, Newton apparently ignored discrepant data that undercut this very assumption. Yet Smith sees a much more complicated affair than Newton's having played fast and loose with data and having abandoned his own methods. Paying close attention to Newton's precise wording, Smith argues that the new vertical-fall experiments were not intended to confirm a theory; rather, the theory (*viz.* the three-term resistance sum) is taken as given. The data are instead used to show that, as a matter of fact, even in air and water resistance reduces effectively to inertia, and that any remaining discrepancies between the inertial term and the measured resistances can therefore be used to evaluate the magnitude of the non-inertial terms. It is not a matter of a tailored confirmation, properly speaking, which might be construed as fitting theory to data; it is instead a matter of discrimination among the magnitudes of terms whose forms are not themselves at issue.

Newton was wrong on both counts: resistance effects due to viscosity and inertia cannot be separated from one another in the way he assumed, and, if estimates of relative contributions are made using modern theory, the inertial contribution does not in general swamp those due to other effects. In which case his argument against the Cartesian plenum has no purchase in his mechanics. But is this a case of bad science? If we put aside the contra-Descartes claim (though this was an extremely important goal for Newton), we may instead see him trying hard to get at the relative contributions of various factors to resistance. Newton did not hide the lack of rigor in

his arguments, and, as a matter of fact, his conclusion that over the range he examined inertia is the predominant factor in resistance is quite true. The problem lies neither in data nor in its manipulation to reach a preconceived result but rather in the assumptions that Newton made regarding the possibility of disaggregating the several resistance effects. Had the data entailed that the inertial term in Newton's expression was not predominant, then it is entirely possible that he would not have used this argument against the Cartesian plenum, though he would have held to his three-sum expression for resistance. For here Newton was faced with an extraordinarily difficult problem, since, Smith remarks, we "still do not have a law for resistance forces of the sort Newton wanted".

It's instructive to compare Newton's work on resistance to what might seem to be an entirely different situation – Fresnel's attempt over a century later to modify his initial theory of diffraction, which was based on the interference of two carefully-chosen rays, to account for empirical discrepancies. Fresnel knew that his assumptions were theoretically questionable, and he never hid the fact. Instead, he took the expressions that resulted from this initial, binary-ray theory as the basis for proceeding further. The discrepancies that new data revealed required Fresnel to increase the complexity of his basic structure by abrogating a basic hypothesis of his earlier physical reasoning, namely that one of the two interfering rays has its origin on the physical diffractor itself; to accommodate the new data he had to shift the ray off the edge into a space where no tangible material object exists. Further data required even more complexity, eventually resulting in the mathematical intricacies of an infinite number of rays and the Fresnel integrals, which can only be computed numerically.

Here we have a situation that differs from Newton on resistance because, unlike Newton's, Fresnel's initial structure does in fact work as a near-enough representation of the complex one that he eventually produced, a fact that was demonstrated at the end of the 19th century by Arnold Sommerfeld using electromagnetic optics, whereas Newton's cannot be obtained by approximation from the modern account. Moreover, although both Newton and Fresnel had begun with what each concluded was an erroneous solution (Newton's for the efflux problem, Fresnel's for diffraction), Newton continued to use his original assumption that resistance terms could be disaggregated, whereas Fresnel complicated his structure. However, in a deeper sense both Newton and Fresnel worked in entirely similar ways. Smith suggests that, in order to understand the depths of Newton's work, we should move away from his critique of the Cartesian plenum to look more closely at how he worked with data and formulae. Similarly, we should move away from Fresnel's concern with the structure and behavior of the ether in relation to matter in order to focus on his moves in the face of discrepant data. And, if we do this, then we can see just how similar his actions were to Newton's. Newton held to his fundamental assumption that resistance involves three summable terms; Fresnel held to his conviction that interference involves a summable set of rays. From that point of view, Fresnel's continuing experiments were intended to evaluate which rays had to be summed, and where they were coming from. Of course, unlike Newton Fresnel did not begin with any such assumption since, like Thomas Young before him, he started with a pair of rays and not an infinite set. Nevertheless, it's quite clear

that once discrepant data emerged Fresnel easily and naturally concluded that, driven by data, he had to widen his choice of rays and their loci. In that sense, both he and Newton were engaged in a process of discovering the relative significance of terms in a set whose basic structure neither of them could actually obtain from prior theory. Fresnel turns out to have been correct in his assumption, and Newton wrong in his, but both worked in entirely similar ways to reach their conclusions.

Xiang Chen considers photometric measurement, beginning with Bouguer in the 18th century, and he concludes with conflicts that arose when a physical comparator, based on thermometric techniques, produced results that differed from those generated by visual comparators during the late 1830s. Richard Potter's visual comparator was developed originally for the practical end of measuring the reflecting quality of metal mirrors. Potter turned his technique to glass, which implicated theoretical issues since Fresnel had produced formulae that permitted the computation of reflecting power for transparent media. Potter obtained results that were uniformly lower that those implied by Fresnel's formulae, which he criticized, eventually becoming perhaps the longest-lasting holdout in Britain against wave optics.

Chen closely examined Potter's experiments, and he discovered that an approximation which Potter had used in calculating the behavior of his device, and which worked well for mirrors placed far away from the observer (as his metal reflectors had been), does not work well for mirrors placed closer in, where Potter had located his glass reflectors. Potter's discrepancies drop to less than half his claims as a result, which might on this score alone have raised substantial questions concerning them – if, that is, someone had been paying enough serious attention to Potter to work through his paper. But no one apparently did, primarily because Potter's critics, including the wave partisans Humphrey Lloyd and Baden Powell (who however misunderstood the technique's requirements), questioned the ability of any intensity-measuring device based on visual judgment to produce accurate results – or, at least, results accurate enough to be used in debates over theory.

The major challenge to Potter's visual method came from James Forbes, who used a photometer based on a thermometric pile. Though Forbes himself found deviations from Fresnel's formulae, these were in the opposite direction from Potter's. Forbes attributed his own deviations from Fresnel to scattered heat, and Potter's to bad work. But the main lessons of Chen's paper concern the range of acceptable accuracy in historical context, for what was good and acceptable, and even desirable instrumentally, for the practical purpose of measuring illumination or even for star categorization did not meet the more demanding tasks of formula-testing. This is of course hardly surprising, but Chen's account further shows that the accuracy of a particular technique may be questioned on grounds that are themselves quite doubtful, since Potter's challengers never did take his work to pieces or provide an acceptable device that avoided Potter's main flaw, the use of the eye. Wrong in a theory-engaged optics laboratory, right in astronomy and the illumination industry, one might say.

Yet it seems reasonably clear from Chen's account that Potter could have been shown that he was mistaken by at least a factor of two in his claim concerning the inadequacy of Fresnel's formulae. This is not however an instance of *intrinsic*

corrigibility, because it's hardly likely that Potter would have caved in, even had this been pointed out, given his later, extended antipathy to wave optics and, especially, its practitioners, whom he felt simply ignored other problems that he pointed out over the years. Moreover, no one ever did provide proper *extrinsic corrigibility* when it mattered, because there just was no way to measure intensities until the entire issue was long dead. Here, then, we have a situation in which one party has a pragmatically-successful device which he applies elsewhere to critique formulae that form part of a new optical system, while the other parties, already committed to the new system, reject the very technique underlying the device as unreliable. Since the only alternative method itself proved inadequate, the debate could not be resolved and eventually just became irrelevant.

The case of Heinrich Hertz examined by Buchwald concerns what occurs when a theory that is altogether wrong in retrospect is applied incorrectly by one scientist, who is then corrected implicitly by another. Hermann von Helmholtz had developed a form of electrodynamics that potentially had a number of novel implications, most of which he eventually decided on the basis of experiment did not occur. Before he had reached a definitive conclusion, however, he envisioned a specific situation that should occur one way according to his new theory, and quite a different way according to others. A decade later he instructed his student Heinrich Hertz to look into ways of testing certain of the theory's implications, and in the analysis that Hertz produced he considered a situation essentially the same as the one that Helmholtz had earlier mentioned. Hertz reached a considerably different conclusion from his mentor's original claim – but he did so by correctly applying Helmholtz's own theory, whereas the master had incorrectly applied it years before. We have here the interesting case of an error within a mistake, with the error having been fixed by a correct application of the mistake.

Turning to a decidedly modern period, Allan Franklin examines the intriguing case of the Konopinski–Uhlenbeck theory of beta decay. Fermi had proposed a Hamiltonian for beta decay whose particular form he justified through relativistic invariance and by analogy with electromagnetism, but the theory allowed a wide field of choices among specific possible mathematical forms. General agreement with the facts held good, but a more detailed examination of decay energy spectra indicated that there were discrepancies. Konopinski and Uhlenbeck proposed a modification to Fermi's Hamiltonian which was specifically targeted to accommodate the discrepancies. As experimental work continued, it was gradually realized that the results of earlier experiments were themselves incorrect, that scattering and energy loss in radioactive sources had distorted the results. Experiments modified to avoid these problems now confirmed the original Fermi theory. In addition, it was realized that an incorrect experiment–theory comparison had been made. Fermi's original theory had been compared to experimental results to which it was not intended to apply. One might expect, given the contemporary emphasis on locality among many historians of science, that Konopinski and Uhlenbeck would have managed their home resources and allies in such a way as to dispute these results. Instead, they pushed calculations into regions where Fermi had

not himself gone, and these calculations supported, as they acknowledged, Fermi's own theory and not theirs. One might say that the two physicists replaced their original analysis with a new one that went deeper when faced with discordant data, and in so doing they substituted a more elaborate computation for their original modification to Fermi's Hamiltonian. And that is little different from Fresnel's method in diffraction, or Newton's in fluid resistance. Like Fresnel, Konopinski and Uhlenbeck altered their original theory as a result of experiment; and, like both Fresnel and Newton, Konopinski and Uhlenbeck can be seen as using new data to choose among the possible terms in a formula. Of course, unlike either Newton or Fresnel, Konopinski and Uhlenbeck were working with someone else's theory, and so their own expression was already a modification to a basic structure. So where Newton held to his structure, and in the end used data to evaluate the relative significance of its terms, and where Fresnel in effect introduced a more general structure in the face of data that encompassed his original one, Konopinski and Uhlenbeck worked from a modification back to the original formula once data became available that indicated the original experiments were faulty, and that new data supported it. Yet in all three cases the method of working with data and formulae scarcely differed; neither do the resulting assertions on the part of the scientists look strikingly different in kind from one another.

California Institute of Technology, USA
and
University of Colorado, USA

NOTES

[1] Nevertheless, it's only fair to say that what Méchain failed to see was that his measuring device (the Borda repeating circle) might develop new systematic errors over time. Newton's optical experiments raised no comparable issue, and neither would most measuring experiments except for those that use one and the same mechanical apparatus over time or in different geographic locations. From Alder's own account of this intriguing story, it seems more likely that Méchain failed to suspect problems with his device because it was widely considered to be so extraordinarily well-built by a famous maker of instruments that to suspect it would have meant challenging not only Borda but also the superiority of French apparatus in comparison to the English Ramsden theodolite during a period of profound cultural tension and war between the two countries. When, many years later, the issue was re-examined by a French astronomer neither of these factors remained.

REFERENCES

Alder, K. (2002). *The Measure of All Things. The Seven-Year Odyssey and Hidden Error that Transformed the World*. New York: The Free Press.

Buchwald, J. Z. (1985). *From Maxwell to Microphysics: Aspects of Electromagnetic Theory in the Last Quarter of the Nineteenth Century*. Chicago: The University of Chicago Press.

Buchwald, J. Z. (1994). *The Creation of Scientific Effects: Heinrich Hertz and Electric Waves*. Chicago: The University of Chicago Press.

Doncel, M. G. (1991). "On the process of Hertz's conversion to Hertzian waves." *Archive for History of Exact Sciences* **41**: 1–27.
Duhem, P. M. M. (1974). *The Aim and Structure of Physical Theory*. New York: Atheneum.
Whewell, W. (1840). *The Philosophy of the Inductive Sciences, Founded upon their History*. London: J. W. Parker.

ALEXANDER JONES

"IN ORDER THAT WE SHOULD NOT OURSELVES APPEAR TO BE ADJUSTING OUR ESTIMATES . . . TO MAKE THEM FIT SOME PREDETERMINED AMOUNT"

Suspicions that something is not quite right about Ptolemy can be traced back to the third century A.D., within a few decades of his lifetime. The astrologers, though happy enough to use his astronomical tables, were not convinced by his observational demonstrations of the theory of precession, and so they introduced a systematic correction of all computed positions of the heavenly bodies and cardinal points. An early commentator or critic, Artemidorus, insinuated that Ptolemy's armillary sphere was too small to justify some of his innovations in lunar theory, and also accused him of inconsistency in his use of uncorrected and corrected mean motions. Porphyry alleges that much in Ptolemy's *Harmonics* was taken over from the writings of his predecessors (a practice that Porphyry regards as unavoidable), but without due credit (which evidently troubles Porphyry).[1] What is most interesting is not that regarding at least two of these specific charges Ptolemy knew better than the cavillers, but that the initial reception of his scientific work was not wholly uncritical. We would never have guessed this from the respectful, voluminous, and dull commentaries of the fourth century and after.

In modern times, discussions of Ptolemy's *bona fides* have focused almost exclusively on his astronomical work, and in particular on the *Almagest*, the treatise in which he attempts to establish a theory of the celestial motions on empirical and rational foundations. The *Almagest* contains more logically interconnected quantitative data than any other work of ancient science, and close study of the mathematical details reveals over and over again that Ptolemy cannot have originally arrived at his theoretical results by the deductive routes that he presents. To put the matter more strongly: not only do the quantified models of the *Almagest* turn out to be prior to the particular deductions that ostensibly lead from the observations to the models, but also at least part of the observational evidence cited in the *Almagest* appears to have been tampered with, or even fabricated, in order to make the deductions work.

In the present article, I am going to examine three specific instances of what one might call Ptolemy's sharp practice in deriving ostensibly definitive results from refractory data. The first two, taken from the *Almagest*, illustrate contrasting approaches. Ptolemy's treatment of the length of the tropical year involves, as is well known, a set of fabricated observation reports that were designed to provide a specious empirical demonstration of a parameter that Ptolemy took over from his predecessor Hipparchus;

J.Z. Buchwald and A. Franklin (eds.), Wrong for the Right Reasons, 17–39.

what is less well known is that Ptolemy also slants his presentation of Hipparchus' work so that Hipparchus falsely appears to endorse this parameter unreservedly. In his investigation of the long-term behavior of the apsidal line of the planets, on the other hand, Ptolemy seems not to have had to address the theoretical work of his predecessors, but since the effect he was in pursuit of was too small to verify, he must force the result that he wants out of ancient observation reports. My third example, from the *Geography*, combines features of the other two: equivocal engagement with the writings of his predecessors, and skewing his calculations to arrive at the desired outcome. But this time, Ptolemy seems to intend that his reader should notice that a trick is being played.

THE LENGTH OF THE TROPICAL YEAR

In *Almagest* 3.1 Ptolemy settles two points: that the number of days in a tropical year, that is, a year reckoned from one solstice or equinox to the next solstice or equinox of the same kind, is constant, and that this constant is 365;14,48 days.[2] The plan of the chapter is as follows:

(A) The length of the solar has been a subject of uncertainty for earlier writers, including Hipparchus. The natural definition of a solar year is the tropical year, since this interval is defined by the sun's own motion without reference to any other heavenly body. Hipparchus, however, suspected that the tropical year is not a constant. Ptolemy argues for a constant tropical year on the following grounds:
  (1) He has observed solstices and equinoxes in successive years, finding no deviations from a $365\frac{1}{4}$ day year beyond what can be attributed to instrumental error.
  (2) Hipparchus' suspicions can be explained as arising from errors associated with the observations he used:
    (a) In *On the displacement of the solstitial and equinoctial points* he first presented observations made by himself and by Archimedes of successive summer and winter solstices, which showed variations from a $365\frac{1}{4}$ day year. He conceded, however, in a passage that Ptolemy quotes that errors of $\frac{1}{4}$ day could arise in these solstices.
    (b) The quotation continues by asserting that discrepancies smaller than $\frac{1}{4}$ can be detected from observations of equinoxes using an equatorial shadow-ring such as the one in the Square Stoa at Alexandria.
    (c) According to Ptolemy (who is no longer quoting), Hipparchus next set out ostensibly accurate observations of autumnal equinoxes for the intervals 162–158 and 147–143 B.C. These are reported to $\frac{1}{4}$-day precision (noon, sunset, midnight, or sunrise) and three of the intervals are $\frac{1}{4}$ day short of the corresponding number of $365\frac{1}{4}$ day years. Hipparchus also set out similar observations of vernal equinoxes for the interval 146–128 B.C., which are all consistent

with the $365\frac{1}{4}$ day year. For the first of these Hipparchus also cited a second observation on the ring at Alexandria which was five hours later than the observation in the consistent series. Ptolemy remarks that errors in equinox observations could reach or even exceed $\frac{1}{4}$ day because of an imperfectly aligned instrument. He mentions in particular two equatorial rings that were in the Palaestra of Alexandria in his time, which sometimes exhibited more than one change of shadow direction on the day of equinox.

(d) Hipparchus used observations of lunar eclipses to determine the elongation of the star Spica from the autumnal equinoctial point, and found that this elongation varied. For example, he measured the interval as $6\frac{1}{2}^\circ$ in 146 B.C., and $5\frac{1}{4}^\circ$ in 135 B.C. Since he did not consider it possible that a star could move this quickly, he inferred that the equinoctial point must have shifted. Ptolemy objects that the calculation of the star's longitude relative to the equinoctial point depended on Hipparchus' observations of the vernal equinoxes in those years, which were consistent with the $365\frac{1}{4}$ day year, as well as on other elements that were subject to error. Hipparchus could not legitimately adopt the equinox observations as accurate and then use them as the basis for concluding that the equinoctial point had moved eastward $1\frac{1}{4}^\circ$ in 11 years (which would imply that the interval between the equinoxes should have been about $1\frac{1}{4}$ days in excess of 11 times $365\frac{1}{4}$ days).

(e) Hipparchus likely was not convinced by his own arguments for a variable tropical year, but reported the observational evidence because of his "love of truth." In his solar and lunar models he hypothesized a single solar anomaly with the tropical year as its period.

(3) An eclipse theory based on the hypothesis of a constant tropical year gives rise to eclipse predictions that do not significantly differ from observations; yet an error of 1° in the calculated solar longitude would result in an error of about two hours in the time of mid-eclipse.

(4) From all these considerations, and from Ptolemy's own observations of successive courses of the sun, he concludes that the tropical year is constant.

(5) Finally, he declares the principle that the phenomena should be explained by the simplest possible model that does not lead to significant discrepancies with observations.

(B) Ptolemy now turns to the specific value of the length of the tropical year:

(1) Hipparchus' demonstrations had shown that the tropical year is less than $365\frac{1}{4}$ day but failed to determine the size of the shortfall.

(2) The shortfall is small, and cannot be measured exactly. The best way to determine an approximate value, as for all the periodic revolutions in the celestial models, is to compare pairs of observations separated by the greatest possible number of periods, since the inaccuracy of the

parameter is less when the observational error in each observation is spread over a larger interval. One should not claim that such parameters are accurate over an interval much longer than that over which the observations have been made.

(3) The oldest solar observations are the summer solstices observed by the "circle of Meton and Euctemon" and subsequently by the "circle of Aristarchus."[3] Solstices are, however, difficult to observe accurately, and these particular ones seem to have been conducted comparatively crudely.

(4) Ptolemy therefore chooses to compare pairs of equinox observations, using observations selected for their particular accuracy from among Hipparchus' and Ptolemy's own. These are Hipparchus' observations of the autumnal equinox in 147 and the vernal equinox in 146 B.C., and Ptolemy's observations of the corresponding equinoxes in A.D. 139 and 140. In each 285-year interval Ptolemy finds an accumulated shortfall of $^{19}/_{20}$ day compared to $365\,^{1}/_{4}$ day years, so that the shortfall per year is $^{1}/_{300}$ day, i.e., the tropical year is 365;14,48 days.

(5) Ptolemy also compares the summer solstice observed by the "circle of Meton and Euctemon" in 432 B.C. with a summer solstice that he observed in A.D. 140. Here he finds an accumulated shortfall of $1\,^{11}/_{12}$ days compared to $365\,^{1}/_{4}$ day years over the 571 year interval, which leads very nearly to the same shortfall of $^{1}/_{300}$ day per year.

(6) He asserts that the same result follows from "a number of other observations of our own."

(7) He quotes several statements in Hipparchus' writings that are in agreement with this result:

  (a) In *On the length of the year* Hipparchus compared the summer solstice observed by Aristarchus in 280 B.C. with one observed by himself in 135 B.C. and found a shortfall of $^{1}/_{2}$ day compared to $365\,^{1}/_{4}$ day years over the 145-year interval.

  (b) In *On intercalary months and days* he stated that he had found that the year is $^{1}/_{300}$ day shorter than $365\,^{1}/_{4}$ days.

  (c) In his list of his own writings he states that in *On the length of the year* he established that the tropical year is $^{1}/_{300}$ day less than $365\,^{1}/_{4}$ days.

The remainder of the chapter concerns the laying out of a mean motion table for solar longitude based on this parameter.

As this summary shows, *Almagest* 3.1 is a densely argumentative chapter, the most so in the entire work. Yet on the face of it the issues do not look as if they should be hard to settle. If, as Ptolemy maintains, the tropical year is constant, this should become apparent from consistent measurements obtained from several widely spaced pairs of equinox and solstice observations. The three comparisons in (B5) and (B6) should suffice to establish this consistency. Why, then, does Ptolemy take so long to get to

them, and, more particularly, why is Hipparchus so prominent in this chapter, first as a target for refutation, then as a supporting authority?

The question is entwined with the authenticity of Ptolemy's equinox and solstice observations. Ptolemy says (3.1, Heiberg 1.203; Toomer 137) that these observations were made using the meridian instruments described in 1.12, which measure the angles of noon shadows. He characterizes these observations as "very precise" (3.4, Heiberg 1.233–234; Toomer 154) and made "with the greatest accuracy" and "very securely" (3.1, Heiberg 1.203–204; Toomer 137–138). The events are reported to a precision of one hour – how could one do this using noon shadows? – and all are in agreement to this precision with the dates and times that Ptolemy would have obtained if he counted off the relevant number of years of 365;14,48 days from the Hipparchian and Metonic observations with which they are paired. Also, each of Ptolemy's observations is approximately one day later than the actual moment of equinox or solstice. By way of contrast the Hipparchian equinoxes cited in 3.1, which have a precision of $^1/_4$ day, are with a single exception accurate within 10 hours. A consistent error of one day in the same direction in both kinds of equinox as well as summer solstices cannot be explained as the result of instrumental misalignment or refraction, even if Ptolemy actually used an equatorial ring instead of a meridian instrument to observe the equinoxes.[4] Since Delambre it has been widely supposed that Ptolemy fabricated his solar observations by extrapolating from Hipparchus' observations, using the very value for the length of the tropical year that these observations are supposed to establish. The conclusion seems inescapable.[5]

Ptolemy, then, did not – and presumably could not – measure the length of the tropical year from his own observations. If we read again his presentation of this topic in the second part of *Almagest* 3.1, rejecting the ostensibly crucial demonstrations in (B4–6), the citations of Hipparchus in (B7) no longer look like supporting testimony but appear to become the actual basis for equating the tropical year with 365;14,48 days. It seems as if Ptolemy simply took over this parameter from Hipparchus, and (if we choose to be cynical) we can interpret his criticisms of Hipparchus in (A2) and (B1) as an effort to undermine Hipparchus' authority so that Ptolemy can take the credit for truly establishing the length of the year. But a closer look at the evidence for Hipparchus' works on solar theory suggests that things would not have been that simple for Ptolemy.

Hipparchus discussed observational evidence for the length of the tropical year in two works: *On the displacement of the solstitial and equinoctial points* (henceforth abbreviated to *Displacement*) and *On the length of the year* (henceforth *Length*).[6] According to *Almagest* 7.2 (Heiberg 2:17–18; Toomer 329), *Displacement* was the earlier work; and we have seen already from *Almagest* 3.1 (A2c) that it cited observations as late as 128 B.C. In *Displacement* Hipparchus investigated the longitudinal intervals between certain fixed stars and the equinoctial points, detecting both the short-term variations that Ptolemy dismisses in (A2d) and the long-term precessional shift. Hipparchus' calculations of these longitudinal intervals for his own time evidently depended fairly directly on his own equinox observations, which are listed by Ptolemy in (A2c), but to calculate the intervals for the time of Timocharis, who

observed roughly 150 years earlier, he probably had to extrapolate from his own equinoxes using some provisional value for the tropical year. The *Almagest* does not tell us what that provisional value was.

In *Length*, Hipparchus attempted to measure the length of the tropical year directly by comparing widely separated observations of solstices. Ptolemy's quotation from this work in (B7a) pertains to one of these comparisons, between Hipparchus' summer solstice observation of 135 B.C. and Aristarchus' of 280 B.C. Ptolemy does not give us the precise dates and times of these observations, although the quotation shows that the interval between them was $52{,}960\frac{3}{4}$ days. This would imply a tropical year of approximately 365;14,47,35 days, which (as Ptolemy remarks) is in good agreement with the value 365;14,48.

Ptolemy leaves us two hints that Hipparchus also brought the Metonic solstice of 432 B.C. into play. One hint is that in (B5) Ptolemy calculates the interval in years between the Metonic solstice and his own of A.D. 140 by adding together 152 years from the Metonic to the Aristarchan solstice and 419 years from the Aristarchian solstice to Ptolemy's; Ptolemy tells us that Hipparchus was responsible for determining that the former interval comprised 152 years. The other hint is in *Almagest* 7.2 (Heiberg 2:15–16; Toomer 328), where Ptolemy quotes a sentence from *Length* in which Hipparchus works out how far the equinoctial points move in 300 years relative to the stars on the hypothesis of a precessional motion of "not less than" $\frac{1}{100}^\circ$ per year; this 300-year interval can be explained as a rounding of the interval of 297 years between the Metonic and Hipparchian summer solstices. (Did Hipparchus have some sort of stellar observations from Meton's time to compare with those of his own time to check for this three degree shift?) But Ptolemy does not refer directly to any use made by Hipparchus of the Metonic solstice.

As it happens, we have independent evidence relating to this question. First, we have other ancient sources that report the date of the Metonic solstice. Ptolemy's version of the date is "the archon-year of Apseudes at Athens, Phamenoth 21 according to the Egyptian calendar." But an Egyptian calendar date cannot be an original feature of the report; and both Diodorus and one of the fragmentary parapegmata inscriptions from Miletus give us an Athenian lunar calendar date (Skirophorion 13) to go with the Athenian lunar archon-year. The Miletus parapegma, which is from a little after 109 B.C. (and hence also after Hipparchus), also has Ptolemy's Egyptian-calendar date.[7]

The Egyptian date must have been determined by someone working later than Meton but not later than Hipparchus (it might have been Hipparchus himself), in one of three possible ways:

(1) The date Skirophorion could have been taken as pertaining to a schematic lunar calendar modeled on the Athenian calendar, e.g., the Callippic calendar or its putative predecessor, the Metonic calendar.[8] This calendar could then have been correlated with the Egyptian calendar – exactly, if the schematic calendar had a rule to determine which months were hollow (29-day) and which were full (30-day).

(2) The Athenian date could have been taken as pertaining to an observational lunar calendar. Using astronomical theory one could calculate the Egyptian date on which the first day of this lunar month should have fallen.
(3) One could ignore the reported Athenian date, and use astronomical theory to calculate the Egyptian date on which the solstice should have been observed.

We cannot now be sure whether the equation Skirophorion 13 = Phamenoth 21 is correct, because, aside from the unsettled question of whether Meton's date followed the actual civil calendar of Athens or his own regulated version of the calendar, we do not even know the answers to such basic questions as whether the Athenian day began at sunrise or sunset, and whether the Athenian month began with the sighting new moon crescent or with an earlier phase, say with the first morning when the waning moon could no longer be seen.[9] Nor do we know how Meton determined that a solstice had occurred, or how accurate his method was. For our present purposes it is not important whether Meton really observed the summer solstice as occurring about daybreak of 432 B.C. June 27 = Phamenoth 21; what does matter is that whoever established the Egyptian date was almost certainly also unsure of this result. The only circumstances under which he can have been sure of the date equation are if he had used method (1) and he *knew* that Meton really employed the same schematic calendar assumed by that method.

The second piece of evidence is an astronomical tablet from Babylon, ACT No. 210 (BM 55555 + 55562). This tablet contains an assortment of data pertaining to solar and lunar theory, including (lines 11–12) the statement, "[1,4]9,34;25,27,18 days of 18 years of the sun returning [to] its [longitude] in 18 rotations" (the restorations in brackets are secure).[10] As Neugebauer noted, this statement implies that one longitudinal (i.e., for the Babylonians, ostensibly sidereal) year equals 365;14,44,51 days. In Babylonian astronomy the fundamental unit of time is the lunar month, not the day, and it is highly unusual to find a parameter for the length of the year expressed in days. Moreover, this parameter is rather small for a sidereal year; indeed it looks rather like a tropical year. It is expressed to a precision of $^1/_{216000}$. Any number having this precision can be generated, with rounding or truncation in the third sexagesimal place, by dividing some whole number of days d by a whole number of years y, where the smallest possible y is generally less than a thousand. The existence of such a fractional approximation does not, of course, in itself assure or even make it probable that the parameter was derived by dividing d by y.

However, Rawlins has discovered that for the parameter 365;14,44,51, the approximate fractional representation with the smallest d and y is 108478/297, where $y = 297$ is precisely the number of years between the Metonic solstice of 432 B.C. and Hipparchus' solstice of 135 B.C., which he is known to have used in *Length* in conjunction with the Aristarchan solstice of 280 B.C. It is very unlikely that this is an accidental coincidence; we have extremely few dates of solar observations of any kind from the period up to the end of the cuneiform record (first century A.D.), and we know that Hipparchus discussed the Metonic solstice somewhere, probably in *Length*.[11]

If $y = 297$ represents the number of years between the Metonic and Hipparchian solstices, then $d = 108478$ must represent the interval between them in days. This is a whole number, but probably should be interpreted as precise to the quarter-day, since that is the precision that Hipparchus assumed for solstices. Hence, as Rawlins shows, Hipparchus' solstice observation turns out to have been 135 B.C. June 26, about daybreak; and from Ptolemy's quotation of *Length* (B7a), he also obtains the date of the Aristarchian solstice observation, 280 B.C. June 26, about midday.[12]

It seems, therefore, that in *Length* Hipparchus made at least two comparisons of widely spaced solstices to determine the length of the tropical year. I would guess that his starting point was the comparison of the solstices of 135 B.C. and 280 B.C. The Aristarchian solstice was apparently reported with an "Athenian" date in the Callippic calendar (Ptolemy mentions its year number in the first Callippic Period), and because of the schematic nature of the Callippic calendar, Hipparchus could reliably calculate the number of days between the Aristarchian solstice and his own as $52960\frac{3}{4}$, to the nearest quarter-day, which is half a day less than 145 years of $365\frac{1}{4}$ days, as Hipparchus remarked in the passage quoted by Ptolemy.

Hence in the 297 years between the Metonic solstice and his own, the shortfall should be about twice as great, i.e., about one whole day, making the expected interval $108478\frac{1}{4}$ days, with an expected margin of error about double the margin of error of the Aristarchian observation. If Hipparchus believed that the Aristarchian observation was fairly accurate (say, good to the nearest $\frac{1}{4}$ day), he may not have thought that he needed prior knowledge of the equivalent of the Metonic date in the Callippic or Egyptian calendars over which he had satisfactory control, because he could now calculate it with a margin of error less than a day. Since his backward projection predicted that the Metonic solstice should have occurred at midnight, whereas it was reported for daybreak, he could, if he wished, apply a small correction to the year length he had obtained from the Aristarchian observation.[13] The parameter 365;14,44,51 of ACT No. 210 might have been explicitly stated in *Length*, or (as Rawlins supposes) it could have been trivially derived from his data by a subsequent reader.

What did Hipparchus conclude in *Length* about the length of the tropical year? Supposing that his treatment of this topic involved only the three solstices of 432, 280, and 135 B.C., he could have arrived at three different values: from the latest two, approximately 365;14,47,35 days (i.e., about $365\frac{1}{4}-\frac{1}{300}$ days); from the earliest two, approximately 365;14,42,14 days (i.e., about $365\frac{1}{4}-\frac{1}{200}$ days); and from the first and last, approximately 365;14,44,51 days (i.e., about $365\frac{1}{4}-\frac{1}{240}$ days). Perhaps he did not commit himself to a specific number; one thing we may be quite sure about is that he did not single out 365;14,48 days as the correct figure, because if he had, Ptolemy would undoubtedly have quoted this rather than the passage discussing the interval between the Aristarchian and Hipparchian solstices, which only approximately and implicitly fits this parameter. On the other hand, in the presumably later work *On intercalary months and days*, which prescribed a small correction to the Callippic calendar (deducting one day from every fourth Callippic cycle, so that the mean calendar year becomes $365\frac{1}{4}-\frac{1}{304}$ days) he stated that the year is $365\frac{1}{4}-\frac{1}{300}$ days,

and in a retrospective summary of his writings he read this parameter back into *Length*.

Let us now return to Ptolemy, and try to imagine his task in constructing a satisfactory solar theory. We may assume that, whatever attempts he made to observe solstices and equinoxes, he knew that he had no observational evidence from his own time that could settle the questions left by Hipparchus concerning the constancy and precise value of the tropical year.[14] On the other hand, he knew that his solar theory would have to be one that could, in principle, be demonstrated from observations, even if those observations were factitious.

The decisive issue would have been the solar anomaly. Hipparchus had hypothesized a simple eccentric circular orbit (or its simple epicyclic equivalent) for the sun, and had determined the size and direction of the orbit's eccentricity from the assumed time intervals $94\frac{1}{2}$ days from vernal equinox to summer solstice and $92\frac{1}{2}$ days from summer solstice to autumnal equinox. If these intervals were obtained from specific observations, as seems likely, they would have been the observations Hipparchus made in 147–146 B.C., since the equinoxes from these years are the only ones separated by the right interval; one may conjecture an observation of a summer solstice not mentioned by Ptolemy, say 146 B.C. June 26 about sunset, to complete the necessary trio.[15] Now, Ptolemy would ask, what has happened in the three centuries since Hipparchus to the solar eccentricity?

Any change in the size of the eccentricity would mean an insufficiency in Hipparchus' simple model. Ptolemy would wish to avoid that unless he had positive evidence for it; and there was none. But there might be a gradual shift in the direction of the eccentricity, analogous to the motion of the moon's eccentricity. We know that in Ptolemy's time there were people who believed in such a motion of the sun's apsidal line, and even made tables based on this model.[16] This would show up in a change in the observed lengths of the astronomical seasons. Ptolemy could have tried to fabricate observations fitting a shifted apsidal line while keeping the size of the eccentricity constant, but this would have added to his burden of calculation while making it much harder to dismiss Hipparchus' suspicions of short-term variations in the length of the tropical year (with a moving apsidal line, the tropical year would vary in length). It was much simpler to retain Hipparchus' season lengths, and assume that the tropical year is exactly equal to the period of the sun's anomaly, which makes the tropical year constant.

In (A3) Ptolemy sets out an indirect empirical argument for the constancy of the tropical year. What may be the model for this argument appears in Hipparchus' extant *Commentary on the Phaenomena of Aratus and Eudoxus* (Manitius 90). Hipparchus maintains that the sun cannot have a significant latitudinal deviation from the ecliptic, because if it did, eclipse theories that do that hypothesize such a deviation would predict incorrect magnitudes for lunar eclipses, whereas the discrepancies between observations and predictions are within the negligibly small margin of two eclipse digits. Ptolemy similarly argues that the sun cannot have a significant *longitudinal* deviation separate from the familiar single anomaly that according to hypothesis has the tropical year for its period, because if it did, eclipse theories that omit this

second anomaly would err in the predicted *times* of eclipses, whereas (he implies) the discrepancies do not get as large as the two hours that would correspond to one degree in unaccounted-for anomaly. Two hours is in fact a fairly good estimate of the maximum periodic error in the times of syzygies predicted by Ptolemy's tables.[17] Nevertheless, the consistency of eclipse times is not a sensitive measure of inequalities in the length of the tropical year: it shows that there are no gross variations on the order of one day (Hipparchus himself had set $^{3}/_{4}$ day as the upper limit), but it would not detect smaller effects such as would arise if the tropical year was not the exact period of solar anomaly.

For the actual length of the tropical year Ptolemy turned *faute de mieux* to the parameter that Hipparchus adopted in his latest writings, a parameter that now looks as if it was supported by only a single pair of observations (the solstices of 135 and 280 B.C.). Ptolemy had merely to add a suitable multiple of this year length to the dates of ancient observations in order to find the dates when he *should* have made the observations that would confirm this year length. But there was a further constraint: Ptolemy's fabricated observations should also agree with the Hipparchian season lengths so that his own calculation of the solar eccentricity will agree with Hipparchus' model. This forced Ptolemy to select carefully which ancient observations he should extrapolate from: only Hipparchus' observations of 147–146 B.C. would yield the desired season lengths. Ptolemy chose to construct his observations for 285 years later (A.D. 139–140; it is not clear to me why he chose those particular years, which are close to the end of his known observational career[18]), so that the moments of observation would, according to his year length, fall $^{19}/_{20}$ days earlier than they would have if the tropical year was exactly $365^{1}/_{4}$ days. He rounded this fraction to the nearest number of equinoctial hours, namely 23 hours, and produced his observation reports accordingly.

We have conjectured that Hipparchus made an observation of the summer solstice in 147–146 B.C., $94^{1}/_{2}$ days after the vernal equinox. Suppose for the sake of argument that this was 146 B.C. June 26 about sunset (i.e., about $7^{1}/_{3}$ equinoctial hours past noon); then, adding 285 tropical years, Ptolemy would have calculated that his own solstice observation in A.D. 140 should fall about two equinoctial hours past midnight, which is precisely the time he gives in *Almagest* 3.1. Ptolemy's astronomical spring is thus actually $94^{13}/_{24}$ days, and his summer $92^{11}/_{24}$ days, apparently because he calculated his solstice assuming Hipparchus' was at true sunset, not 6 p.m. Hipparchus legitimately neglected such finesses, but Ptolemy is not really entitled to do so, since he claims to be able to observe to the nearest hour.

But Ptolemy does not compare this solstice of A.D. 140 with the Hipparchian solstice. He has something better up his sleeve: for it turns out that the Metonic solstice also lines up rather nicely with Ptolemy's, providing the appearance of a nearly six-century interval of confirmation for his year length. The fit is not perfect. The Metonic solstice was reported for sunrise, i.e., about $4^{3}/_{4}$ hours past midnight, which would lead using Ptolemy's year length to a time of solstice in A.D. 140 just one hour after midnight. Ptolemy removes this blemish by treating the time of the Metonic solstice as 6 a.m.

Finally, Ptolemy goes out of his way to distract his readers' sight as much as possible from observational evidence that might undermine confidence in his 365;14,48 day tropical year. He suppresses the dates of the solstices of 135 and 280 B.C., since although the interval between them agrees with his year length, they are $^1/_4$ day out of line with his own solstice. He conceals the shorter year length that Hipparchus obtained from the Metonic solstice. He gives only the Egyptian date for the Metonic solstice, perhaps out of a sense that there was something shaky about the Athenian date. The cat only peeps out of the bag in his discussion of Hipparchus' speculations about the variability of the year, when Ptolemy lists Hipparchus' observations of autumnal equinoxes, with their annoying quarter-day slippages out of synchronization. On the whole his tidying up of Hipparchus' legacy is efficient enough so that a reflective reader might by puzzled by his perfectly true statement that Hipparchus had failed to establish the precise amount by which the tropical year is less than $365^1/_4$ days.

THE PRECESSIONAL MOTION OF MERCURY'S APOGEE

Ptolemy locates the apsidal line of Mercury's model by examining observations of the planet when it is at its greatest elongation from its mean position.[19] The principle is that these elongations will be equal for pairs of observations that are symmetrically situated relative to the apsidal line, and Ptolemy assumes – falsely, as it turns out – that the converse is also valid. He gives us two such pairs of observations made by himself. In each pair the elongations are *exactly* equal, and each pair leads to the same apsidal line, with a negligible $^3/_8^\circ$ discrepancy. This is too good to be true. Real observations probably lie behind Ptolemy's choice of apsidal line, but he has surely tampered with them to obtain a fictitious tidiness and consistency.

But then he repeats the procedure with a second set of observations from the third century B.C. in order to show that the apsidal line has a slow shift corresponding to the precessional motion of the fixed stars. Once more he uses two sets of observations, and each set leads him to the same apsidal line with a mere $^1/_6^\circ$ discrepancy, this line being shifted with respect to the one obtained from his own observations by the 4° that corresponds to four centuries of precessional motion. This is again too good to be true. But this time it is a more serious matter. It is one thing to fiddle with one's data, or the analysis of it, so as to make a plausible result come out more cleanly and unambiguously. It is quite another to fiddle with a second analysis of a second set of data so as to demonstrate a small, slow effect that could never be detected from the unmanipulated data. Ptolemy's symmetry method for locating the apsidal line is so sensitive to small errors in the data that it would have been useless for trying to measure shifts as small as 4°; and Ptolemy must have known that. As a matter of fact his apsidal line for Mercury is more than 30° off.[20]

Let us see how he does it. The old observations that he uses came ultimately from two sources: the observers of Babylon who are known more directly to us through the surviving cuneiform tablets of their astronomical archive, and an anonymous observer or set of observers working probably in Egypt.[21] In both cases the observations that Ptolemy uses are dated sightings of a planet relative to one or more stars. A Babylonian

observation would have originally been dated according to the Babylonian lunar calendar, and the planet's distance "above" or "below" the star was given in units called "cubits" and "fingers." The Greco-Egyptian observations are dated according to the "calendar of Dionysius," which is only known to us from these observations in the *Almagest*, and the planet is described sometimes as having passed in front of a star, sometimes as being certain distances expressed in "lunar diameters" north or south or "to the rear" of a star or an imaginary line drawn through two stars.

Ptolemy does not say how he knew that these ancient observations caught the planet at about the moment of greatest elongation. It is just possible, though I think it unlikely, that the Greco-Egyptian observations were intended in the first place to be greatest elongations; more probably, however, they were selected by Ptolemy – or some intermediary? – from a more extensive series of planetary observations. This was certainly the case with the Babylonian stellar passages. Ptolemy might have searched for observations that were made half way between two other observations for which the elongations were equal (a rough way of doing this would be to find observations half way through the period of visibility of the planet as morning or evening star). Alternatively, Ptolemy might have used a provisional version of his planetary model to calculate the approximate dates of greatest elongation, and searched the records for observations falling close to these predicted dates.[22]

Ptolemy therefore has to go through the following steps to render the observations usable: (1) He has to find an equivalent date for each observation in his own chronology, which is based on the Egyptian calendar. (2) He has to deduce Mercury's longitude from the information given in the report, using the longitudes of the stars in his own star catalogue (*Almagest* 7–8). (3) He has to calculate the mean longitude of the sun for the date in question, since it is the difference between this and the planet's longitude that he treats as the greatest elongation. (4) Since he does not find pairs of equal but opposite elongations among these old observations, he uses linear interpolation to find the approximate place where symmetric greatest elongations would be expected to occur.

If we repeat Ptolemy's calculations, we find that the numbers do not always come out the way he says they do. Part of this is due to his practice of rounding the positions to simple fractions, sometimes not the closest ones; in some instances he attributes to Mercury a longitude that is not quite what he should have obtained from the wording of the report. For one observation he gives us not the original report but Hipparchus' reduction of it to a longitude, which might make us suspect that there was something about this one that might have given trouble. I doubt whether he tampered with the reports themselves, a proceeding that would have been dangerous if, as seems likely, Ptolemy had taken them from older books that were generally available. Had he dared to rewrite the observations, there would have been no need to adjust the arithmetic to make them fit.

But some of his Egyptian calendar dates are pretty clearly wrong by one or two days. If we use modern theory to calculate where Mercury really was on the dates Ptolemy gives, we find that for two of the reports its situation relative to the stars does not at all fit the description in the text, but a shift of one or two days makes

the description appropriate. Now it is possible that Ptolemy came by these errors of calendar conversion honestly: they may have been made by someone before him, say Hipparchus, or he may have been mistaken about the workings of the Dionysian calendar.[23] The unsettling fact, however, is that these shiftings of one or two days have an even larger effect on Ptolemy's determination of the apsidal line than his little numerical fudgings, and without them he would not have obtained the result that he does.

The effects of the apparent misdatings and Ptolemy's inexact arithmetic can be easily seen if we recalculate the position of Mercury's apsidal line, accepting the validity of Ptolemy's general method, his stellar positions, and his general interpretations of the observation reports. Ptolemy finds from the first set of three observations that the line of symmetry in the perigeal region of the ecliptic[24] falls at Aries 5;50° (which Ptolemy regards as negligibly different from his "goal" location, Aries 6°). If we accept his dates but recompute all the remaining steps, we obtain Aries 3;9°. Using a date that better fits the second of the three observations, the midpoint shifts significantly, to Aries 5;26°. Again, from the set of three observations that he uses to locate the apogee, Ptolemy finds the line of symmetry at Libra 6°, which is precisely his expected position. Recomputation using his dates leads instead to Libra 7;21°; and with the first of the three observations moved to a more satisfactory date, the apogee leaps to Libra 20;42°.

It is worth remarking that, while shifting the dates of the observations obviously leads to unusable results, even the numerical inexactnesses in Ptolemy's calculations make the difference between an ostensible demonstration that the apsidal line moves at the rate of precession and a more doubtful indication that the apsidal line seems not to be tropically fixed. These discrepancies start out quite small, generally less than a quarter of a degree; and the errors of measurement in the observation reports could be much larger than that, as Ptolemy must have realized. Ptolemy's own tables for Mercury yield significantly different longitudes from the ones he forces out of the reports (by more than a third of a degree in one case).

We can form a plausible guess as to the reason why Ptolemy went to so much trouble to put together a specious proof of the precessional shift of Mercury's apsidal line. If he was to present a complete model for each of the planets, he had to decide what to do with the apsidal line in the long term: was it to remain tropically fixed as in the solar model, or was it to move? Obviously if it moved, the effect had to be gradual. But Ptolemy knew that his procedures for locating the apsidal line were not robust (even if he had no suspicion that the error could be as great as 30°), and hence he also knew that he had absolutely no observational resources for testing for the presence of a slow shift or for measuring it. So what he did was to acquiesce in a traditional assumption. Older models for describing or predicting planetary phenomena, in particular the Babylonian ones, use a sidereal frame of reference for the longitudes, and the zodiacal anomaly, that is the effects that Ptolemy would associate with the eccentricity, are supposed to be sidereally fixed. So Ptolemy has made the conservative choice, though this is disguised by the translation of everything into a tropical frame of reference. Instead of admitting that he was following tradition where there was no evidence to overturn it,

Ptolemy creates the evidence; but he does this only for Mercury, appealing to analogy of models to attribute the same apsidal shift to the remaining four planets.

In passing, it deserves pointing out that the conservative rule of thumb that I conjecture that Ptolemy tacitly followed here was treated as an explicit methodological principle by other Greek scientific writers, in particular Hipparchus. As Hipparchus applied it in his book attacking Eratosthenes' geography, it takes the strong form that the opinion of the older writers on a subject, even if unsupported by empirical arguments, is to be preferred over a newer hypothesis that is supported by faulty empirical arguments.[25] In other words, according to the standards of the scientific community of his time, there was no need for Ptolemy to justify the precession of the planetary apsidal lines by an analysis of observations.

THE LONGITUDINAL EXTENT OF THE KNOWN WORLD

Ptolemy seems to have written the *Geography* after the *Almagest* but before the *Handy Tables*.[26] In it, he sets out to provide the intellectual materials for constructing a map of the part of the world that was known to the Greco-Romans of Ptolemy's time, a region that extended from Britain and the Baltic southward to nearabouts Zanzibar, and from the Canary Islands eastward to central China and southeast Asia. As well as supplying the reader with suitable map projections and numerous items of practical advice, Ptolemy presents a list of about 8,000 localities with their longitudes and latitudes; the map is to be drawn by inscribing and, where appropriate, joining up the localities to form the outlines of coasts, rivers, mountains, and borders. It is, as it were, a digital representation of a map that Ptolemy himself prepared on what must presumably have been a vast scale.

In the introductory part of the *Geography* (1.6 and 1.19) Ptolemy disclaims most of the credit for the geographical data in his map. It is, he tells us, a revision of a map that a certain Marinus of Tyre produced, cleaned up to eliminate certain inconsistencies, illogicalities, and other errors but only intermittently brought up to date in the light of other information that Ptolemy independently obtained. Inspection of his geographical catalogue of places bears this out: for the most part, it reflects the state of the Roman Empire and its neighbours about A.D. 110, which is presumably when Marinus was active, rather than about A.D. 160, which is about when we believe Ptolemy was working on the *Geography*.[27]

Ptolemy makes two large-scale corrections to Marinus' conception of the known world. For Marinus and Ptolemy, the southern limit of the known world was marked in the interior of Africa by a country called Agisymba, and on the continent's east coast by a promontory called Cape Prason. Marinus situated these places on the Tropic of Capricorn; Ptolemy moves them further north, to a little over 16° south of the equator. Marinus' eastern limit is marked in the interior of Asia by the principal city of the "Seres", the "silk people," and on its south coast by a port city called Kattigara and another city slightly inland which was the principal city of the "Sinai." The meridian delimiting the eastern end of the known world, according to Marinus, was 225° east

of the westernmost meridian; Ptolemy reduces this interval to 180°, so that his map wraps exactly half way around the globe.

To find the latitude of the southern limit, Marinus used certain travellers' reports to estimate the distance south from localities of which the latitudes were known. For the interior route, he had accounts of two Romans, Septimius Flaccus and Julius Maternus, who had made expeditions south from Leptis Magna on the north coast of Africa to Garame, and then onwards to Agisymba, "where the Rhinoceros congregate" (*Geography* 1.8 and 1.10). For the part of the route from Leptis to Garame, Marinus had the travellers' statements of how far each stage was; for the rest, he had only the number of months and days of travel. Marinus disregarded the reported distances, however, and simply multiplied the total time in days by a rule-of-thumb estimate of how far one would walk in a day. What he found was that Agisymba was 24,680 stades south of the equator, which is nearly 50° south according to Marinus' assumption that 1° was equivalent to 500 stades. Seeing that this result was unacceptable, Marinus reduced it by about half, putting Agisymba on the Tropic of Capricorn.

Ptolemy agrees with the need for the reduction, but objects to Marinus' halving as capricious. He would prefer to have some sort of astronomical observation to determine Agisymba's latitude, but since of course none was available, Ptolemy employs a climatic argument: similar animals and similar people should flourish in symmetrically situated latitudinal belts north and south of the equator. Since the people of Agisymba have very dark skins and rhinoceros are found, we see how far north of the equator the same conditions apply, and find that the limit is close to Meroe, a place on the upper Nile that was believed to be $16^{5}/_{12}$° north of the equator.

Marinus' handling of the Indian Ocean coast of Africa was similar (*Geography* 1.9). He had the stories of two mariners, Diogenes and Theophilus, who were blown by storms between Aromata at the tip of the Horn of Africa and a place further south called Cape Rhapton; he derived the distance by multiplying the number of days of sail by a rule-of-thumb estimate of a day's sail. For the further interval from Cape Rhapton to Cape Prason, he had the report of a certain Dioscorus that the distance was 5,000 stades. He then projected these distances as if they were oriented north-south from his assumed latitude of Aromata, and ended up with Cape Prason lying 27,800 stades south of the equator, which was even further south than his first estimate of the latitude of Agisymba. Again he corrected this unacceptable result by simply shifting Cape Prason north to the Tropic of Capricorn. Ptolemy does not discuss this case separately, but rather arbitrarily moves Cape Prason north by the same amount as he moved Agisymba north.

The inland route that Marinus used to estimate the distance to the city of the Seres (shown as a broken line in Figure 1) is one of the variants of the Silk Road, which led from the crossing of the northern Euphrates near Hierapolis across Assyria and northern Persia to Bactria, and then to a trading post called the "Stone Tower" in the mountain system west of China (*Geography* 1.11–1.12). He had reported distances in stades all along this route. For the remainder of the route from the Stone Tower to the city of the Seres, he merely had the statement of a merchant named Maes Titianus that the journey took seven months, which Marinus converted to stades in

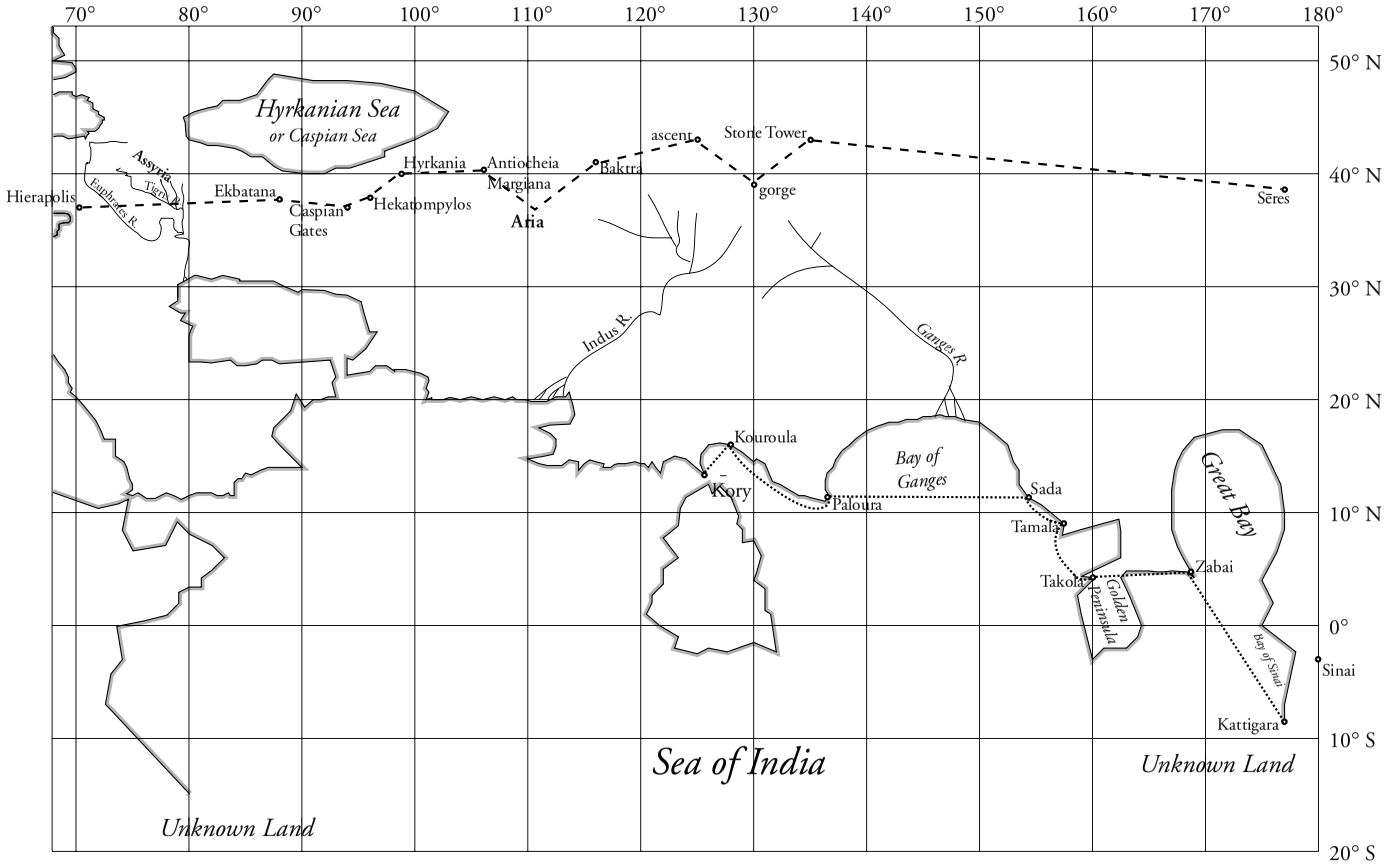

Figure 1. Ptolemy's Asia, showing the inland and coast routes to the eastern limits of the known world.

his usual way. Ptolemy trims a bit off Marinus' calculation of the east–west distance from Hierapolis to the Stone Tower, without justifying the amount but merely pointing out that Marinus had ignored the fact that parts of the route were not due east. The distance from the Stone Tower to the city of the Seres Ptolemy divides by two. He justifies this by arguing that the same factors must have been slowing down travel from an ideal pace on this route as on the route from Garame to Agisymba, if not more so since the weather on the parallel in question would supposedly be harsh. So since Marinus reduced the southward distance to Agisymba by half, he ought to have done the same here. Ptolemy does this, and quite remarkably, when he has added together the unaltered interval from the western limit of the world to Hierapolis, the slightly reduced interval from Hierapolis to the Stone Tower, and the precisely halved interval from the Stone Tower to the city of the Seres, he ends up with a total interval of $177\frac{1}{4}^\circ$ of longitude, that is, within easy rounding distance of 180°.

For the coastal route, again, Marinus starts from a place the longitude of which seemed secure, Cape Kory, on the mainland side of the channel separating India from Sri Lanka (*Geography* 1.13–1.14). Marinus situated Cape Kory a little to the east of the meridian through the source of the Indus, which itself is 125° east of his westernmost parallel the Islands of the Blest (the Canaries). In his analysis Marinos then proceeds along what was evidently a trade route (the solid line in Figure 1) by successive stages, for each of which his source gave a distance in stades. But when he got as far as a place called the Golden Peninsula, which seems to be a shortcut by land across the Malay Peninsula, the final stages to Kattigara no longer had stade figures assigned to them. Instead, his source said that there was a sail of 20 days eastward to Zabai, and a further sail of some days south and a little bit east to Kattigara. It is not clear from Ptolemy's account how Marinus translated this information into a longitudinal interval; all that he tells us is that Marinus questionably interpreted the "some days" characterizing the sail from Zabai to Kattigara as meaning "many days." My guess is that Marinus converted the 20 days from the Golden Peninsula to Zabai into stades in the same way as he did for the distances along the African coast, and neglected the remainder of the voyage because it was directed almost due south. Kattigara was the final port on this Indian Ocean trade route; but it gave access to goods from an inland city, the metropolis of the Sinai, which therefore was a bit further east.

Ptolemy's reworking of the data is a *tour de force* of numerical juggling. To start with, he accepts Marinus' longitude of 125° for Cape Kory, conveniently forgetting that Marinus had said that the cape was actually a bit east of this meridian. In the major part of the trade route, the part consisting of five separate reported stade distances, he deducts from each number in various ways, sometimes because the coastline is not straight, sometimes because the direction is not due east, sometimes merely because he thinks that any actual voyage will be slower than the ideal speed of sail assumed by Marinus. Each reduction is individually justified, and he is consistent in the fractions deducted for each cause. In this way he brings down the interval from Cape Kory to the Golden Peninsula, which Marinus must have calculated as approximately $61\frac{1}{6}^\circ$, to just $34\frac{4}{5}^\circ$. (Thus the Golden Peninsula is $20\frac{1}{5}^\circ$ short of the 180° meridian.)

But for the last part, from the Golden Peninsula *via* Zabai to Kattigara, Ptolemy plays an audacious trick (*Geography* 1.14). The passage in question deserves to be quoted:[28]

> But in order that we should not ourselves appear to be adjusting our estimates of the distances to make them fit some predetermined amount, let us treat the sail from the Golden Peninsula to Kattigara, which comprises twenty days as far as Zabai and "some" more as far as Kattigara, in the same way as we treated the sail from Aromata to Cape Prason, which also consists of the same number of days (twenty) to Rhapta according to Theophilos, plus "many" more to Prason according to Dioskoros. Hence we, too, in Marinos' manner, shall make the expression "some days" equivalent to "many days." Now we have shown from reasonable arguments and on the basis of the [natural] phenomena themselves that Cape Prason is on the parallel that is $16\frac{5}{12}^\circ$ south of the equator, while the parallel through Aromata is $4\frac{1}{4}^\circ$ north of the equator, so that the distance from Aromata to Cape Prason amounts to $20\frac{2}{3}^\circ$. Hence it would be reasonable for us to make the voyage from the Golden Peninsula to Zabai and from thence to Kattigara the same number [of degrees' worth of distance].
>
> There is no need to diminish [the sail] from the Golden Peninsula to Zabai, since it is parallel to the equator because the country in between lies facing the south; but that from Zabai to Kattigara should be reduced to get the direction parallel to the equator, because the sail is toward the south wind and to the east. If we assign half the degrees to each of the intervals, because the difference between them is not clear, and again subtract a third of the $10\frac{1}{3}^\circ$ from Zabai to Kattigara on account of the inclination [from due east], we will get the distance from the Golden Peninsula to Kattigara, as [measured] in the direction parallel to the equator, as approximately $17\frac{1}{6}^\circ$.

In other words, Ptolemy manages to get a longitude of just over 177° for Kattigara (leaving a comfortable 3° for the inland road to the metropolis of the Sinai, which he places right on the meridian 180° east of the Islands of the Blest) out of the absurd hypothesis that we can presume two distances to be equal if both are described by some source or combination of sources as consisting of a sail of 20 days followed by a sail of many days. Ptolemy certainly knows better than to expect his reader to take this kind of argument seriously, and the first sentence quoted above is pure magician's patter.

Not surprisingly, he has indulged in a bit of sleight-of-hand along the way. First, the distance he calculates for the sail from Aromata to Cape Prason is simply the latitudinal interval in degrees, disregarding the considerable addition to the distance resulting from the curving outline of the African coast. Ptolemy has no right to disregard this journey's deviations from due south if he is going to correct for the sail to Kattigara's deviation from due east; if he had calculated the actual distances, he would have ended up with a longitude for Kattigara several degrees east of the 180° meridian. Secondly, in his own coordinate lists for East Africa, Ptolemy assigns Aromata a latitude 6° north of the equator (*Geography* 4.7); yet here he says that it

is $4\frac{1}{4}^\circ$ north. The smaller latitude may in fact have been Marinus'; but more to the point, Ptolemy has tacitly stolen $1\frac{3}{4}^\circ$ that he wants for the interval from Kattigara to the metropolis of the Sinai. Again, his decision to put Zabai halfway along the trip from the Golden Peninsula to Kattigara is disingenuous. Why does he not make the part from the Golden Peninsula to Zabai equal to the 20 days' sail from Aromata to Rhapta, and the part from Zabai to Kattigara equal to the "many days'" sail from Rhapta to Cape Prason? Clearly, because on his map the latitudinal interval from Aromata to Rhapta is almost twice as long as that from Rhapta to Cape Prason, and if he situated Zabai along the same lines, Kattigara would once more be pushed beyond the 180° meridian unless Ptolemy placed it almost due south of Zabai.[29]

This 180° is obviously a preconceived figure, making meridians bounding the known part of the world neatly bisect the globe. Ptolemy has forced his arithmetic to obtain this number, twice, by a succession of fudges. This all looks very like the sort of manipulations we found in the theory of Mercury in the *Almagest*, and that could be paralleled in a number of other passages of that work. But there is a difference. In the *Geography* Ptolemy is quite frank about what he is doing. If he repeatedly objects to Marinus' procedures as arbitrary or careless, he likewise repeatedly draws attention to the inexact, stopgap nature of his own corrections. It is as if Ptolemy wants the reader to see what an inexact thing cartography is.

In both the *Almagest* and *Geography* we can see certain prevailing habits of thought: (1) his belief, expressed in the first chapter of the *Almagest* and in the work's very title ("Mathematical Composition"), that astronomy is "mathematics," that is, a science in which by the application of reason to the evidence of our senses we can achieve unshakable knowledge of unchanging reality; (2) his awareness, displayed in several *obiter dicta* in both treatises, that the power of deductive proof can be limited by deficiencies in the available observations or by the complexity of nature; (3) his unwillingness to leave loose ends, so that every element of the celestial models has a definite size, motion, or position and every locality on the earth has a definite longitude and latitude; (4) a tendency to accept parameters from his predecessors; and (5) a tendency to clean up results, hiding discrepancies, and assimilating numbers to round figures or mathematical patterns.

The first three, I suggest, in combination gave Ptolemy much trouble in the *Almagest*. Ptolemy's plan, and in particular the decision to include predictive tables for the full range of celestial phenomena, required him to have an opinion on practically every aspect of celestial mechanics.[30] The requirements of drawing a map have a similar role in the *Geography*; but the positions of localities on the terrestrial globe are, for Ptolemy, a fit subject for approximation, not demonstrative knowledge, so that he is willing in that work to adjust his standard of evidence to the quality of the data. In the *Almagest*, notwithstanding his intermittent protestations that certain kinds of thing, especially mean motions, cannot be measured exactly, Ptolemy adheres to a standard that he often could appear to meet only by resorting to the kinds of deception we have seen in the passages discussed above.

Department of Classics, University of Toronto, Canada

NOTES

[1] Astrologers' longitudes: Jones (1999), 1:343. Artemidorus: Jones (1990), 28–31. Porphyry: Düring (1932), 4–5.

[2] To represent the sexagesimal (base 60) fractions of ancient astronomy I employ the standard notation in which a semicolon separates the whole number from the fraction, and commas separate sexagesimal "digits."

[3] Ptolemy uses the expression "hoi peri *n*," here translated "the circle of *n*," when speaking of observations associated with a few older Greek astronomers (Meton and Euctemon, Timocharis, Aristarchus, but not Hipparchus). He may intend a certain vagueness about who, precisely, performed the observations in question. The sole specific observation attrbuted to the "circle of Aristarchus" (Heiberg 1.206; Toomer 138) is ascribed unambiguously to Aristarchus himself a page further on (Heiberg 1.206; Toomer 139). Observations of the "circle of Timocharis" (Heiberg 2.18; Toomer 329–330) turn out to include specific ones ascribed to Timocharis and to Aristyllus (Heiberg 2.19–21; Toomer 331). The solstice of the "circle of Meton and Euctemon" (Heiberg 1.203 and 1.205; Toomer 137–138) is also described as of the "circle of Euctemon" (Heiberg 1.206; Toomer 139). Clearly Ptolemy does not use the phrase "hoi peri *n*" as many later Greek writers do, as a periphrasis for "*n*" unqualified; on the other hand, the conventional translation "the school of *n*" misleadingly implies an organized group distinct from the person named.

[4] Britton (1992), 24–37.

[5] Delambre (1817), 1: xxvi and 2: 107–114; Delambre (1819), lxvii–lxix. See also Newton (1977), 87–94, and Britton (1992), 24–37. Evans (1998), 209 suggests as a "less radical hypothesis" that Ptolemy selected out of a larger but discordant body of his own observations those that agreed with Hipparchus' observations and the 365;14,48 day year. This seems rather improbable, particularly given that the three observations in 3.1 are conveniently close to each other in time, facilitating their reuse in the establishment of the solar eccentricity in 3.4.

[6] The most satisfactory survey of Hipparchus' astronomical writings is Toomer (1978).

[7] Toomer (1984), 9–14 gives a useful and accurate brief description of the chronological systems of the observations in the *Almagest*.

[8] On the Callippic calendar see Jones (2000a).

[9] The scholarly literature on the Athenian calendar and on the Metonic and Callippic calendars is vast and cannot be reviewed here. Even the most judicious discussions of Greek calendrics tend to be more confident in their assumptions than the evidence warrants. An admirable exception is Depuydt (1996), which also refers to earlier work. Depuydt (1997), 127 suggests that, contrary to the prevailing hypothesis, the Greek day began at sunrise and the month with first invisibility. Depuydt's model is compatible with the four equations of "Athenian" (Callippic) and Egyptian dates in *Almagest* 7.3, the equation of an Athenian and Egyptian date of the summer solstice of 109 B.C. in the Miletus parapegma fragment, and the assignment of the battle of Arbela, which took place on 331 B.C. October 1 (Sachs and Hunger, 1998, 179) to Boedromion 26 (Plutarch, *Vita Camilli* 19). For the Metonic solstice it suggests the equation Skirophorion 13 = 432 B.C. June 28, but June 27 would be possible if the ideally still visible waning crescent was not seen on the morning of June 15.

[10] Neugebauer (1955), 1:271–273.

[11] Rawlins (1990), 49–51. If the parameter 365;14,44,51 had been found in a classical source, I expect few would question its derivation from the Metonic and Hipparchian solstices. Transmission of such information to Babylon in the late second or early first century B.C. is historically very interesting but not at all implausible. Dates of solstices and equinoxes in Babylonian observational texts were always computed using the 19-year "Uruk scheme," not observed.

[12] Rawlins (1990), 51–53.

[13] I suspect that Hipparchus used this strategy of examining successively longer intervals between observations to check or determine the dates of other observations that were recorded in

calendars over which he had less than perfect independent control, in particular the Babylonian lunar calendar. The alternative, that he possessed a complete and perfectly accurate list of all the Babylonian full and hollow months over a span of more than six centuries, strikes me as implausible.

[14] This means not only that he had not himself made accurate observations, but that, so far as he was aware, no one else had. If there had been a recent tradition of good solar observations, Ptolemy's one-day errors would be inexplicable.

[15] Ptolemy does not say that Hipparchus used specific dated observations (*Almagest* 3.4, Heiberg 1:233; Toomer 153), and the word he uses for the assumed season-lengths is the more general term *phainomena*. It has been speculated that Hipparchus took over his estimate of 187 days for the interval from vernal to autumnal equinox from earlier Greek astronomy (Britton, 1992, 23–24) or derived his estimate of $94\frac{1}{2}$ days from vernal equinox to summer solstice from the Babylonian System A lunar theory (Bowen and Goldstein, 1988, 68–69; asserted as a fact in Gingerich, 1980; cf. Kugler, 1900, 83–87). However, *Almagest* 4.11 shows that Hipparchus adduced specific dated observations when carrying out the analogous determination of the moon's eccentricity; and since he had equinox and solstice observations, it is hard to see why he should not have used them here.

[16] Jones (2000b). Interestingly, the model as it is described in the papyrus PSI XV 1490 (contemporary with Ptolemy) does not use Hipparchus' value for the size of the solar eccentricity. The apsidal line is presumed to advance $\frac{1}{4}^\circ$ per year.

[17] Only part of this error is directly attributable to the periodic errors in Ptolemy's solar model; see Britton (1992, 42–47).

[18] I also do not see the reason why, when Ptolemy calculates an epoch position for the solar model in *Almagest* 3.7, he bases it, not on one of these dates, but on an equinox observation (presumably also fabricated) from A.D. 132, which he describes as "among the first of the equinoxes observed by us, one of the most accurately determined" (Heiberg 1:256; Toomer 168).

[19] For general discussion of Ptolemy's method of determining the apsidal line, see Swerdlow (1989, 56–58 for the specific problem discussed here) and the works cited in his nn. 5 and 7 (p. 59).

[20] Mercury's apogee in Ptolemy's time was near Scorpio 10°, whereas Ptolemy finds Libra 10°. A Greco-Egyptian papyrus horoscope from about A.D. 100 (P. Lond. 130 col. vii 157–162; Neugebauer and van Hoesen, 1959, 22) states that Mercury, at Aries 10° at the time of nativity, was "near perigee" (*perigeios*, in this context not to be understood as the precise perigee); this is diametrically opposite Ptolemy's apogee, suggesting that he was adopting a traditional alignment (van der Waerden, 1988, 293). In the *Canobic Inscription*, composed a few years before the *Almagest*, Ptolemy situated Mercury's apogee for his own epoch at Libra 6°; the change shows that he should have been aware that he was unable to determine the apsidal line accurately enough to test whether it was sidereally or tropically fixed (Hamilton et al., 1987).

[21] The reason for suspecting that the latter observers worked in Egypt is that the epoch year of their "calendar of Dionysius" was 285 B.C., which was also the year from which Ptolemy II Philadelphus counted his regnal years.

[22] The mean difference between the dates of the observations according to Ptolemy and the actual greatest elongation from the mean sun, whether calculated by his models our modern theory, is less than one day. The maximum difference is five days with respect to Ptolemy's models, and three days with respect to modern theory.

[23] We lack exact knowledge of the structure of the Dionysian calendar, for which the only evidence is seven date equations with Egyptian calendar dates in the *Almagest*. A hypothetical reconstruction by Boeckh (1863, 286–340; cf. van der Waerden, 1984) matches six out of seven of these equations. If some of Ptolemy's equations are incorrect by one or two days, Boeckh's reconstruction ceases to be viable; but another plausible reconstruction can be made to fit the revised dates about as well as Boeckh's fits Ptolemy's, as I hope to show elsewhere.

[24] I employ this awkward phrase because Ptolemy's model for Mercury has *two* perigees, each about 120° from the apogee.
[25] Dicks (1960), 64–65 and 80–81 (Strabo 2.1.4 and 2.1.38). Ptolemy takes a similar stance in *Geography* 7.7 and 8.1 on the question of whether the "Sea of India" (Indian Ocean) is landlocked (as he says the earlier writers maintained) or connected to the Atlantic Ocean (as he alleges the more recent cartographers drew it on unscientific grounds).
[26] Berggren and Jones (2000), 17–20.
[27] Berggren and Jones (2000), 23–24.
[28] Berggren and Jones (2000), 76.
[29] When in *Geography* 7.2–7.3 Ptolemy gives the actual coordinates for this part of the South Asian coast, he reduces the longitudinal interval between the Golden Peninsula and Zabai by another $1\frac{1}{6}^\circ$, while keeping Kattigara at 177°.
[30] Practically the only question that Ptolemy refuses to answer in the *Almagest* is the order of the spheres of the planets (*Almagest* 9.1).

REFERENCES

Berggren, J. L. and A. Jones (2000). *Ptolemy's Geography: An Annotated Translation of the Theoretical Chapters*. Princeton: Princeton University Press.
Boeckh, A. (1863). *Ueber die vierjährigen Sonnenkreise der Alten, vorzüglich den Eudoxischen*. Berlin: Georg Reimer.
Bowen, A. C. and B. R. Goldstein (1988). "Meton of Athens and astronomy in the late fifth century B.C." *A Scientific Humanist: Studies in Memory of Abraham Sachs*, eds. E. Leichty, M. deJ. Ellis, and P. Gerardi. Philadelphia: Occasional Publications of the Samuel Noah Kramer Fund 9, pp. 39–81.
Britton, J. P. (1992). *Models and Precision: The Quality of Ptolemy's Observations and Parameters*. New York: Garland.
Delambre, J. B. J. (1817). *Histoire de l'astronomie ancienne*, 2 vols. Paris: Courcier.
Delambre, J. B. J. (1819). *Histoire de l'astronomie du moyen âge*. Paris: Courcier.
Depuydt, L. (1996). "The Egyptian and Athenian dates of Meton's observation of the summer solstice (–431)." *Ancient Society* **27**: 27–45.
Depuydt, L. (1997). "The time of death of Alexander the Great: 11 June 323 B.C. (–322), ca. 4:00–5:00 P.M." *Die Welt des Orients* **28**: 117–135.
Dicks, D. R. (1960). *The Geographical Fragments of Hipparchus*. London: University of London, The Athlone Press.
Düring, I. (1932). *Porphyrios Kommentar zur Harmonielehre des Ptolemaios*. Göteborg: Göteborgs Högskdas Årsskrift 38.
Evans, J. (1998). *The History and Practice of Ancient Astronomy*. New York: Oxford University Press.
Gingerich, O. (1980). "Was Ptolemy a fraud?" *Quarterly Journal of the Royal Astronomical Society* **21**: 253–266.
Hamilton, N. T., N. M. Swerdlow, and G. J. Toomer (1987). "The Canobic Inscription: Ptolemy's earliest work." *From Ancient Omens to Statistical Mechanics*, eds. J. L. Berggren and B. R. Goldstein. Copenhagen: Acta historica scientiarum naturalium et medicinalium 39, pp. 55–73.
Heiberg, J. L. (1898–1903). *Claudii Ptolemaei Opera quae exstant omnia*, Vol. 1 (2 parts): *Syntaxis Mathematica*. Leipzig: Teubner.
Jones, A. (1990). *Ptolemy's First Commentator*. Transactions of the American Philosophical Society 80.7.
Jones, A. (1999). *Astronomical Papyri from Oxyrhynchus*, 2 vols. in 1. Philadelphia: Memoirs of the American Philosophical Society 233.

Jones, A. (2000a). "Calendrica I. New Callippic dates." *Zeitschrift für Papyrologie und Epigraphik* **129**: 141–158.
Jones, A. (2000b). "Studies in the astronomy of the Roman Period IV: solar tables based on a non-Hipparchian model." *Centaurus* **42**: 77–88.
Kugler, F. X. (1900). *Die Babylonische Mondrechnung*. Freiburg: Herder.
Neugebauer, O. (1955). *Astronomical Cuneiform Texts*, 3 vols. London: Lund Humphries.
Neugebauer, O. and H. B. van Hoesen (1959). *Greek Horoscopes*. Philadelphia: Memoirs of the American Philosophical Society 48.
Newton, R. R. (1977). *The Crime of Claudius Ptolemy*. Baltimore: Johns Hopkins University Press.
Rawlins, D. (1990). "Hipparchos' ultimate solar orbit & the Babylonian tropical year." *Dio* **1.1**: 49–66.
Sachs, A. J. and H. Hunger (1998). *Astronomical Diaries and Related Texts from Babylonia*, Vol. 1. Wien: Österreichische Akademie der Wissenschaften, Philosophisch Historische Klasse, Denkschriften 195.
Swerdlow, N. M. (1989). "Ptolemy's theory of the inferior planets." *Journal for the History of Astronomy* **20**: 29–60.
Toomer, G. J. (1978). "Hipparchus." *Dictionary of Scientific Biography* **15**: 207–224.
Toomer, G. J. (1984). *Ptolemy's Almagest*. London: Duckworth.
van der Waerden, B. L. (1984). "Greek astronomical calendars. III. The calendar of dionysios." *Archive for History of Exact Sciences* **29**: 125–130.
van der Waerden, B. L. (1988). *Die Astronomie der Griechen: Eine Einführung*. Darmstadt: Wissenschaftliche Buchgesellschaft.

N. M. SWERDLOW

PTOLEMY'S THEORIES OF THE LATITUDE OF THE PLANETS IN THE *ALMAGEST*, *HANDY TABLES*, AND *PLANETARY HYPOTHESES*

The theory of planetary latitude in Book 13 of the *Almagest* is known, if at all, for its complexity. This has the pleasant result that there is only a small literature on it and that literature is on a high level of technical competence. The same, by the way, is true of latitude theory in general. There are recent expositions by Pedersen and Neugebauer, earlier ones by Delambre and Herz, and a few briefer treatments. Paradoxically, the complexity of Ptolemy's theory is both its strength and its weakness, its strength because he reached it by doing everything right, at least in principle, its weakness because it is ultimately wrong, as was later recognized by Ptolemy himself, who went on to remedy its deficiencies. It is, as we may say, wrong for the right reasons. And since being wrong for the right reasons is more or less the subject of this collection – for is not most interesting older science wrong for the right reasons? – Ptolemy's latitude theory seems quite appropriate. Our object here is to explain the latitude theory, first its original form in the *Almagest*, then its later modifications in the *Handy Tables* and *Planetary Hypotheses*, each of which shows improvements, and to investigate its observational foundation, for it is the observations that are the cause of both its strength and its weakness. It is unusual to find any revisions in the work of an ancient scientist, but in the case of Ptolemy's latitude theory three distinct stages are known, which may be unique, showing that he himself knew something was wrong and twice set out to correct it.

The latitude theory of the *Almagest* is complex because it is so strictly empirical, which is true of all of Ptolemy's mathematical astronomy, and empiricism, we all agree, is a good thing. Every *hypothesis*, a technical term meaning 'model', is either derived or confirmed by observation, and every numerical parameter is derived directly and uniquely from observation. There is, however, a large range of precision in Ptolemy's observations, from positions and times measured to within a few minutes for the derivation of parameters, although their accuracy is more variable, to rough, qualitative observations for demonstrating the applicability of models. The observations upon which Ptolemy founds the theory of latitude fall somewhere in between these, and he uses them to derive both the model and its parameters. As crucial as these observations are, he gives no information about how they were made – he never mentions using an armillary, which could measure latitude – and many could be conventional estimates rather than his own measurements. It is this strict adherence to the requirements of the observations that makes the latitude theory complex, so complex that even

J.Z. Buchwald and A. Franklin (eds.), Wrong for the Right Reasons, 41–71.

Ptolemy remarks on it in a famous passage (13.2) on complexity and simplicity in astronomical hypotheses. He says, in essence, that we should seek the simpler hypotheses for the motions in the heavens, but failing that, any hypotheses that fit the phenomena. We must do the best we can with observation as our foundation and confirmation. And we must remember that our own ideas of complexity and simplicity may not be applicable to the heavens, which are eternal and unchanging in their motions, something not merely difficult, but impossible to us, meaning that *nature is not necessarily simple according to our way of thinking*, a lesson taught again and again by the science of every age including, or especially, our own. Further, the phenomena of latitude are distinctly different for the superior and inferior planets, and thus they require distinctly different models, another kind of complexity. We shall therefore consider the superior and inferior planets separately.

SUPERIOR PLANETS

The apparent motion in latitude of the superior planets (13.1) is as follows: (1) When the planet is near the apogee of the eccentric, it reaches its greatest northern latitude, and when near the perigee its greatest southern latitude, indicating that the eccentric is inclined to the north in the direction of the apogee and to the south in the direction of the perigee. (2) In the northern and southern limits, the latitudes are greater at opposition when the planet is at the perigee of the epicycle than near conjunction near the apogee – true conjunction itself is invisible – indicating that the epicycle is inclined with its perigee in the same direction, north or south, as the eccentric. In the case of Mars, the planet cannot be seen near conjunction because of its long period of invisibility, but the same conditions are assumed to hold. (3) When the center of the epicycle is a quadrant from the limits and the planet a quadrant from the apogee of the epicycle, it has no latitude, indicating that the epicycle then lies in the plane of the ecliptic. It is this condition that allows the direction of the nodal line, and thus of the limits, to be found from the computed longitude of the center of the epicycle when the planet has no latitude.

The model to account for these three conditions (13.2) is shown in Figure 1. (1) The earth is at O, through which passes the nodal line ☊☋ of the eccentric, which is inclined to the plane of the ecliptic at an angle i_1; N is the northern limit, near apogee, S is the southern limit, near perigee, and the midpoint of NS is M' at an eccentricity e' from O. M' and e' are the center of the eccentric and eccentricity projected into the line joining the limits. (2) When the center of the epicycle is at N or S, it is inclined to the plane of the eccentric in the line of sight, with the perigee to the north at N and to the south at S, so that the latitude β_o at opposition P_o is greater than the latitude β_c at conjunction P_c. It is found from observation that the difference between β_o and β_c is so large that the epicycle is also inclined to the plane of the ecliptic by i_2 and thus to the plane of the eccentric by $i_1 + i_2$. (3) When the center of the epicycle is at the ascending node ☊ or descending node ☋, it lies in the plane of the ecliptic so the planet has no latitude wherever it is located. Hence as the epicycle moves from

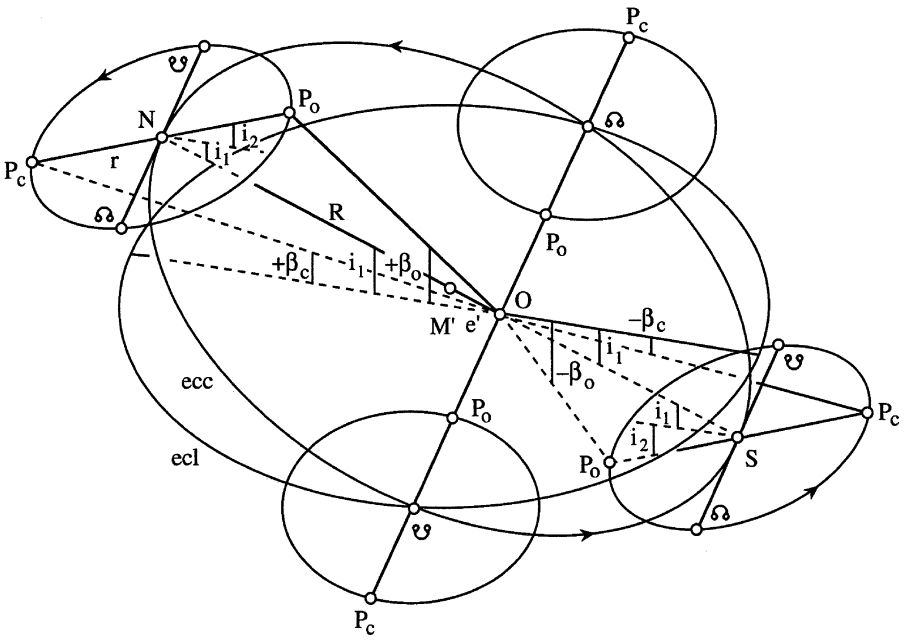

Figure 1. Superior planets.

the limit to the node, i_2 decreases from its maximum to zero, and as it moves to the next limit i_2 again increases to its maximum. Ptolemy treats $i_1 + i_2$ as a single inclination of the epicycle to the plane of the eccentric. But since i_1 may be taken as a fixed inclination, holding the epicycle parallel to the plane of the ecliptic, leaving i_2 alone variable, which we believe a clearer way of showing the variable inclination, we have divided the inclination of the epicycle into two components, the fixed i_1 and the variable i_2.

The derivation of the parameters (13.3) is rigorously empirical. Ptolemy derives i_1 and $i_1 + i_2$ from β_o and β_c using an ingenious method of interpolation in the correction tables for longitude, as though $i_1 + i_2$ were the anomaly on the epicycle measured from apogee or perigee and $\beta_o - i_1$ and $i_1 - \beta_c$ the equation of the anomaly. The derivation for Saturn and Jupiter is shown in Figure 2, in which the earth is at O, the center of the epicycle C is at either limit of latitude, as the eccentricity is neglected, and the planet is at opposition P_o with the larger latitude β_o and at conjunction P_c with the smaller latitude β_c where the difference $\beta_o - \beta_c = \delta$. The plane of the eccentric OC is inclined to the plane of the ecliptic by i_1 and the plane of the epicycle P_cCP_o is inclined to the plane of the eccentric by $i_1 + i_2$ and to a plane parallel to the ecliptic by i_2. We are given by observation β_o and β_c, and we wish to find i_1 and i_2. Now imagine the epicycle rotated into the plane perpendicular to the planes of the eccentric and the ecliptic. We may thus regard $i_1 + i_2$ as the 'anomaly' measured from the apogee

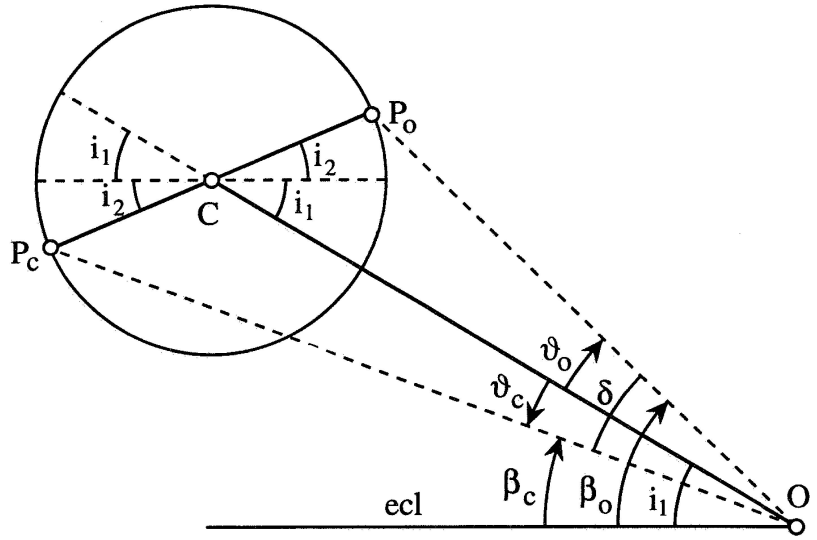

Figure 2. Inclinations i_1 and i_2 of Saturn and Jupiter.

or perigee of the epicycle – they are far smaller than they appear in the figure – and $\vartheta_{\mathrm{o}} = \beta_{\mathrm{o}} - i_1$ and $\vartheta_{\mathrm{c}} = i_1 - \beta_{\mathrm{c}}$ as proportional to the equations of the anomaly c_{o} and c_{c} at some small arc from apogee and perigee respectively. We now have the relations

$$\vartheta_{\mathrm{o}} + \vartheta_{\mathrm{c}} = \beta_{\mathrm{o}} - \beta_{\mathrm{c}} = \delta, \qquad \frac{\vartheta_{\mathrm{c}}}{\vartheta_{\mathrm{o}}} = \frac{c_{\mathrm{c}}}{c_{\mathrm{o}}},$$

from which,

$$\frac{c_{\mathrm{o}}}{c_{\mathrm{c}}}\vartheta_{\mathrm{c}} + \vartheta_{\mathrm{c}} = \vartheta_{\mathrm{c}}\left(\frac{c_{\mathrm{o}} + c_{\mathrm{c}}}{c_{\mathrm{c}}}\right) = \delta, \qquad \vartheta_{\mathrm{c}} = \left(\frac{c_{\mathrm{c}}}{c_{\mathrm{o}} + c_{\mathrm{c}}}\right)\delta,$$
$$\vartheta_{\mathrm{o}} = \delta - \vartheta_{\mathrm{c}}, \qquad i_1 = \beta_{\mathrm{o}} - \vartheta_{\mathrm{o}} = \beta_{\mathrm{c}} + \vartheta_{\mathrm{c}}.$$

And letting the 'anomaly' at opposition, angle $OCP_{\mathrm{o}} = \alpha'$,

$$\frac{i_1 + i_2}{\alpha'} = \frac{\vartheta_{\mathrm{o}}}{c_{\mathrm{o}}}, \qquad i_1 + i_2 = \frac{\vartheta_{\mathrm{o}}}{c_{\mathrm{o}}}\alpha', \qquad i_2 = (i_1 + i_2) - i_1.$$

Since the extreme latitudes are the same at both limits, they are unaffected by the eccentricity. Hence, Ptolemy finds the equations of the anomaly c_{c} and c_{o} for mean distance, c_6 in the correction tables for longitude (11.11), which are described in the Appendix, taking c_{c} for 3° from apogee and c_{o} for 3° from perigee, and lets $\alpha' = 3^\circ$. The latitudes β_{c} near conjunction and β_{o} at opposition, found by observation, simple

integers, are obviously estimates. The computations of i_1 and i_2 are summarized as follows:

	$\pm\beta_c$	$\pm\beta_o$	δ	c_c	c_o	ϑ_c	ϑ_o	i_1	i_1+i_2	i_2
Saturn	2°	3°	1°	0;18°	0;23°	0;26°	0;34°	2;26° $\approx 2\frac{1}{2}$°	4;26° $\approx 4\frac{1}{2}$°	2°
Jupiter	1	2	1	0;29	0;43	0;24	0;36	$1;24 \approx 1\frac{1}{2}$	$2;31 \approx 2\frac{1}{2}$	1

In the case of Mars, the period of invisibility near conjunction is so long, from 90 to more than 200 days, that a nearby latitude cannot be observed, so Ptolemy uses latitudes at opposition at the northern and southern limits, which differ greatly because the limits are exactly at apogee and perigee of the eccentric, which is very nearly true, and because of the large eccentricity, an effect not noticeable for Saturn and Jupiter. The principle of the derivation is nevertheless the same. In Figure 3 the earth is at O, the center of Mars's eccentric is M, the planet in the epicycle is at opposition in the perigee of the epicycle, P_n with latitude β_n at the apogee of the eccentric and northern limit N, P_s with latitude β_s at the perigee of the eccentric and southern limit S, and considering absolute values $\beta_s - \beta_n = \delta$. The plane of the eccentric is inclined to the plane of the ecliptic by i_1, and the plane of the epicycle is inclined to the plane of the eccentric by $i_1 + i_2$ and to a plane parallel to the ecliptic by i_2. Considering $i_1 + i_2$ as the 'anomaly' measured from the perigee of the epicycle, $\vartheta_n = \beta_n - i_1$ and $\vartheta_s = \beta_s - i_1$ are proportional to the equations of the anomaly c_n and c_s at a small arc from perigee. Consequently,

$$\vartheta_s - \vartheta_n = \beta_s - \beta_n = \delta, \qquad \frac{\vartheta_n}{\vartheta_s} = \frac{c_n}{c_s},$$

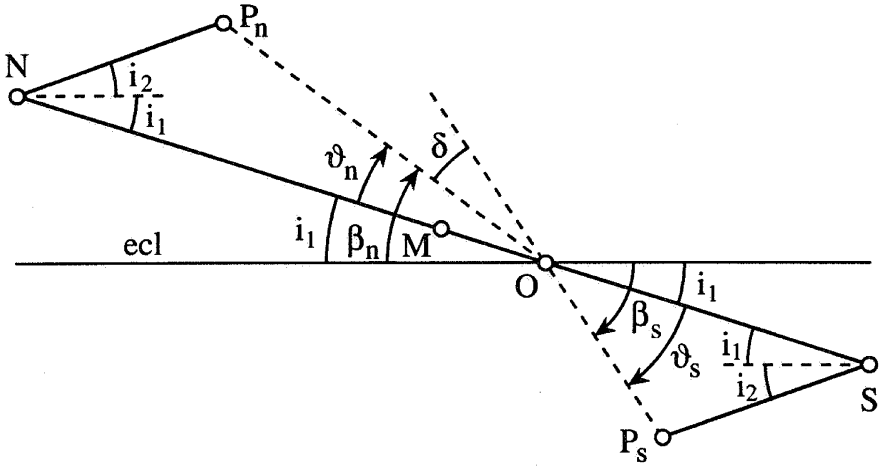

Figure 3. Inclinations i_1 and i_2 of Mars.

so that

$$\frac{c_\mathrm{s}}{c_\mathrm{n}}\vartheta_\mathrm{n} - \vartheta_\mathrm{n} = \vartheta_\mathrm{n}\left(\frac{c_\mathrm{s}-c_\mathrm{n}}{c_\mathrm{n}}\right) = \delta, \qquad \vartheta_\mathrm{n} = \left(\frac{c_\mathrm{n}}{c_\mathrm{s}-c_\mathrm{n}}\right)\delta,$$
$$\vartheta_\mathrm{s} = \vartheta_\mathrm{n} + \delta, \qquad i_1 = \beta_\mathrm{n} - \vartheta_\mathrm{n} = \beta_\mathrm{s} - \vartheta_\mathrm{s}.$$

Again letting the 'anomaly' at opposition at the northern limit, angle $ONP_\mathrm{n} = \alpha'$,

$$\frac{i_1+i_2}{\alpha'} = \frac{\vartheta_\mathrm{n}}{c_\mathrm{n}}, \qquad i_1+i_2 = \frac{\vartheta_\mathrm{n}}{c_\mathrm{n}}\alpha', \qquad i_2 = (i_1+i_2) - i_1.$$

For the computation Ptolemy takes the equation of the anomaly c_n at the apogee of the eccentric and c_s at the perigee, that is, from the equation tables for longitude, $c_\mathrm{n} = c_6 - c_5$ and $c_\mathrm{s} = c_6 + c_7$, each at 3° from the perigee of the epicycle, and likewise $\alpha' = 3^\circ$. Hence, $c_\mathrm{n} = 4;29^\circ$ and $c_\mathrm{s} = 8;5^\circ$, from which he takes the ratio $\vartheta_\mathrm{n}/\vartheta_\mathrm{s} = c_\mathrm{n}/c_\mathrm{s} \approx 5/9$, which is quite accurate, so that $c_\mathrm{n}/(c_\mathrm{s} - c_\mathrm{n}) \approx 5/4$. The observed latitudes at opposition, β_n and β_s, are again obviously estimates. The computation of i_1 and i_2, in which the roundings are very close, is summarized as follows:

	$+\beta_\mathrm{n}$	$-\beta_\mathrm{s}$	δ	c_n	c_s	ϑ_n	ϑ_s	i_1	i_1+i_2	i_2
Mars	$4\frac{1}{3}^\circ$	7°	$2\frac{2}{3}^\circ$	4;29°	8;5°	$3\frac{1}{3}^\circ$	6°	1°	$2\frac{1}{4}^\circ$	$1\frac{1}{4}^\circ$

With these inclinations, Ptolemy can compute the latitudes at conjunction at the limits, which could not be estimated from nearby observations, as was done for Saturn and Jupiter, because of Mars's long periods of invisibility on either side of conjunction. He places in the table of latitude (13.5) for 6° from apogee of the epicycle $\beta_\mathrm{n} = +0;8^\circ$ and $\beta_\mathrm{s} = -0;4^\circ$, which also apply very nearly to conjunction itself at apogee. These are virtually in the plane of the ecliptic and are erroneous, for as we shall see, correctly both latitudes are slightly over 1°.

The problem in Ptolemy's model is the variable inclination of the epicycles, which should always be parallel to the plane of the ecliptic, that is, correctly $i_2 = 0^\circ$, for the epicycles of the superior planets are transformations of the heliocentric motion of the earth, which is always in the plane of the ecliptic. The variable inclinations in turn lead to yet more complications, as Ptolemy also describes (13.2) how they may be produced by small vertical circles upon which the inclining diameters move, with equant motion no less. These small circles, dismissed by Neugebauer with some justice as 'a feeble attempt' – they are a curious cross between a mathematical and mechanical model – have been described in detail for Venus and Mercury by Riddell. It is after describing the operation of these devices that Ptolemy writes his remarks on simplicity and complexity, and with good reason as the models for latitude are kinematically the most complex in the *Almagest*. Ptolemy himself obviously realized that there was something implausible about them. Our concern here, however, is not the complexity or implausibility of the models, but only to inquire into the

reason for the variable inclinations of the epicycles, which is what makes them so complex.

It has been said that the reason is their equivalence to heliocentric models with the nodal line passing through, not the true sun, but the mean sun, the center of the earth's orbit, which is the principle of Copernicus's transformation of them to a heliocentric form, in which the plane of the eccentric has a variable inclination. However, the plane of the eccentric passing through the mean sun introduces a variation of inclination that is only a small fraction of that in Copernicus's model and takes place in different directions on either side of the ecliptic, as has been shown by Swerdlow and Neugebauer, and the same is true of the equivalent small variation of inclination of the epicyclic plane in Ptolemy's model. Thus, it is not the cause of the large variation of the inclination considered here. Rather, the cause is the very strength of Ptolemy's mathematical astronomy, its rigorous empiricism, for the variable inclinations are directly determined, indeed dictated, by the very observed extreme latitudes at opposition and near conjunction from which the inclinations are derived. Likewise, the inclination of the eccentric in Copernicus's model varies, not because its plane passes through the mean sun, but because the model is a transformation of Ptolemy's based upon the same extreme latitudes, although for Mars Copernicus was forced to make small adjustments because the models are not exactly equivalent.

Ptolemy's derivations require observations of the planet at opposition and as near as possible to conjunction, with the center of the epicycle at each of the limits of latitude, but these conditions occur simultaneously only rarely. There is an opposition or conjunction *near* each limit once in 30 years for Saturn, once in 12 years for Jupiter, once in 15 or 17 years for Mars, but the distance from the limit may be quite large for Jupiter and Mars, and finding any of these *at* each limit is much less frequent. And while observations at opposition may be made with the planet well above the horizon, with clearly visible reference stars if such were used in any way, observations near conjunction, thus shortly after first and before last visibility, must be made low on the horizon, possibly without suitable reference stars, and affected by refraction. As it turns out, none of the apparent latitudes Ptolemy cites, without specific information, without any information, about how they were found, is particularly accurate, and the latitudes at opposition are not really better than near conjunction. They were surely not derived strictly from these conditions, but were probably only conventional estimates in integer degrees, with a single simple fraction for Mars, and are insufficiently accurate to find the correct inclinations, which would show that always $i_2 = 0^\circ$. Further, the method of deriving the inclinations, the computation itself, is so sensitive to small imprecisions and roundings that even with accurately observed latitudes, it would still be difficult to find the inclinations exactly.

The clearest way to show this is to begin with correct apparent latitudes according to modern theory and use them to compute the inclinations by Ptolemy's method with the rest of his parameters, the equations of the anomaly c at apogee and perigee of the epicycle, the same. We have mentioned that the simultaneous conditions for extreme

latitudes occur rarely; by inspection of Tuckerman's tables for much of the second century, a rather long period, we find the following values for latitudes at conjunction, which strictly could not be observed, and opposition at the limits of latitude rounded to the nearest $0;3^\circ = 0.05^\circ$, which we compare with Ptolemy's:

	$\pm\beta_c$	P. $\pm\beta_c$	$\pm\beta_o$	P. $\pm\beta_o$
Saturn	$2;18^\circ$	2°	$2;51^\circ$	3°
Jupiter	1;9	1	1;45	2
Mars	+1;9	–	+4;36	+4;20
Mars	−1; 6	–	−6;54	−7

There are differences in the tables of less than 0.05° in the positive and negative latitudes of Saturn and Jupiter. But even $0.05 = 0;3^\circ$ is far less than anything Ptolemy could measure, which he seems to have believed was $\frac{1}{6}^\circ = 0;10^\circ$, so these results show that his assumption that β_c and β_o are the same on either side of the ecliptic was correct even if the errors in his integer values of β_c and β_o reach $0.3^\circ = 0;18^\circ$. For Mars, however, although the difference in $\pm\beta_c$ is very small, the difference in $\pm\beta_o$ is very large, and thus for this reason as well as the invisibility of $\pm\beta_c$, it is also correct that he based his derivations solely on observations at opposition. We can also clearly see the error of the computation in his tables of β_c for Mars as $+0;8^\circ$ and $-0;4^\circ$, for correctly both exceed 1°.

The recomputations have been done in two ways: (1) In keeping with a maximum precision of Ptolemy's observations of $\frac{1}{6}^\circ$, which we note is very optimistic for latitudes near conjunction, we round the modern computed values of $\pm\beta$ to the nearest $0;10^\circ$. (2) To show the extreme sensitivity of the computation, we also take $\pm\beta$ to the nearest $0;3^\circ$, as given above, although this is far beyond the precision of any observation possible to Ptolemy. We then repeat the computations carried out before, using the same values of the equations of the anomaly c at apogee and perigee of the epicycle.

	$\pm\beta_c$	$\pm\beta_o$	δ	c_c	c_o	ϑ_c	ϑ_o	i_1	$i_1 + i_2$	i_2
Saturn (1)	$2;20^\circ$	$2;50^\circ$	$0;30^\circ$	$0;18^\circ$	$0;23^\circ$	$0;13^\circ$	$0;17^\circ$	$2;33^\circ$	$2;13^\circ$	$-0;20^\circ$
Saturn (2)	2;18	2;51	0;33	$0;18^\circ$	$0;23^\circ$	0;14	0;19	2;32	2;29	−0; 3
Jupiter (1)	1;10	1;50	0;40	0;29	0;43	0;16	0;24	1;26	1;40	0;14
Jupiter (2)	1;9	1;45	0;36	0;29	0;43	0;15	0;21	1;24	1;28	0;4

	$+\beta_n$	$-\beta_s$	δ	c_n	c_s	ϑ_n	ϑ_s	i_1	$i_1 + i_2$	i_2
Mars (1)	$4;40^\circ$	$6;50^\circ$	$2;10^\circ$	$4;29^\circ$	$8;5^\circ$	$2;42^\circ$	$4;52^\circ$	$1;58^\circ$	$1;48^\circ$	$-0;10^\circ$
Mars (2)	4;36	6;54	2;18	4;29	8;5	2;52	5;10	1;44	1;55	0;11

Note that for Saturn and Jupiter i_1 is nearly the same as found by Ptolemy, 2;30° and 1;30° – correctly Saturn is 2;33° and Jupiter 1;25°, very close to these calculations – but i_2 is reduced by nearly 1° for Jupiter and more than 2° for Saturn, for which it is here even slightly negative, meaning that the epicycle is inclined in the opposite direction. Since correctly $i_2 = 0°$, what this shows is that, although the method of derivation, including the use of the correction tables for longitude, is satisfactory for finding i_1, it is too sensitive to small errors in β and c to find i_2 with great accuracy. For Mars, i_1 is increased from 1° to nearly its correct value 1;52° and i_2 is reduced from 1;15° nearly to 0° and is both positive and negative; correctly both i_1 and i_2 are about midway between (1) and (2). In fact, as uncertain as these results may be for i_2, for all three planets i_1 is close to its correct value and i_2 is at least close to 0°. Hence, the problem in Ptolemy's latitude theory is not the model itself, in which the inclination of the epicycle is an independently derived parameter, nor the method of deriving the parameters, as sensitive as it is for i_2, but the observations, rough estimates of latitude at opposition and near conjunction, which, by Ptolemy's rigorously empirical method, require the variable inclination of the epicycle. We shall return to this subject in considering the revised latitude theory of the *Planetary Hypotheses*.

There is one further parameter in Ptolemy's latitude theory of the superior planets, the distance of the northern limit of latitude from the apogee, which is used to find the argument of latitude measured from the northern limit. He first locates the northern limits rather roughly (13.1) as near the beginning of Libra for Saturn and Jupiter and near the end of Cancer for Mars, almost exactly at the apogee. Then, taking the longitudes of the apogees, with slight rounding, he gives distances from the apogee to the northern limit ω_A (13.6): −50° for Saturn, +20° for Jupiter, 0° for Mars. (Correctly for A.D. 140 these are about: Saturn −42°, Jupiter +28°, Mars +11°. The distances from the aphelia, apsidal lines through the true sun, are about: Saturn −49°, Jupiter +7°, Mars +3°, two of which are, by coincidence, closer to Ptolemy's values.) This parameter is difficult to find with accuracy. At the limits the latitude is highly variable and the maximum latitude at opposition, which could in principle locate the limit, very seldom occurs exactly at a limit. Whenever the planet crosses the ecliptic, the center of the epicycle is in the nodal line ±90° from the limits. But this too is not easy to observe as the latitude of the planet changes most rapidly when crossing the ecliptic and the chance of catching it exactly in the ecliptic is slight. Still, some kind of interpolation between small latitudes on either side of the ecliptic is probably the most reasonable way of finding the longitude of the nodes and thus, by ±90°, of the limits. Whether this explains the errors in Ptolemy's locations of the limits, I do not know. It does not help in finding this parameter that the period of Saturn is nearly 30 years and of Jupiter nearly 12 years, so for long periods no useful observations can be made for finding either limits or nodes; and although the period of Mars is less than two years, its long periods of invisibility and rather large and irregular synodic arcs also make it difficult to observe exactly at a limit or crossing the ecliptic.

The tables for latitude of the superior planets (13.5) are very easy to use although at the cost of some precision. An example from the table for Mars is given here at intervals of 6°:

1	2	3	4	5
Argument		$+\beta(\alpha)$	$-\beta(\alpha)$	Inter. (ω_C)
6°	354°	0;8°	0;4°	0;59,36
12	348	0;9	0;4	0;58,36
18	342	0;11	0;5	0;57,0
...	...	...	...	...
84	276	0;46	0;42	0;6,24
90	270	0;52	0;49	0;0,0
96	264	0;55	0;52	0;6,24
...	...	...	...	...
168	192	4;0	5;53	0;58,36
174	186	4;14	6;36	0;59;36
180	180	4;21	7;7	1;0,0

Columns 1 and 2 are arguments of entry for 6°–180° and 180°–354° at intervals of 6° for 270°–90° and 3° for 90°–270°. Columns 3 and 4 are latitudes as a function of true anomaly α on the epicycle computed for the maximum inclination $i_1 + i_2$ and the center of the epicycle at the limits of latitude, column 3 northern, column 4 southern. The differences in the two columns are due to the different distances of the limits on the eccentric; these are large for Mars – at opposition 2;46°, at conjunction 0;4°, but note that $+\beta$ is twice $-\beta$ – since the eccentricity is large and the limits are in the apsidal line, but small for Jupiter and Saturn at opposition and conjunction – from 0;2° to 0;4°, each a very small fraction of β – since their eccentricities are smaller and their limits are removed from the apsidal line, for Jupiter by +20° and for Saturn by −50°. This is a rather crude way of handling the effect of distance, and Ptolemy developed a more accurate method in the *Handy Tables*. Column 5 is a coefficient of interpolation for locations of the center of the epicycle other than the limits as a function of the distance of the center of the epicycle from the northern limit, $\omega_C = \lambda_C - \lambda_N = \kappa - \omega_A$, where κ is the true eccentric anomaly. It is used as a cosine since both the latitude on the eccentric and the inclination i_2 of the epicycle vary nearly as cos ω_C, that is $c_5(\omega_C) = \cos\ \omega_C$; it is, however, computed by multiplying the tabulated lunar latitude for each entry, with a maximum of 5°, by 0;12. The computation of the latitude from the table is simply

$$+\beta = c_5(\omega_C) \cdot c_3(\alpha), \qquad \text{if } 270^\circ \le \omega_C \le 90^\circ,$$
$$-\beta = c_5(\omega_C) \cdot c_4(\alpha), \qquad \text{if } 90^\circ \le \omega_C \le 270^\circ.$$

INFERIOR PLANETS

The apparent motion in latitude of the inferior planets is entirely different and even more complex. Because their orbits are inside the heliocentric orbit of the earth, their

motions in latitude can be seen in two ways, both in the line of sight and across the line of sight, and these appear to behave quite differently. It happens that the nodal lines of the inferior planets are rather close to the directions of the apsidal lines found by Ptolemy – $+5^\circ$ to the ascending node for Venus, $+16^\circ$ to the descending node for Mercury – and his description of the latitudes (13.1) follows from assuming that the apsidal and nodal lines coincide. (1) When the center of the epicycle is a quadrant from the apsidal line, the greatest differences in latitude occur, on the same side of the ecliptic, near superior and inferior conjunction, the larger near inferior conjunction, the smaller near superior conjunction, and when the planet is a quadrant from either conjunction it has no latitude. (2) When the center of the epicycle is in the apsidal line, the greatest differences in latitude occur, on opposite sides of the ecliptic, at opposite greatest elongations, differing by approximately equal amounts from the latitudes at apogee and perigee. (3) When the center of the epicycle is in the apsidal line and the planet is near superior or inferior conjunction, it has a small latitude, to the north for Venus and to the south for Mercury.

The model to account for these latitudes (13.2) is shown in Figure 4. The earth is at O, the center of the eccentric at M, and the epicycle is shown at apogee at 0°, and at 90°, 180°, and 270° from apogee. The eccentricity itself has a small effect on latitude for Mercury and virtually none for Venus, and is shown here to distinguish the direction of the apsidal line. Consider the configuration at 90° and 270°. (1) The epicycle is inclined in the line of sight at an angle i_1 so the greatest differences in latitude are on the same side of the ecliptic, β_a at the apogee of the epicycle P_a and β_b at the perigee P_b, and because the distance OP_b is less than OP_a, β_b is greater than β_a. This component of latitude, which we call β_1, in the line of sight, is called the 'inclination' (*enklisis*), the same term used for the latitude of the superior planets, which is also

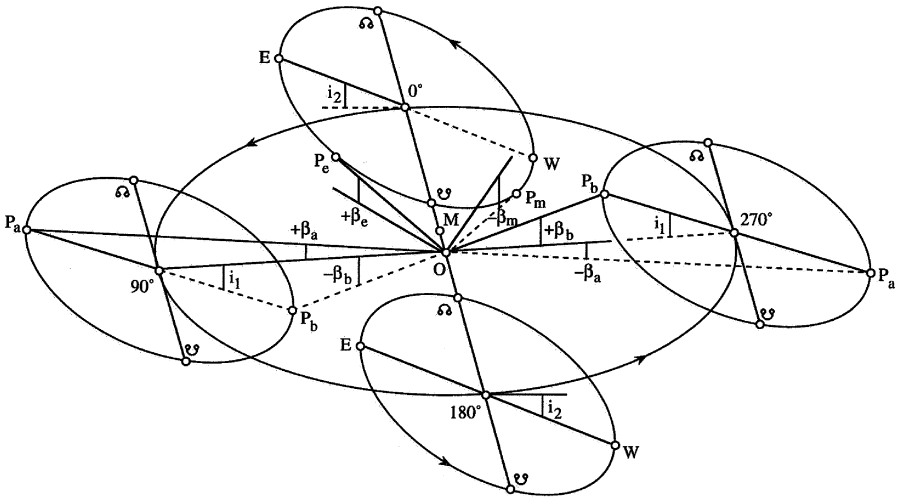

Figure 4. Inferior planets.

in the line of sight, and i_1 is the inclination of the epicycle. When the planet is in the nodal line across the line of sight a quadrant from P_a and P_b, it is in the plane of the ecliptic without latitude. As the epicycle moves from 270° to 0°, the inclination i_1 decreases to zero and the epicycle takes on a second inclination i_2 along a nodal line in the line of sight. (2) At 0° the epicycle is inclined across the line of sight and the greatest differences in latitude are on opposite sides of the ecliptic, β_e at greatest evening elongation P_e and β_m at greatest morning elongation P_m. This component of latitude, which we call β_2, across the line of sight, is called the 'slant' or 'obliquity' (*loxosis*), and i_2 the slant or obliquity of the epicycle. As the epicycle continues to 90°, i_2 goes to zero and i_1 increases to its maximum, and so on. (3) Finally, and this is not illustrated, to account for the small latitude β_3 near conjunction, when the center of the epicycle is at 0° and 180°, the plane of the eccentric has a small inclination i_3 on a nodal line perpendicular to the apsidal line, thus passing through 90° and 270°, moving the epicycle and the planet to the north for Venus and to the south for Mercury, an inclination that also goes to zero as the epicycle moves to 90° and 270°. Ptolemy's calls this the 'inclination of the eccentric' as it is also in the line of sight.

The inclination i_1 is found from β_1 using the correction table for longitude, in much the same way as for the superior planets (13.3). In Figure 5, the earth is at O and the plane of the epicycle is inclined to the plane of the eccentric OC, which lies in the plane of the ecliptic, by i_1 such that the planet P_a at superior conjunction at apogee has latitude β_a and P_b at inferior conjunction at perigee has latitude β_b. Neither location can be directly observed since the center of the epicycle lies *nearly* in the direction of the mean sun ($\bar{S}$), and thus the planet is too close to the true sun to be visible, so β must be inferred from nearby observations, a difficulty to which we shall return. Now imagine the epicycle rotated into a plane perpendicular to the plane of the ecliptic; then i_1 may be taken as proportional to the anomaly at a small arc α_a from apogee and α_b from perigee, and β_a and β_b as proportional to equations of the anomaly c_a and c_b corresponding to these arcs, that is, in each case $i_1/\alpha = \beta/c$. Since the maximum effect of i_1 takes place $\pm 90°$ from the apsidal line, the equations for

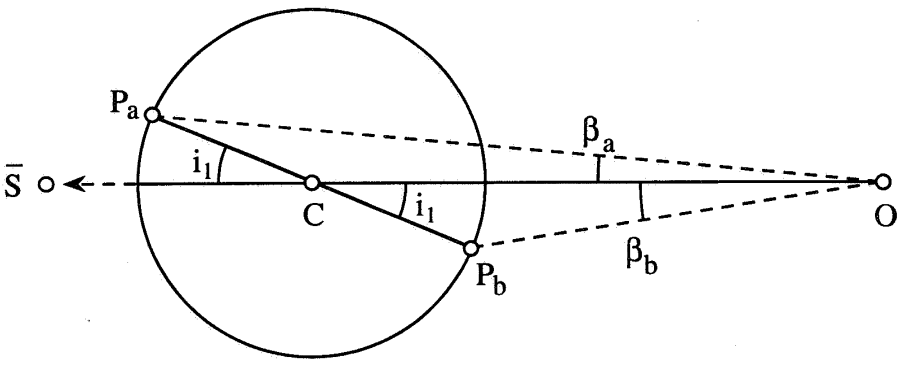

Figure 5. 'Inclination' i_1 of inferior planets.

Venus are taken for mean distance, $c = c_6$ in the correction tables, but for Mercury, which is not at mean distance, $c = c_6 + c_8 \cdot c_7$ for 90° of mean eccentric anomaly for which $c_8 \approx 0;40$. From the tables (11.11), letting $\alpha_a = 6^\circ$ and $\alpha_b = 3^\circ$, we find two values, i_{1a} and i_{1b}, from $i_1 = \beta(\alpha/c)$, which we compare with Ptolemy's i_1 and his β_a and β_b computed for confirmation from $\beta = i_1(c/\alpha)$, three of which have discrepancies of $\pm 0;1^\circ$ compared to an accurate calculation.

	β_a	β_b	c_a	c_b	i_{1a}	i_{1b}	P.i_1	P.β_a	P.β_b
Venus	1°	$6\frac{1}{3}^\circ$	2;31°	7;38	2;23°	2;29°	$2\frac{1}{2}^\circ$	1;2°	6;22°
Mercury	$1\frac{3}{4}$	4	1;41	1;57	6;14	6;9	$6\frac{1}{4}$	1;46	4;5

Ptolemy's i_1 is very close to i_{1b} for Venus and i_{1a} for Mercury, and his recomputations of β, which are also the extreme latitudes in his tables in 13.5, are sufficiently close to the observed latitudes that the selection and rounding of i_1 are reasonable. Still, there are problems with β_1 and i_1 that we shall take up after considering β_2 and i_2.

The slant of the epicycle i_2 is found (13.4) from the maximum apparent latitude β_2, which takes place at greatest elongation (from the mean sun) where Venus is always visible and Mercury usually visible, although near apogee of the eccentric only its morning elongation and opposite to apogee only its evening elongation are visible. (In fact, the maximum latitudes β_2 do not take place at greatest elongation from the mean sun, but at maximum equation of the anomaly, when the line from the earth to the planet is tangent to the epicycle. However, in the apsidal line the equation of center is zero and these are the same.) In Figure 6 the planet P_e is shown at greatest evening elongation, to the east of the mean sun $\bar{S}$, with latitude β_2. Since CP_eO and CGP_e are right angles, triangles CP_eO and CGP_e are similar so that $a/\rho = r/R'$ where $R' = OC$. It follows that

$$d = \rho \sin \beta_2 = a \sin i_2 = \rho \frac{r}{R'} \sin i_2, \qquad \sin i_2 = \frac{R'}{r} \sin \beta_2.$$

Ptolemy notes (13.1, 13.4) that the maximum opposite apparent latitudes, on either side of the ecliptic, at greatest elongations at apogee and perigee of the eccentric differ, for Venus by slightly under 5° at apogee and slightly over 5° at perigee, and for Mercury by about $5^\circ - \frac{1}{2}^\circ$ at apogee and about $5^\circ + \frac{1}{2}^\circ$ at perigee, the differences from 5° due to the effect of the greater distance at apogee and lesser distance at perigee. For the derivation of i_2, he ignores the variation and takes a difference of about 5°, and for the latitudes on each side of the ecliptic he takes the arithmetic mean so that $\beta_2 \approx \pm 2\frac{1}{2}^\circ$ for both planets. This, as we shall see, is a considerable simplification although it leads to excellent results. The distance of the center of the epicycle of Venus in the apsidal line is $R + e$ at apogee and $R - e$ at perigee, and since the effect of change of distance is small, he takes $R' = R = 60$. The distance of the center of the epicycle of Mercury at apogee is $R + 3e = 69$ and at the opposite end of the apsidal line $R - e = 57$, which is not its least distance, and for the derivation he takes the

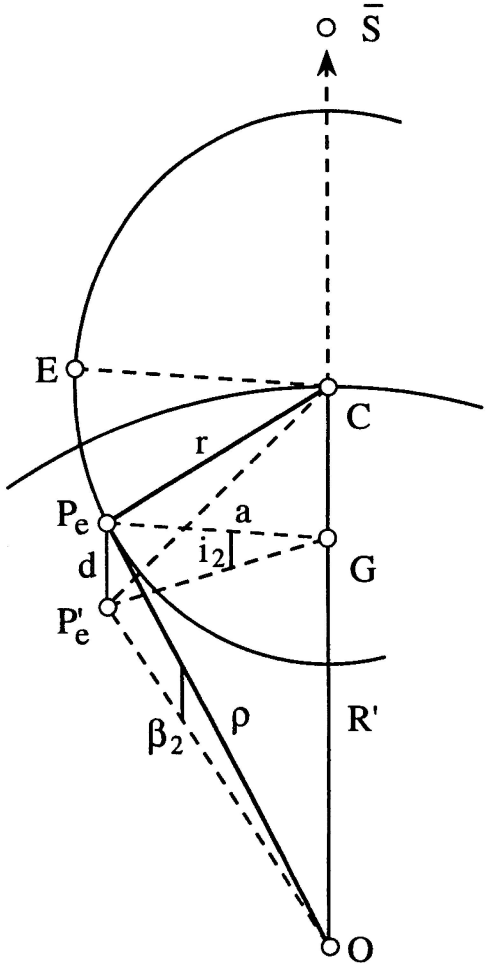

Figure 6. 'Slant' i_2 of inferior planets.

mean $R' = R + e = 63$. With these values, the computation, from $\sin i_2 = (R'/r) \sin \beta_2$, is straightforward.

	R'	r	β_2	i_2
Venus	60	43;10	2;30°	3;28° $\approx 3\frac{1}{2}°$
Mercury	63	22;30	2;30	7

In order to analyze empirically the 'inclination', latitude β_1 and inclination i_1, we find the apparent latitudes as close as possible to superior and inferior conjunction

at the limits of latitude according to modern theory, using inspection in Tuckerman's tables for the second century. Since these change in the course of the century, particularly for Venus, as the conjunction points shift with respect to the limits, we cite them only to tenths of a degree. The inclination i_1 is found from β_a at superior conjunction and apogee of the epicycle or from β_b at inferior conjunction and perigee. These occur in the invisible arc and must be inferred from observations before last visibility and after first visibility when the planet is near the horizon, affected by refraction, and possibly with no reference stars. For Venus the period of invisibility around superior conjunction is quite long, from 55 to 69 days at Ptolemy's latitude, and around inferior conjunction, entirely within the retrograde arc, from 1 to 18 days. Mercury's period of invisibility at superior conjunction is from about 27 days to an entire invisible evening phase and at inferior conjunction, at least in part within the retrograde arc, from about 13 days to an invisible morning phase. The strictly invisible extreme values of β_a and β_b to $0;6^\circ = 0.1^\circ$ are as follows along with the values cited by Ptolemy:

	$+\beta_a$	$-\beta_a$	P. $\pm\beta_a$	$+\beta_b$	$-\beta_b$	P. $\pm\beta_b$
Venus	$1;30^\circ$	$1;30^\circ$	1°	$8;36^\circ$	$8;48^\circ$	$6;20^\circ$
Mercury	1;48	2;0	1;45	3;48	4;42	4

The more serious errors, more than $\pm 2^\circ$ for β_b, with more serious consequences, are for Venus. (I do not believe that modern computations and graphs of the motion of the planet from before last to after first visibility give any idea of just how difficult it is to measure these latitudes accurately.) We may use a selection of these values of β_a and β_b to $0;6^\circ$, more precise than Ptolemy could reach, to compute i_1 by Ptolemy's method, in Figure 5, $i_1 = \beta(\alpha/c)$, taking c from the correction tables (11.11) and letting $\alpha_a = 6^\circ$ and $\alpha_b = 3^\circ$. In this way, we find i_{1a} and i_{1b} and compare them with i_1 from Ptolemy and modern theory.

	β_a	β_b	c_a	c_b	i_{1a}	i_{1b}	P. i_1	M. i_1
Venus	$1;30^\circ$	$8;36^\circ$	$2;31^\circ$	7;38	$3;34^\circ$	$3;23^\circ$	$2\frac{1}{2}^\circ$	$3;22^\circ$
Mercury	1;48	4;42	1;41	1;57	6;25	7;14	$6\frac{1}{4}$	6;58

Ptolemy's i_1 for Venus is erroneous, following from his erroneous value of β_a and β_b, while for Mercury it is close to i_{1a} as his $\beta_a = 1;45^\circ$ is close to $\beta_a = 1;48^\circ$ here; the modern i_1 is close to i_{1b} for both planets. A different selection of β_a and β_b would produce different results – for example, for Mercury $\beta_a = 2^\circ$ gives $i_{1a} = 7;8^\circ$, close to i_{1b} – and more precise values would give better agreement with modern theory. So again we see that the problem in Ptolemy's derivation is the inadequate observations, particularly for Venus, as it is difficult to estimate the invisible latitudes at superior and inferior conjunction. The derivation itself of i_1 by this method is about as sensitive to inaccuracies as finding i_1 for the superior planets, that is, moderately sensitive.

The difficulties of the 'slant', latitude β_2 and inclination i_2, are entirely different, and there are difficulties even though Ptolemy's results are excellent. Latitude β_2 is observed at greatest elongation, where Venus is always visible and Mercury usually visible, as far above the horizon as they can be seen, which appears promising. But for i_2 to have its greatest effect, and be isolated from i_1, the heliocentric orbit of the planet must be seen across the line of sight, which occurs when the earth is in the planet's nodal line, meaning, in Ptolemy's theory, when the center of the epicycle is in the planet's apsidal line, which happens to be close to the heliocentric orbit's nodal line. The difficulty here, for Venus above all, is that the planet is seldom at greatest elongation when the center of the epicycle is in the apsidal line, and small departures from these two conditions can noticeably affect the apparent latitude. The same problem occurs in Ptolemy's determination of the parameters for longitude, the eccentricity and radius of the epicycle, which require the same strictly unobtainable conditions, greatest elongation with the center of the epicycle in the apsidal line and at other specified locations. For Mercury, as noted before, it happens that near apogee only morning elongation and opposite to apogee only evening elongation are visible.

Ptolemy was doubtless aware of these difficulties, although he does not mention them, for they may explain the way he describes the behavior of the slant (13.1, 13.4). He does not say directly that at greatest elongation in the apsidal line $\beta_2 = \pm 2\frac{1}{2}^\circ$, even though he uses that value for finding i_2 for both planets. Rather, he says that the *total* variation in latitude, north and south of the ecliptic, for Venus is slightly under 5° at apogee of the eccentric and slightly over 5° at perigee, for Mercury is about $5^\circ - \frac{1}{2}^\circ$ at apogee and $5^\circ + \frac{1}{2}^\circ$ at perigee, and that he will use $\beta_2 \approx \pm 2\frac{1}{2}^\circ$ as a *mean* value. There is good reason for his description, for if we use Tuckerman's tables to examine the apparent latitudes *near* greatest elongation with the earth or sun *near* each planet's nodal line, where the slant is isolated from the inclination, the apparent latitude is *almost never* $\pm 2\frac{1}{2}^\circ$, but varies quite widely. As close as we can come to these conditions in the period A.D. 130–146, we find for Venus positive latitudes restricted to about $+2.1^\circ$ and $+3.1^\circ$, the mean of which is $+2.6^\circ$, and a single negative latitude of -2.4°. For Mercury we find a positive range of about $+2.4^\circ$ to $+2.7^\circ$ with a mean of $+2.55^\circ$ and a negative range of -2.5° to -3.2° with a mean of -2.85°, although I am not certain how many of these are actually visible. Hence, Ptolemy's mean latitudes for both planets, $\beta_2 \approx \pm 2\frac{1}{2}^\circ$, even if theoretical, and his inclinations, for Venus $i_2 = 3\frac{1}{2}^\circ$ and for Mercury $i_2 = 7^\circ$, are better than can be reached from observations during this period. This is also true of some of his other parameters.

Ptolemy describes the third component of latitude β_3 (13.1, 13.3) as an *equal* latitude from the ecliptic, reaching $+\frac{1}{6}^\circ$ for Venus and $-\frac{3}{4}^\circ$ for Mercury, at *both* apogee and perigee of the epicycle, thus at superior and inferior conjunction, as inferred from nearby observations, when the center of the epicycle is in the apsidal line. Since the latitude is the same at apogee and perigee of the epicycle, it is not affected by distance on the epicycle. Ptolemy therefore attributes β_3 to a variable inclination i_3 of the eccentric on a nodal line passing through the earth perpendicular to the apsidal line, hence through 90° and 270° in Figure 4; i_3 is maximum when the

center of the epicycle is in the apsidal line, goes to zero at a quadrant from the apsidal line, and then returns to maximum in the same direction, to the north for Venus, to the south for Mercury.

It is not at all obvious how Ptolemy found β_3, but it does seem to be his own discovery rather than a conventional value. I know of no correct explanation of just what it is that he observed, what accounts for β_3, and will not trouble the reader with my own attempts. There are discussions by Pedersen and Neugebauer. In any case, inferring a latitude of $+0;10^\circ$ for Venus or even $-0;45^\circ$ for Mercury at conjunction when the planet can only be observed many days before or after, the latitude is changing the most rapidly across the line of sight, and is strongly affected by refraction near the horizon, seems very insecure. And just as for observing greatest elongations with the center of the epicycle in the apsidal line, the simultaneous conditions for observing the planet near conjunction with the center of the epicycle in or near the apsidal line occur rarely, especially for Venus. Since the effect of i_3 is the same wherever the planet is on the epicycle, β_3 could be related to a variation in observed latitudes of β_2 near greatest elongation when the center of the epicycle is near the apsidal line and the effect of β_2 is greatest. Ptolemy does describe effects of this sort, opposite for Venus and Mercury, at greatest elongations (13.1), but he also describes the latitudes near apogee and perigee, and I would not doubt his report of the kind of observations by which he discovered it. As he changed i_3 to a fixed inclination in the *Handy Tables* and the *Planetary Hypotheses*, he himself must have come to doubt these observations.

The tables for latitude of the inferior planets (13.5) are more complex than those for the superior planets, and the computation is considerably more complex since three components, the inclination and slant of the epicycle and the inclination of the eccentric, must be computed separately and added together. An example from the table for Venus is given here at intervals of 6°:

1	2	3	4	5
Argument		$\pm\beta_1(\alpha)$	$\pm\beta_2(\alpha)$	Inter. (κ)
6°	354°	$1;2^\circ$	$0;8^\circ$	0;59,36
12	348	1;1	0;16	0;58,36
18	342	1;0	0;25	0;57,0
...	...	...	...	...
84	276	0;8	1;50	0;6,24
90	270	0;0	1;57	0;0,0
96	264	0;10	2;3	0;6,24
...	...	...	...	...
126	234	1;18	2;27	0;35,12
132	228	1;38	2;30	0;40,0
138	222	1;59	2;30	0;44,24
...	...	...	...	...
168	192	5;13	1;27	0;58,36
174	186	5;52	0;48	0;59;36
180	180	6;22	0;0	1;0,0

Columns 1 and 2 are arguments of entry for 6°–180° and 180°–354° at intervals of 6° for 270°–90° and 3° for 90°–270°. Column 3 is the inclination β_1 computed from i_1 and column 4 the slant β_2 computed from i_2, both functions of the true anomaly α on the epicycle. There is also a rather crude correction for distance in computing β_2 for Mercury, taking $\frac{9}{10}c_4(\alpha)$ in the semicircle of the eccentric around apogee and $\frac{11}{10}c_4(\alpha)$ in the semicircle around perigee. Since $\frac{1}{10} \cdot 2\frac{1}{2}^\circ = \frac{1}{4}^\circ$, this gives a range of β_2 of $5 - \frac{1}{2}^\circ$ at apogee and $5^\circ + \frac{1}{2}^\circ$ at perigee, as Ptolemy reported. Column 5 is the same coefficient of interpolation tabulated for the superior planets, used as a cosine, as a function of the true eccentric anomaly $\kappa = \lambda_C - \lambda_A$ since the variations of i_1 and i_2 are both functions of the distance κ of the center of the epicycle from the apogee of the eccentric; hence $c_5(\kappa) = \cos\ \kappa$. There are rather complex rules (13.6) for how $c_5(\kappa)$ is applied to β_1 and β_2 because the inclinations i_1 and i_2 are maximum and zero 90° apart, so that one uses both $c_5(\kappa)$ and $c_5(\kappa \pm 90^\circ)$, and the rules are reversed for Venus and Mercury since the inclinations of their epicycles are in opposite directions. Column 5 is also used to compute the latitude β_3 due to the inclination of the eccentric i_3, +0;10° for Venus and −0;45° for Mercury; it is applied as $c_5(\kappa)^2 = \cos^2\kappa$ in order to compute both the variation of i_3 and the change of latitude on the eccentric, each separately a function of $\cos\ \kappa$. A complete statement of the rules for the application of $c_5(\kappa)$ is given by Neugebauer. Here we note only that one forms, with the proper signs and the correction in $c_4(\alpha)$ for Mercury,

$$\pm\beta_1 = c_5(\kappa) \cdot c_3(\alpha), \quad \pm\beta_2 = c_5(\kappa) \cdot c_4(\alpha), \quad \pm\beta_3 = c_5(\kappa)^2 \cdot i_3,$$
$$\beta = \beta_1 + \beta_2 + \beta_3.$$

Because of the errors in i_1 and β_1 in particular, the computed latitudes of the inferior planets are not at all accurate, with errors for Venus reaching over 2° near inferior conjunction, as we have seen. Ptolemy must have become aware of these problems, for the latitude theory of the inferior planets receives notable correction in the *Handy Tables*, to which we now turn.

HANDY TABLES

It is possible that Ptolemy's latitude theory differs from the *Almagest* in the *Canobic Inscription*, which shows a still earlier stage of his work, but the numbers in the text appear so corrupt that no conclusions can be drawn. Hence, we can say nothing about this earliest latitude theory if it did in fact differ. There is, however, no doubt that the latitude theory later received important modifications, improvements, in the *Handy Tables*, although with a curious error not present in the *Almagest*. The tables for computing latitude, which are entirely different from those in the *Almagest*, are similar in form to the correction tables for longitude, described briefly in the Appendix, and the computation is now the same for superior and inferior planets although it is also somewhat more laborious. There is a detailed examination by Neugebauer, in part following an analysis by van der Waerden, showing how the tables are computed from the underlying model, which is nowhere explained by Ptolemy. The latitude is computed as the sum of two components, one due to the inclination of the eccentric, the other due to the inclination of the epicycle, and the most notable result is a better

control over the effect of the distance of the center of the epicycle. This is done in the same way as in the correction tables for longitude, that is, there is a column for latitude at mean distance of the center of the epicycle, two additional columns, a subtraction for greatest distance and an addition for least distance, and a coefficient of interpolation for intermediate distances. An example from the table for Mars is given here at 6° intervals:

1	2	3 (κ)	4 (α)	5 (α)	6 (α)	7 (ω)
6°	354°	−0; 60	0;3°	0;54°	0;3°	0;60
12	348	−0; 59	0;3	0;55	0;3	0;59
18	342	−0; 57	0;3	0;56	0;4	0;57
...	...	...	...	...	...	...
84	276	−0; 4	0;7	1;7	0;9	0;6
90	270	+0;3	0;8	1;11	0;11	0;0
96	264	+0;8	0;9	1;15	0;13	0;6
...	...	...	...	...	...	...
168	192	+0;58	0;51	3;46	1;29	0;59
174	186	+0;59	0;56	4;6	1;39	0;60
180	180	+0;60	0;59	4;20	1;46	0;60

Columns 1 and 2 are arguments for 3°–180° and 180°–357° at intervals of 3°. Column 7 is a coefficient of interpolation, a cosine, $c_7 = \cos\ c_{1,2}$, rounded from c_5 in the *Almagest*, which is used in two ways. In the first, it is used to compute the latitude β_1 due to the inclination of the eccentric i_1 by $\pm\beta_1 = c_7(\omega_\mathrm{p}) \cdot i_1$, where ω_p is the distance of the planet from the northern limit, $\omega_\mathrm{p} = \lambda_\mathrm{p} - \lambda_\mathrm{N}$, which also determines the sign of β_1. For the superior planets, this is close to the heliocentric latitude of the planet, which does not appear independently in the computation from the tables in the *Almagest*. Column 5 is the latitude due to the epicycle, under the assumption that the planet is always at greatest distance from the plane of the eccentric, as though the epicycle is parallel to the eccentric and raised from the eccentric by $i_1 + i_2$, as shown in Figure 7A, in which P_o is the planet at opposition, P'_c the projection of the planet P_c at conjunction, and P' the projection of an arbitrary position P. Column 5 is a function of the true anomaly α measured from the apogee of the epicycle when the center of the epicycle is at mean distances. Column 4 is the subtraction for the epicycle at apogee of the eccentric, column 6 is the addition for the epicycle at perigee of the eccentric, both also functions of α. Column 3 is a coefficient of interpolation for the distance OC of the center of the epicycle on the eccentric, a function of the true eccentric anomaly $\kappa = \lambda_\mathrm{C} - \lambda_\mathrm{A}$. To combine the effect of the two variables α and κ, one computes either of

$$\beta'_2 = c_5(\alpha) + c_3(\kappa) \cdot c_4(\alpha), \qquad \text{if } c_3(\kappa) < 0,$$
$$\beta'_2 = c_5(\alpha) + c_3(\kappa) \cdot c_6(\alpha), \qquad \text{if } c_3(\kappa) > 0.$$

It is this calculation, the principal innovation of the latitude tables in the *Handy Tables* and a great improvement over the treatment of the effect of distance in the *Almagest*,

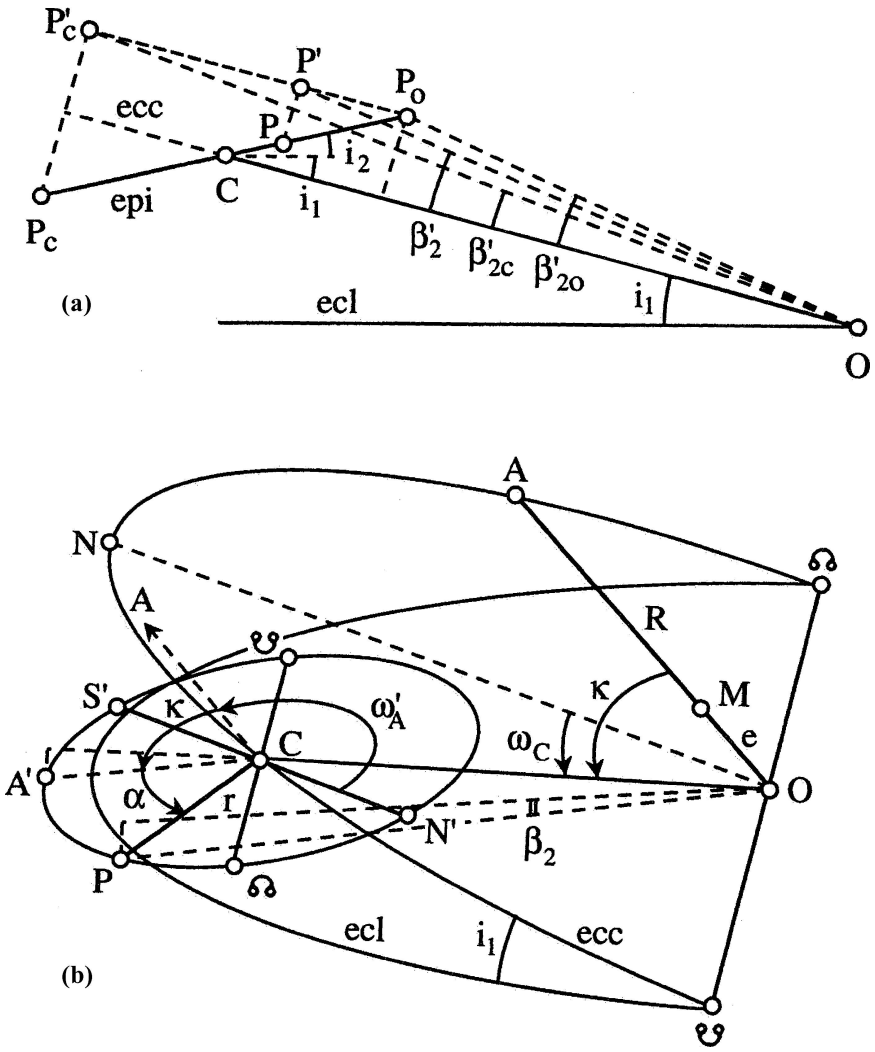

Figure 7. (A) Preliminary latitude β_2' in *Handy Tables*. (B) Inclination of epicycle and final latitude β_2 in *Handy Tables*.

that corresponds to the calculation of the second inequality in the correction tables for longitude. Finally, column 7 is again used as a coefficient of interpolation for the angular distance of the planet on the epicycle – now regarded as inclined to the eccentric *with a fixed inclination* – from the northern limit of the epicycle, $\omega_p' = \omega_A' + \kappa + \alpha$, as shown in Figure 7B. ω_A' is an invariable distance from the northern limit of the epicycle N', which holds a fixed direction, to a direction CA on the epicycle parallel to the direction OA of the apogee of the eccentric. Hence, one finds $\pm\beta_2 = c_7(\omega_p') \cdot \beta_2'$,

which also determines the sign of β_2, and then the final latitude $\beta = \beta_1 + \beta_2$. (The configuration and notation here differ slightly from Neugebauer's.)

The inclinations of the planes of the eccentric and the epicycle of the superior planets are the same as in the *Almagest*, and the latitudes at opposition and conjunction at the limits differ by not more than $\pm 0;1^\circ$. But there is no doubt that the inclination of the epicycle is now fixed, as has been noted by van der Waerden, and this seriously affects latitudes when the center of the epicycle is near the nodes, a problem that seems thus far to have gone unnoticed. Were the inclination variable as in the *Almagest*, for the superior planets from i_2 at its greatest value when the center of the epicycle is at the limits to $i_2 = 0^\circ$ at the nodes, one would have to multiply β_2' by a further coefficient using $c_7 = \cos c_{1,2}$, the distance of the center of the epicycle from the northern limit of the eccentric $\omega_C = \lambda_C - \lambda_N$, that is, $c_7(\omega_C) = \cos \omega_C$, the same coefficient used in the *Almagest*, so that $\beta_2 = c_7(\omega_C) \cdot c_7(\omega_p') \cdot \beta_2'$. But nothing of the kind is mentioned in Ptolemy's instructions, nor in Theon's instructions, in which β_2 is determined only by $\beta_2 = c_7(\omega_p') \cdot \beta_2'$ for a fixed inclination of the epicycle. Although this is correct for the inferior planets, for which the epicycle is always inclined to the ecliptic, it is an error for the superior planets, for it means that when the center of the epicycle is in the nodal line, the epicycle is still inclined to the ecliptic and the planet may have latitude. One may say that the superior planets now behave like the inferior planets, showing (1) an 'inclination' in the line of sight when the center of the epicycle is at the limits, $\pm 90^\circ$ from the nodal line, and the epicycle is inclined in the line of sight, and (2) a 'slant' across the line of sight when the center of the epicycle is in the nodal line and the epicycle is inclined across the line of sight. This is wrong because heliocentrically the center of the epicycle is the planet and the epicycle is the orbit of the earth, which is in the plane of the ecliptic; when the planet is in the nodal line, it can have no latitude no matter where the earth is in its orbit. Thus, geocentrically when the center of the epicycle is in the nodal line, the planet should have no latitude no matter where it is on the epicycle, which must therefore lie in the plane of the ecliptic. Ptolemy himself had stated this condition in the *Almagest*, and it is hard to know why he would wish to change it in the *Handy Tables*. Perhaps he meant to write also to use $c_7(\omega_C)$ for the superior planets in computing β_2, and just nodded – as we should prefer to believe – but there is no way of confirming this, and that Theon's instructions are the same as Ptolemy's probably rules out a textual error. Hence, as the tables and instructions have been transmitted, the inclination of the epicycle is fixed.

For the inferior planets, with the inclination of the epicycle fixed, the nodal line of the epicycle is always parallel to the apsidal line. Hence, (1) when the center of the epicycle is in the apsidal line, the nodal line of the epicycle coincides with the apsidal line in the line of sight, the fixed inclination of the epicycle is observed across the line of sight and one sees the 'slant'; (2) when the center of the epicycle is $\pm 90^\circ$ from the apsidal line, the nodal line of the epicycle is across the line of sight, the fixed inclination of the epicycle is observed in the line of sight and one sees the 'inclination'. Thus, both components are now a single latitude resulting from one fixed inclination seen in two ways. So the fixed inclination of the epicycle, which is incorrect for the superior planets, is correct for the inferior planets.

We may investigate the consequences of fixed inclinations by computing from the tables the maximum latitude when the center of the epicycle is at the ascending and descending node of the eccentric for the superior planets, and at apogee and the point opposite apogee, which for Mercury is not perigee, for the inferior planets. Distances from the apogee to the northern limit ω_A are now -40° for Saturn, which in the *Almagest* was -50°, $+20^\circ$ for Jupiter, and 0° for Mars, which are unchanged; hence the eccentric anomaly κ of the ascending node $\kappa_{☊} = \omega_A - 90^\circ$ and of the descending node $\kappa_{☋} = \omega_A + 90^\circ$. For the inferior planets $\kappa = (0^\circ, 180)$. The following table gives κ, the approximate range of β_2', of distances ω_p' from the northern limit of the epicycle at which the maximum latitude occurs – the true anomaly $\alpha = \omega_p' + 90^\circ$ – of the coefficient $c_7(\omega_p')$, and the resulting latitude β_2 computed from the tables by the procedure just explained. For the superior planets, the first computation is for the descending node and the second for the ascending node of the eccentric; for the inferior planets, the eccentric is taken to be in the plane of the ecliptic.

	κ	β_2'	ω_p'	$c_7(\omega_p')$	β_2	κ	β_2'	ω_p'	$c_7(\omega_p')$	β_2
Saturn	50°	0;30°	0–3°	1;0	0;30°	230°	0;32°	3–6°	1;0	0;32°
Jupiter	110	0;31	0–6	1;0	0;31	290	0;31	0–6	1;0	0;31
Mars	90	1;55–2;0	39–42	0;46–0;44	1;28	270	1;55–2;0	39–42	0;46–0;44	1;28
Venus	0	3;31–3;54	45–51	0;42–0;38	2;28	180	4;22	51	0;38	2;46
Mercury	0	2;7–2;9	12–15	0;59–0;58	2;5	180	2;52–2;56	24–27	0;55–0;54	2;38

In fact β_2 may stay within $0;1^\circ$ of these maximum values for a considerable range of ω_p', and roundings to minutes in the tables introduce small irregularities by which the latitude can decrease and again increase. But the important point is that the superior planets may have latitude, a notable latitude, when the center of the epicycle is at a node, when the latitude should be zero. This is a significant error, an error not present in the *Almagest*. For the inferior planets, the maximum latitude is no longer $\pm 2;30^\circ$, but varies with distance as it should, something treated only roughly by the correction to β_2 for Mercury of $\pm\frac{1}{10}c_4(\alpha)$ in the *Almagest*, and the maximum latitudes do not occur together with the maximum equations of the anomaly, as was true for the method of computing β_2 in the *Almagest*. (For this reason, one cannot compute the maximum latitude here as one computes the maximum slant in the *Almagest*, from $\sin\beta_2 = \sin i_2 (r/R')$, as in Figure 6, which is close for the inferior planets, but much smaller for the superior planets.) Neugebauer has computed 22 latitudes for Mars and 17 latitudes for Venus at 10-day intervals to show a more general comparison between the *Handy Tables* and the *Almagest*. The differences in these reach nearly 1° for Mars and are greatest near the nodal line for reasons just explained. The differences for Venus reach about $1;30^\circ$, are greatest where the 'inclination' has its greatest effect, $\pm 90^\circ$ from the apsidal line, and agree better with modern theory for reasons we shall now show.

For the inferior planets, as noted, the method of computation no longer distinguishes the inclination and slant, which nevertheless occur where and as they should, but gives a single latitude due to the inclination of the epicycle, which is added to the small

latitude due to the inclination of the eccentric. The latitude just computed, when the center of the epicycle is in the apsidal line, is the 'slant'. The following table shows the inclination of the epicycle i_1 and the resulting latitude, the 'inclination', at superior conjunction at apogee of the epicycle β_a and at inferior conjunction at perigee β_b from the *Almagest*, *Handy Tables*, and a modern computation to tenths of a degree.

	Almagest			*Handy Tables*			Modern				
	i_1	$\pm\beta_a$	$\pm\beta_b$	i_1	$\pm\beta_a$	$\pm\beta_b$	i_1	$+\beta_a$	$-\beta_a$	$+\beta_b$	$-\beta_b$
Venus	2;30°	1;2°	6;22°	3;30°	1;29°	8;52°	3;22°	1;30°	1;30°	8;36°	8;48°
Mercury	6;15	1;45	4;5	6;30	1;50	4;14	6;58	1;48	2;0	3;48	4;42

The inclination of the epicycle of Venus in the *Handy Tables* is that of $i_2 = 3;30°$ in the *Almagest*, which is about correct and produces far better results for β_a and β_b. The inclination for Mercury, 6;30°, is closer to $i_1 = 6;15°$, although $i_2 = 7;0°$ in the *Almagest* is preferable; the resulting β_a and β_b differ only slightly from the *Almagest*. (In the smaller commentary to the *Handy Tables*, Theon gives β_b for Venus as 8;56° and for Mercury as 4;18°.)

The inclination of the eccentric i_3 is now a fixed inclination of 0;10° for both planets, with the northern limit at the apogee for Venus and the southern limit at the apogee for Mercury. Here too, as for the variable inclination in the *Almagest*, I know of no correct explanation of what Ptolemy actually observed that could account for i_3. The remarks made earlier about the difficulty of observing these latitudes at all still apply, and it is notable that each is at the limit of precision of Ptolemy's observations. They could perhaps have been derived from an effect seen near greatest elongation, as Ptolemy also mentioned in the *Almagest*, rather than near inferior or superior conjunction, but that too would be difficult and the *Handy Tables* contain no explanation.

The latitude tables of the *Handy Tables* seem to have been of little influence, even in tables otherwise based upon Ptolemy's models and parameters, which is of some interest. The correction tables for longitude of the planets in the *Handy Tables* are, directly or indirectly, the basis of many, perhaps most, later tables following Ptolemy, in Greek, Arabic, eventually Latin, differing for the most part only in textual errors or adjustments of interpolation for the intervals of 1° and one commonly altered parameter: Venus is given the equation of center of the sun, as is also found in tables based upon Indian models and parameters, although without adjustment for the effect of distance on the equation of the anomaly, which is so small as to be of no consequence. (The *Alfonsine Tables* also have a different equation of center for Jupiter, of unknown origin, likewise without adjustment of the equation of the anomaly.) However, later Ptolemaic tables for latitude are overwhelmingly based upon the tables in the *Almagest*, some with various modifications, mostly to facilitate computation, that have been described by van Dalen. There are two known partial and curious exceptions. Kennedy (2) lists maximum latitudes of the *Mumtaḥan Zīj*

(*Az-Zīj al-Ma'mūnī lī'l Mumtaḥan*) of the early ninth century, which agree with those given by Theon in his smaller commentary to the *Handy Tables*, although the latitude tables themselves are based upon nothing more than a sine function, which is quite primitive. The latitudes listed by Kennedy, some of which differ slightly from direct computation from the *Handy Tables*, are as follows:

	$+\beta_o$	$-\beta_o$		$\pm\beta_b$
Saturn	3;1°	3;6°	Venus	8;56°
Jupiter	2;3	2;9	Mercury	4;18
Mars	4;23	7;6		

These are maximum latitudes at opposition for the superior planets – Theon gives Saturn $+\beta_o = 3;2°$ – and at inferior conjunction for the inferior planets, although it is not clear just how they are applied in these tables. The same maximum latitudes are attributed by Ibn Hibintā of the mid-tenth century to either the *Zīj as-Sindhind* or the *Zīj ash-Shāh*, with $+\beta_o = 5;23°$ for Mars, doubtless a textual error; since the former is based on Indian models and parameters and the latter on a Pahlavi translation of an Indian original, latitudes from the *Handy Tables* seem out of place. Nevertheless, the tables and report do show that maximum latitudes from Theon's commentary were known and used, here it appears in rather crude adaptations.

The other example comes from a very different period and region, the *Zīj al-Muqtabis* of Ibn al-Kammād, who lived in Andalusia in the twelfth century, which survives in a Latin version made in Palermo in 1260 by one John of Dumpno, of which there is a study by Chabás and Goldstein. The latitude tables for the superior planets are those of the *Almagest*, but those for the inferior planets are from the *Handy Tables*, at 6° intervals and with a maximum for Venus in $c_5(\alpha)$ of 8;35° instead of 8;51°, presumably a textual error. The same tables are found in the *Tables of Barcelona* of the fourteenth century, published by Millás Vallicrosa and analyzed by Chabás; here the maximum for Venus in $c_5(\alpha)$ is 8;55°. These sources show that, in addition to Theon's commentary, the latitude tables themselves were known in Arabic, and it was the choice of Ibn al-Kammād, or of his source whatever that may have been, to use only those for the inferior planets. Why? Could someone have understood that the tables are in error for the superior planets, but not for the inferior planets, because the fixed inclination of the epicycle gives the planet latitude even when the center of the epicycle is in the nodal line? One would not need observations to understand this, just an understanding of the latitude theory in the *Almagest* and an ability to figure out that the computation in the *Handy Tables* implies a fixed inclination of the epicycle, which is not exactly obvious as the model is not explained. Still, the way astronomical tables were often haphazardly and inconsistently thrown together for hundreds of years, this is a higher level of understanding than one is accustomed to. It is obvious, for example, that Theon did not recognize a problem. Another possible explanation is that the latitude tables for the inferior planets in the *Almagest* were considered too complicated, which is true enough with all the rules for applying the coefficient $c_5(\kappa)$,

although those in the *Handy Tables* are also complicated. Perhaps it is safer to confess that we have no idea why only the latitude tables for the inferior planets have been found, and someday a source containing the tables for the superior planets may be discovered. Benno van Dalen, who, in answer to my inquiry about latitude tables from the *Handy Tables* in Arabic referred me to Ibn al-Kammād's tables, informs me that thus far he knows of none for the superior planets, so if they exist at all they must be very uncommon.

PLANETARY HYPOTHESES

The *Planetary Hypotheses* may be considered Ptolemy's last word on latitude theory, in which he finally got nearly everything right. The models in the *Planetary Hypotheses* have the same inclinations as the *Handy Tables* for the inferior planets, but the superior planets differ from both the *Almagest* and the *Handy Tables*, and are definitely improved, one reason for believing that the *Hypotheses* is later than the *Handy Tables*. For the superior planets the epicycles now have fixed inclinations to the eccentric parallel to the ecliptic, that is, the inclination of the epicycle to the eccentric $i_1 + i_2 = i_1$, and thus correctly $i_2 = 0^\circ$. The inclinations of the eccentric, for Saturn $i_1 = 2;30^\circ$, for Jupiter $i_1 = 1;30^\circ$, are unchanged, but for Mars $i_1 = 1;50^\circ$, which is correct. It is easy to show that these inclinations produce quite accurate latitudes at conjunction and opposition at the limits. We have only to reverse Ptolemy's procedure for finding the inclinations from the correction tables for longitude in *Almagest* 11.11. Thus, in Figure 8, where i_1 is proportional to the 'anomaly' α measured from apogee or perigee of the epicycle, and ϑ is proportional to the equation of the anomaly c, so that $\vartheta/c = i_1/\alpha$, we have at both conjunction and opposition $\vartheta = (c/\alpha)i_1$, from

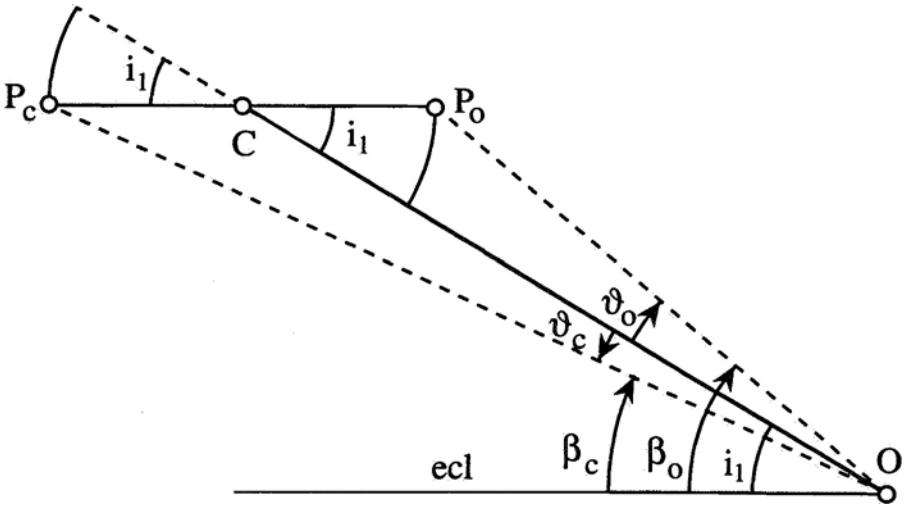

Figure 8. Latitudes β_c and β_o from inclination i_1 in *Planetary Hypotheses*.

which $\beta_c = i_1 - \vartheta_c$ and $\beta_o = i_1 + \vartheta_o$. In the computation we take the distance of the northern limit from apogee of the eccentric as: Saturn -40°, Jupiter $+20^\circ$, Mars 0°, and we take c from the correction tables for $\alpha = 3^\circ$. The results are shown along with modern inclinations and latitudes to $0;3^\circ$.

	i_1	$+\beta_c$	$-\beta_c$	M. $\pm\beta_c$	$+\beta_o$	$-\beta_o$	M. $\pm\beta_o$	M. i_1
Saturn	2;30°	2;16°	2;14°	2;18°	2;48°	2;51°	2;51°	2;33°
Jupiter	1;30	1;16	1;15	1;9	1;50	1;53	1;45	1;25
Mars	1;50	1;9	1;3	1;9–1;6	4;34	6;46	4;36–6;54	1;52

The most notable improvement over Ptolemy's earlier theories is that β_c for Mars is now correctly just over 1° where before it was very close to 0°; every other value is also improved, with no error exceeding $0;8^\circ$, and of course when the center of the epicycle is at the nodal line the latitudes are correctly zero. (Had Ptolemy not rounded i_1 for Jupiter from $1;24^\circ$ to $1;30^\circ$ in the *Almagest*, its latitudes would be better, but the rounding of i_1 for Saturn from $2;26^\circ$ to $2;30^\circ$ is an improvement.) Intermediate latitudes, not at conjunction or opposition at the limits, or at the nodes with zero latitude, may not be as accurate as these, but the problems are due to errors in longitude theory and distance, not to the latitude theory itself. The latitudes of the inferior planets are the same as those of the *Handy Tables* tabulated earlier. One may now say that Ptolemy has corrected both the variable inclinations of the epicycles in the *Almagest* and, for the superior planets, the fixed inclinations not parallel to the ecliptic in the *Handy Tables*.

The subject of the first part Book I of the *Hypotheses*, where these parameters are given, is the construction of instruments, analogue computers consisting of graduated circles, eccentric and epicyclic, within a concentric graduated zodiac, for finding longitudes of the sun and longitudes and latitudes of the moon and planets without correction tables. They are the earliest known examples of what were later called 'equatoria'; later specimens, and there are many, generally do not include latitude. Kennedy (1) has described an instrument of the early fifteenth century for computing latitudes by the ingenious Jamshīd Ghīāth ad-Dīn al-Kāshī, based upon the latitude theory of the *Almagest*, in which the latitudes are projected into a plane. Ptolemy's instruments in the *Hypotheses*, however, actually have inclined rings. How one could make these things, presumably of metal, with tiny inclinations of about 2° to 6°, even $0;10^\circ$ for the eccentrics of Venus and Mercury, is not a trivial question, and perhaps their use for latitude was only theoretical. One might suggest that the fixed inclinations of the epicycles in the *Hypotheses* were a simplification for making such instruments, as who could possibly construct the variable inclinations of the *Almagest* with their small vertical circles? But this cannot be so, first because the latitude theory of the fixed inclinations, with the new and correct inclination for Mars, really is superior to the variable inclinations in the *Almagest*, and to the fixed inclinations in the *Handy Tables*, surely deliberately so, and second because the spherical models described in

Book II, which are supposed to be the physical mechanisms in the heavens responsible for the motions of the planets, have the same fixed inclinations, for the superior planets parallel to the ecliptic. Of course historically the more important parts of the *Planetary Hypotheses* are the distances and sizes of the planets and the spherical models, the cosmology and physical astronomy, on which the principal studies are by Goldstein and Murschel, and the latitude theory, for all its ingenuity, seems to have been without influence, for it appears that no one made use of Ptolemy's final, corrected latitude theory.

How Ptolemy made these changes, he does not say except to remark that by continuous observations he has made corrections, compared to the *Almagest*, of the hypotheses themselves or of their proportions or periodic times, and in fact the *Planetary Hypotheses* contains various changes in the hypotheses and their parameters, including periods, of which the changes in the hypotheses and inclinations for latitude are the most significant. So one must conclude that he corrected the theory of latitude of both superior and inferior planets from his own observations, improving upon the rough values of extreme apparent latitudes, which we believe to have been conventional estimates, used earlier in the *Almagest* when they were all he had. Since it was not possible in the years of his observations to observe all the planets under the special conditions used to derive the inclinations in the *Almagest*, for the superior planets at opposition and near conjunction at the limits, he must have derived the inclinations from other sorts of observations. One possibility is a series of oppositions over several years with a large although not maximum latitude, showing by computation both the inclination i_1 and, from $i_1 + i_2 \approx i_1$, that $i_2 = 0^\circ$, that the epicycle remains parallel to the ecliptic, which would require two oppositions for each planet. In this way he could correct both the *Almagest* and the *Handy Tables*. But on such things one can only speculate, fully aware of the difficulty of finding these parameters even from accurate observations. However Ptolemy did it, he got it right.

It is of interest that John Bainbridge, Henry Savile's successor as professor of astronomy at Oxford, who edited and translated the part of Book I of the *Planetary Hypotheses* surviving in Greek in its first edition in 1620, specifically noticed the changes in the theory of latitude, which otherwise remained unknown until our own time. After stating that the three manuscripts of the *Hypotheses* he used were corrupt and incomplete, requiring much emendation by comparing the text with the *Almagest*, he remarks: 'In the hypotheses of latitudes I desired to change nothing. For wise (*prudens*) Ptolemy, if I judge this matter, departed from the hypotheses established in the *Syntaxis* and proposed other easy, convenient, and far truer (*longeque veriores*) hypotheses.' This observation is interesting, not only for noticing the differences in the hypotheses, but in recognizing that those in the *Planetary Hypotheses* are 'far truer' than those in the *Almagest*. The only way he could know this in 1620 was by having read, and approved, Kepler's *Astronomia nova* (1609), in which it is shown (13–14) using Tycho's observations that the inclination of the plane of Mars's eccentric to the ecliptic is about 1;50° and fixed, implying in a geocentric model a fixed inclination of the epicycle parallel to the ecliptic, just as Ptolemy found. And Kepler computes (65) the extreme latitudes: northern limit, opposition +4;31,45°, conjunction +1;8,30°;

southern limit, opposition $-6;52,20^\circ$, conjunction $-1;4,20^\circ$, all very close to the extreme latitudes from Ptolemy's theory. Remarkably, Ptolemy's latitude theory in the *Planetary Hypotheses* anticipates Kepler, as Bainbridge recognized. In a (lost) letter of 24 February 1625 Bainbridge brought the revised latitude theory to Kepler's attention, and later that year Kepler received Bainbridge's publication and noted the fixed inclination of $1;50^\circ$ for Mars.

We conclude with a table of the inclinations in the *Almagest*, *Handy Tables*, and *Planetary Hypotheses* with modern values for A.D. 100 from P.V. Neugebauer, the same ones quoted earlier, which are also given by O. Neugebauer. In the *Almagest*, Ptolemy gives the inclination of the epicycle of the superior planets as $i_1 + i_2$, and the same inclination applies in the *Handy Tables* although it is not explicitly given. The inclination of the epicycle i_2, to a plane parallel to the plane of the ecliptic, is variable in the *Almagest* from its given value to 0° but fixed in the *Handy Tables*; in the *Planetary Hypotheses* the epicycle for the superior planets is parallel to the ecliptic so that $i_2 = 0^\circ$. The two variable inclinations of the epicycle for the inferior planets i_1 and i_2 in the *Almagest* are replaced with a single fixed inclination we call i_1 in the *Handy Tables* and *Planetary Hypotheses*. The inclination of the eccentric for Venus and Mercury i_3 is variable in the *Almagest* from its given value to 0° but fixed in the *Handy Tables* and the *Planetary Hypotheses*. We have marked variable inclinations v; all others are fixed. Aside from the component i_3 for the inferior planets, it is obvious that in the *Planetary Hypotheses*, Ptolemy has reached something close to modern latitude theory, with a single defective inclination for Mercury, for which i_2 from the *Almagest* would have been correct. No one did better until Kepler in the *Epitome of Copernican Astronomy* and the *Rudolphine Tables*.

	Almagest			Handy Tables			Planetary Hypotheses		Modern
Planet	i_1	$i_2 v$	$i_3 v$	i_1	i_2	i_3	i_1	i_3	i_1
Saturn	$2;30^\circ$	2°	–	$2;30^\circ$	2°	–	$2;30^\circ$	–	$2;33^\circ$
Jupiter	1;30	1	–	1;30	1	–	1;30	–	1;25
Mars	1;0	1;15	–	1;0	1;15	–	1;50	–	1;52
Venus	$2;30v$	3;30	$+0;10^\circ$	3;30	–	$0;10^\circ$	3;30	$0;10^\circ$	3;22
Mercury	$6;15v$	7;0	$-0;45$	6;30	–	0;10	6;30	0;10	6;58

APPENDIX

CORRECTION TABLES FOR LONGITUDE IN THE *ALMAGEST* AND *HANDY TABLES*

The correction tables for longitude in the *Almagest* are used to derive the inclinations of the planes of the eccentric and epicycle in the theory of latitude, and the latitude tables in the *Handy Tables* control the effect of distance of the center of the epicycle in the same way as the correction tables for longitude. For these reasons, we include brief

descriptions of the correction tables for longitude and of the formation of equations between mean and true motions. The reader is referred to Pedersen or Neugebauer for more detailed treatment of the tables in the *Almagest*, including the method of computation, and to Neugebauer for the *Handy Tables*. We show an excerpt from the table for Mars from *Almagest* 11.11.

1	2	3 ($\bar{\kappa}$)	4 ($\bar{\kappa}$)	5 ($\bar{\alpha}$)	6 ($\bar{\alpha}$)	7 ($\bar{\alpha}$)	8 ($\bar{\kappa}$)
6°	354°	1;0°	+0;5°	0;8°	2;24°	0;9°	−0; 59, 53
12	348	2;0	+0;10	0;16	4;46	0;18	−0;58,59
18	342	2;58	+0;15	0;24	7;8	0;28	−0; 57, 51
...	...	...	...	...	...	...	...
96	264	11;29	−0; 4	2;42	35;6	3;6	−0; 3, 3
99	261	11;32	−0; 8	2;49	35;56	3;15	+0;0,5
102	258	11;32	−0; 12	2;56	36;43	3;25	+0;3,13
...	...	...	...	...	...	...	...
174	186	1;30	−0; 10	2;27	11;15	4;26	+0;59,43
177	183	0;45	−0; 5	1;16	5;45	2;20	+0;59,52
180	180	0;0	−0; 0	0;0	0;0	0;0	+1;0,0

Columns 1 and 2 are arguments of entry for 6°–180° and 180°–354° at intervals of 6° for 270°–90° and 3° for 90°–270°. Column 3, a function of the mean eccentric anomaly $\bar{\kappa}$, the uniform motion of the center of the epicyle from apogee of the eccentric measured at the equant point, is the equation of center for an eccentric circle of ecentricity $2e$, from the earth to the equant point. Column 4, also a function of $\bar{\kappa}$, is a correction for the bisection of the eccentricity for an eccentric circle of eccentricity e, from the earth to the center of the eccentric. Columns 3 and 4 are added to find the equation of center $c_3' = c_3 + c_4$. The true eccentric anomaly κ is computed from $\bar{\kappa}$ by $\kappa = \bar{\kappa} + c_3'$, where $c_3' < 0°$ for $\bar{\kappa} < 180°$ and $c_3' > 0°$ for $\bar{\kappa} > 180°$. The true anomaly on the epicycle α is computed from the mean anomaly $\bar{\alpha}$ by $\alpha = \bar{\alpha} + c_3'$, where $c_3' > 0°$ for $\bar{\kappa} < 180°$ and $c_3' < 0°$ for $\bar{\kappa} > 180°$.

Columns 5–7, all functions of the true anomaly α, are used to compute the equation of the anomaly c due to motion of the planet on the epicycle: c_6 is the equation for the center of the epicycle at mean distance on the eccentric R, c_5 a subtraction for greatest distance $R + e$, and c_7 an addition for least distance $R - e$. For Mercury the distances are $R + 3e$ and $R- \sim \frac{3}{2}e$ respectively. Finally, column 8, a function of $\bar{\kappa}$, is a coefficient of interpolation for intermediate distances of the center of the epicycle, extending from -1 at greatest distance to $+1$ at least distance. The equation of the anomaly c is computed from either of

$$c = c_6(\alpha) + c_8(\bar{\kappa}) \cdot c_5(\alpha), \quad \text{if } c_8(\bar{\kappa}) < 0, \qquad c = c_6(\alpha) + c_8(\bar{\kappa}) \cdot c_7(\alpha), \quad \text{if } c_8(\bar{\kappa}) > 0.$$

In the *Handy Tables*, the entries in c_1 and c_2 are at intervals of 1°; the equation of center, $c_3 + c_4$ in the *Almagest*, is combined into a single c_3; c_4 is the coefficient of interpolation for computing the equation of the anomaly, computed as a function of

the true eccentric anomaly $\kappa = \bar{\kappa} \pm c_3$ and rounded to one fractional place; columns c_5, c_6, c_7 for the equation of the anomaly are the same as in the *Almagest*. The contraction of intervals to 1°, which facilitates the use of the tables, is done by linear interpolation in the tables in the *Almagest* although there are small discrepancies of no consequence.

Department of Astronomy and Astrophysics, University of Chicago, USA

REFERENCES

Ptolemy: Texts and Translations

Claudii Ptolemaei Opera quae exstant omnia. Ed. J. L. Heiberg. I. *Syntaxis mathematica*. 2 pts. 1898, 1903. II. *Opera astronomica minora*. 1907.

Ptolemäus. *Handbuch der Astronomie*. Trans. K. Manitius. 2 vols. Leipzig. 1912–13.

Ptolemy's Almagest. Trans. G.J. Toomer. New York. 1984.

Tables manuelles astronomiques de Ptolemée et de Theon. Ed. N. Halma. 3 pts. Paris. 1822, 1823, 1825.

The Astronomical Tables of Codex Vaticanus Graecus 1291. Ed. W.D. Stahlman. Brown University Dissertation. 1959.

Le "Petit Commentaire" de Théon d'Alexadrie aux Tables Faciles de Ptolémée. Ed. and trans. A. Tihon. *Studi e Testi* 282. Città del Vaticanó. 1978.

Procli sphaera, Ptolemaei de Hypothesibus Planetarum liber singularis. Ed. and trans. J. Bainbridge. London. 1620.

The Arabic Version of Ptolemy's Planetary Hypotheses. Ed. and trans. B.R. Goldstein. *Trans. Am. Phil. Soc.* 67. Philadelphia. 1967.

Ptolemy: Studies

Delambre, J.B.J. *Histoire de l'astronomie ancienne*. 2 vols. Paris. 1817. 2. 393–408.

Hamilton, N.T., N.M. Swerdlow, G.J. Toomer. The Canobic Inscription: Ptolemy's Earliest Work. *From Ancient Omens to Statistical Mechanics. Essays on the Exact Sciences Presented to Asger Aaboe*, Ed. J.L. Berggren and B.R. Goldstein. Copenhagen, 1987, 55–74 (67–68).

Herz, N. *Geschichte der Bahnbestimmung von Planeten und Kometen*. 2 vols. Leipzig. 1887–1894. 1. 143–159.

Murschel, A. The structure and Function of Ptolemy's Physical Hypotheses of Planetary Motion. *Journal for the History of Astronomy* 26 (1995), 33–61.

Neugebauer, O. *A History of Ancient Mathematical Astronomy*. 3 pts. New York. 1975. 206–226, 1006–1016.

Pedersen, O. *A Survey of the Almagest. Odense*. 1974. 355–386.

Riddell, R.C. The Latitudes of Venus and Mercury in the Almagest. *Archive for History of Exact Sciences* 19 (1978), 95–111.

van der Waerden, B. L. Bemerkungen zu den Handlichen Tafeln des Ptolemaios. *Sitzungsber. d. Bayerische Akad. d. Wiss*. Math.–Natur. Kl. (1953), 261–272 (265–270).

Later Texts and Studies

Chabás, J. Las Tablas de Barcelona. *From Baghdad to Barcelona. Studies in the Islamic Exact Sciences in Honour of Prof. Juan Vernet*. ed. J. Casulleras and J. Samsó. 2 vols. Barcelona. 1996. 1. 477–525 (504–505).

Chabás, J. and B. R. Goldstein. Andalusian Astronomy: *al-Zīz al Muqtabis* of Ibn al-Kammād. *Archive for History of Exact Sciences* 48 (1994), 1–41 (31–32).

Kennedy, E. S. (1). An Islamic Computer for Planetary Latitudes. *Journal of the American Oriental Society* 71 (1951), 13–21. Reprinted in E. S. Kennedy. *Studies in the Islamic Exact Sciences*. Beirut. 1983. 463–471.

Kennedy, E. S. (2). *A Survey of Islamic Astronomical Tables*. *Trans. Am. Phil. Soc.* 46 (1956), 24, 50–51.

Kepler, J. *Astronomia nova*. *Johannes Kepler Gesammelte Werke* 3. Ed. M Caspar. Munich. 1937, 133–142, 396–397. *Briefe 1620–1630*. *Gesammelte Werke* 18. 1959, 244, 253.

Millás Vallicrosa, J. M. *Las Tablas Astronómicas del Rey Don Pedro el Ceremonioso*. Madrid. 1962. 150–151, 226–227.

Swerdlow, N. M. and O. Neugebauer. *Mathematical Astronomy in Copernicus's De Revolutionibus*. 2 pts. New York. 1984. 489–491.

van Dalen, B. Tables of Planetary Latitudes in the *Huihui Li*. *Current Perpsectives in the History of Science in East Asia*. Ed. Y. S. Kim and F. Bray. Seoul, 1999, 316–29.

Auxiliary Sources

Neugebauer, P. V. *Tafeln zur astronomischen Chronolgie*. II. *Tafeln für Sonne, Planeten und Mond*. Leipzig. 1914.

Neugebauer, P. V. Tafeln zur Berechnung der jährlichen Auf- und Untergänge der Planeten. *Astronomische Nachrichten* 264 (1938), Nr. 6331, 313–322.

Tuckerman, B. *Planetary, Lunar, and Solar Positions A.D. 2 to A.D. 1649 at Five-day and Ten-day Intervals. Mem. Am. Phil. Soc.* 59. Philadelphia. 1964.

WILLIAM R. NEWMAN AND LAWRENCE M. PRINCIPE

ALCHEMY AND THE CHANGING SIGNIFICANCE OF ANALYSIS

It is hardly necessary to emphasize the fact that alchemy has long held a privileged position in the popular mind as the epitome of "wrong" and misguided science. Whether viewed as the product of an ignorant and clownish empiricism, the embodiment of dishonest greed, or the vehicle of attempts to attain a mystical union with divinity, alchemy has only rarely received praise for its scientific content. Until quite recently, the subject much more commonly served moderns as an object of ridicule. The seeds of this usage were already planted in the Middle Ages themselves, when such hapless alchemists as Dante's Griffolino d'Arezzo and Capocchio da Siena were consigned to the rigors of the Inferno and Chaucer's Canon's Yeoman to the infernal duplicity of alchemical charlatans. Later historians have, more often than not, drawn upon such negative images of alchemy and its practitioners for a foil against which more "progressive" historical developments could be made to stand in higher relief. This is a historiographical motif that extends back through Charles Mackay's famous *Memoirs of Extraordinary Popular Delusions and the Madness of Crowds* to the Enlightenment in one direction, and forward to Shapin and Schaffer's *Leviathan and the Air Pump* in the other.[1]

On the other hand, there have also been some historians of chemistry and chemists who have themselves maintained that alchemy was a fruitful ancestor of their subject. One thinks immediately of Justus Liebig, Hermann Kopp, Marcelin Berthelot, and J. R. Partington, who were all significant chemists as well as participants in the historiography of chemistry.[2] But here too a problem emerges, in that such authors have tended to extract what they viewed as chemically significant from its historical context, and to ignore the vast reams of text that contained no material of obvious relevance to their views on the development of chemistry. Any attempt to isolate what was scientifically "right" about alchemy will run full force into the same problem – the positivist tendency to extract precious nuggets of fact while ignoring such problems as historical context, the role of theory in its interaction with practice, and the overall purpose of the works in question. Nonetheless, the historian of this subject must confront a stark and ineluctable choice between adopting the picture of alchemy as folly that derives from the tradition already begun by Dante and Chaucer and passed on by Pieter Bruegel the Elder and his heirs among the Netherlandish genre painters, or arguing that the field contributed something to the science and technology of its time.[3] Since the default picture of alchemy in the mind of most scientists, philosophers, and general historians as well belongs more to the first category than the second, a strategy of silence entails a tacit capitulation to the received view.

J.Z. Buchwald and A. Franklin (eds.), Wrong for the Right Reasons, 73–89.

It is possible, however, to illuminate some trends in the history of alchemy that were significant for later science without engaging in the special pleading of the positivists. The key feature to keep in mind, in our view, is what the theories and practices meant to those who were employing them, rather than isolating them from their original context and seeing them only from the privileged perspective of historical hindsight. A particular and relevant example lies in the area of chemical analysis, which occupied an increasingly important position in alchemy and which became ever more sophisticated and quantitative during the early modern period. The subject is of interest not only because of its obvious significance for later chemistry, but because it reveals that alchemy was itself a dynamic field evolving under the influence of laboratory practices, and undergoing metamorphoses that paralleled similar moves towards greater emphasis on probatory experiments and on quantification in other fields during the 17th century.

In the following we will argue that the emphasis on analysis in alchemy originated in the dual context of mimicking and assaying the precious metals. Gradually developing in sophistication over a number of centuries, analytical processes began to assume a much more quantitative character in alchemical treatises after the popularizing of the field by Paracelsus von Hohenheim in the early 16th century. It is precisely in the ambit of Paracelsus's followers and revisers that one begins to see a new emphasis on analysis – and importantly synthesis as well – as the defining operations of the chymical art.[4] By the time of Joan Baptista Van Helmont in the first third of the 17th century, one can speak of a consciously employed model in which chymistry had become the art of analysis and synthesis par excellence, viewed in explicitly gravimetric terms. This is not to say, of course, that every chymist adhered to the Helmontian model employing quantitative analysis and synthesis, but the fact that some of the most influential, such as the Harvard-educated George Starkey, did do so, even in his chrysopoetic writings penned under the *nom de guerre* of Eirenaeus Philalethes, is highly significant for our understanding of alchemical practice and thought in this period.

Interestingly, much of the traditional concern with analysis in alchemy occurred in the context of assaying processes. It may seem paradoxical at first glance that alchemists would have wished to develop and transmit tests that held every promise of revealing the products of transmutation as false, but the irony is only apparent. If one can pass beyond the stereotype of alchemists as deceitful charlatans, then an alchemist's inclusion of tests for the determination of the accuracy of his or her results becomes an obvious desideratum. Indeed, once these techniques are seen as tests for chemical identity, quality, or purity, their existence and deployment among hopeful transmuters of metals casts the latter into a new light, not simply as perpetrators of fraud but as seekers after what amount to technological processes. Furthermore, numerous chrysopoetic and argyropoetic recipes require the presence of real gold or silver in order to succeed, since many of these processes involve "seeding" a base metal or mixture with a precious metal in the hope that this would produce more of the precious metal. From a modern perspective such recipes actually involved debasing precious metals, and could have worked only if a certain percentage of genuine gold

or silver entered into the desired alloy. It is understandable, then, that docimastic recipes often appear in alchemical texts with the expressed purpose of eliminating fraudulent or debased gold or silver at the outset as ingredients of recipes. A third area where assaying processes appear in alchemical texts occurs in the very context of making artificial gold or silver. The two classic tests of cupellation and cementation, which both worked by eliminating base metals (and in the case of cementation, by eliminating silver as well) from an alloy, could themselves be viewed in such a way that they appeared to be "graduating" or improving an imperfect metal rather than separating impurities from it.

Cupellation worked by placing the alloy to be tested in a porous vessel, made primarily of bone-ash in most cases, and heating the metal to the point of fusion. Lead was added to the fused metal, and blasted with a current of air. The lead would oxidize, and the resulting yellow litharge would partly sink into the cupel walls, and partly float to the surface, where it could be skimmed off or blown away by the current of air. If the test was successful, a button of relatively pure gold and silver would be left at the bottom of the cupel. Since cupellation alone could not separate gold from silver, a further process was required for this purpose in the days before the discovery of the mineral acids. This additional process, cementation, worked by placing leaves of the gold–silver alloy in a crucible with alternating layers of the "cement," often a mixture of brick-dust and salt. The sealed crucible was heated to a high temperature, but one beneath the melting point of the alloy, whereupon the silver was corroded but the gold left unscathed. In order to make cupellation and cementation meaningful as assaying tests, one obviously had to weigh the initial and final samples of the metals undergoing the assay, for it was the very loss of weight during the test that revealed the impurity or imperfection of the metal. If, on the other hand, one did not concern oneself with weight measurements, it was possible to view these processes as means of "ennobling" poor-quality gold or silver in the case of cupellation or gold alone in that of cementation.

Thus chrysopoetic and argyropoetic texts from late antiquity onward prominently display the assaying processes of cupellation and cementation. Cupellation appears already in one of the foundation texts of alchemy, the Leiden Papyrus, a collection of recipes dating from the 4th century A.D. devoted mostly to the manufacture and improvement of dyes, ersatz precious stones and metals.[5] In the midst of 10 recipes devoted to such topics as softening silver, whitening copper, hardening tin, turning copper objects into silver ones, making copper look like gold, and manufacturing ersatz silver, one finds a recipe for cementing gold with *misy* (iron pyrite), alum, and salt.[6] The recipe, which is presented as a means of "purifying the brilliance of gold" (*diakatharsis lamprotētos chrysou*) is immediately followed by a "purification of silver" (*argyrou katharsis*) by means of cupellation with lead (Recipe 25). The two recipes could succeed, respectively, in eliminating the major impurities from genuine gold and silver, and yet both are found in the immediate context of recipes for producing fraudulent or debased gold and silver. It is more than likely that the compilers of the Leiden papyrus saw no irony in this, but viewed all these processes as being similar in kind. Just as one argyropoetic recipe "improved" copper and tin

by giving them the appearance of "fusible silver" (Leiden papyrus recipe 28), so the recipes for cementation and cupellation "improved" the precious metals by giving them more shininess, ductility, and other properties associated with the best gold and silver. Later recipes carry this tendency even further, when cementation, for example, is used to induce the surface enrichment of a gold alloy by eliminating the base metals from the outer layers of an artifact.[7]

In other alchemical texts, as we have mentioned, assaying processes are used not only to enhance the appearance of precious metals and their alloys, but to detect fraud. They appear, for example, in the *De anima in arte alkimiae* of pseudo-Avicenna, translated into Latin in the 12th or 13th century, and employed extensively by the well-known scholastic Roger Bacon. The *De anima* lists seven tests for distinguishing natural gold from "gold of the magistery," that is, the product of chrysopoeia. Interestingly, the *De anima* is unclear as to whether chrysopoetic gold – even that which has been produced by the very means prescribed in the text – can ever pass these tests. The tests include fusion with unspecified salts that will attack alchemical gold but leave the natural variety untouched, use of the touchstone, determination of specific weight, exposure to high temperature in order to detect color changes in artificial gold, determination of volatility in a vessel of sublimation (natural gold is not volatile), presence or absence of boiling upon fusion, and taste.[8]

Similar tests are found in the "Bible" of the medieval alchemists, the *Summa perfectionis* of Geber, composed around the end of the 13th century. They are again presented with the explicit goal of distinguishing false from genuine gold, but the *Summa* differs from the *De anima* in explicitly arguing that properly manufactured alchemical gold and silver behave the same upon testing as do the naturally generated precious metals. The *Summa* first advises "manifest" determinations of weight, color, and malleability, that is, tests requiring no chemical reactions. But this is only the beginning – the *Summa* then prescribes that the alchemist see if his metal incandesces before fusion (a crude measure of melting point), blackens after fusion, effloresces upon exposure to acid vapors, changes color when quenched in aluminous water or sulfur, alters in weight or color when heated with sulfur, changes weight, volume, or color when repeatedly calcined and reduced, and amalgamates easily with quicksilver.[9] In addition to this battery of texts, the *Summa* gives extremely clear descriptions of cupellation and cementation, again with the explicit purpose of distinguishing failed attempts at transmutation from genuine gold and silver (whether natural or artificial).

The *Summa*, unlike most previous alchemical texts, goes into considerable detail in explaining how cupellation and cementation work, and the author unequivocally presents these tests as examples of the analysis of an alloy rather than the ennoblement or graduation of an imperfect precious metal. Employing a thoroughly corpuscular theory of matter, the *Summa* argues that the mercurial particles making up gold and silver are of sufficient size that they do not sublime upon intense heating, but rather only separate somewhat from one another so that the metal fuses in an incandescent heat. This is related to the author's theory of *mediocris substantia*, according to which

particles of excessively small size are too volatile to withstand long heating without passing off, and particles of excessively large size introduce gross interparticular pores. Unlike the situation with the base metals, the micro-structure of gold and silver is not interrupted by large particles of an earthy impurity, which would impede the cohesion of the metallic particles and introduce excessive porosity. This earthy purity abounds most of all in lead, which means that it is the quickest metal to separate from an alloy. The rapidity of its separation also means that it can be added to other metals during the test, for it will draw their impurities off with itself. The *Summa* gives essentially the same explanation for cementation that it gave for cupellation, except that the earthy impurity is now explicitly linked with the alchemical principle sulfur. The author clearly says that the effectiveness of both tests results, ultimately, from the fact that there is a correlation between metallic "perfection," exemplified above all in gold, and homogeneity – both of substance and particle size. It is the heterogeneity of particles due to the presence of earthiness or sulfur that results in the passage of the base metals from the cupel into its porous walls, and it is the same heterogeneity that leads to the corrosion of all but gold in the crucible of cementation.[10]

Geber's *Summa* clearly employed cupellation and cementation as means of analyzing alloyed metals into their constituents. Equally manifest is the fact that the author weighed his samples before and after testing – as he explicitly says, "both in the tests described by us and in those to be described, if the altered body should exchange any of the differences (i.e., qualities) of perfection – namely of weight or color – the artifex has not rightly investigated the work."[11] What we do not find Geber doing, however, is extending his gravimetric and analytical project to materials beyond the metals and a few minerals, or attempting to resynthesize the starting material from the separated products of his analysis. The restricted scope of analysis in the *Summa perfectionis* is a direct reflection of the purpose that Geber has in mind for his battery of tests: on the one hand, they serve as indicators for the alchemist of the success or failure of his transmutational techniques, while on the other, they help to reveal the constituents of the metals, thus providing a basis for Geber's theory of metallic generation within the earth. Because he hopes to model his own transmutational efforts on the example of nature, Geber's theory of metallic generation in turn guides his choice of materials for converting the base metals into precious ones. Since nature employs mercury and sulfur rather than "a quickly terminable humidity" such as oils derived from plants and animals, there is no reason for Geber to extend his analytical project beyond the mineral realm.[12] At the same time, Geber explicitly rejects the project of synthesizing the metals directly from mercury and sulfur themselves. Although art mimics nature "insofar as it can," there are limits to what man can achieve – he can purge the base metals of their unfixed, dirty sulfur and mercury and replace these defective principles with mercury and sulfur that have been corrected in the laboratory, but he cannot simply make metals *de novo*. Hence Geber is largely unconcerned with resynthesis, just as he finds no reason to push his analytical project beyond the mineral realm. In both cases, the interaction of theory and practice and Geber's specific goals determine the limits of analysis in the *Summa perfectionis*.

THE WIDER CONTEXT OF ANALYSIS IN ALCHEMY

Up to now we have restricted our discussion to the presence of metallurgical processes and tests in alchemy. In order to understand the eventual centrality that analysis and synthesis acquired in early modern chymistry, however, one must cast a wider net, both in terms of the materials analyzed and the various meanings attached to their decomposition. Already in the 10th century, if not earlier, Arabic writers employing the *nom de plume* of Jābir ibn Ḥayyān were describing their attempts to separate a wide variety of substances into the four elements or elemental qualities by means of laboratory processes such as distillation and sublimation, largely as a means of acquiring the *iksīr* (elixir), an agent of transmutation.[13] The same project was adopted and transmitted to the West both in Latin translations taken from the Jābirian corpus and by works such as the pseudo-Avicennian *De anima in arte alkimiae*. Pseudo-Avicenna speaks, for example, of the fractional distillation of milk. The first product to pass over will be the element water, which is "clearer than tears," then a less clear, yellow fluid distills – this is air, while the black earth that remains in the bottom of the flask will be earth, and the vapors and fumes that pass off are fire.[14]

The Jābirian (and pseudo-Avicennian) expansion of alchemical analysis into non-metallurgical realms received additional impetus in the *Testamentum* of pseudo-Ramon Lull, a work apparently composed in the 1330s. This important text forms the nucleus of a vast pseudo-Lullian corpus of over one hundred texts, thus rivalling Geber's *Summa perfectionis* in terms of its influence.[15] The *Testamentum* assimilates alchemical analysis to the very creation of the universe by God – the Creator first made the world by fabricating a quintessence, which he then divided into three portions of descending purity. The first became the habitation of the angels, the second the realm of the planets and stars, and the third, most impure part became the stuff of the four terrestrial elements. In turn, the Creator divided the elemental Ur-stuff into five parts, one also called "quintessence," and the other four the traditional elements – fire, air, water, and earth. As a result of Adam's fall, pseudo-Lull continues, the four elements, which were originally pure and clear, have become corrupted, and they grow more corrupt daily. Only at the end of time will they be restored to their original state of clarity, when a vast conflagration will purge the elemental world of its dross and restore it to its prelapsarian perfection. The alchemist, however, can enact a sort of a local version of the conflagration in his furnaces by subtly burning out the impurities of the elements in material substances. By this means he can arrive at the great transmutational agent of the chrysopoetic art – the philosophers' stone. As pseudo-Lull puts it, the elements must first receive a "philosophical purgation" which will allow the alchemist to extract a "crystalline and resplendent" matter corresponding to the pure material out of which they were first made.[16] Pseudo-Lull's strange assimilation of Christian eschatology and alchemy reveals the power that processes considered quite mundane today, such as sublimation, distillation, and calcination, held for the premodern mind. These operations offered both a way of understanding the *modus operandi* of the Creator and of imitating His works on a local level.

However unusual the pseudo-Lullian assimilation of laboratory analysis to the creation and purgation of the world may sound to the modern ear, it provided a remarkably persuasive theme to subsequent alchemists. It is likely that Paracelsus von Hohenheim read the pseudo-Lullian *Testamentum* as part of his alchemical education, and its emphasis on the separation of a primordial first matter is replayed in Paracelsus's understanding of Genesis 1.[17] In his magnum opus, the *Astronomia magna*, Paracelsus develops an extended comparison between the extraction of essences by means of distillation and God's creation of man.[18] The Paracelsian theme of Genesis 1 as an alchemical separation acquired fantastic popularity in the 16th and 17th centuries, when it became grist for the mills of philosophers, poets, and divines.[19] Additionally, Paracelsus seems to recapitulate pseudo-Lullian themes when he stresses that the conflagration predicted by the Biblical book of Revelations will act to produce a *Scheidung* or separation, after which the world will be clear and pure, like the egg white within an egg.[20] From this topos an ideology emerges in Paracelsus' work according to which the chymist is seen as modeling his analytical processes on those of the Creator Himself. It is no surprise, then, that analysis or *Scheidung* would form one of the most fundamental pillars of the Paracelsian system.

Despite its debt to pseudo-Lull and others, however, Paracelsian analysis differs from that of earlier alchemists in several significant ways. First, Paracelsus added the new principle "salt" to the pre-existing pair made up of sulfur and mercury, which found its earliest clear expression in Arabic alchemy. Second, Paracelsus extended the domain of the principles to cover not only the metals and minerals, but literally all substances. Third, Paracelsus and his followers laid the foundations for viewing analysis as only half of the equation – as a necessary preliminary to resynthesis. The first two of these developments are clearly illustrated in Paracelsus' *Opus paramirum* (1531), where Paracelsus argues that the original "material" out of which God made the world was the *fiat lux* – the divine expression by which the Creator fashioned his product by voice alone. But because the Creator is a trinity, his product is also necessarily threefold, and its tripartite nature is revealed by analysis in the fire. In combustible materials the sulfur appears in the form of flame, the mercury in the volatile components driven off, and the salt in the ash that remains.[21] As Paracelsus stresses in his meteorological works, it is not only such mundane materials as wood and minerals that consist of the three principles, but the planets and stars themselves. In the hands of Paracelsus, the sulfur–mercury theory of the medieval alchemists has become a universal explanatory tool allowing him to explain everything from the powers of the sun and moon to the subterranean actions of vapors within mines in terms of his chymical principles.

The analytical focus of Paracelsian chymistry appears also in the term *spagyria*, a neologism that seems to have meant more or less the same thing as *Scheidung* to Paracelsus.[22] Around the beginning of the 17th century, chymists began to interpret this strange term of art as a combination of two Greek works – *span* (to pull apart) and *ageirein* (to bring together).[23] This reflects a growing realization that analysis was only half of the story – that a cycle of analysis and synthesis could reveal new and important facts about material substances. The roots of this perspective may

already be found in the work of Paracelsus himself, though without the demonstrative focus that it would later attain. In his *De renovatione et restauratione* (1526), for example, Paracelsus claims that copper may be analyzed into its mercury, sulfur, and salt, and that the latter may be recombined in order to produce the metal anew. The result, according to Paracelsus, will be a "reborn" and "perfected" metal, presumably because the copper has been deprived of its erstwhile impurities. But Paracelsus leaps immediately into a discussion of ways to restore and renovate the human body rather than developing the implications of the analysis–synthesis cycle.[24] While the theme of "rebirth" after purgation by fire appears repeatedly in the Paracelsian corpus, it is usually seen by him and his pseudonymous copyists as a means of attaining purification and spiritualization of substances rather than serving to reveal new truths about the nature of matter. Indeed, the generally non-quantitative character of the Paracelsian corpus would have made it difficult for synthesis to provide a great deal in the way of new knowledge. This would not be the case for Paracelsus's most significant exponent and reformer in the 17th century, however, namely the Belgian chymist Joan Baptista Van Helmont.

THE HELMONTIAN MODEL AND QUANTITATIVE SPAGYRIA

It would be impossible here to do justice to Van Helmont's chymistry *in toto*, and instead we will focus on the twin themes of gravimetric analysis and the demonstrative use of synthesis insofar as they can be illustrated within the scope of this paper. Despite a vociferous dislike of the scholastic tradition focusing on the degrees and quantities of humors and their constituents, Van Helmont explicitly argued for a quantitative approach to knowledge in general. In his *Opuscula medica inaudtia* of 1644, Van Helmont expresses what might be considered the chymist's creed,

> We read in our furnaces that there is no more certain genus of acquiring knowledge (*sciendi*) for the understanding of things through their root and constitutive causes than when one knows what is contained in a thing and how much of it there is.[25]

One would find it difficult to overstate the importance of this credo for understanding Van Helmont's way of thinking, and for pointing to the analytical ideal which was developing within chymistry. Unlike his more traditional scholastic contemporaries, Van Helmont explicitly argued that the weight of matter in a given reaction is fixed. Although it was nothing new to say that matter as such can neither be destroyed nor created (for the maxim *ex nihilo nihil fit* was a tenet extending back to the origins of Greek philosophy), Aristotle had argued that the four elements can be transmuted into one another. Owing to the innate tendency of the elements to move to their natural places – the heavy elements earth and water towards the center of the universe, and the light elements fire and air towards the sphere of the moon – it was possible for a heavy substance to be transmuted into one having no weight at all, hence making the very possibility of gravimetric analysis untenable. For example, as water is converted into air by boiling, all of the original weight of the water is wholly lost since air has no weight in the Stagirite's system. In this context, what sense would it make

to compare the initial and final weights of a material being analyzed if there was no expectation that the two would be the same? One could learn nothing about the amount of constituents from their weights, since those weights were not tied to the amount of matter at hand.

It would, however, be a mistake to think that alchemists working within a scholastic context were unconcerned with the weights of their materials. To the contrary, Geber's *Summa*, to name a source that Van Helmont openly used, explicitly advised that "if the altered body should exchange any of the differences (i.e., qualities) of perfection – namely of weight or color – the artifex has not rightly investigated the work."[26] But in fact the Aristotelian theory of four elements plays only a tiny part in Geber's own material explanations. It appears in the famous 24th chapter of the *Summa perfectionis*, where Geber states that the four elements combine "through their smallest particles" (*per minima*) in order to compose the two principles, sulfur and mercury.[27] Throughout the rest of the *Summa*, the Aristotelian quaternity of fire, air, water, and earth is scarcely mentioned, for their role as explanatory constituents of matter has been replaced by that of the Islamic conception of sulfur and mercury. When Geber wishes to explain the differences between the precious and base metals, his discussion is couched solely in terms of the respective quantity, volatility, color, and purity of the sulfur and mercury involved, without the slightest appeal to the four elements. For the author of the *Summa perfectionis*, the four elements were not a tool to be utilized, but an inadequate account to be tacitly set aside.

The Helmontian rejection of Aristotelian elemental theory may be seen, then, as bringing an old and implicit tension to the foreground of discussion. This observation does not, however, diminish Van Helmont's importance in the slightest. On the contrary, Van Helmont openly affirmed the principle that modern chemists refer to as "mass balance" – the fact that the weight of initial ingredients is preserved in the outcome of any reaction. As Van Helmont put it, "Nothing comes into being from nothing. Hence weight comes from another body weighing just as much."[28] What Van Helmont has in mind is that the weight of ingredients going into a reaction must come out in the products, regardless of any transformations that have taken place. Needless to say, he did not articulate the concept of mass in the modern, Newtonian sense, but the fact remains that he clearly believed the gross weight of initial and final states to remain constant, a fact that he linked to the indestructability of matter. Additionally, Van Helmont seems to have believed that ordinary analysis by fire or acids did not normally resolve materials into their ultimate material, which he believed to be water. Hence chymical analysis provided a means for him to remove the "masks" (*larvae*) that clothed compounds and hid their immediate constituents, rather than reducing compounds into the uniform substrate from which all things were ultimately composed. This view, coupled with his insistence on mass balance, provided him with a powerful means of providing "a mathematical demonstration stronger than any syllogism," in the form of gravimetric analysis.[29]

Van Helmont's emphasis on gravimetric measurement allowed him, on occasion, to provide quantitative demonstrations employing *spagyria*, now seen as the discipline of combined analysis and synthesis. One such demonstration occurs in his *Ortus*

medicinae, where Van Helmont attacks the Aristotelian theory of *mixis*, according to which the four elements were thought capable of combining to form a perfectly homogeneous mixture. Intent on showing that such seemingly homogeneous materials are really composed of minute, juxtaposed particles, Van Helmont adduces the example of glass. Although glass appears to be perfectly homogeneous, he can show by means of analysis and synthesis that it is not. For, "by means of art glass returns to its original ingredients (*pristina initia*) once the bond holding them together is broken: the sand can even be regained in the same number and weight [as before]."[30] Employing a scholastic term, Van Helmont says that it is the self-same sand (*eadem numero*) as it was before it went into the glass, not newly created sand. What is the basis for this unlikely-sounding claim? Van Helmont describes that

> If one melts a fine powder of glass with a large amount of alkali and exposes it in a humid place, one will presently find that all the glass dissolves into a water. If *chrysulca* [mainly nitric acid] is poured on in a quantity sufficient to saturate the alkali, one will at once find that the sand sinks to the bottom [of the vessel] in the same weight as it was before it was used in making the glass.[31]

Van Helmont's experiment presupposes that one makes the glass himself, first weighing the sand that goes into it. Once the glass is made, it is melted with additional "alkali," probably soda or salt of tartar (sodium or potassium carbonate, respectively). This results in the formation of sodium or potassium silicates, which are then allowed to deliquesce into the so-called "oil of glass" of early modern chymistry. Nitric acid is then added to the dissolved silicate, resulting in the formation of sodium or potassium nitrate and the precipitation of silicon dioxide – sand – in the same weight used to make the glass.

A similar, if more complicated, methodology underlies Van Helmont's experiments for proving his theory of "exantlation" – the belief that acids do not become neutralized by combining with other substances but are rather infeebled and weakened by their efforts in the same way that an athlete might be tired out by his exertions. Although this belief may seem intuitively wrong-headed to anyone with a knowledge of modern chemistry, Van Helmont had experimental evidence employing, once again, a combination of synthesis and analysis linked with weight measurements.

> If you distill oil of vitriol [mostly sulfuric acid] from running mercury, the oil is coagulated with the mercury, and they both remain in the bottom in the form of a snow. Whatever you distill thence is mere water. But this snow, if it is washed, becomes a yellow powder, which is easily reduced into a running mercury in just the same weight as before. But if you distill off the wash water, you have a pure alum in the bottom, from the acid salt in the vitriol. Therefore dissolvents are mutated even if the dissolved lose nothing of their substance or matter.[32]

As in the case of the glass-experiment, Van Helmont first advises that a substance be synthesized, in this case the white "snow" that results from distilling sulfuric acid with a weighed quantity of mercury. The snow is then washed, whereupon it turns yellow (due to the hydrolysis of white mercuric sulfate into a yellow basic sulfate).

The wash water is in turn saved and distilled, revealing a residue of a crystalline "alum." Meanwhile, Van Helmont has reduced the yellow powder back into mercury, and recovered the entire weight of the mercury employed. He can thus conclude that the "alum" found in the wash water must contain no mercury at all – instead it is the acid itself in exantlated, enfeebled form; the acid has "congealed" into the crystalline salt he calls alum as a result of its efforts in corroding mercury.

Van Helmont's results in the exantlation experiment are clearly flawed, perhaps as a result of his using an inaccurate balance. He could not have retrieved all of his initial mercury from the yellow powder, since the "alum" found in the wash water is also a mercury compound.[33] Nonetheless, the experiment reveals, once again, the tremendous emphasis that Van Helmont placed on the combination of synthesis and analysis in a gravimetric context for the study of chymical processes and procedures. The Helmontian emphasis on comparing the weights of initial and final products in a process would bear important fruit as the 17th century progressed. It would lead to a nascent form of tallying initial and final weights that has been famously called "the balance-sheet" method when it emerges fully formed in the celebrated work of Antoine Laurent Lavoisier.[34] In this paper we shall not attempt to follow the Helmontian emphasis on mass balance as far as Lavoisier, but now show only how it functioned in the work of a convinced Helmontian of the mid-17th century, namely George Starkey.[35] Although Starkey's techniques were also influenced by other sources in addition to Van Helmont, important parts of Starkey's *modus operandi* are to be seen in the context of his professed master, "the noble Bruxellian."

Starkey is well known to modern scholarship as the chymical tutor of Robert Boyle, and as the fabulous Eirenaeus Philalethes, probably the favorite alchemist of Isaac Newton. His most famous process, without doubt, is an operation for making a "sophic mercury," intended to be the initial ingredient of the philosophers' stone. This special mercury was supposed to dissolve gold into its primordial matter, and to be fertilized by the "seed" of the gold, whereon the impregnated mercury would (at least in theory) pass through a succession of regimens or color changes culminating in the summum bonum of the alchemists – the philosophers' stone. In fact, the process of making Starkey's mercury has been replicated recently, and it does indeed result in the striking formation of a lovely metallic "tree," although without leading to the great transmutational agent of chrysopoeia.[36] As we have shown elsewhere, Starkey's process was heavily indebted to an earlier writing on antimony and its mysteries by the Prussian Paracelsian, Alexander von Suchten. Suchten reveals in his *Tractatus secundus de antimonio vulgari* (1606) that mercury can be amalgamated with the star regulus of antimony (a form of metallic antimony bearing a striking crystalline surface pattern) in order to form a special mercury of antimony. Suchten accompanies this process with an interesting quantitative analysis of the antimony regulus and mercury that reveals how Paracelsian matter-theory was becoming integrated with the more metallurgical alchemy descending from such writers as Geber.

Suchten begins by amalgamating quicksilver with a silver–antimony alloy, whereupon a black powder is expelled. Collecting this powder by washing the amalgam, Suchten then dries it and warms it in order to remove any residual mercury, apparently

then weighing the latter. He then burns the powder to an ash in a crucible, weighs it again, and reduces it to a regulus. At this point, Suchten is able to reach some conclusions. First, since sulfur is the Paracelsian principle responsible for flammability, the fact that the black powder was combustible means that it contained that principle. Second, since burning the black powder consumes its sulfur entirely, a comparison of the powder's weight before and after burning will reveal the amount of sulfur that came out of the regulus when it was combined with quicksilver. Similarly, by weighing the powder before and after its mercury is evaporated off, one can determine how much of the quicksilver was lost mechanically during amalgamation. One can then compare the weight of the original quicksilver employed and the weight of the quicksilver that was in the black powder in order to arrive at the weight of quicksilver present in the product, the mercury of antimony. The net result of Suchten's operations, then, will be a quantitative analysis revealing the amount of sulfur lost by the regulus and the amount of common mercury present in the final "antimonial" mercury.

Suchten's analysis forms the partial inspiration to a more elaborate operation carried out by George Starkey, parts of which he recorded in a 1651 letter to Boyle and in his Philalethan *Sir George Riplye's Epistle to King Edward Unfolded*.[37] But Starkey's analysis incorporates the concept of mass balance in a more explicit and sophisticated way than does Suchten's. Unlike Suchten, Starkey raises the Helmontian issue of "missing mass" in a desire to account after analysis for the full weight of the mercury employed, and the American chymist also reveals his Helmontian heritage by determining exact proportions, which again distinguishes him from Suchten. We argue that Starkey's process reveals the development within chrysopoetic alchemy of the essential features of the so-called balance-sheet method of 18th-century chemistry. Starkey first weighs all his starting materials: a quantity of common quicksilver, some regulus of antimony, and a pint of water. He then amalgamates the mercury with the regulus (which has been fused with two parts of silver) and then washes the amalgam repeatedly in the water in order to remove the black powder that emerges. After this, he distills off the mercury and then amalgamates it again with fresh silver–regulus alloy, repeating the same process of amalgamation, grinding, washing, and distilling from seven to nine times. After the completion of this reiterated procedure, he weighs the resulting "sophic mercury" produced by the last distillation. He then separates the black powder from the wash water by decanting, dries and weighs it, and gently evaporates some residual common mercury from it, as did Suchten. Starkey weighs the black powder after the removal of this common mercury, then ignites the powder to a slow burn, reduces the ash to a regulus, weighs this, and discovers that it weighs two-thirds as much as the original regulus employed. This means, by mass balance, that the "missing" one-third of the regulus is now in the sophic mercury.

Starkey then adds the weight of the common mercury evaporated from the black powder to the weight of the sophic mercury (minus what it acquired from the regulus) and finds that the combined weights do not add up to that of the common mercury employed initially. Where is the missing mass of the mercury? Starkey has to go in

search of it, so he presumably weighs the wash water again and discovers its increased weight.[38] He then evaporates the wash water and finds it to contain a crystalline salt, or as Starkey calls it, "the Salt of *Mercury* Crude." This salt accounts for the missing weight of mercury, and thus Starkey has carried out a complete gravimetric analysis of the common quicksilver employed. Hence Starkey concludes happily, "it is a content for the Artists to see how the Heterogeneities of *Mercury* are discovered."[39]

One could find further elaborations of Starkey's gravimetric analysis among the French Academicians of the 18th century, in particular, with Wilhelm Homberg (1652–1715).[40] In this article, however, our point is not to show the continuity of Enlightenment chemistry and the Helmontian projects of the 17th century, but rather to argue that the analytical concerns of the earliest alchemists bore fruit in the 17th century with the explicit invocation and deployment of mass balance by Van Helmont and his followers. Indeed, the example of Alexander von Suchten shows that Paracelsian *spagyria* could have a quantitative emphasis even before Van Helmont came on the scene, a fact that is hardly surprising when one considers the gravimetric tendencies already prominent in the Geberian tradition and the long emphasis on assaying techniques that extends back to the very origins of alchemy in the Leiden and Stockholm papyri. Starkey's later expansion of Suchten's procedures, however, illustrates with particular clarity the dominant Helmontian theme that "weight comes from another body weighing just as much."

The gradual metamorphosis of analytical techniques in alchemy to the proto-balance-sheet method found in the work of George Starkey reveals an evolutionary process that had been occurring over the course of many centuries. We have charted the skeletal framework in which this evolution occurred – beginning with the Leiden and Stockholm papyri's interest in assaying operations that could also be employed for mimicking the precious metals, we passed to a brief discussion of medieval alchemy, where the techniques of cupellation and cementation were coupled with a host of operations for determining changes in weight and color. Although the alchemy of Geber is remarkably dry and sober, we pointed to other alchemists, such as pseudo-Ramon Lull, who envisioned the creation and final purgation of the world in terms of analytical techniques involving distillation and calcination. These motifs were subsequently adopted in the 16th century by Paracelsus, who also expanded on the Jābirian and pseudo-Avicennian program of subjecting plant and animal products, as well as minerals and metals, to analysis. Finally we passed to the Joan Baptista Van Helmont and his follower George Starkey, who coupled the Paracelsian theme of *spagyria* – now interpreted to mean the twin processes of analysis and synthesis – with the consciously gravimetric concept of mass balance.

As it turns out, Van Helmont's mercurial "snow" experiment (unlike his glass analysis) was flawed, and Suchten's analysis of antimony and Starkey's analysis of mercury were "wrong" insofar as we now recognize these substances as elemental, and thus not susceptible to chemical analysis. Yet all of these experiments further refined and elaborated the analytical methods which had been developing for centuries within the alchemical tradition. This cumulative work thus produced by the early

modern period a generalized technique for chymical study, and perhaps more importantly, an operative mindset towards analysis/synthesis and weight-determination as the basic implements for the investigation and control of material substances. These techniques – worked out in the context of preparing grand chymical arcana – would be inherited and deployed famously by Lavoisier and his successors (like Liebig and other 19th century analytical chemists) down to the present day.

It must furthermore be borne in mind that chymists of Starkey's stamp were still actively searching out such grand arcana as the philosophers' stone and the Helmontian alkahest or universal dissolvent, yet this in no way prevented, but instead arguably encouraged the growth of the discipline in richness of technique, precision, and quantitative accuracy over the *longue durée*. A lesson to be drawn from this is that we should not allow the exaggerated promise of chrysopoeia to blind us to the real capabilities and accomplishments of the premodern chymical laboratory. Indeed, alchemy has been treated far more harshly in this regard than other pre- and early modern practices whose goals have since been revealed as unattainable. While transmutational efforts, and with them a vast amount of pre-18th century activities labeled as alchemy received scorn and rejection owing to the imputation of fraud and outlandish (unfulfilled) promises, other fields which made similar grandiose claims were treated far more leniently, during both the Enlightenment and the early years of the professionalized discipline of the history of science. For example, medieval and early modern physicians routinely made equally grandiose claims for the curative properties of such products of their art as theriac, spirit of wine, and innumerable other nostrums, without the reputation of their entire field being relegated to the status of "pseudo-science." Engineers dreamed of impossible ornithopters and huge battlefield tanks that could not have moved an inch, astronomers cast natal charts, answered horary questions, and predicted the weather, and perpetual motion machines were not only designed by shadowy figures such as Cornelius Drebbel but were devised by the likes of Newton and Leibniz.[41] We should not judge the premodern scientific disciplines by their failure to attain the ultimate goals that they set for themselves, any more than we condemn contemporary physics for failing (so far) to arrive at nuclear fusion, modern medicine for being unable to "cure cancer," or economics for so manifestly lacking the ability to predict or forestall recessions or even market fluctuations. Such overarching projects, even if they prove fully unattainable, may yet provide a theoretical and practical framework in which more quotidian results can indeed be realized. Although the old positivist inclination to hack out fruits more digestible to moderns from their thorny, "archaic" surroundings has long been abandoned, it does not follow that either the supporting framework or its products can be ignored by the historian who is genuinely concerned with capturing the nature of change over time.

Department of History and Philosophy of Science, Indiana University, USA
and
Department of the History of Science and Technology and Department of Chemistry, The Johns Hopkins University, USA

NOTES

[1] Charles Mackay, *Memoirs of Extraordinary Popular Delusions and the Madness of Crowds* (New York: L.C. Page and Company, 1932), pp. 98–256. The *Memoirs* were originally published in 1841 in London, and then in a second edition in 1852. For the use of alchemy as a foil in *Leviathan and the Air Pump*, see Lawrence M. Principe, *The Aspiring Adept: Robert Boyle and His Alchemical Quest* (Princeton: Princeton University Press, 1998), pp. 107–111.

[2] For a brief treatment of Kopp and Berthelot as historians, see Robert Halleux, *Les textes alchimiques* (Turnhout: Brepols, 1979), pp. 52–54. For Liebig's views on the history of his field see Justus Liebig, *Chemische Briefe* (Heidelberg: C. F. Winter, 1851). J. R. Partington was the author of the influential *A History of Chemistry* (London: Macmillan, 1961), in four volumes.

[3] The "spiritual" interpretation of alchemy presents another distinct set of historiographical problems, which we have dealt with elsewhere. See Lawrence M. Principe and William R. Newman, "Some Problems with the Historiography of Alchemy," in *Secrets of Nature: Astrology and Alchemy in Early Modern Europe*, eds. Newman and Anthony Grafton (Cambridge, MA: MIT Press, 2001), pp. 385–431. For a brief, general introduction to the image of alchemy in Netherlandish and later art, see Lawrence M. Principe and Lloyd DeWitt, *Transmutations: Alchemy in Art, Selections from the Eddleman and Fisher Collections* (Philadelphia: Chemical Heritage Foundation, 2001). For additional considerations on the place of alchemy in early modern visual art, see William R. Newman, *Promethean Ambitions: Alchemy and the Quest to Perfect Nature* (Chicago: University of Chicago Press, 2004), Chapter 3.

[4] Note that we use the terms "chymistry" and "chymical" to include the entirety of the subject retroactively collected under the terms "alchemy" and "chemistry"; these latter two terms were used largely interchangeably during the early modern period. See William R. Newman and Lawrence M. Principe, "Alchemy vs. Chemistry: The Etymological Origins of a Historiographic Mistake," *Early Science and Medicine* 3 (1998): 32–65.

[5] Robert Halleux (ed. and tr.) *Les alchimistes grecs: papyrus de Leyde, papyrus de Stockholm, recettes* (Paris: Belles Lettres, 1981), pp. 22–24.

[6] Rijksmuseum van Oudheden, *Papyrus Leidensis X*, see recipes 20–29, esp. 24, of the Leiden papyrus, presented in *Les alchimistes grecs*, Tome I, ed. Robert Halleux (Paris: Les Belles Lettres, 1981) pp. 90–92. See also Halleux, "L'alchimiste et l'essayeur," in *Die Alchemie in der europaeischen Kultur- und Wissenschaftsgeschichte*, ed. Christoph Meinel (Wiesbaden: Otto Harrasowitz, 1986) and "Methodes d'essai et d'affinage des alliages auriferes dans l'Antiquité et au Moyen Age," *Cahiers Ernest Babelos* 2 (1985): 39–77.

[7] William J. Wilson, "An Alchemical Manuscript by Arnaldus de Bruxella," *Osiris* 2 (1936): 220–405; see pp. 316 and 319.

[8] Pseudo-Avicenna, "*De anima in arte alchemiae*," in *Artis chemicae principes, Avicenna atque Geber* (Basel: Petrus Perna, 1572), pp. 125–126.

[9] William R. Newman, *The Summa Perfectionis of Pseudo-Geber* (Leiden: Brill, 1991), pp. 590–627 (Latin), and pp. 769–783 (English).

[10] *Ibid.*, pp. 591–608 (Latin), and pp. 769–776 (English).

[11] *Ibid.*, p. 780.

[12] *Ibid.*, p. 718.

[13] Paul Kraus, "*Jābir ibn Ḥayyān, contribution à l'histoire des idées scientifiques dans l'Islam*," in *Mémoires presentés à l'Institut d'Egypte*, vol. 45 (Cairo: Institut Francais de l'Archéologie Orientale, 1942), vol. 2, pp. 1–18.

[14] Pseudo-Avicenna, "*De anima*," in *Artis chemicae principes* (Basel: Petrus Perna, 1572), pp. 14–15.

[15] Michela Pereira, "The Alchemical Corpus Attributed to Raymond Lull," in *Warburg Institute Surveys and Texts*, vol. 18 (London: University of London, 1989), pp. 1–20.

[16] Michela Pereira and Barbara Spaggiari, *Il "Testamentum" alchemico attribuito a Raimondo Lullo* (Florence: SISMEL-Edizioni del Galluzo, 1999); cf. pp. 12–24, 248–256.
[17] Wilhelm Ganzenmüller, "Paracelsus und die Alchemie des Mittelalters," in Ganzenmüller, *Beiträge zur Geschichte der Alchemie* (Weinheim: Verlag Chemie, 1956), pp. 300–314, esp. 311–312.
[18] Paracelsus, *Astronomia magna*, in Karl Sudhoff, *Sämtliche Werke* 12: 31–51.
[19] For an introduction to these and other alchemical themes in English poetry, see Stanton J. Linden, *Darke Hieroglyphicks: Alchemy in English Literature from Chaucer to the Restoration* (Lexington: University Press of Kentucky, 1996).
[20] Paracelsus, *Astronomia magna*, in Karl Sudhoff, *Sämtliche Werke* 12: 322. For a more detailed description of pseudo-Lull's cosmological speculations, see Newman, *Gehennical Fire*, pp. 98–103.
[21] Paracelsus, *Opus paramirum*, in Karl Sudhoff, *Sämtliche Werke* 9: 46–48. See also Paracelsus, *Das Buch Meteororum*, in Sudhoff, *Sämtliche Werke* 13: 134–136.
[22] Paracelsus, *Opus paramirum*, in Karl Sudhoff, *Sämtliche Werke* 9: 55: "darumb so lern alchimiam die sonst spagyria heisst, die lernet das falsch scheiden von dem gerechten."
[23] This etymology appears in the work of Andreas Libavius, for example. See Libavius, *Commentariorum alchymiae . . . pars prima*, in Libavius, *Alchymia* (Frankfurt: Joannes Saurius, 1606), p. 77.
[24] Paracelsus, *Liber de renovatione et restauratione*, in Sudhoff, *Sämtliche Werke* 3: 203–204.
[25] Joan Baptista Van Helmont, *De lithiasi*, chap. 3, no. 1, p. 20, in *Opuscula medica inaudita* (Amsterdam, 1648; reprint ed., Brussels: Culture et Civilisation, 1966).
[26] Newman, *Summa perfectionis*, p. 780; for Van Helmont's use of Geber, see Newman, *Gehennical Fire*, pp. 145–149.
[27] Note however that Geber occasionally treats arsenic as well as a principle.
[28] Joan Baptista Van Helmont, *Ortus medicinae* (Amsterdam, 1648; reprint ed., Brussels: Culture et Civilisation, 1966), "Progymnasma meteori," no. 18, p. 71.
[29] Van Helmont, *Opuscula*, *De lithiasi*, chap. 1, no. 2, p. 10.
[30] Van Helmont, *Ortus*, "Terra," no. 14, p. 56.
[31] *Ibid.*
[32] Van Helmont, *Opuscula*, *De febribus*, chap. 15, no. 20, p. 57.
[33] This "alum" is actually mercuric sulfate which, although normally hydrolyzed into the insoluble yellow sulfate by pure water, was able to dissolve in the wash water because the latter is rendered acidic by the sulfuric acid released during the hydrolysis of the major part of Van Helmont's "snow."
[34] On Lavoisier's balance-sheet method see Douglas McKie, *Antoine Lavoisier* (New York: Henry Schuman, 1952), esp. pp. 283–284.
[35] For a fuller treatment of this topic than is possible here, see our *Alchemy Tried in the Fire*.
[36] For the philosophical tree (with photograph) see Lawrence M. Principe, "Apparatus and Reproducibility in Alchemy," in *Instruments and Experimentation in the History of Chemistry*, eds. Trevor Levere and Frederic L. Holmes (Cambridge, MA: MIT Press, 2000), pp. 55–74.
[37] Starkey to Robert Boyle, April/May 1651; *The Correspondence of Robert Boyle*, eds. Michael Hunter, Antonio Clericuzio, and Lawrence M. Principe, 6 vols. (London: Pickering and Chatto, 2001), 1: 90–103, on pp. 97–98; [George Starkey], *Sir George Riplye's Epistle to King Edward Unfolded*, pp. 19–47 in Samuel Hartlib, *Chymical, Medicinal, and Chyrurgical Addresses* (London, 1655), a slightly different version appears in Eirenaeus Philalethes [i.e., George Starkey], *Ripley Reviv'd* (London, 1678), pp. 1–29; on the evolution of this important Philalethes text and for two related but previously unpublished texts by Starkey, see George Starkey, *Alchemical Laboratory Notebooks and Correspondence*, eds. Newman and Principe (Chicago: University of Chicago Press, 2004), pp. 309–317.
[38] It is hard to imagine that Starkey could actually have measured an increase in the gross weight of the water, since a considerable amount would surely be lost from evaporation and

non-quantitative transfers during the manifold washings. But since he explicitly notes the volume ("a Pint of water") he weighed at the outset, it is possible that he actually compared the weight of a standard volume of the wash water before and after washing (rather than the entire quantity), that is, he checked for an increase in specific weight, which water containing a dissolved solute would show.

[39] Philalethes [Starkey], *Ripley Reviv'd*, p. 14.

[40] On Homberg's use of Starkey's method see Lawrence M. Principe, "Wilhelm Homberg: Chymical Corpuscularianism and Chrysopoeia in the Early Eighteenth Century," in *Late Medieval and Early Modern Corpuscular Matter Theories*, eds.Christoph Lüthy, John Murdoch, and William Newman (Leiden: Brill, 2001), pp. 535–556, and for the possible link through Homberg to Lavoisier see *Alchemy Tried in the Fire*, pp. 303–311.

[41] For a convenient illustration of the combined treatment of useful and fantastic machines in the Renaissance, see Paolo Galluzzi, *The Art of Invention: Leonardo and Renaissance Engineers* (Florence: Giunti, 1999). For the interest of Newton and Leibniz in perpetual motion machines, see Alan Gabbey, "The Mechanical Philosophy and its Problems: Mechanical Explanations, Impenetrability, and Perpetual Motion," in *Change and Progress in Modern Science*, ed. Joseph C. Pitt (Dordrecht: D. Reidel, 1985), pp. 9–84, esp. 38–67.

MARJORIE GRENE

DESCARTES AND THE HEART BEAT: A CONSERVATIVE INNOVATION

William Harvey's *Exercitio Anatomica de Motu Cordis et Sanguinis* was published in 1628. It was not well received. Indeed, after its publication, Harvey informed John Aubrey, "he fell mightily in his practize, and . . . 'twas believed by the vulgar that he was crack-brained; and all the physicians were against his opinion and envyed him."[1] Nine years later, in his *Discourse on Method*, René Descartes was one of the first who publicly applauded Harvey's discovery of the circulation.[2] At the same time, Descartes emphatically disagreed with the English physician in his account of cardiac motion. It had in fact been Harvey's careful investigation of the heart beat in numerous animals that had led him, first to describe cardiac motion as he had observed it, and from this to infer the blood's circular course. Descartes, on the other hand, accepted the circulation but contradicted its discoverer on his first and fundamental discovery, the motion of the heart. Harvey saw the heart as a hollow muscle, by its beat propelling the blood out through the body. Descartes found it to be a kind of bladder, containing a hidden "fire without light" by which entering drops of blood were subject to rarefaction, so that they escaped as the organ swelled and hardened. Harvey, Descartes remarked, would need some obscure faculty to induce the heart beat and another to "attract" blood into the heart.[3] His own view, he was confident, was much simpler and more straightforward, following only the laws of nature, which are in fact the laws of mechanics. Harvey, it appeared, was caught up in a confused and confusing tradition; Descartes was putting forward a new and more economical approach to the phenomena.

On the face of it, it is true that, of the two, Harvey was the more conservative thinker, while Descartes favored – and was a major initiator of – the new (or newly fashionable) interpretation of living things, including the human body, as machines, and no more than machines. Acknowledging this contrast on one level, however, I want to suggest that, paradoxically, some aspects of Harvey's account, and in particular one prominent aspect of it, would in fact have struck his readers as more radical than the corresponding features of the Cartesian story. Thus Descartes's view of the heart beat may have been acceptable to many physicians and physiologists, not only because they were attracted by the new mechanism, but also because the furnace theory allowed them to keep beliefs they had been taught to accept as fact, and which Harvey's interpretation would force them to reject. In his discussion of differing thought styles, Ludwig Fleck suggested that there were sometimes "pre-ideas" contained in a traditional conception that would make a new approach easier

J.Z. Buchwald and A. Franklin (eds.), Wrong for the Right Reasons, 91–97.

to accept.[4] In the case before us, it seems to me, there were also what one might call "post-ideas," or better, post-beliefs, seemingly factual beliefs one had been brought up on, so to speak, which the new radicalism allowed one to retain, while what seemed in general a more conservative point of view would entail rejecting those "facts" that one had seen, or touched, for oneself.

As to the general contrast: to put it crudely, Harvey espoused a kind of 'vitalism' as against the beast–machine theory of Descartes. He is often called an Aristotelian, in contrast with Descartes's more innovative position. Maybe so, in a sense. He was by no means a scholastic thinker, as Descartes, trained by scholastics, and speaking to scholastics, still partly was. Nor was he an Aristotelian cosmologist, as some have taken him to be. He did not discover the circulation because, like Aristotle in the *De Gen. et Corr.*, he admired circles as such and thought that vital rhythms should imitate the circling of the heavens.[5] What he discovered, by observation and experiment – both dissection and extensive and varied vivisection – was the motion of the heart. And then he wondered, given the heavy pulsing of the heart, whether the blood so forced out might be going round in a circle. But he was an Aristotelian in his fascination with the living, and his attention to the details of living forms and their development. Until the rise of microscopy, no one since Aristotle had studied the embryology of the chick as carefully as he did. As we shall see, he differed from the Greek biologist in some matters, but not in his devotion to the study of living phenomena as such. It is true that in describing cardiac motion he used some mechanistic analogies: it was like the firing of a gun.[6] But so did Aristotle. That by no means implies that either of these biological investigators thought their subjects *were* machines.

In relation to this interest in vital processes, and in particular in development, Harvey also followed Aristotle, and most thinkers in the tradition, in applying to his study the distinction of four kinds of causality, finding the material cause everywhere subordinate to the efficient, formal and final causes, with special emphasis on the last of these. "Nature does nothing in vain" was for him no empty formula, but the expression of a deep-seated belief in the directedness, almost an autonomous directedness, of each organ or organ system's activity. Be it noted: this is not Galenic teleology, where ends come from outside (as they do also for Descartes, namely, in his case, from God, the maker of those "animate" machines); this is genuinely internal, Aristotelian teleology. "Nature" is what by its very nature contains in itself a principle or change, a principle of change well-regulated, for the most part, for the good of the organism in question. There are things that happen by chance, but they are leftovers, so to speak, in a harmonious, well-ordered world.

Within such a world, finally, the natural tendencies displayed by its components are to be contrasted with the "violent" motions sometimes induced by human or other intervention. Heavy things try to reach the earth's center, but my arm – or, by Harvey's time, a gun's firing – can propel heavy objects up or out instead of down. Such a violent motion, in fact, is the heart's beat, pushing the blood outward against its natural centripetal inclination.

Descartes's account of cardiac motion, and of vital phenomena in general, differs from Harvey's in all these respects. As we have already noticed, and as he himself

emphatically boasts, both the structures and functions of living things could be adequately explained, or so he believed, in terms that would apply to the non-living as well – particularly to artifacts. Living bodies are like clocks: once wound up, they run on purely mechanical principles. Descartes compared them to the amusing automata found in the gardens at Versailles: for instance, a Diana who retreats when you approach her or a Neptune who defends her with his trident.[7] Animals, made by God, are more ingenious machines than we can make, but machines all the same. Any talk or vegetative or sensitive souls is nonsense. We have souls, quite separate (or at least separable) from our bodies; and God of course is an infinite soul, beyond our comprehension. Otherwise there is just spread-outness, to be understood by geometers, though misunderstood by the many, or by the schools, who believe what their senses seem to tell them. (Incidentally, that is another difference from Harvey, though it must be accepted with caution. Although Descartes wanted to guide the mind away from the senses, he was by no means an *a priorist*. He saw the value of experiments, and performed a number, including some he thought refuted Harvey. But Harvey was first and foremost an anatomist, for whom sense was the primary and ultimate teacher, while Descartes's chief aim was to initiate a mathematical physics, to apply the clear and distinct ideas of the geometer, or of a thinker as rational as a geometer, to the world of bodies beyond us (or beyond our minds) out there. But that is by the way.)

Given his concept of body as extension, further, and of the identity of the living with the "merely" bodily, Descartes could tolerate no intrusion of final causality into the study of life. We must have a purely when–then, mechanically reactive account of all vital processes, including the heart beat. And that, he believed, was what he was offering. Nor, thirdly, would there be any remaining contrast between natural and violent motion. Motion is motion, that's all there is to it.

So much for the general contrast between Harvey the traditionalist and Descartes the radical innovator. Moreover, as the idea of the circulation came to be accepted, it was in fact Descartes's view of the heart's motion that was, in many cases, accepted as the more persuasive account. And it seems to have been his thoroughgoing mechanism that was appealing. Among those who followed him, for example, were Hencricus Regius, Cornelis von Hoghelande, Franciscus Sylvius, Theodor Craanen on the continent, or George Ent and Thomas Willis in England.[8] Thus von Hoghelande wrote in his *Cogitationes* (1646):

> . . . we are of the opinion, that all bodies, however they act, are to be viewed as machines, and their actions and effects . . . are to be explained only in accordance with mechanical principles.[9]

That he attributed this viewpoint to Descartes is clear from the reply he gave to skeptical writers, with their perverse desire to spread doubt and ignorance:

> But truly the incomparable Descartes has favored our century with the splendor of his mind, so that any one who follows in his footsteps easily distinguishes the doubtful from the certain, the true from the false. And beyond that it is the good fortune of this century, that further, what is evident and what is obscure

> in material things is now in general at all events apparent to every one: namely, quantity and figure . . . and local motion.[10]

In 1659, in his *De Fermentione* and his *De Febribus*, Thomas Willis was still supporting the Cartesian theory.[11] Even when features of Descartes's explanation proved untenable, his general, mechanistic position was considered appropriate. Thus in 1666 Nicolaus Steno wrote:

> No one but he has explained mechanically all the actions of man, and especially those of the brain; others describe for us man himself; M. Des Cartes speaks to us only of a machine, which, however, makes us see the inadequacy of the others' doctrines, and teaches us a method of investigating the functions of the other parts of the human body, with the same evidence with which he demonstrates to us the parts of his man–machine, something no one has done before him. Thus we should not condemn M. Des Cartes if his system . . . does not prove to be entirely in agreement with experience; the excellence of his mind . . . excuses the errors of his hypotheses.[12]

In this context, then, Descartes appears as the more advanced thinker, while Harvey, enmeshed in traditional obscurities, lags behind him. As I suggested at the start, however, this is by no means the whole story. In terms of traditional medical views of the heart and blood, it was Harvey who was making an unprecedented move, and Descartes who was more conservative, allowing his readers to retain beliefs they had been trained to accept, but which Harvey would urge them to discard. When we look at the two accounts now, we find three traditionally accepted features that Descartes allows his readers to retain, while Harvey's version would call them in question. Of the three items, only two are expressed in the *DMC*, but let me list them all, with appropriate qualifications.

To begin with, as Descartes points out, physicians all observe that the heart is hotter than the rest of the body.[13] Just touch it (during vivisection; in this respect, the history of biology is an unpleasant subject, but there it is: Descartes mentions cutting open dogs, for example, or rabbits,[14] and Harvey of course prided himself on the scope and variety of his vivisections on all sorts of animals) – just touch it, and you can tell. Every one in those days, Harvey as well as Descartes, thought there must be some source of vital heat to distinguish the living body from a cold corpse, and in most cases – including Aristotle – that source was supposed to be the heart. That was the first part Aristotle saw beating in the chick's egg, and that was for him the center of life, even of thought. Even when the role of the nervous system had long been distinguished and the brain had been associated with the higher faculties, the heart was still thought to be the site of the vital heat. Harvey, following Aristotle in many things, disagreed with him in this: he found the first pulsation in the chick's egg to be that of a spot of blood, before the formation of the heart. So he insisted it was the blood itself, not the heart, that contained the vital heat.[15] True, in the *DMC* he also speaks of the blood's being warmed as it passes through the heart. But this may be simply its return to its proper state after cooling in the periphery; and the heart

itself is warmed by the coronary arteries, so need not have any special "fire" within it. In any case, Harvey does insist explicitly in the *De Generatione* that it is the blood in general, not specially the heart, that provides the warmth necessary to life, and he anticipates this doctrine in the *DMC*. At least he suggests that a speck of blood is the first to live and the last to die:

> I have . . . observed the first rudiments of the chick in the course of the fourth or fifth day of the incubation, in the guise of a little cloud, the shell having been removed and the egg immersed in clear tepid water. In the midst of the cloudlet in question there was a bloody point so small that it disappeared during the contraction and escaped the sight, but in the relaxation it reappeared again, red and like the point of a pin; so that betwixt the visible and the invisible, betwixt being and non-being, as it were, it gives by its pulses a kind of representation of the commencement of life. (W. 30–31)

And it would be the "primum vivens, ultimum moriens" that would have to contain the special warmth necessary to life.

Secondly, to orthodox physicians, with their view of the bipartite blood system, arterial on the left and venous on the right, it appeared obvious that the blood changed color in the heart. And Descartes's hidden fire could easily produce that change. Now from Harvey's lectures it appears that he observed the change of color in the lungs, although he had no very good way of accounting for it. (The lungs might be a cooling organ – as the brain had been for Aristotle!) Descartes, however, would not have known the lectures; this is simply a retroactive comparison.

Thirdly, however, there is the truly shocking allegation prominent in Harvey's argument, that what appears to be the diastole of the heart is in fact its systole. Traditionally, not only in the writings of Galen, but, as it appeared, in the experience of anatomists, the heart when it hardened and struck the chest, was stretching, and this is called diastole. Then when it collapsed, correspondingly, it narrowed, and that was called its systole. Descartes accepted this doctrine and in fact performed some experiments which seemed to him to confirm it.[16] Harvey, on the contrary, itemized in Chapter Two of the *DMC* the observations that had led him to deny this long accepted doctrine, and to declare

> . . . that the motion of the heart consists in a certain universal tension – both contraction in the line of its fibres, and constriction in every sense. It becomes erect, hard, and of diminished size during its action; the motion is plainly of the same nature as that of the muscles when they contract.

Continuing his description, he concludes:

> Hence the very opposite of the opinions commonly received, appears to be true; inasmuch as it is generally believed that when the heart strikes the breast and the pulse is felt without, the heart is dilated in its ventricles and is filled with blood; but the contrary of this is the fact, and the heart, when it contracts, is emptied. Whence the motion which is generally regarded as the diastole of the heart is in fact its

> systole. And in like manner the intrinsic motion of the heart is not the diastole but the systole; neither is it in the diastole that the heart grows firm and tense, but in the systole, for only then, when tense, is it moved and made vigorous.[17]

Now, given the long history of medical and anatomical authority and experience, this is a truly shocking statement. By Harvey's, and Descartes's, time, every one who knew anything knew about the valves in the veins, which kept the blood from flowing back into the heart, as Galenic ebb and flow would have it. So some kind of circular motion seems to follow reasonably enough. But to reverse diastole and systole is another matter. Even Harvey's admiring English disciple, George Ent, could not follow the master in this, but held that he had confused systole with diastole.[18] Like Descartes, he believed that some obscure "faculty" would be needed in the heart to explain Harvey's contraction, whereas, given the usual view of the heart's innate heat, the notion of that organ's warming and expanding the blood seemed much less mysterious.

In short, even if one happily accepts the new notion of the blood's circulation, it would be much more comfortable to retain the old conception of the heart's beat as its diastole. Besides, in general, muscles are structures controlled by the will; even though one ancient Hippocratic writer, in the *de Corde*, considered the heart to be a muscle, that was a strange and unlikely view. If it was a sac that swelled (a little) when it hardened, this would correspond to its innate heat, which every anatomist could feel, and to the familiar stretching (or diastole) identified with its beat. So it was possible to be fashionably up-to-date, accepting the identification of bodies with machines, and their activities with easily intelligible mechanisms, and still retain those long-accepted facts, the heart's special heat, the color change in it from veins to arteries, and its diastole when beating, its systole when flaccid and collapsed. The more archaic, vitalistic and teleological, approach, on the other hand, with its location of vital heat in the blood as such and especially with its alarming reversal of diastole and systole, would demand the abandonment of data learned from experience and authority, which the Cartesian reading would allow one to retain. So the more innovative thought style permits the conservation of features that observations made from what looks like a more conservative point of view would force one to sacrifice.

That is not to say that Descartes's mechanism was not in itself attractive; yet his account of the heart beat was at the same time less radical than Harvey's, and was, one would conjecture, more acceptable for that reason as well. It would allow one to make a great step forward intellectually, while at the same time happily assimilating details one had learned from one's masters and confirmed by the most direct, literally tangible, experience.

Department of Philosophy, Virginia Tech, USA

NOTES

[1] John Aubrey, *Brief Lives*, vol. I, p. 130.
[2] René Descartes, *Oeuvres*, vol. VI, eds. Adam and Tannery (Paris: Vrin, 1996), p. 50 (referred to as AT).

[3] AT XI, pp. 243–244.
[4] Ludwig Fleck, *Genesis and Development of a Scientific Fact* (Chicago: University of Chicago Press, 1979), p. 23, 100.
[5] This is the view supported chiefly by Walter Pagel; see his *Harvey's Biological Ideas* (Basel, 1967).
[6] W. 31–32.
[7] AT XI, p. 131.
[8] See the account by Thomas Fuchs in his *Mechanization of the Heart: Harvey and Descartes* (Rochester, NY: University of Rochester Press, 2001), pp. 141–175.
[9] Cornelis von Hoghelande, *Cogitationes* . . . (Leyden, 1676) (1st ed. 1646), p. 137, 124.
[10] *Ibid.*, p. 135f.
[11] See *ibid.*, pp. 160–164.
[12] Nicolaus Steno, *Discours sur l'Anatomie du Cerveau*, 1666, p. 8.
[13] *Discours*, vol. 5, AT VI, p. 48.
[14] *Description of the Human Body*, Part 2, At XI, pp. 239–245.
[15] *DMC*, chap. 4; *The Works of William Harvey, M.D.*, Trl. R. Willis (London: Printed for the Sydenham Society, 1847) (Johnson reprint, 1965), pp. 30–31 (referred to henceforth as W.).
[16] *Description of the Human Body*, AT XI, pp. 242–244.
[17] W. 22.
[18] George Ent, *Apologia pro Circulatione Sanguinis*, . . . (London, 1641), p. 137. For an account of Ent and other early commentators on Harvey, see Thomas Fuchs, *Mechanization of the Heart: Harvey and Descartes* (Rochester, NY: University of Rochester Press, 2001).

ALAN E. SHAPIRO

SKATING ON THE EDGE: NEWTON'S INVESTIGATION OF CHROMATIC DISPERSION AND ACHROMATIC PRISMS AND LENSES

In Experiment 8 of Book I, Part II of the *Opticks* (1704) Newton set forth two experiments from which he deduced a dispersion law that implied that it is impossible to construct an achromatic system of lenses or prisms. Although this law was first published in the *Opticks*, over thirty years earlier Newton had in his private papers proposed it independently of the problem of achromatism, so that it was not in fact derived from these experiments. In 1672, however, Newton had not yet accepted the law as a true one, nor had he related it to the impossibility of constructing an achromatic prism or lens. Indeed, in this period he did not believe that it was impossible to construct an achromatic lens and attempted to design and construct achromatic refracting systems. Likewise, in 1672 he described the two experiments with contrary results. How did Newton arrive at such dramatically different positions thirty years apart?

Experiment 8 is no minor matter, for by denying the possibility of correcting lenses for chromatic aberration, it had a pernicious influence on the development of optics. After John Dollond found both the experiment and dispersion law to be erroneous and succeeded in constructing achromatic lenses in 1758, Newton's pronouncements on optics were no longer taken as virtually infallible. In the nearly two hundred and fifty years since Dollond's announcement Newton's experiment has attracted much attention, and various explanations have been proposed for how he could have so erred.[1] The focus of these inquiries has been on the experiments Newton used to deduce his dispersion law, but I will focus on the law itself and attempt to uncover its origins by treating it as a quantitative mathematical law that can be considered independently of the problem of achromatism.

In order to understand what may have led him to adopt these positions, I will first examine the law itself, setting it within Newton's program to develop a mathematical theory of color, and analyzing its physical meaning and relation to other issues such as the musical division of the spectrum and achromatism. A primary concern here will be to study the approximations Newton used to connect the dispersion law to observable phenomena, for the relation between the law's truth and the impossibility of constructing an achromatic lens or prism is not at all straightforward and depends crucially on a small angle approximation. I will then examine the various experiments and observations bearing on it, together with their experimental errors, and show that at best they provided ambiguous support for his law. While the approach taken in

J.Z. Buchwald and A. Franklin (eds.), Wrong for the Right Reasons, 99–125.

this paper will clarify many problems in Experiment 8, critical ones will nonetheless remain: why Newton chose to obscure the real status of his law, fabricate the case in its behalf, and clothe Experiment 8 in an aura of superficial certainty.

To avoid unduly interrupting the historical narrative, I will first briefly define some optical terms as I shall be using them. Since Newton always expressed the index of refraction n as a ratio of two numbers, the sine of refraction R to the sine of incidence I, and generally considered refractions from a dense to a rare medium, we can write the index of refractions as

$$n = \frac{\sin r}{\sin i} = \frac{R}{I},$$

where i and r are the angles of incidence and refraction. For the refraction of rays of different color in the same medium, Newton adopts a natural physical interpretation: since all the rays of an incident beam of white light have the same angle of incidence but are refracted unequally, he always considers I to be a constant and R to vary. The term *dispersion*, or *chromatic dispersion*, will refer to the difference of the index of refraction Δn for the rays at the extreme ends of the spectrum, and in Newtonian notation will be expressed by ΔR, that is,

$$\Delta R = R_{\mathrm{p}} - R_{\mathrm{t}} \quad \text{and} \quad \Delta n = n_{\mathrm{p}} - n_{\mathrm{t}}.$$

The subscripts p and t will always refer to the most refrangible or extreme blue rays and to the least refrangible or extreme red rays respectively. The term *partial dispersion* refers to Δn or ΔR for some smaller part of the spectrum, such as for each of the principal colors, red, orange, yellow, and so forth. The *angular dispersion* $\Delta r (= r_{\mathrm{p}} - r_{\mathrm{t}})$ is the angle which the extreme blue and red rays make with one another after a ray of white light is refracted. The *dispersive power*, $1/\nu$, is the quantity

$$\frac{1}{\nu} = \frac{\Delta R}{R - I} = \frac{\Delta n}{n - 1},$$

where an n or R without a subscript represents the value for a mean refrangible ray. Finally, the *angle of deviation* $D(= r - i)$ represents the amount a ray is deviated by refraction from its initial path, and its measure is the angle contained between the incident and refracted ray.

EXPERIMENT 8

A proper analysis of Experiment 8 involves three distinct components, the experiments, the deduction, and the theorems. Before presenting such an analysis, however, let us first consider Experiment 8 in Newton's own words, for it is surprisingly brief:

> I found moreover, that when Light goes out of Air through several contiguous refracting Mediums as through Water and Glass, and thence goes out again into Air, whether the refracting Superficies be parallel or inclin'd to one another, that Light as often as by contrary Refractions 'tis so corrected, that it emergeth in Lines parallel to those in which it was incident, continues ever after to be

white. But if the emergent Rays be inclined to the incident, the Whiteness of the emerging Light will by degrees in passing on from the Place of Emergence, become tinged in its Edges with Colours. This I try'd by refracting Light with Prisms of Glass placed within a Prismatick Vessel of Water. Now those Colours argue a diverging and separation of the heterogeneous Rays from one another by means of their unequal Refractions, as in what follows will more fully appear. And, on the contrary, the permanent whiteness argues, that in like Incidences of the Rays there is no such separation of the emerging Rays, and by consequence no inequality of their whole Refractions. Whence I seem to gather the two following Theorems.

1. The Excesses of the Sines of Refraction of several sorts of Rays above their common Sine of Incidence when the Refractions are made out of divers denser Mediums immediately into one and the same rarer Medium, suppose of Air, are to one another in a given Proportion.

2. The Proportion of the Sine of Incidence to the Sine of Refraction of one and the same sort of Rays out of one Medium into another, is composed of the Proportion of the Sine of Incidence to the Sine of Refraction out of the first Medium into any third Medium, and of the Proportion of the Sine of Incidence to the Sine of Refraction out of that third Medium into the second Medium.[2]

Newton then continues with numerical examples of the theorems, and concludes with an encomium on the value of these mathematical theorems to the science of optics.

The experiments show that when a refracted beam of light emerges parallel to the incident beam (i.e., when there is no net refraction), there will be no dispersion or separation of the colors, and the beam will ever remain white; but if the emergent beam is not parallel to the incident beam, then colors will always be generated. The first theorem can be formulated as

$$\frac{R_\mathrm{p} - I}{R_\mathrm{t} - I} = \frac{R'_\mathrm{p} - I'}{R'_\mathrm{t} - I'} = \text{constant},$$

where R and I represent the sines of refraction and incidence, and the primes a second medium; the constant is independent of the media. This entails the dispersion law

$$\frac{\Delta R}{R - I} = \frac{\Delta R'}{R' - I'},$$

where $\Delta R = R_\mathrm{p} - R_\mathrm{t}$. I shall refer to this as the linear dispersion law. In modern notation the theorem becomes

$$\frac{n_\mathrm{p} - 1}{n_\mathrm{t} - 1} = \frac{n'_\mathrm{p} - 1}{n'_\mathrm{t} - 1},$$

and the dispersion law becomes

$$\frac{\Delta n}{n - 1} = \frac{\Delta n'}{n' - 1}.$$

In crude physical terms the law states that the chromatic dispersion ΔR or Δn in all substances is always proportional to what can be called the "amount of refraction" $R - I$ or $n - 1$. I call this simple physical interpretation "crude," because it depends on an implicit small angle approximation to get from $R - I$ to $r - i$, that is, the deviation, which is a true measure of the "amount of refraction" or what Newton calls the "whole refraction" or the "quantity of refraction."

The second theorem states that $n_{12} = n_{13} \cdot n_{32}$ or, transforming it into a more familiar form, that $n_{23} = n_{13}/n_{12}$ where the numbers 1, 2, 3, represent different media. The theorem then states that the index of refraction for a ray of any particular color passing from medium 2 to 3 is determined by the ratio of the indices of refraction from medium 1 to 3 and from 1 to 2. Since this law is correct and was proved by Newton independently of these experiments in his *Optical Lectures*, it need no longer concern us.[3]

Newton has described *two* experiments, one with parallel surfaces, and one with prisms. In general neither of them will turn out as described, and historical debate has centered upon explaining how he arrived at the contrary, particularly for the compound glass–water prism. Let us first consider the experiment with parallel plates, for it can be speedily dispatched. In the revised version of his *Optical Lectures*, he devoted an entire section to "the phenomena of light transmitted through a refractive medium bounded by parallel planes." He wrote disdainfully that,

> philosophers have hitherto believed that no colors are generated in this way, for they suppose that by a contrary refraction the second surface destroys all the effects in the rays that the first one induced. Instead of testing this they consider it as certain . . . But they are deceived in this . . .

Since the colors produced by refraction in parallel surfaces depend on the thickness of the refracting substance, colors are not seen in ordinary glass plates like windows, because "the colors are so fine and subtle and contained within such a narrow space that they escape the senses; but when thicker glasses are used, or preferably little parallelepipedal glass vessels full of very clear water, then colors are clearly perceived."[4] He then demonstrated that a narrow band of colors is produced on each edge because of the unequal refractions in the water; after leaving the parallelepiped all rays become parallel to the incident beam, and so to one another, independent of color. Strictly considered this experiment is not identical to the one described in Experiment 8, because Newton ignores the refractions in the glass plates forming the aqueous parallelepiped since they are so thin; but he is able to ignore them precisely because after passing through the plates the rays become parallel to the incident beam and thus do not alter the effect produced by the water alone. If the glass plates were thicker, the basic phenomena would not change, only the extent and quality of the colors. Newton's outright misrepresentation of his earlier experiments in the *Opticks* is puzzling, since it is superfluous, for the experiment with the compound prism is sufficient to establish his case.

The experiment with the compound water–glass prism has been the center of attention, since it was by repeating this experiment with flint glass in 1758 that John Dollond found Newton's claim in Experiment 8 to be wrong.[5] Moreover, it is

also possible to account for Newton's experimental result. Peter Dollond, the son of John, recounted in 1789 that when he repeated Newton's experiment with some old Venetian glass which appeared to correspond to the glass used by Newton, he was able to confirm his result.[6] In this century Boegehold calculated that if one assumes glass similar to Newton's, it is possible, when there is no net refraction, to eliminate the colors almost entirely.[7] Other explanations have been set forth, but these need not concern us, for there is no doubt that an experiment such as Newton described can very nearly yield his results under suitable conditions.[8] Any studies of his dispersion laws must accept as a fact that the dispersive powers of water and Newton's glass were nearly the same and that Newton knew this. The partial dispersions of water and glass are not, however, the same, so that a perfectly achromatic compound glass-water prism cannot be constructed; but Boegehold's calculations show that the residual colors can be made sufficiently small to attribute them to the roughness of the glass or some other disturbing cause.

It was inevitable, and not unreasonable, that it would be asked whether Newton actually performed his experiment with the compound prisms. Bechler's answer is that it is "an armchair experiment."[9] He bases his claim on two pieces of evidence: First, Newton has falsified the results of the experiment with parallel surfaces; and second, in an unsent draft of a letter to Hooke, written in the spring of 1672, he describes an experiment with a compound glass–water prism identical to Experiment 8 of the *Opticks*, but with contrary results. While I would not go as far as Bechler in denying that Newton did this experiment, I do consider Newton's account to be utterly unreliable; and indeed I can add one more piece of evidence to make it still more suspect. Around 1687 Newton began a treatise in Latin entitled *Fundamentum opticae* that is essentially a first draft of Book I of the *Opticks*. He then decided to revise this work and write it in English. Experiment 8 is in the *Fundamentum* and is essentially the same as in the *Opticks*, but with one crucial exception. The sentence that asserts that he did the experiment – "This I try'd by refracting Light with Prisms of Glass placed within a Prismatick Vessel of Water" – is lacking.[10] Evidently, in revision Newton realized that though this section is entitled "Experiment," he had in fact failed to describe an actual experiment and so added this one sentence to fill the gap.

Nonetheless, if it is granted that Newton did perform his experiment with a compound glass–water prism and observed what he claims to have observed, it would still not suffice to establish his dispersion law. First, he claims that this is a universal law applying to all transparent media, though he tested it for only one pair, water and glass. While one could not reasonably expect Newton to have tested every transparent substance available to him, he could have tested it for at least a few more pairs. Second, and far more crucial, Samuel Klingenstierna in 1754 rigorously demonstrated that Newton's experiment is in general incompatible with his law, and that for any pair of substances, the law and experiment will agree at only one set of angles.[11] Klingenstierna concluded his paper by observing that the two agree only for small-angled prisms. Newton says nothing about such a restriction to small angles. Indeed, he is surprisingly vague about his deduction of the dispersion law, just a "Whence I seem to gather." This inclines one to believe that all that is involved here is some physical intuition and an implicit small-angle approximation, but no rigorous deduction.

THEORETICAL BACKGROUND

Much of the difficulty in analyzing Experiment 8 stems from looking at the problem the wrong way, namely, the way Newton has set it up in the *Opticks* moving from experiment through deduction to a theorem. If, however, we look at the problem the other way, by moving from the theorem to the experiment, then the problems of non-experiment and vague deduction become tractable. I will show that Newton initially arrived at his dispersion law independent of achromatism, then became convinced of its validity while recognizing its relation to achromatism, and finally in writing the *Opticks* sought a justification for it. The theorem has been too closely associated with achromatism, rather than being viewed more generally as a dispersion law.

Both of the theorems of Experiment 8 first appeared in a draft of Newton's reply to Hooke's critique of his new theory of color in spring 1672. On a single folio separated from the rest of the letter Newton set forth two "Rules" which are essentially the same as the two theorems in the *Opticks*. They are presented without demonstration, and with no reference to the impossibility of achromatism. Such a reference, in fact, would be quite surprising, since one of the main thrusts of this letter was to show that contrary to Hooke's claim, colors can be generated even when there is no net refraction, and to this end he presented the first version of the compound-prism experiment.

This new dispersion law posed a problem for Newton, since he had been using a different one since 1666 and had structured most of the mathematical part of his *Optical Lectures* around that law. From 1672 onwards he was confronted by a choice between these two dispersion laws and was unable to choose decisively between them. In the *Optical Lectures* he presented his dispersion law as a mathematical construction without any physical interpretation, though it is based on Descartes' "velocity" model of refraction from *La Dioptrique*.[12] Newton's construction in modern notation yields the refraction law

$$\frac{n_{\mathrm{p}}^2 - 1}{n_{\mathrm{t}}^2 - 1} = \frac{n_{\mathrm{p}}'^2 - 1}{n_{\mathrm{t}}'^2 - 1} \approx \text{constant},$$

where the constant is independent of the media. This entails the dispersion law,

$$\frac{\Delta n}{\Delta n'} = \frac{(1/n)(n^2 - 1)}{(1/n')(n'^2 - 1)} = \frac{(n - 1)(1 + (1/n))}{(n' - 1)(1 + (1/n'))},$$

which I shall refer to as the quadratic dispersion law. By the factorization on the right we can see that quantitatively this law differs littles – only by the ratio of the factor $(1 + 1/n)$ – from the later, linear law of dispersion. The indices of refraction of the only two substances whose dispersion Newton also considered were $4/3$ for water and $17/11$ for glass, to use his own convenient ratios. With these values we find that the two laws differ by only $17/16$ or about 6%, which, as we shall see, would be essentially undetectable with Newton's method of measurement.

While I cannot explain how Newton arrived at his linear dispersion law, it is possible to trace its gradual emergence in his draft reply to Hooke. The new dispersion law is on a single folio separate from the rest of the letter.[13] Judging by the handwriting and

content, there is little doubt that it belongs with this letter as a late insertion. Although Newton does not indicate where he intended to insert it, it seems to belong in the draft after he explains to Hooke that,

> The law of refractions on w^{ch} I proceeded was that of Des-Cartes applyed to every particular sort of rays. Namely that rays of divers sorts or colours are refracted according to divers ratios of the sines, & that these ratios in the same sorts of rays are constantly the same.[14]

The sine law of refraction, he states, is an exact law, i.e., if the ratio of sines is known for the extreme rays, then "the refractions of those rays may be easily determined in all cases by the Rule & Compasse." But, as the draft stands without the intended insertion, he continued:

> If no accurate determination be intended, but onely a rude conjecture at the circumstances so far as not sensibly to vary from the truth, it may suffice to estimate them by assuming the greatest difference of the refractions of rays alike incident to beare a certain proportion to the whole refractions. As if the refraction be out of Air into glasse the difference is about a 36^{t} part of the whole, & if out of glasse into Air it is about a 24th part, & in other Mediums it hath other proportions. But this is not accurate enough to determin the circumstances of all experiments.[15]

Despite some difficulties of interpretation, Newton has here invoked the approximation:

$$\frac{\Delta r}{r-i} \approx \frac{\Delta R}{R-I} \approx \text{constant} \tag{1}$$

where I have set the constant, which depends on the medium, equal to the dispersive power, $\Delta R/(R-I)$, both on physical grounds, and because it is consistent with the *Optical Lectures*' numerical values and its approximation for the refraction in a lens.[16] The right-hand side of the equation would be the linear dispersion law if the constant were the same for all media, although Newton explicitly asserts the contrary here. It is more important to note, however, that he has recognized that the ratio of the angular dispersion Δr to the deviation $r-i$ is proportional to the dispersive power, and that this is only an approximation. Newton's deduction in Experiment 8 involves, as I have already observed, three distinct elements, the experiments, the dispersion law, and some relation linking the law to the experiments. It is this approximation which links them and allows him to "seem to gather" his dispersion law.

The approximation is the standard small-angle approximation of replacing the sines by their angles, and when the angles are truly small, it is a perfectly valid one for refractions in prisms and lenses. Newton does not tell us how he arrived at this proportion, but there can be little doubt that he simply replaced the sines by their angles. In the second version of his *Optical Lectures*, Part II, Proposition 37, he explicitly used just this approximation of replacing the sines by their angles to arrive at equation (1) for the refraction of a ray traversing a lens.[17] He had calculated the

chromatic aberration in a plano-convex lens in order to show that it is more than 1500 times greater than the spherical aberration: "so exceedingly great a disproportion" that the spherical aberration "can be considered as nothing" with respect to the chromatic aberration.[18] This impediment to the further development of the refracting telescope is a constantly recurring theme in Newton's optical writings, starting with his first publication, "New theory about light and colors" (1672).[19] Newton's derivation of equation (1) for lenses was a significant result, for it expresses the chromatic aberration in terms of the dispersive power. By a straightforward physical argument, it showed that if the chromatic aberration is to be eliminated by a compound lens, i.e., if the angular dispersions Δr of two lenses are to cancel and so be equal and opposite, then when each lens had the same dispersive power, there would be no net refraction, for then the deviation $r - i$ of each lens would also cancel. At this time, though, Newton neither asserts that the dispersive powers of all substances are equal nor explicitly invokes equation (1) for prisms.

It is only in the *Opticks* that we find the explicit conjunction of all these factors, the small-angle approximation and linear dispersion law applied to prisms and lenses. It is not, however, to be found in the inscrutable Experiment 8 where these factors are only implicit, but in Book I, Part I, Proposition 7, "The Perfection of Telescopes is impeded by the different Refrangibility of the Rays of Light." Newton begins this extended discussion of lens aberrations with an account of his measurement of the index of refraction and dispersion of three different glass prisms. He found "in the least round Numbers" that the ratio of the sines of refraction R to the sine of incidence I for the most and least refrangible rays to be as 78 and 77 to 50. He then states that,

> Now, if you subduct the common Sine of Incidence 50 from the Sines of Refraction 77 and 78, the Remainders 27 and 28 shew, that *in small Refractions* the Refraction of the least refrangible Rays is to the Refraction of the most refrangible ones, as 27 to 28 *very nearly*, and that the difference of the Refractions of the least refrangible and most refrangible Rays is *about* the $27\frac{1}{2}$th part of the whole Refraction of the mean refrangible Rays.[20]

For small refractions Newton correctly asserts that

$$\frac{R_t - I}{R_p - I} = \frac{27}{28} \approx \frac{r_t - i}{r_p - i}$$

and

$$\frac{R_p - R_t}{R - I} = \frac{\Delta R}{R - I} = \frac{1}{27\frac{1}{2}} \approx \frac{\Delta r}{r - i}.$$

Bearing in mind that R and I represent the sines of refraction and incidence, $\sin r$ and $\sin i$, Newton has simply replaced the sines by their angles. Passing from prisms to lenses, he then specifies both the longitudinal and lateral chromatic aberrations in terms of the dispersive power. Since in Experiment 8 Newton was in fact invoking an approximation, and one which he had noted in his draft to Hooke "is not accurate enough to determine the circumstances of all experiments," we can now appreciate

his vague "Whence I seem to gather." Nonetheless, despite the approximation, his position was not as tenuous as it may at first seem so long as it turned out that the law was valid for small angles, for, as Klingenstierna observed, this still would entail the impossibility of constructing an achromatic lens, since the angles involved here are truly small.

Let us now leave the *Opticks* and return to 1672 and Newton's draft reply to Hooke. Sometime after he formulated equation (1) in which the dispersive power differs for different media, on a separate sheet he drew up his new dispersion law that asserted that the dispersive powers of all substances are equal. He here couples the dispersion law, "the other Rule," with a "first Rule" for determining relative indices of refraction, as he was later to do in Experiment 8:

> To these I added two other Rules, whereof one was to know the proportion of the sines measuring y^{e} refractions of homogeneall rays made out of one Medium into another, by knowing the proportions of y^{e} sines measuring the refractions of those rays made out of Air or any third Medium into those two. And y^{e} other was to know the difference of the refractions of heterogeneall rays alike incident out of any Medium into any other Medium, by knowing the difference of their refraction out of glasse into Air.
>
> The first Rule is, that as y^{e} ratio of y^{e} given sines of incidence is to y^{e} ratio of y^{e} given sines of refraction, so are y^{e} desired sines of incidence & refraction to one another . . .
>
> The other Rule is, that if refractions be made out of divers Mediums into one common Medium wth equall incidence, the differences between the common sine of incidence & y^{e} sines of y^{e} refractions of differm rays shall have a given ratio.[21]

Newton presents no demonstration, experimental or theoretical, for these rules, nor does he attempt to justify them in any way. He illustrates them with specific numerical examples (to be discussed shortly) calculating, for instance, the indices of refraction of the extreme rays for refractions from water to air and water to glass, but that is all. The dispersion law is essentially identical to that in the *Opticks.*

These two "Rules" together with Descartes' sine law of refraction were to serve as the foundation of Newton's new mathematical science of color. The sine law defines the refraction for each ray at any incidence in a given medium; a dispersion law relates the refractions of rays of each color to one another in any given medium; and the law of relative refractions relates the refractions in one medium to those in any other. Given these rules, if the refraction of a single ray at one angle of incidence is known in any medium, then the refractions of all other rays in that medium may be determined for any angle of incidence without any additional measurements. This is the ideal of a rational optics. In the *Optical Lectures*, Newton had attempted to develop his new science on the same family of three laws, except that the dispersion law differed. The particular form of the dispersion law is irrelevant in this respect, although it is essential that one exists, for without a universal dispersion law a rational optics cannot be constructed. Newton seems almost as concerned in his optical work to provide the science of color with a mathematical foundation as he is to establish his

own theory of color, and near the beginning of the *Lectures*, he informed his audience of his new program:

> the generation of colors includes so much geometry, and the understanding of colors is supported by so much evidence, that for their sake I can thus attempt to extend the bounds of mathematics somewhat . . . Thus although colors may belong to physics, the science of them must nevertheless be considered mathematical, insofar as they are treated by mathematical reasoning. Indeed, since an exact science of them seems to be one of the most difficult that philosophy is in need of, I hope to show – as it were, by my example – how valuable mathematics is in natural philosophy.[22]

In his "New Theory" Newton alluded to the certainty of his mathematical science of color, only to have that contention summarily dismissed by Hooke.[23] Newton reiterated his claim in his reply to Hooke, and then immediately turned to the laws of refraction that he had used, i.e., Descartes' sine law of refraction for each sort of ray and the approximation of equation (1). It is at this point that I believe that he intended to insert his two new rules, while deleting, or at least modifying, the paragraph with equation (1). This location for the insertion is supported by a reference later in the draft to the sine law together with these two rules – his "three Principles" – which, he tells us, were to serve as the foundation of a new mathematical science of color. After presenting his calculations for the colors generated by a compound glass–water prism in his first version of this experiment, and observing that other problems, such as the colors produced in a compound glass–water lens "may come under the same rule of computation," he asserts:

> And all this I suppose would afford subject & scope enough for any Mathematician that should undertake to compose a Science & gradually demonstrate it from *the first three Principles* w^{ch} *I layd down for* y^{e} *Rules whereby to measure refractions*; & together w^{th} the Theoremes determin such Problems as may occurr concerning the greatest or least extent of colours in any case, or concerning y^{e} ways whereby their quantity may be augmented or diminished or absolutely destroyed in Optique Instruments, or in y^{e} Eye it selfe, to y^{e} intent that if Theories can w^{th} sufficient exactnesse be put in practise, o^{r} vision may be perfected . . . And I see no reason to doubt why such a science being well composed should not prove as certain pleasant & usefull as that part of Optiques w^{ch} is already in being.[24]

The "three Principles layd down" for "y^{e} Rules whereby to measure refractions," it seems to me, can refer to nothing else but the sine law of refraction *and* the two new "Rules" which were to be inserted here.[25]

There is one more piece of evidence that Newton had intended to insert his new refraction rules in this draft. It is, rather surprisingly, to be found in Experiment 9, where Newton describes an achromatic compound glass-water prism. Experiment 9, like the refraction rules, is a late addition, but it is not separated from the rest of the letter.

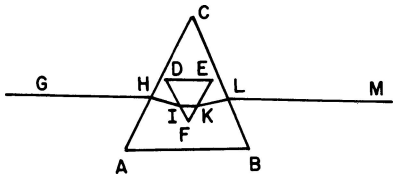

Figure 1. Newton's compound glass–water prism, Experiment 9 in his draft reply to Hooke (1672).

The experiment begins with a general description of the experimental arrangement and the phenomena observed:

> Let *ABC* [Figure 1] represent a Prismatic Box made of polished plates of glasse cemented together at their edges: & this being filled with water, put into it another Prism of solid glasse or crystall *DEF*. And the light transmitted through them both will variously exhibit colours according to their severall positions. But that w^{ch} I would here observe is that if their Bases *DE* & *AB* be parallel & opposite to one another so that y^{e} refractions of the glasse within the water be contrary to those of y^{e} concave Prism conteining the water, [1] there will be colours generated when the contrary refractions are as neare as may be equall to one another. But [2] if the refractions of interior Prism be something lesse then those of the exterior the transmitted light may be wthout any colour at all.[26]

These two claims are directly contrary to those of Experiment 8 some three decades later: (1) asserts that when the emergent beam was parallel to the incident one, colors were observed; while (2) asserts that he had observed no colors when these beams were not parallel, or that he had constructed an achromatic prism.

Newton then describes his observations in some detail:

> Thus I have observed that when the angle *DFE* of the glasse-Prism was 64^{degr} 8^{min}; & y^{e} angle *ACB* of the water-Prism 49^{degr} 28^{min}; if I so placed these Prisms that y^{e} light *GHIKM* was equally refracted on either side y^{e} water Prism at *H* & *L*, & also on either side the glasse-Prism at *I* & *K*; the emerging light *LM* was inclined to y^{e} incident light *GH* at an angle of about 50^{min} being more refracted at *H* & *L* then at *I* & *K*, & yet there was no appearance of any colour at its edges. Whereas by turning the Prism *DEF* a very little way about its axis to augment y^{e} summe of y^{e} refractions at *I* & *K* & make them equall to the summe of those at *H* & *L*, there both manifestly appeared a blew & purple colour at that side of y^{e} transmitted light towards *C* the verticall angle of the water Prism, & a red & yellow on the other side. The same thing also happened by diminishing the angle *ACB* of y^{e} water Prism so much as without varying the position of the glasse within it to make the refractions at *H* & *L* equall to those at *I* & *K*. For y^{e} Prism *ACB* was so made that I could at pleasure alter any of its angles.[27]

Table 1. The index of refraction from the refraction rule and Experiment 9 of Newton's draft reply to Hooke spring (1672)

	Experiment 9	Refraction rule
From air to water	$n_{\mathrm{p}} = \dfrac{425056}{317157} = 1.34021$	$\dfrac{97\frac{1}{2}}{72\frac{3}{4}} = 1.34021$
	$n_{\mathrm{t}} = \dfrac{425055}{320443} = 1.32646$	$\dfrac{96\frac{1}{2}}{72\frac{3}{4}} = 1.32646$
From water to glass	$n_{\mathrm{p}} = \dfrac{619900}{530906} = 1.16763$	$\dfrac{264\frac{9}{3}}{236\frac{9}{13}} = 1.16763$
	$n_{\mathrm{t}} = \dfrac{617177}{530906} = 1.16250$	$\dfrac{264\frac{9}{3}}{227\frac{9}{13}} = 1.16250$

Newton writes here as if he had performed this experiment and made these observations; after considering his calculations I will return to examine this point.

"Because y$^{\mathrm{e}}$ calculation for this Experiment is more necessary & something more troublesome then for any of those that have preceded," Newton presents only one of the cases, that where no colors are generated when the emergent beam makes an angle of 50′ with the incident one. The calculation is carried out only for the extreme rays, "And for determining their refractions made in their passage out of any one into any other of these three medium[s], glasse water & Air, I make use of those proportions of the sines w$^{\mathrm{ch}}$ I have already mentioned."[28] These "already mentioned" proportions of the sines, or indices of refraction, are not to be found anywhere in the draft except in the refraction rules. Newton's calculations are straightforward enough, and so need not detain us other than to extract the values he used for the ratios of sines. Table 1 shows at a glance that the values from the calculation of Experiment 9 and the refraction rules are identical; they were calculated by the linear dispersion law using the *Optical Lectures*' $v = 24$. Newton, however, made an arithmetical error in originally calculating the index of refraction from water to glass and carried the error into Experiment 9;[29] the correct values should be

$$n_p = \frac{268\frac{7}{26}}{230\frac{7}{26}} = 1.16502 \quad \text{and} \quad n_{\mathrm{t}} = \frac{268\frac{7}{26}}{231\frac{7}{26}} = 1.15999.$$

It is therefore certain that the values for the refractive indices which he used in his calculations were not experimentally determined, but calculated ones taken from the refraction rules. Using these erroneous values, Newton computed that when no colors were observed, the refracted beam made an angle of 50′ 30″ with the incident one: and he found "50 1/2$^{\mathrm{min}}$ according to y$^{\mathrm{e}}$ Experiment."[30] The properly calculated values show that this angle should be about 67′, while the emergent beam is no longer strictly achromatic, since the extreme rays now diverge 25″ from each other. To be sure, these differences would not be observable, but then neither would the initial deviation of 50 1/2′. Newton himself was clearly aware of both the delicacy of the observations and sensitivity of the calculations, for he concluded this experiment by observing:

> In like manner may be computed the Phaenomena of these Prisms in any other position to one another, as also those of the sucessive refractions of any other Mediums. But there is requisite a very great exactnesse in making the Experiments, & in measuring the angles of the Prisms & y^{e} refractive de[n]sities of y^{e} Mediums. For a very small error in any of those will cause a very great disagreement in the computation.[31]

Experiment 9 was probably not originally formulated with the problem of achromatism in mind, but rather, as Bechler has observed, to refute Hooke's color theory and to establish the power of his mathematical science of color.[32] Hooke's theory of color asserted the contrary to the initial two claims of Experiment 9. The outcome of Newton's experiment, however, was perilously close, a mere 50 1/2′, to supporting Hooke. If the observations and calculations were ever so slightly different, the opposite conclusion – that of Experiment 8 – could follow. This is in fact the actual situation, for with water and Newton's glass it is possible, as Boegehold showed, to make a nearly achromatic combination, but only nearly. In short, this experiment is ambiguous and inconclusive. Thus the observational claims of Experiments 9 and 8 are not as opposed as they at first seem, differing by less than 1°: it is rather their conclusions that are so radically different. It is possible, then, that Newton performed the compound-prism experiment, though it is clear that neither account is an honest one: Experiment 9 relates a calculated result as an observed one, and Experiment 8 is contradicted by Experiment 9.

This problem serves to bring into sharp relief the decisions, indeed dilemmas, confronting Newton. For, even granting the truth of the linear dispersion law, it could not rigorously exclude the possibility of constructing a compound achromatic prism, for they are related only by an approximation. It would, however, exclude the possibility of constructing a compound achromatic lens, for the angles involved here are truly small. But was it impossible to construct an achromatic lens? Newton did not yet think so, though, as he reported to Hooke, he appears to have attempted to construct one without great success.[33] Moreover, was the linear dispersion law true? It had apparently been derived, as his earlier quadratic dispersion law, on theoretical grounds and so remained to be confirmed (or disproved) experimentally. And in any case, at most one of these laws could be true.

Newton's views were in a state of flux in this period beginning in 1672, and he was never able to choose decisively and confidently between many of these alternatives. Bechler has shown that his commitment to his velocity model of refraction, upon which his quadratic dispersion law was based, began to wane during this period, and I suspect that the appearance of his new, linear dispersion law was not unrelated to this lessened commitment.[34] It is, furthermore, important to bear in mind that Newton did not at this time make either of his dispersion laws public; the *Optical Lectures* remained unpublished, and this draft to Hooke was suppressed. To complicate matters further, during this same period Newton was revising his *Optical Lectures* and in revision had introduced his musical division of the spectrum.

Newton's musical division of the spectrum defines how the spectrum is to be divided and therefore is itself a partial dispersion law with major implications for a dispersion law. Underlying both the partial dispersion law of the musical division

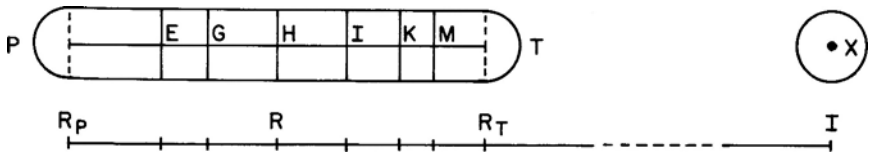

Figure 2. The partial dispersion law of the musical division of the spectrum.

and the linear dispersion law is Newton's idea that the spectrum is always divided in the same way independent of the refracting medium. Indeed, with a few simple and natural assumptions the partial dispersion law implies his linear dispersion law.

Newton cast a spectrum from a glass prism onto the opposite wall and divided this image into seven principal portions corresponding to seven principal colors. In Figure 2 the portion *TM* represents the red, *MK* the orange, *KI* the yellow, *IH* the green, *HG* the blue, *GE* the indigo, and *EP* the purple. When the spectrum was so marked off, Newton found that, "everything appeared just as if the parts of the image occupied by the colors were proportional to a string so divided so it would cause the individual degrees of the octave to sound." Yet, after specifying the particular scale adopted, he was forced to admit that,

> I could not, however, so precisely observe and define these without being compelled to admit that it could perhaps be constituted somewhat differently. For instance, if . . . one takes eleven mean proportionals . . . this distribution of the image will also seem to fit the colors' expanses sufficiently well. For such quite minute differences that occur between this and the above distribution can produce errors hardly visible to the keenest judge.

The two proposed distributions of the string, whose length is assumed to be 720 parts, yielded the following divisions:

360 . 320 . 300 . 270 . 240 . 216 . $202\frac{1}{2}$. 180 Musical division

360 . 321 . 303 . 270 . 240 . 214 . 202 . 180 Geometrical division

The difference between these two divisions could cause the space occupied by any color to vary by more than 10%; for instance, blue occupies 33 geometrical divisions but only 30 musical ones. Newton forthrightly admits that although the difference between these divisions was unobservable, "I have, to be sure, preferred to use the upper distribution, not only because it agrees with the phenomena very well, but also because it perhaps involves something about the harmonies of colors . . . perhaps analogous to the concordances of sounds."[35]

Thus far Newton was merely adding a new twist to the millennia-old quest to uncover a relation between musical harmonies and colors. He adds, however, that by a "mechanical procedure," or an approximation, the refractive index for each color can be determined by dividing the interval between the two extreme colors in the same proportion as the spectrum. Since he had earlier found for glass that $I = 44\,1/4$, $R_t = 68$ and $R_p = 69$, "divide the intermediate unit in the ratio of the parts of this image; and there will result 68, 68 1/8, 68 1/5, 68 1/3, 68 1/2,

68 2/3, 68 7/9, 69 for the sines pertaining to the boundaries and ends of the seven individual colors." Newton intends his rule to apply universally to all substances; the numerical examples are merely illustrations. In conclusion he admonishes us to "remember here that these determinations are not precisely geometric but still as nearly accurate as practical matters of this kind require."[36] The approximation that the angular dispersion ΔD is proportional to the dispersion Δn or ΔR is a valid first order approximation introducing an error far smaller than the observational error in determining the boundaries of the colors.

Newton's partial dispersion law in essence asserts that the expanse of each color is always a fixed proportion of the spectrum's length, independent of that length and the refracting substance. Let us now imagine in Figure 2 that the spectrum seen on a wall is just the visible portion of a spectrum which extends from purple through red to still less refrangible rays all the way to rays with zero degree of refrangibility which pass through the prism unrefracted and fall at X. Now, just as the red rays always occupy some fixed proportion of the "visible" spectrum PT, conceive the "visible" spectrum itself always to occupy a fixed proportion of the entire spectrum, "visible" and "invisible," extending from P to X, independent of its total length and the refracting substance. By this natural extension of the ideas underlying the partial dispersion law of the musical division, we have arrived at the linear dispersion law; for taking the length of the "visible" spectrum proportional to ΔR, as Newton himself does, and that of the entire spectrum proportional to $R - I$, we find $\Delta R/(R - I) = \text{constant}$. Considering the lengths of the "visible" and of the entire spectrum to be respectively proportional to ΔR and $R - I$ (which are equivalent to small-angle approximations), and *imagining* that there is an "invisible" portion of the spectrum composed of rays with a continually decreasing degree of refrangibility, both have the warranty of Newton's own usage elsewhere in the *Optical Lectures*.[37] This is not intended to be a reconstruction of Newton's derivation of the linear dispersion law. Rather it is intended to show that the same physical conception underlies both the partial dispersion and the linear dispersion laws. There can be no doubt that Newton himself considered his partial and linear dispersion laws to be closely related, for in the *Opticks* they compose the two experiments, 7 and 8, of Book I, Part II, Proposition 3, "To define the Refrangibility of the several sorts of homogeneal Light answering to the several Colours."[38]

Before leaving the theoretical aspect of Experiment 8, we should not forget that Newton suppressed much of the draft to Hooke, in particular, the ninth experiment with the compound glass–water prism, and the two refraction rules. Indeed, all references to dispersion laws were eliminated from the letter he finally sent to Hooke, as were the remaining nine experiments some of which showed – contrary to Hooke's belief and the Newton of Experiment 8 – that colors were generated when a beam of white light traversed parallel refracting surfaces. At roughly the same time, in revising the *Optical Lectures* Newton kept his quadratic dispersion law, the foundation of much of the mathematical portion of the *Lectures*, while he added the musical division of the spectrum. The quadratic dispersion law and the partial dispersion law derived from the musical division are strictly incompatible, but in the *Lectures* this caused no problem, since the derivation of the partial dispersions from the musical division was only considered to be "mechanical" or approximate. By 1672 Newton's optical

investigations had led him to a set of related problems which he had yet to resolve. When he returned to them in the early 1690s in composing the *Opticks*, he would frequently arrive at very different conclusions.

EXPERIMENTAL EVIDENCE

In his investigations of refraction and dispersion Newton relied primarily on theory and turned to empirical measurements sparingly. In the *Optical Lectures* he readily conceded that he did not derive his quadratic dispersion law from experiment and had not attempted to confirm it experimentally:

> I have not yet, to be sure, derived the certainty of this theorem from experiments, but since it appears scarcely to differ much from the truth, for the present I have not hesitated to assume it gratuitously. In the future perhaps I will either confirm it by experiment or, if I will find it to be false, correct it.[39]

This attitude is a far cry from the Newton who throughout his career publicly rejected a science based on hypotheses. Newton did of course make some measurements, but these (or at least the ones he chose to make public or preserve in his papers) were just sufficient to derive the fundamental parameters for his dispersion model. He devoted a portion of his *Lectures* to the measurement of refraction, but the only experimentally derived results he presents there are the mean refraction of glass and water, and the dispersion of glass. All other values, such as the dispersion of water and the indices of refraction from water to glass are calculated. At various times he made other measurements, such as the chromatic aberration of a lens, which directly bear on the linear dispersion law. I will analyze these measurements to see how much support they provided for his linear dispersion law. We shall see that it was minimal.

Newton measured indices of refraction and dispersion by the method of minimum deviation, an elegant method entirely new with him and now commonly accepted as the most accurate means of measuring indices of refraction. He placed a prism (Figure 3) with a refracting angle C of about 60° in a sun beam admitted through

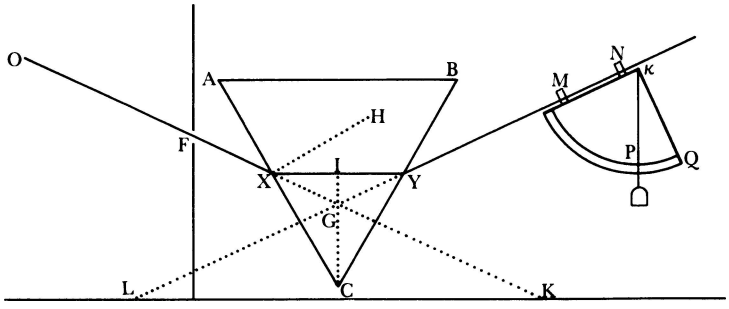

Figure 3. Newton's method of minimum deviation for measuring the index of refraction, which is described in the *Optical Lectures*.

a small hole F, and so positioned the prism that the central ray passed through it symmetrically; in this position the deviation, or the angle made by the incident and the twice-refracted rays, is a minimum, and the angles of incidence and refraction are equal. Then using a quadrant to measure the sun's altitude, angle OKL, and that of the center of the image, angle MLK, he calculated the index of refraction by means of an expression he had derived that is essentially equivalent to that found in any modern optical textbook:

$$n = \frac{\sin[(C + D)/2]}{\sin(C/2)},$$

where C is the prism's refracting angle and D the deviation.[40] He measured all three angles with a quadrant having a one-foot radius.

To determine the dispersion Newton did not reset the prism to minimum deviation for each of the extreme rays and then measure its index of refraction as he did for the mean refrangible one. Rather he directly measured the angular dispersion when the prism was set at minimum deviation for the mean refrangible rays. This required three additional measurements, the spectrum's length and breadth and its distance from the hole. The length of the sun's spectrum minus its breadth to correct for the sun's finite size – its corrected length – divided by its distance from the hole is the angular dispersion; when its half is added and subtracted to the mean deviation, the index of refraction for the extreme rays is then calculated by equation (2).[41] The errors introduced by not making independent measurements of each refractive index at minimum deviation are quite small, as Newton himself recognized, and are far more than compensated by eliminating the effect of systematic errors, since in this way all three indices of refraction will be uniformly affected and tend to increase and decrease together.[42]

We are now confronted by two separate issues: What is Newton's assessment of his measurements, and what is our modern reassessment of them?[43] Newton did not, of course, express his judgement in terms of percentage of error and mean deviations, an approach foreign to 17th-century science. He did, however, consider the issue. In the *Lectures*, after presenting his method of minimum deviation and some illustrative measurements, he offered the following assessment of it:

> a small departure from the [prism's] required position is almost inconsequential, insofar as the deviation MGK will not be perceptibly changed from this, as will be clear to anyone trying it. Namely, that angle is then a minimum, and when quantities generated by motion are either a maximum or minimum – that is, in the moment of regression – their motions are generally infinitely small. . . . But if the prism were placed in any other position than that described here . . . the least departure from that required position would greatly change the deviation, and the experiment would thus be much more liable to uncertainty and errors.[44]

Newton, as we should expect, has grasped the essential virtue of his method of minimum deviation: the errors too are at a minimum.

Newton's judgement that the errors due to misplacing the prism are "inconsequential" is valid. One way to appreciate this is "by trying it," as he suggests. Another way, in the absence of a formal theory of error analysis, is simply to perform the requisite calculations using the experimentally determined values but altering them one by one by a slight amount equivalent to the estimated experimental error in each value.[45] For instance, if we vary the prism angle by 1°, while keeping the deviation fixed, the index of refraction differs from Newton's measured value by less than 0.8%. Likewise, if it is assumed that the deviation, which involves two independent angular measurements, errs by as much as 2°, this will cause a difference of 1.25% with respect to the actually measured value. These calculations clearly show the advantage of using a large-angled prism at minimum deviation, just as Newton had claimed. The angular errors of a degree or two that I have substituted are probably realistic evaluations of the precision that is to be expected of measurements made with a one-foot quadrant and in placement of the prism.[46]

The technique of varying each parameter slightly, while keeping the others fixed, to see how it affects the observed result can be applied experimentally, and this was undoubtedly one of the principal means of estimating errors in the era before formal error analysis had developed. Newton frequently employed this method and invoked it in his most extensive analysis of his measurements in a letter to Anthony Lucas in August 1676. Newton's aim was to defend his claim in the "New Theory" that he found the spectrum's length to be more than five times its diameter, against Lucas' charge that it was never greater than three or three and a half times greater. Newton argued that much of the difference could be attributed to the prism angle, since Newton's angle was over 3° larger, and if Lucas had only estimated the angle or measured it inexactly it could easily be more than 5° larger.[47] After repeating his measurements a number of times, Newton was convinced that "my own observation was exact enough." Nonetheless, "that it might appear experimentally how y^{e} increase of y^{e} angle increases y^{e} length of y^{e} Image . . . I tryed y^{e} experiment wth divers angles, and have set down my trialls in y^{e} following table."[48] Newton's table with measurements of the spectrum made with all three angles on three prisms clearly indicates the sensitivity of the spectrum's length to changes of the prism angle. Thus an increase in the angle of a bit over 1 1/2° causes the length to increase 5/8 inch, which is almost an 8% increase in the corrected length. His table also showed that measurements made on a clearer day are about 5% larger and that the curvature of the prism's faces affected the lengths.

In contrast to the index of refraction, where the deviation and so the errors are at a minimum, the spectral length is not a minimum but continually increases.[49] Since the dispersion Δn depends directly on the corrected spectral length, we can therefore appreciate why Newton's measurements of the index of refraction could be so precise – at least for the 17th century – while his measurements of dispersion could be subject to errors an order of magnitude larger. Newton has already introduced a sequence of factors – mismeasuring the prism angle, variations of the sun's brightness, and curvature of the prism's faces – which could cause an error on the order of 5% each.

Altogether these factors, as well as still others, could cause his measurement of the dispersion to be in error by 10% or even 15%.

In fact, in measurements on two different glass prisms Newton found dispersive powers that differed by as much as 13%. In the *Optical Lectures* using a prism with a refracting angle of $63^\circ\,12'$ he found the dispersive power $v = 24\,1/2$ and the mean index of refraction to be $69/44\frac{1}{2}$; whereas in the *Opticks* with a prism whose refracting angle was $62\,1/2^\circ$ he found $v = 27\,1/2$ and the mean index of refraction to be $77\frac{1}{2}/50$.[50] One might interpret this as simply the substitution of *better* values, but I believe that they must be considered *different* values. In the *Fundamentum opticae* and in the manuscript of the *Opticks*, Newton had already substituted a different, and presumably better, measurement of the dispersion of the same prism used in the *Optical Lectures*.[51] Thus, Newton had made at least two independent measurements of this prism's dispersive power and was content with the results, as he should have been, since they were virtually identical. In the manuscript of the *Opticks* one can see where he later struck out the remeasured values and inserted the new series of measurements that appear in all the published editions.

Thus, Newton had measured the dispersive power of two different kinds of glass and found values differing by 13%, a discrepancy which, while not actually disproving his linear dispersion law, put it into serious doubt. The immediate cause for his adopting this new value and compromising his dispersion law was his frustrating attempt to measure the chromatic aberration of a lens (*Opticks*, Book I, Part I, Proposition 7, Experiment 16). A lens's longitudinal chromatic aberration (the difference of focal length for the extreme colors) varies directly as its dispersive power and focal length. Newton's initial aim in this experiment was not to determine the lens' dispersive power, but rather assuming it to be known, to show that the chromatic aberration is indeed as large and hence as serious a problem as he had persistently claimed. As a draft of this proposition shows, he initially was prepared to accept a measured value of the chromatic aberration substantially smaller than the predicted value, because the extreme red and violet were "too faint & thin to make them sufficiently visible."[52] Newton, however, did not long remain content to accept this discrepancy and instead adopted new values for both the lens' dispersive power and focus, thereby bringing the measured and calculated values of chromatic aberration into near perfect agreement. Consequently, he substituted the new set of measurements yielding the smaller dispersive power for glass and then changed the value for the dispersion everywhere else it appeared in the *Opticks*, such as in the calculation of the rainbow's diameter.[53] Newton evidently considered this to be a better value for the dispersive power of glass rather than a different one, for – entirely consistent with his linear dispersion law – he assumed that the measurements with the glass prism applied to the lens, as well as all the different glass that he used in other experiments elsewhere in the *Opticks*. The new value gave much better experimental agreement for the chromatic aberration of a lens and the diameter of the rainbow, reinforcing Newton's judgement on the validity of the new value.[54] A little later in his revision of the *Opticks* he would make an equally dramatic 11% change in the value for the fits (corresponding

to the wavelength in modern theory of light) that gave better agreement for an independent set of measurements of the colored circles seen in thick plates. Newton's quest for experimental consistency in his measurements was an innovation that he introduced into physics from astronomy and should not be viewed as merely a cosmetic touch.[55]

Newton's measurements of the dispersive power of two different kinds of glass brought him, as I see it, perilously close to disproving his linear dispersion law, but we should investigate whether he had additional evidence from measurements in other substances, such as water. Newton never presented a direct measurement of the dispersive power of water, or any other substance besides glass, but always calculated it from his dispersion laws, using in the *Optical Lectures* the quadratic dispersion law, and in the *Opticks* the linear dispersion law.[56] Since the breadth of the colors of the rainbow depends on the dispersion of water, it may seem as if this could serve as an empirical test for its value. It turns out, however, that these were insufficiently sensitive to test his dispersion law.[57]

Newton reports another observation relating to the dispersion of water that, I believe, may provide a key to understanding Experiment 8. In the account in the *Opticks* of his basic experiment of allowing a narrow sun beam to pass through a prism and cast its elongated spectrum on the wall, he considers the various factors that could cause it to lengthen, only to reject them all in favor of unequal refrangibility. Varying any of these factors, such as the size of the hole,

> made no sensible changes in the length of the Image. Neither did the different matter of the Prisms make any: for in a Vessel made of polished Plates of Glass cemented together in the shape of a Prism and filled with Water, there is the like Success of the Experiment according to the quantity of the Refraction.[58]

This is a very strong claim, nearly equivalent to Experiment 8, for it asserts that for prisms of any substance when the deviations ("the quantity of Refraction") are equal, then the angular dispersions or the lengths of the spectra will be equal; and that this was found to be true for a glass-prism and a water-filled one. From this observation, the impossibility of constructing an achromatic glass–water prism follows immediately, for if the dispersions of contrary refractions cancel one another so will the deviations, and the emergent rays will be parallel to the incident ones.

We may ask whether Newton actually carried out this experiment? While the observation is in general invalid, it is true or very nearly true for the observation he actually claims to have made, as a calculation using a modern value for water's dispersive power shows. In his notes summarizing his measurements of indices of refraction for the table of refractive powers in the *Opticks*, the angle of the hollow prism used for measuring all the fluids is $78^\circ\ 50'$.[59] Since he also gives the observed deviation for the water-filled prism ($37^\circ\ 12'$), there is sufficient information to calculate the angular dispersion of the beam emerging from his prism; the result is $1^\circ\ 54'$.[60] We can then set up a proportion to determine what the angular dispersion of the water-filled prism would be if its deviation were the same as that of his glass prism. We find that it would yield a spectrum of virtually the same length as the 13 inch spectrum he

observed with his first glass prism ($C = 63^\circ\ 12'$) just as he had claimed. Although this is just a calculation that indicates what Newton could have observed, it is entirely consistent with his observation.

CONCLUSION

After pursuing the twists and turns in Newton's investigations of dispersion over a period of more than three decades, what are we to conclude? Experiment 8 is undoubtedly not a reliable record of Newton's experiments and deduction of his dispersion law. There is simply too much contrary evidence: both the experiment on the colors generated when light traverses parallel surfaces and that with a compound glass–water prism had earlier yielded opposite results; the late introduction in the manuscript of one sentence describing the experiment with a compound prism; the lack of a rigorous deduction of the theorems from the experiments; and the evidence that he had already formulated the dispersion law (and also Theorem 2 on relative indices of refraction) by 1672, when he still believed, it would be possible to construct an achromatic lens, so that it was clearly derived independently of the described experiments (as was Theorem 2).

Let us first try to get a handle on the experimental evidence. The description of the colors generated by parallel plates contradicts the earlier, long exposition in the *Optical Lectures*, yet in Experiment 8 Newton explicitly and inexplicably claims to have done only the compound prism experiment, "This I try'd." Using contemporary terminology, we would have to say that Newton was guilty of falsification here.[61] This is particularly puzzling, since we know that the crucial experiment with the compound glass-water prism will turn out about as Newton claimed. His calculation in his draft reply to Hooke would have taught him that such a prism is very nearly achromatic and sensitive to slight changes in parameters. His observation in the *Opticks* that the dispersion of a glass and a water-filled prism is the same when their deviations are the same is nearly equivalent to that of a compound prism and would have confirmed the experiment and calculation with the compound prism.

Even more important than the compound glass-water prism experiment, which has been the focus of attention but which can turn out very much as Newton claimed, are the experiments that he did not do. In particular, we can ask why he measured the dispersion of glass alone, for had he made measurements on other substances, he surely would have recognized that his linear dispersion law was not true. David Brewster perceptively observed that Newton's "opinion" that when the deviations produced by any prism are equal, then the dispersions will also be equal, "seems to have been impressed on his mind with all the force of an axiom."[62] This question (as well as the related one of why from the measurement of the spectrum of but one glass prism Newton asserted that the spectrum cast by any substance will be divided musically) has, I suggest, the simple answer Brewster proposed: it was axiomatic. For Newton colors and *degree* of refrangibility were innate, immutable properties of light. These were properties of light and not matter. To admit otherwise would undermine the very foundation of his theory of color: the one-to-one correspondence between

color and degree of refrangibility, and their immutability. The *index* of refraction alone, according to Newton, depended on the properties of matter and indicated how strongly a particular substance acted on light to bend it. Newton accepted these principles as axioms, and so he never doubted that his dispersion laws and musical division of the spectrum applied to all substances, for they were laws about light and not refracting substances. Conversely, he assumed that the index of refraction was a property of matter, and so for his table of refractive powers in the *Opticks* he systematically measured it in twenty-two substances.

The crucial role that the principle of immutability played in determining the direction of Newton's experimental program leads us to Newton's goal of creating a mathematical science of color. There is a fundamental difference between the theory of color in Book I of the *Opticks* and the rest of the book, namely, Book I articulated a highly developed theory, whereas the rest of the book essentially related the discovery of a number of properties of light, most notably periodicity. The theory of color is formulated as a sequence of propositions in imitation of a mathematical work, unlike nearly all the rest of the book. Newton had always desired to create a predictive science of color founded on a fully developed mathematical theory. Experiment 8 concludes with an encomium to a mathematical theory of color and helps us to understand Newton's aim – first expressed in his *Optical Lectures* – of creating a rational science of color:

> And these Theorems being admitted into Opticks, there would be scope enough of handling that Science voluminously after a new manner, not only by teaching those things which tend to the perfection of Vision, but also by determining mathematically all kinds of Phaenomena of Colours which could be produced by Refractions.[63]

The two theorems of Experiment 8 were essential to completing the mathematical theory of color. He had to have a dispersion law, and he at last chose the linear dispersion law, which brings us back to the principle of immutability. The partial dispersion law of the musical division implies, as we saw, the linear dispersion law with a few simple and natural assumptions. Physically, the musical division can be interpreted as a statement of the immutability of color and degree of refrangibility, for it asserts that the proportion of the spectrum occupied by each color is precisely the same in all substances and so is an innate property of light. Newton considered the musical division and the linear dispersion law to be so closely related that they form the two experiments, 8 and 9, of one proposition.

In analyzing Experiment 8 undue, but quite understandable, attention has been given to the role of Newton's belief in the impossibility of constructing an achromatic lens. Historians have never been able to explain why Newton became convinced of the impossibility of constructing an achromatic lens. I strongly suspect that it was his commitment to the linear dispersion law that lead to his belief in the impossibility of constructing an achromatic lens or prism. As late as the *Principia* (1687), he still would not commit himself to its impossibility. In the Scholium to Book I, Proposition 98, he wrote "the differing refrangibility of different rays prevents optics from being

perfected by spherical or any other shapes. Unless the errors arising from this source can be corrected, all labor spent in correcting the other errors will be of no avail."[64] Yet, just a few years later, when he was composing the *Opticks*, he had become convinced that "the Improvement of Telescopes . . . by Refractions is desperate."[65] It was precisely in this period that Newton adopted the linear dispersion law, and it is a specific and plausible cause for his dramatic change of position.

Experiment 8 seems as if it were constructed out of a number of elements that were established independently of one another. It is rather loosely formulated: The experiments are described generically with no details that lend a sense of verisimilitude, and there is no formal deduction. Nonetheless, from Newton's perspective the components of Experiment 8 all seem to follow so naturally from one element to another and fit so harmoniously. The principle of the immutability of color demands that the spectrum be divided in the same way in all substances, and this in turn yields the linear dispersion law. From his mathematical investigation of chromatic aberration of lenses, he knew the condition for achromacy, namely, the linear dispersion law. Finally, from his calculations and experiments, he knew that a compound glass–water prism was achromatic, or very nearly so. Every step of this sequence is so evident that to Newton it may have seemed unnecessary to test or rigorously derive them. Newton frequently pushed his scientific understanding to the edge, but he knew when to pull back and test his conclusions. Perhaps the various elements of Experiment 8 just looked too evident and straightforward to bother, especially when they agreed with fundamental principles of his optical science and allowed "determining mathematically all kinds of Phaenomena of Colours which could be produced by Refractions."[66]

School of Physics and Astronomy, University of Minnesota, USA

NOTES

[1] The studies of Newton's investigations of achromatism and dispersion by H. Boegehold (1928, 1943) and Zev Bechler (1973, 1975) are valuable; the former stresses experiment and pursues these problems through the 18th century, and the latter focuses on theory in Newton's own work. For the invention of the achromatic lens, see Nordenmark and Nordstrom (1938, 1939), and on Dollond, see Sorenson (2001).
[2] Newton (1952, pp. 129–130).
[3] Newton (1984, pp. 183–185). Newton delivered his *Optical Lectures* between 1670 and 1672, after he was appointed Lucasian Professor at Cambridge. In the winter of 1671/2 he began a major revision which was completed by October 1674. My edition contains both versions.
[4] Newton (1984, pp. 563–565).
[5] John Dollond (1758).
[6] Peter Dollond (1789, pp. 14–15).
[7] See Boegehold (1928, pp. 12–13).
[8] See, for instance, Priestley (1772, pp. 805–807), for the oft-repeated account of the possible influence of the *saccharum saturni* (lead acetate) that Newton occasionally added to his water.
[9] Bechler (1975, pp. 113–116, 122–125); see also Bechler (1973, p. 31).
[10] University Library Cambridge, MS Add. 3970, ff. 411r, 412r.
[11] Klingenstierna (1754), which is a translation by A. G. Kästner from the Swedish original that appeared in the same year. See Boegehold (1928, pp. 15–18), for an account of this paper

and, in particular, Boegehold's calculations showing that a small-angle approximation is still a reasonably good one for a glass prism as large as 60°.

[12] Newton (1984, pp. 199–201, 335–337). On this refraction model see Bechler (1973, pp. 3–6); and Lohne (1961, pp. 397–398).

[13] MS Add. 3970, f. 529r. Although separated from the remainder of the draft letter (ff. 433r–444v), this folio immediately follows the "Discourse of Observations" (ff. 519–528), which Westfall (1963, no. 28) has established also belongs with this draft.

[14] MS Add. 3970, ff. 438v–439r.

[15] *Ibid.*, f. 439r; the final sentence was added to the rest of the paragraph.

[16] Since Newton had not yet developed an entirely consistent terminology, from the language alone it is not clear whether he is referring to the angles or their sines:

$$\frac{\Delta r}{r-i} \approx \text{constant} \quad \text{or} \quad \frac{\Delta R}{R-I} \approx \text{constant.}$$

The latter interpretation can be eliminated solely on physical grounds, since the dispersive power, or any other measure of dispersion, is always strictly, not approximately, constant in any given medium. "The difference of the refractions" always seems to refer to the angle Δr; see, for instance, Newton (1984, Pt. 11, Props. 14–17) The term "the whole refraction" is once used in the "New theory about light and colors" to refer to the difference of the sines, $R - I$; see Newton (1959, vol. 1, p. 95). However, in the *Opticks* it is used to refer to the difference of the angles, $r - i$, both in Bk. 1, Pt. I, Prop. 7 (quoted below at no. 20) and in Experiment 8 itself (see the penultimate sentence of the first paragraph quoted above at no. 2). The most convincing argument for the interpretation I have adopted is that Newton had already adopted it for the refraction in a lens; see the next paragraph.

[17] Newton (1984, p. 425).

[18] Newton (1984, p. 429).

[19] Newton (1959, vol. 1, p. 95).

[20] Newton (1952, pp. 84–85); italics added.

[21] MS Add. 3970, f. 529r.

[22] Newton (1984, p. 87); see also p. 439. Lohne (1961), stresses the importance of the mathematical ideal for Newton's theory of color.

[23] Newton (1959, vol. 1, pp. 96–97); Oldenburg eliminated this passage from the paper published in the *Philosophical Transactions*. For Hooke's rejection, see p. 111.

[24] MS Add. 3970, ff. 443v–444r; italics added.

[25] That part of the *Optical Lectures* which presents the three equivalent principles is entitled "The measure of refractions," Newton (1984, p. 311); see also p. 169. In the *Lectures* he refers to the sine law and his dispersion law as a "rule" for the "measures of refractions" Newton (1984, pp. 169, 199, 311, 335).

[26] MS Add. 3970, f. 443r.

[27] *Ibid*.

[28] *Ibid*. I have added some punctuation.

[29] Newton forms the following proportion in which the fourth number is to be determined: "as 13 . . . to 93. . . so 37 1/2 . . . to a fourth number 264 9/13, w^{ch} being put the common sine of incidence . . . " (*ibid.*, f. 529r). The correct number is 268 7/26, with which I then recalculated the sines of refraction.

[30] *Ibid.*, f. 443v.

[31] *Ibid*.

[32] Bechler (1975, pp. 113–116).

[33] In the four months that it took Newton to respond to Hooke he drafted a series of replies. It seems that in this period Newton moved from the theoretical study of a compound glass-water achromatic lens to its actual construction, and that he could not successfully correct for chromatic aberration by this means; see MS Add. 3970, ff. 433r, 445r, 447r.

[34] See Bechler (1973, pp. 6–8).

[35] Newton (1984, pp. 543–547).
[36] Newton (1984, pp. 547–549).
[37] See Shapiro (1979, p. 111).
[38] It should be noted that the relation between the two laws does not at all depend on the musical division *per se*, but only on its linearity. The same relation would follow for any other division of the spectrum, such as the rejected geometrical division, provided that each color always occupies some fixed proportion of the spectrum.
[39] Newton (1984, p. 339); see also p. 201 for a similar argument in the first version.
[40] The demonstration is presented in two lemmas, *ibid.*, pp. 177–179, 319.
[41] The angular dispersion $\Delta r \equiv \Delta D$ *is* given by $\Delta D = (l - b)/s$, where l is the spectrum's length, b its breadth, and s its distance from the hole; $l - b$ is the corrected length. The indices of refraction for the extreme rays, n_{p} and n_{t}, are then calculated using equation (2) with $D = D \pm \Delta D/2$.
[42] Newton observes that by using this method his calculation "although no longer absolutely true, will still so closely approach the truth that it may be taken as true with respect to sense and a mechanical calculation" (*ibid.*, p. 333; see also p. 193). An exact calculation bears out Newton's claim that the error introduced in this way may be safely ignored, since it changes the dispersive power by less than 0.1%, orders of magnitude less than his experimental error.
[43] See Laymon (1978). Laymon based his account of Newton's precision on the "New theory" and was unaware of Newton's measurements in the *Optical Lectures* from which those in the "New theory" derive.
[44] Newton (1984, pp. 321–323; see also pp. 181–183).
[45] Newton undoubtedly performed such a series of calculations for the elements of the achromatic prism in Experiment 9. See the quotation at no. 31, above, where he calls attention to the sensitivity of the computations to small errors in the parameters.
[46] Lohne (1977, p. 241). For a further analysis of the precision of Newton's measurements and consideration of additional data see Shapiro (1979, pp. 116–124).
[47] Newton (1960, vol. 2, p. 76).
[48] Newton (1960, pp. 76–77). Laymon (1978), calls attention to the importance of this technique of varying the parameters to estimate errors.
[49] Newton (1952, Bk. I, Pt. I, Prop. 2, p. 29).
[50] Newton (1984, pp. 329–333); and Newton (1952, Bk. I, Pt. I, Prop. 7, pp. 82–84). Although Newton does not calculate it, in the *Opticks* he gives sufficient data to find the dispersive power for the prism with angle 63°. I found $v = 28$, which agrees quite well with the other prism.
[51] For the *Fundamentum opticae* see MS Add. 3970, ff. 421r, 411r; and for the *Opticks*, ff. 96r–8r.
[52] MS Add. 3970, f. 388r.
[53] It is evident from the manuscript of the *Opticks* that Prop. VII was added to the end of Pt. I, Bk. I after the rest of that book was composed, and that it was this measurement of chromatic aberration that caused him to adopt a new value for the dispersive power, after he had already committed himself to the linear dispersion law of Experiment 8; see Shapiro (1979, p. 121, no. 65).
[54] I must confess that this improved agreement seems spurious, because a smaller value just compensates for the difficulty of measuring the faint, extreme ends of the spectrum.
[55] Westfall (1973).
[56] Newton (1984, pp. 203, 339); and Newton (1952, Bk. I, Pt. II, Expt. 8, pp. 130–131).
[57] See Shapiro (1979, p. 122).
[58] Newton (1952, Bk. I, Pt. I, Prop. II, pp. 30–31).
[59] MS Add. 3970, f. 304v; published in Lohne (1977, p. 243).
[60] For water's dispersive power I used Boegehold's value, $v = 24.10$, which agrees well with more recent values; Boegehold (1928, p. 12). The choice of wavelengths equivalent to Newton's extreme red and violet is somewhat arbitrary.

[61] It is difficult to believe that Newton forgot about his earlier results, for as A. Rupert Hall (1994, p. 21), has put it, "Newton forgot nothing!"
[62] Brewster (1855, vol. 1, p. 110).
[63] Newton (1952, Bk. I, Pt. II, Prop. 3, Expt. 8, p. 131).
[64] Newton (1999, p. 629).
[65] Newton (1952, Bk. I, Pt. I, Prop. 7, p. 102).
[66] Quoted in full at note 63.

REFERENCES

Bechler, Z. (1973). "Newton's search for a mechanistic model of colour dispersion: A suggested interpretation." *Archive for History of Exact Sciences* **11**: 1–37.
Bechler, Z. (1975). " 'A less agreeable matter': The disagreeable case of Newton and achromatic refraction." *British Journal for the History of Science* **8**: 101–126.
Boegehold, H. (1928). "Der Glas–Wasser-Versuch von Newton und Dollond." *Forschungen zur Geschichte der Optik* **1**: 7–40.
Boegehold, H. (1943). "Zur Vor- und Frühgeschichte der achromatischen Fernrohrobjektive." *Forschungen zur Geschichte der Optik* **3**: 81–114.
Brewster, David. (1855). *Memoirs of the Life, Writings, and Discoveries of Sir Isaac Newton*, 2 vols. Edinburgh: Thomas Constable and Co. [Facsimile reprint. New York: Johnson Reprint, 1965].
Dollond, J. (1758). "An account of some experiments concerning the different refrangibility of light." *Philosophical Transactions* **50**: 733–743.
Dollond, P. (1789). "Some Account of the Discovery, Made by the Late Mr. John Dollond, F.R.S." which Led to the Grand Improvement of Refracting Telescopes, in Order to Correct Some Misrepresentations, in Foreign Publications, of that Discovery: With an Attempt to Account for the Mistake in an Experiment Made by Sir Isaac Newton; on which Experiment, the Improvement of the Refracting Telescope Entirely Depended." London: J. Johnson.
Hall, A. R. (1994). "The contributions of science and technology to the design of early optical elements." In: *Proceedings of the Eleventh International Scientific Instrument Symposium, Bologna University, Italy, 9–14 September 1991*, eds. G. Dragoni, A. McConnell, and G. L'E. Turner. Bologna: Grafis, pp. 19–25.
Klingenstjerna, S. (1754). "Anmerkung über das Gesetz der Brechung bey Lichtstrahlen von verschiedener Art, wenn sie durch ein durchsichtiges Mittel in verschiedene andere gehen." *Abhandlungen aus der Naturlehre, Haushaltungskunst und Mechanik* **16**: 300–309.
Laymon, R. (1978). "Newton's advertised precision and his refutation of the received laws of refraction." In: *Studies in Perception: Interrelations in the History of Philosophy and Science.* eds. P. K. Machamer, and R. G. Turnbull. Columbus: Ohio State University Press, pp. 231–258.
Lohne, J. A. (1961). "Newton's 'proof' of the sine law and his mathematical principles of colors." *Archive for History of Exact Sciences* **1**: 389–405.
Lohne, J. A. (1977). "Newton's table of refractive powers: Origins, accuracy, and influence." *Sudhoffs Archiv für Geschichte der Medizin und Naturwissenschaften* **61**: 229–247.
Newton, I. (1952). *Opticks: Or, a Treatise of the Reflections, Refractions, Inflexions and Colours of Light. Based on the Fourth Edition London*, *1730.* New York: Dover Publications.
Newton, I. (1959–1977). *The Correspondence of Isaac Newton*, 7 vols. eds. H. W. Turnbull, J. F. Scott, A. Rupert Hall, and Laura Tilling. Cambridge: Cambridge University Press.
Newton, I. (1984). *The Optical Papers of Isaac Newton*. *Volume 1. The Optical Lectures, 1670–1672*, ed. Alan E. Shapiro. Cambridge: Cambridge University Press.
Newton, I. (1999). *The Principia: Mathematical Principles of Natural Philosophy,* trans. I. Bernard Cohen, and A. Whitman. Berkeley: University of California Press.

Nordenmark, N. V. E. and J. Nordström. (1938, 1939). "Om uppfinningen av den akromatiska och aplanatiska linsen." *Lychnos* **4**: 1–52, 5: 313–384.
Priestley, J. (1772). *The History and Present State of Discoveries Relating to Vision, Light, and Colours.* London: J. Johnson.
Shapiro, A. E. (1979). "Newton's 'achromatic' dispersion law: Theoretical background and experimental evidence." *Archive for History of Exact Sciences* **21**: 91–128.
Sorenson, R. (2001). "Dollond & Son's pursuit of achromaticity," 1758–1789. *History of Science* **39**: 31–55.
Westfall, R. S. (1963). "Newton's reply to Hooke and the theory of colors." *Isis* **54**: 82–96.
Westfall, R. S. (1973). "Newton and the fudge factor." *Science* **179**: 751–758.

GEORGE E. SMITH

WAS WRONG NEWTON BAD NEWTON?

WRONG SCIENCE IN BOOK 2

As the late Clifford Truesdell pointed out forcefully, Book 2 of Newton's *Principia* is less known for the lasting contributions it made to science than for its errors:

> It was these two flimsy and unmathematical props, the "cataract" and the "solid particles of the air", no better than the vortices of the Cartesians, that drew the strongest criticism upon the *Principia*. However successful was Book I, Book II was a failure as an essay toward a unified, mathematical mechanics.
>
> With its bewildering alternation of mathematical proof, brilliant hypotheses, pure guessing, bluff, and plain error, Book II has long been praised, in whole or in part, and praised justly, as affording the greatest signs of Newton's genius. To the geometers of the day, it offered an immediate challenge: to correct the errors, to replace the guesswork by clear hypotheses, to embed the hypotheses at their just stations in a rational mechanics, to brush away the bluff by mathematical proof, to create new concepts so as to succeed where Newton had failed.[1]

Given Newton's proclaimed aversion to hypotheses in "experimental philosophy," Truesdell is no less accusing Book 2 of departing from Newton's standards for good science. The fact that Truesdell himself did not necessarily view such a departure as bad science is beside the point.[2] It is still appropriate to ask whether the errors of Book 2 stem from a failure by Newton to adhere to his own standards for science.

Without question Book 2 puts forward a number of empirical claims that subsequent science has concluded are wrong. Some of these, however, have little to do with the book's central concern, the theory of resistance forces on bodies moving in fluid mediums. Let me dispense with these first so that we can then focus on errors that were central to the theory construction task of the book. The three most notable errors that have nothing as such to do with resistance forces occur in the last two sections of the book:

1. At the end of Section 8 Newton derives from first principles a value for the speed of sound in air of 979 ft/sec, which by the time of the second edition he realized is well below a properly measured value of "more or less" 1142 ft/sec; before mentioning this measured value, however, he introduces a pair of *ad hoc* correction factors – fudge factors, if you will – that transform his theoretical value into 1142 ft/sec.[3]

J.Z. Buchwald and A. Franklin (eds.), Wrong for the Right Reasons, 127–160.

2. As a first step toward showing that Cartesian celestial vortices are incompatible with Kepler's 3/2 power rule, Newton mistakenly balances forces instead of torques in deriving the velocity variation in a vortex maintained by a rotating cylinder; as a consequence, Proposition 51 concludes that the times for one revolution of fluid particles in such a vortex vary linearly with radius, instead of as the radius squared.[4]
3. In the next proposition Newton again balances forces instead of torques in extending his solution for a rotating cylinder to a rotating sphere, concluding that the times for one revolution in a vortex maintained by the latter vary as the radius squared; but, as Stokes later showed, in contrast to the case of a rotating cylinder, no permanent, stable vortex is generated and maintained by a sphere rotating in a viscous fluid.[5]

These last two errors are not the only place in the *Principia* where Newton shows that he never saw how to reformulate his second law of motion for the case of angular motion.[6] Still, the error in Section 9 of Book 2 was of little historical moment, for comet trajectories, and not the 3/2 power rule, ended up providing the main argument against Cartesian vortices.

The error in the speed of sound, by contrast, reflects nothing more than the impossibility of Newton recognizing the distinction between adiabatic and isothermal expansion in an era before thermometry. Newton's derivation of the speed of sound from first principles is correct save for the need to replace Boyle's law with the corresponding adiabatic relationship for expansions and contractions in sound waves.[7] Moreover, however critical one chooses to be of Newton's fudge factors in presenting his value, they do call attention to the importance of reconciling the discrepancy.

Turning now to the *Principia*'s errors about resistance forces, the two most famous are of less significance to Newton's theory of these forces than their fame would suggest. Both are claims about forces on a body moving in what Newton calls a "rarified" fluid, a limiting case in which the fluid consists of solid particles that in no way interact with one another, but instead act like debris in empty space impacting on the moving body:

4. In such a fluid, Newton concludes, the resistance force on the surface of a moving body varies as the square of the sine of the incidence angle at which the particles impact the surface;[8] this "served, at the beginning of the XIX century, to demonstrate mathematically the impossibility of flying and by reason of [it] Newton has been blamed for having delayed aviation at least for half a century."[9]
5. In a scholium to the proposition in which this sine-squared principle is derived, Newton defines a surface contour of least resistance, remarking, "I think that this proposition will be of some use for the construction of ships;" [10] in the 1760s de Borda produced experimental proof to the contrary.[11]

Neither of these is truly an error, for both make rigorously correct claims about such rarified fluids. The only "error" lay in Newton's suggestion that results for such fluids

might be of some practical use in ship design. As we shall see below, the experimental results presented in the second edition of the *Principia* make clear that not just water, but air too does not produce resistance forces in the manner of a rarified fluid; rather, resistance forces in both air and water correspond to those in what Newton called a "continuous" fluid, a limiting case in which the fluid consists of solid particles effectively in contact with their immediate neighbors. The only bad science associated with Newton's results for rarified fluids was the tendency of too many in subsequent generations to view Newton as somehow infallible in physics, even when he was openly doing nothing more than mathematically exploring possibilities with the goal of enabling the empirical world to choose among them.[12]

The real errors in Book 2 concern resistance forces in "continuous" fluids. The first edition presents a sequence of claims that Newton himself decided were wrong within a few years after its publication:[13]

6. Proceeding from an erroneous solution to the efflux problem – the problem of the velocity of a fluid discharging through a hole in the bottom of a container – Newton concluded that the non-dimensionalized component of the resistance force on the front face of a moving sphere arising purely from the inertia of the fluid amounts to what we would now say is a drag coefficient of 2.0.[14]
7. From pendulum-decay experiments Newton concluded that the total non-dimensionalized resistance force on a moving sphere in both air and water (and mercury too) amounts to a drag coefficient around 0.7.
8. Based on these same experiments, Newton inferred that the purely inertial component of this measured total resistance force also amounts to a drag coefficient around 0.7.
9. From the first and third of these claims he then concluded that the motion of the sphere induces a counteracting effect of the fluid on its rear face that cancels roughly two-thirds of the inertial component of the resistance force on the front face, but never all of this component.

This last conclusion enabled him to imply that the inertia of the fluid of Descartes's vortices would have a clearly detectable effect on the motion of comets, if not the planets and their satellites as well. This was the bottom-line point that explains why in the first place Newton thought it important to include a theory of resistance forces in the *Principia*.

When Newton was told that his value of the efflux velocity was in error, he turned to experiment to determine whether and why he was wrong, pointing the way to the new "cataract" efflux solution of the second edition that Truesdell derides. He also conducted some vertical-fall experiments in water that gave a non-dimensionalized total resistance force amounting to a drag coefficient around 0.5, from which he concluded that the pendulum-decay experiments had given a value 40% too high. Thus, by the early 1690s, some two decades before the second edition finally appeared, new experimental results had convinced Newton to reject all four of the claims listed above.

These experimental results, however, did not prevent his new theory of resistance forces in continuous fluids in the second (and third) edition from including a pivotal error:[15]

10. Proceeding from his new "cataract" solution of the efflux problem, Newton now concluded that, in the absence of possible secondary effects, the non-dimensionalized inertial component of the resistance force on the front face of a moving sphere amounts to a drag coefficient of exactly 0.5 – a value that closely approximates the total resistance force measured in his vertical-fall experiments.

In the process of reaching this conclusion, Newton found it necessary to make three claims about incompressible, inviscid fluids that turned out not to be even roughly true of either water or air, the two fluids to which he applied the conclusion:

11. The disturbance of the fluid that the motion of a body produces ahead of it and on its sides has negligible (or at most second-order) effects on the resistance on the body.
12. The change in the force that the fluid exerts on the rear face of a body as a consequence of its motion is negligible (or at most second-order) except perhaps when the body is moving at very high velocity; in other words (contrary to the first edition) the motion of the body induces no (new) force at all on its rear face, and hence nothing acts to cancel the inertial resistance on the front face.
13. So long as the frontal area remains fixed, the shape of a moving body has a negligible (or at most second-order) effect on the resistance on it; hence, the non-dimensionalized resistance force is the same on, for example, a sphere and a disk moving in the direction of its axis of revolution.

The theory of resistance in continuous fluids of the second edition is *prima facie* the strongest candidate for bad science in Book 2 because of the extent to which it appears to have been devised *ad hoc* to match the resistance forces Newton had found in his initial vertical-fall experiments.

This list of explicit errors about resistance forces in Book 2 still does not include the deepest – and methodologically most interesting – error Newton made throughout his efforts on resistance. Newton assumed that resistance forces can be mathematically represented by means of a sum of terms representing distinct, independent physical mechanisms contributing to resistance: one term for the effects of the inertia of the fluid, another for the effects of viscosity, and so on. In fact, we still do not have laws for resistance forces of the sort Newton was pursuing precisely because fluid resistance involves a complex interaction of inertial and viscous effects that cannot be represented by any such sum. In particular, there is no such thing as a component of resistance resulting *purely* from the inertia of the fluid! As d'Alembert showed a quarter century after Newton died, the resistance force a body would encounter when moving in an incompressible, inviscid fluid is *exactly zero*, regardless of its shape![16] I will return to this fundamental error at the end of the paper, asking whether

it represents a case of bad science. First, however, we should ask this question of the errors listed above.

NEWTON'S STANDARDS FOR GOOD SCIENCE

The issue of what constitutes good and bad science is sure to be contentious. Newton had his own distinctive standards for science, ones that he considered more demanding than those of his contemporaries, almost all of whom favored the "method of hypotheses" that he held in low regard. Asking first whether the wrong science of Book 2 was bad science by Newton's standards will allow us to postpone to the end the thorny question of what the standards should be. To do this, however, we need to be clear about what Newton's standards were.

Newton expressed his low regard for the method of hypotheses in an unpublished portion of his initial letter to Oldenburg on refraction and color of 6 February 1672, more than a decade before he began work on the *Principia*:

> For what I shall tell concerning them [colors] is not an Hypothesis but most rigid consequence, not conjectured by barely inferring 'tis thus because not otherwise or because it satisfies all phænomena (the Philosophers universal Topick) but evinced by the mediation of experiments concluding directly and without any suspicion of doubt.[17]

During the subsequent dispute over the conclusions he drew from his experimental results, he added: "For if the possibility of hypotheses is to be the test of the truth and reality of things, I see not how certainty can be obtained in any science; since numerous hypotheses may be devised, which shall seem to overcome new difficulties."[18] In particular, as Newton showed in a response to Hooke during this dispute, both a particle and a wave theory of light were compatible with his experimental results on refraction.[19] Thus, on Newton's view, confirmation of an observable claim deduced from a broad explanatory hypothesis is at most a very weak form of evidence. Indeed, the second quotation seems to indicate that Newton viewed any such observable consequence as confirming only the *possibility* of the truth of the hypothesis from which it is deduced; and verifying a large number of distinct deduced claims does not transform the logic of the situation because alternative broad explanatory hypotheses can too readily be adapted to them.

By the time he started writing the *Principia* a decade later, Newton had come upon an important further reason to view theories advanced solely on the basis of such hypothetico-deductive evidence as bad science. In an unreleased, amended version of the tract *De Motu Corporum in Gyrum* Newton had concluded:

> By reason of the deviation of the Sun from the center of gravity, the centripetal force does not always tend to that immobile center, and hence the planets neither move exactly in ellipses nor revolve twice in the same orbit. There are as many orbits of a planet as it has revolutions, as in the motion of the Moon, and the orbit of any one planet depends on the combined motions of all the planets, not

> to mention the actions of all these on each other. But to consider simultaneously all these causes of motion and to define these motions by exact laws admitting of easy calculation exceeds, if I am not mistaken, the force of any human mind.[20]

But then hypothetico-deductive evidence for any theory of the physics underlying planetary motion would suffer from not one, but two defects. First, as above, deriving Kepler's orbital rules from such a theory would not constitute compelling evidence for the theory, for (just as Leibniz showed two years after publication of the *Principia*) these rules could be derived from competing theories as well. Second, and more important, given the complexity of the true motions, Kepler's orbital rules hold only to a certain approximation, and in principle any number of alternative descriptions of the orbits could hold to the same level of approximation. As Newton was aware while writing the *Principia*, several such alternatives to Kepler were already in place.[21] A derivation of Kepler's rules from a theory of orbital physics therefore carried even less weight insofar as these rules do not strictly describe the planetary trajectories, but are only one among several alternative approximate descriptions of them, none of which should be expected to describe the true complex motions exactly. The new approach to marshalling evidence that Newton offers as superior to the method of hypotheses in the *Principia* was a response to complexity of this sort.

Good science for Newton was science in which theoretical propositions "are gathered from phenomena by induction" in a manner that rules out alternatives to them – at least "until yet other phenomena make such propositions either more exact or liable to exceptions."[22] Newton's approach to meeting this demand in the *Principia* involves two stages. The aim of the first is to derive from the laws of motion a generic mathematical theory consisting of "if–then" propositions that relate forces to motions. Thus the theory of centripetal forces presented in Book 1 is *generic* in the sense that it is not confined to inverse-square forces, but includes forces varying linearly with distance, inverse-cube forces, and ultimately forces varying as any arbitrary function of distance; and the theory of mathematically characterized resistance forces of Book 2 is not confined to forces that vary as the first and second powers of velocity, but again ultimately as any power whatever. The second stage then proceeds from generic mathematical theory to physical theory:

> Mathematics requires an investigation of those quantities of forces and their proportions that follow from any conditions that may be supposed. Then, coming down to physics, these proportions must be compared with the phenomena, so that it may be found out which conditions of forces apply to each kind of attracting bodies. And then, finally, it will be possible to argue more securely concerning the physical species, physical causes, and physical proportions of these forces.[23]

The challenge is, first, to derive an adequate range of "if–then" propositions that can serve as "inference tickets" tying different possible conditions of forces to contrasting phenomena; and, second, to find phenomena either in nature or through experiment that can combine with these "if–then" propositions to yield laws characterizing forces

physically. Success typically requires both mathematical and, in a broad sense of the term, experimental advances.

The way in which this approach can provide a response to the complexity of real-world phenomena lies in the details of the "comparison" between the mathematically characterized proportions of forces and the phenomena. The best way to see this is to consider carefully the logical force of Newton's reasoning to the law of gravity. The phenomena of orbital motion from which he starts he expressly identifies as holding at least *quam proxime* – that is, very nearly; and the "if–then" propositions that he uses in inferring the law of force are ones which hold not only strictly, but also continue to hold in an "if *quam proxime*, then *quam proxime*" form.[24] The initial logical force of his reasoning from orbital phenomena accordingly gives the inferred conclusion only a *quam proxime* status. Thus, for example, from the fact that Kepler's area rule holds at least to a high approximation, Newton concludes that the governing force is centripetal, at least to a high approximation; and from the fact that Kepler's 3/2 power rule holds at least to a high approximation and the fact that the perihelia of the orbits are stationary at least to a high approximation, he concludes that the exponent of r, the distance to the center, in the law characterizing the physical force that retains the planets in their orbits is, at least to a high approximation, -2. In short, the initial derivation of the law of gravity establishes that it holds at least *quam proxime* for the known orbital motions, but only that.

The further step to the claim that the law of gravity holds exactly for these motions amounts to a research strategy in response to the complexity of the real motions. Newton appears to require that two conditions be met in order for this step to be appropriate. First, forces in the case of macroscopic bodies must be composable from forces involving only their parts in a way that yields the force law exactly for some, perhaps idealized, macro-scopic configuration. Thus, Newton shows that if the gravitational force toward a sphere of uniform (or spherically symmetric) density is composed of gravitational forces toward its microphysical parts, then his law of gravity holds exactly in the space surrounding it.[25] Second, the force law, taken to hold exactly, must yield identifiable physical conditions under which the phenomena from which it was derived would hold exactly. Thus, for example, Kepler's area and 3/2 power rules and the lack of motion of the orbital perihelia *would* hold exactly if no force were acting on the orbiting body other than the gravitational force directed toward the central body. When these conditions are met and the force law is taken to hold exactly, one can deduce from it an *idealized* first approximation to the real motions – that is, an approximation that would hold exactly in the indicated physical conditions. Thus, the law of gravity entails that Keplerian orbital motion with stationary perihelia would hold exactly were it not for forces acting on the bodies in addition to the dominant centripetal gravitational force retaining them in their orbits.

Taking the derived force law as holding exactly therefore carries with it an implication: *any real-world deviation from the idealized approximation to the phenomena must be physically significant* – this in contrast, for example, to being nothing more than a reflection of the specific mathematical framework used to curve-fit the phenomena. Thus, Newton's law of gravity, taken as holding exactly, implies that every

deviation from Keplerian orbital motion with stationary orbits results from some further physically characterizable force or forces acting on the orbiting bodies. The research focus then shifts to the deviations from the idealized orbital phenomena and to the problem of identifying further forces at work and the contributions they make to the real motions. The overall approach to the complexity of the real motions thus becomes one of successive approximations in which theory implies that each approximation would hold exactly in the identified idealized circumstances. The force law itself is, strictly speaking, only provisional, but it should continue to be taken as exact "until yet other phenomena make such propositions either more exact or liable to exceptions." The critical long-term question is whether continuing research yields increasingly smaller discrepancies between theory and observation without requiring anything more than mere refinement of the force law – for example, refinement to compensate for some parochial feature in the phenomena from which the law was originally derived.[26]

To summarize, good science for Newton is science that responds to the complexity of real-world phenomena by requiring theoretical propositions to be determined by suitably idealized approximations to these phenomena. The demand is to proceed (1) from *quam proxime* descriptions of phenomena, (2) to a theory that holds at least *quam proxime*, (3) to a theory that is taken to hold exactly provided certain requirements are met, (4) to a sequence of idealized versions of all the relevant real-world phenomena, (5) with continuing research focusing on the discrepancies between these idealizations and the real world, under the demand that every such discrepancy be physically significant, as judged from the point of view of the theory.[27] The initial evidence for the theory establishes its promise; further evidence accrues to it, entrenching it, from continual improvement in the agreement between theory and observation. In the case of gravitation, Newton appears also to have demanded as part of his initial evidence for the promise of the theory some immediate success in addressing discrepancies between the actual phenomena and the initial idealized approximation to them. Specifically, he took it upon himself to show that some deviations of the Moon from Keplerian motion could be accounted for to a high approximation by the perturbing effect of the Sun's gravity.[28]

Newton had good reasons to think that motion in resisting media is also complex. Whether he knew it or not, both Descartes[29] and Galileo[30] had concluded that a science of resistance was impossible because too many independent factors affect the strength of the resistance, including several different features of both the fluid and the moving body. More directly, Newton himself ultimately thought that the resistance introduced by a fluid medium arises from as many as three independent physical mechanisms: the inertia of the fluid, its internal friction or viscosity, and surface friction on the moving body.[31] He further thought that these three mechanisms would vary with the velocity of the moving body in entirely different ways, so that the net resistance force would result from a complex combination in which different mechanisms dominate under different conditions. He therefore viewed the problem as one of establishing separate force laws for the different mechanisms of resistance, and this could be done only by experimentally *disaggregating* the contributions made

by each mechanism to the motions observed in resisting media. He was ultimately inclined to think that the contribution made by surface friction would be independent of velocity, the contribution made by the viscosity of the fluid would vary linearly with velocity, and the contribution made by its inertia would vary as velocity squared.[32] In other words, he was inclined to think that the total resistance force can be represented as a sum:

$$f_{\text{resist}} = a_0 + a_1 v + a_2 v^2$$

The problem thus amounted one of finding phenomena of motion that would allow him to either confirm or replace these terms and exponents and then would allow him to determine the physical parameters and constants of proportionality governing each of the a_i, at least to a high approximation. Based on his rarified fluid model, he concluded that the density of the fluid, ρ_{f}, and the frontal area of the moving body, A_{front}, were the principal parameters governing a_2, yielding

$$f_{\text{resist}} = a_0 + a_1 v + b_2 \rho_{\text{f}} A_{\text{front}} v^2$$

where b_2 was hopefully a constant for bodies of any given shape. The problem then became one of finding phenomena of motion that would further allow him to confirm this dependency of a_2 or replace it with a more appropriate dependency if needed.

NEWTON'S FIRST APPROACH TO RESISTANCE

The difficult problem with resistance forces, from Newton's point of view, was to find a combination of theorems and experiments that would allow him to disaggregate the different mechanisms contributing to resistance so that, at the very least, he could establish a law for the separate contribution made by the inertia of the fluid in the case of spheres. As remarked before, he took very different approaches to this problem in the first and second editions. The question to begin with is whether the approach he took in the first edition conformed with his standards for good science.

This approach was, by any conceivable standards, ingenious. If resistance forces are the sum of three independent components, then the total arc lost by a (cycloidal) pendulum in any one swing as a consequence of resistance is also the sum of three terms. In particular, with the three terms above, the arc lost per swing can be represented by the following sum:

$$\delta_{\text{arc}} = A_0 + A_1 V_{\text{max}} + A_2 V_{\text{max}}^2$$

where V_{max} is the maximum velocity of the bob during the arc and the A_i's are constant for any given bob and fluid medium. By starting the pendulum bob at different points, Newton could experimentally obtain large variations in the peak velocity. Consequently, so long as the peak velocity could be determined for any one starting point, starting the pendulum at three distinct points – ranging from a small arc and hence a small peak velocity to a large arc with high peak velocity – would allow a simultaneous algebraic equation solution for the three A_i's from the measured values of the arc lost. Thus, staying with the three expected exponents, he could determine the

individual contributions to the total resistance force made by the three mechanisms:[33]

$$F_{\text{resist}} = \left[\left(\frac{1}{2}\right) A_0 + \left(\frac{2}{\pi}\right) A_1 V_{\max} + \left(\frac{3}{4}\right) A_2 V_{\max}^2\right]\left(\frac{W}{l}\right)$$

where W is the weight of the bob, corrected for buoyancy, and l, the length of the pendulum.

In fact, Newton was not restricted to these three exponents. He managed the mathematical feat of solving the pendulum decay problem not only for resistance varying as velocity to the 0, 1, and 2 powers, but also for any power whatever![34] This solution allowed him to try any combination of powers that the experimentally measured lost arcs required. He could carry out lost arc measurements starting the pendulum at, say, six points, use three of these to determine the constants A_n corresponding to a choice for the exponents, and then assess the adequacy of this choice by comparing calculated and observed arcs lost for the other three starting points. The ultimate choice of three (or two or even four) exponents would be dictated by requiring the residual discrepancies to be comparatively negligible. Admittedly, this is a form of curve-fitting, but it is curve-fitting under constraints that allow the comparison with experiment to determine, in principle, specific values for the exponents, *so long as the total resistance force involves a small number of independent mechanisms*. The approach is thus in full accord with Newton's standards for deriving force laws from phenomena. For, his mathematical solution for pendulum-decay covers the "quantities of forces and their proportions that follow from any conditions that may be supposed," and comparison with experiment can then in principle determine, at least to high approximation, which quantities and rules of proportion hold physically.

For his baseline experiment Newton employed a 126 inch long pendulum in air, with initial arcs of 2, 4, 8, 16, 32, and 64 inches.[35] Because the arc lost on any one swing was too small to measure, he counted the number of swings required for the pendulum to lose one-eighth of its initial arc, dividing to obtain the average arc lost per swing; and then, in a complementary baseline experiment, he counted the number of swings required for it to lose one-fourth of its initial arc, thereby safeguarding against being misled by his having to resort to an inexact measure of the arc lost per swing. In subsequent experiments he changed the diameter of the bob in air with the intent of verifying that the inertial component of resistance varies as the frontal area of the moving body; and he immersed the moving bob in troughs of water and mercury with the intent of verifying that the inertial component varies as the density of the medium. Had these experiments succeeded, they would surely be among the legendary experiments in the history of physics.

Unfortunately, as Newton fully realized, they did not succeed. He tried at least four different combinations of exponents of velocity, ultimately settling on 1, 3/2, and 2. The residual discrepancies remained disappointingly large in every case. Depending on which combination of initial arcs he used in determining the A_i's, he could obtain radically different values for those corresponding to exponents below 2, often indeed impossible negative values. In other words, the v^2 term was dominating to the point

of masking the other terms. Moreover, even the value of A_2 varied by much more than Newton wanted, depending on which combination of initial arcs was used to determine the A_i's in any one of the two baseline experiments, and it differed by 14% between the one-eighth-arc-lost and one-fourth-arc-lost experiments.

The only strong conclusion that could be drawn from the experiments was that no power of velocity greater than two is needed. Beyond this, Newton had to settle for rough values for the magnitudes of the total resistance and of the v^2 component. With a fair amount of handwaving to account for problems in the data obtained from the experiments with the smaller bob in air and the experiments in water and mercury, he concluded that the results did not falsify the proposal that the v^2 component varies as the frontal area of the bob and the density of the fluid. This, in turn, allowed him to conclude that the v^2 term represents the contribution made by the inertia of the fluid, and this contribution dominates at least over the range covered by his pendulum-decay experiments. But the experiments yielded no conclusions at all about the contributions made by the viscosity of the fluid and surface friction, nor about whether any other mechanisms contribute. Simply put, the phenomenon of pendulum decay was turning out not to be adequate for establishing a full law of resistance force; at best it provided weak evidence for a law of the supposed purely inertial component in the case of spheres, and even in this case the law had more the character of an engineering curve-fit than of a provisionally established law of physics.

Newton was in some respects candid in presenting the shortcomings of the pendulum-decay experiments, and in other respects less than candid. He never expressly indicated how ill-behaved the inferred values of the A_i's were, nor did he offer any explanation whatever for introducing a $v^{3/2}$ term. His invocation of the neglected resistance force on the string first to explain away and then to provide an *ad hoc* correction for the results with a different size bob, as well as the results in water and mercury, is a further example of "fudge factors" in the *Principia*. Nevertheless, Newton did present his raw data from the experiments in air employing one-eighth and one-fourth of the arc lost, and he gave sufficiently detailed instructions to allow readers to calculate for themselves the A_i's in order to decide whether the experiments were yielding sufficiently stable values. In this regard Newton was far more open in putting readers in a position to evaluate the evidence for his theoretical conclusions than, for example, was Huygens, who typically published no data at all, instead simply stating that the proposition "agrees exactly with experience."[36]

As remarked in the first section, Newton's primary goal in Book 2 seems to have been to show that the inertia of any fluid matter like Descartes's celestial vortices would have a detectable effect on the motions of comets, if not the planets as well. Clearly, he thought, the inertia of the fluid would introduce some resistance on the front face of a moving body. The question was whether a change in fluid pressure on the rear face from the motion of the body might cancel the resistance on the front face. Based on the results from the one-eighth-arc-lost experiments in air and the corrected results for water, Newton concluded that the total resistance on a sphere amounts, in modern terms, to a drag coefficient around 0.7, almost all of which results from the inertia of the fluid. From theory alone he had concluded that the inertial resistance

force on the front face amounts to a drag coefficient of 2.0. Specifically, his mistaken solution to the efflux problem gave him this value for a continuous fluid, and impact forces from perfectly rigid fluid particles gave him this value for a rarified fluid. From this he concluded that "about two thirds of that total motion which the front parts of the globe impress upon the medium while moving is restored to the rear parts of the globe by the medium as it returns in a circle and rushes into the space that the globe would otherwise leave empty behind itself."[37] Consequently, no matter how finely divided the parts of the fluid and how little its internal friction, its inertia must still produce a substantial resistance force at high velocities.[38] Thus, while the data were disappointing in many respects, the value for A_2 that they gave was sufficiently stable for Newton to feel that he could draw his desired bottom-line conclusion.

As already pointed out, all these conclusions were wrong. The resistance on a sphere over the range of Newton's pendulum-decay experiments amounts to a drag coefficient around 0.5, and the inertial force imposed by an incompressible, inviscid fluid on the front face is exactly canceled by the induced action of the fluid on the rear face. These mistakes of the first edition, however, were not a product of bad science, whether by Newton's standards or anyone else's. The idea of using pendulum-decay to disaggregate the different mechanisms contributing to resistance, and thus to isolate the contribution made by the inertia of the fluid, was, on its face, well-conceived science. Newton met his own standards by not relying on a hypothesis that the inertial component varies as velocity squared, but by instead deriving a general solution to the pendulum-decay problem for any exponent of velocity whatever, and then allowing the experiments to establish that the v^2 term dominates. Bothered by the vagaries in the data from the pendulum-decay experiments, Newton even carried out conical pendulum experiments to measure the level of total resistance, concluding that "although experiments of this kind allowed less accurate tests, the resistance that was found nevertheless agreed excellently with the preceding."[39] The principal conclusion to draw from Newton's efforts on resistance in the first edition is that an ingeniously conceived approach, careful experimentation, and rigorous mathematical derivation of propositions allowing inferences to be drawn from experimental results are not enough; the empirical world must cooperate as well.

THE MISTAKES EXPOSED

As his trying conical pendulum experiments indicates, Newton knew from the data that something was wrong with the pendulum-decay experiments. Within a few years after the publication of the first edition, Newton had carried out two further sets of experiments, one on efflux from a container and the other on vertical fall in water. The first of these led him to modify his theoretical claims about the inertial resistance on the front face of a sphere, and the second showed him what was limiting the accuracy of the pendulum-decay experiments. The precise dating of these two is unclear.

Newton was told first by Fatio de Duillier and then by Huygens that his published solution for the efflux velocity was wrong – that in fact the velocity is sufficient to raise the escaping fluid to the height of the fluid in the container, and not just

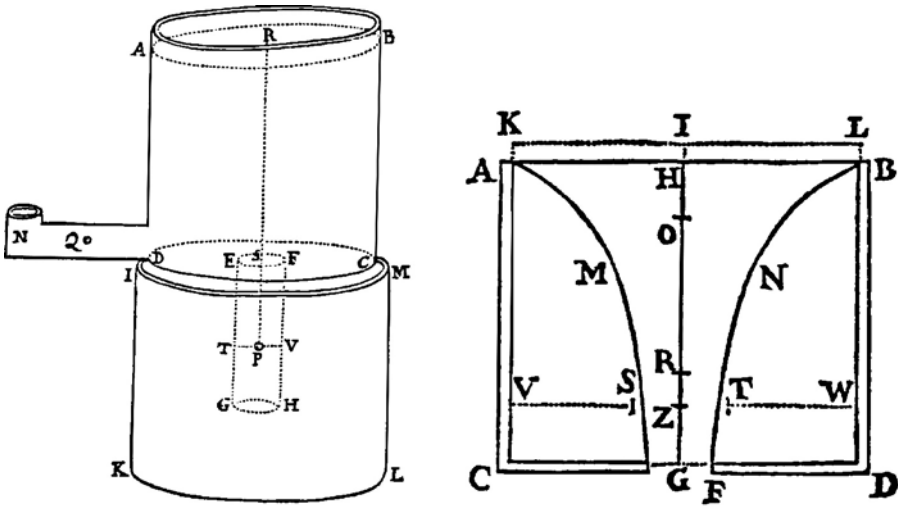

Figure 1. The contrasting diagrams for Newton's original solution to the efflux problem in the first edition of the *Principia* and the "cataract" solution in the second and third editions.

half this height, as Newton had said.[40] Newton had most likely reached his original conclusion by measuring the volume of flow out of the container in a given time and then inferring the velocity from the cross-sectional area of the orifice. Fatio claims to have convinced Newton to carry out experiments to see for himself.[41] Whether Newton actually carried out an efflux experiment in which he determined how far the velocity would carry the escaping fluid is unclear, but the second edition (and third) does report experiments of this sort carried out by others. Newton himself, however, did carry out a careful experiment to determine why a measurement of the volume of escaping fluid does not straightforwardly give the true velocity. What he found was that the width of the efflux stream was around $1/\sqrt{2}$ of the width of the orifice.[42] (The modern version of this so-called efflux coefficient is 0.62, indicating an error in Newton's measurement of the width of the stream of less than 1/40 of an inch.) In other words, Newton concluded that the fluid leaving the container does not flow vertically through the orifice, but enters the orifice at an acute angle, causing the width of the stream to neck-down in comparison to the orifice. This then led Newton to the new "cataract" solution to the efflux problem, the solution Truesdell ridicules. The contrasting diagrams for the efflux solutions of the first and second edition are shown in Figure 1. Newton showed no sign of ever having realized that the necking-down of the efflux stream is caused by a vortex at the mouth of the orifice generated by the viscosity of the fluid, the vortex we all observe as bath water runs out a drain. Regardless, the experiments led him to correct the value of the efflux velocity.

Newton's initial vertical-fall experiments in water were carried out in a specially built 9 1/2 foot deep trough with a glass window at the bottom. These initial

experiments were ended after a few trials when the window broke, inundating him with water.[43] Still, sufficient trials were carried out to reveal that the total resistance during vertical fall in water amounts to a drag coefficient around 0.5, well below the value obtained in the pendulum-decay experiments. Newton obtained this new value for the total resistance by using a quarter-second pendulum to measure the time of fall of different spheres in the trough and then comparing it with his exact theoretical solution for vertical-fall in a medium in which the resistance force is equal to $0.25\rho_{\mathrm{f}}A_{\mathrm{front}}v^2$. In two of the three experiments that Newton managed to carry out before the window broke, the comparison showed differences of less than 2%; the larger differences in the third he attributed to using spheres whose buoyant weight in water was too near zero to allow accurate measurement.[44]

Once Newton knew that the pendulum-decay experiments were giving too high a value for the total resistance, he was quick to see why: the motion of the pendulum bob gives rise to a to-and-fro motion of the ambient fluid, and hence the relative velocity between the bob and the fluid is not the velocity of the bob, but some unknown larger value. Worse, the effects of the induced to-and-fro motion of the fluid differ in an irregular fashion among trials with different initial arcs, explaining the disappointingly large discrepancies when the simultaneous equation solutions for the A_i's were checked against the other data points. The fundamental problem with Newton's pendulum-decay experiments was the need to count the number of swings before a measurable fraction of the initial arc was lost. If he had been able to measure the fraction of the arc lost on the initial swing, before flow patterns are set up in the surrounding fluid, the phenomenon of pendulum-decay would have provided more accurate results for the total resistance.

Of course, even had the arc lost in the initial swing been measurable, the pendulum-decay experiments still would not have succeeded in disaggregating the contributions made to resistance by different fluid mechanisms, for the v^2 term would have remained overwhelmingly dominant; and, regardless, the different contributions are not independently additive in the manner Newton's approach had presupposed. I will return to this last point at the end of the paper.

Newton downplayed the pendulum-decay measurements in the second edition of the *Principia*, shifting the presentation of the results from the end of Section 7 to the end of Section 6, in the process dropping some of the conclusions he originally drew from them. The new version of Section 7 ends with a series of vertical-fall experiments. In addition to the initial set of vertical-fall experiments, it includes a new set consisting of nine experiments involving different size spheres falling in a 15 foot trough of water, and a set of vertical-fall experiments in air carried out by Hauksbee under Newton's direction in the newly completed St. Paul's Cathedral. These additional vertical-fall experiments were carried out within four or five years before the publication of the second edition in 1713. A still further set of experiments in St. Paul's was subsequently carried out by Desaguliers and included in the third edition. In point of fact, these vertical-fall experiments gave more accurate values for the resistance on spheres than Newton ever realized, as accurate as any experimentally determined values before the 20th century.[45] Any shortcoming in Newton's treatment

of resistance in the second and third editions therefore arose not, as in the first edition, from limitations in the experiments.

Before turning to his new theory of resistance, I should point out that Newton's response to the recognizable shortcomings of the data from the pendulum-decay experiments following the publication of the first edition was, on the whole, good science. He carried out careful experiments that not only exposed whether the shortcomings were leading to erroneous conclusions, but also indicated the likely sources of error; these he reported in the second edition, deterring others from making the same mistakes. True to the standards of the new "experimental philosophy" of the *Principia*, Newton was still insisting, at least during the years between the first and second editions, that experiment dictate theory, and to this end he was still taking the trouble to carry out demanding experiments.

THE NEW APPROACH TO RESISTANCE

From Newton's point of view, the problem he faced with resistance forces remained one of disaggregating the contributions made by different mechanisms, for only then could he experimentally establish laws characterizing each of these contributions, at least for spheres. At an absolute minimum, he wanted to isolate the contribution made by the inertia of the fluid sufficiently to establish a law for it. The realization that pendulum-decay experiments offered no hope led him to adopt an entirely new approach to this problem in the second edition. The pendulum-decay experiments had shown that the total resistance varies roughly as v^2 for everyday size spheres at moderate velocities. Newton took this to show that the contribution made by the inertia of the fluid is dominant over this range of size and velocity. His new approach was to try first to establish a law for the inertial contribution purely by itself. More precisely, with the second edition he introduced the notion of a continuous fluid with no internal or surface friction – what we term an incompressible, inviscid fluid – thinking he could first establish a law of resistance for bodies moving in such fluids. His hope was then to use deviations between measured resistance in actual fluids and calculated resistance in these idealized fluids in order to experimentally isolate the contributions made by such mechanisms as fluid viscosity and surface friction.[46] In other words, his hope was that resistance in the ideal case of an incompressible, inviscid fluid would provide at least a first approximation in a sequence of successive approximations.

Newton's attempt to establish a law of resistance in an idealized inviscid fluid appears at first to present the strongest candidate for bad science in Book 2, if not in the entire *Principia*.[47] I mean here not just bad science by his standards, but by those of anyone who thinks that experimental evidence is central to establishing claims in science. His reasoning involved three steps, in each of which he makes moves that seem open to serious objection.

The first step was to obtain a new result for purely inertial resistance on the front face of a body in a continuous fluid. The revised solution for the efflux velocity by itself cut this resistance in half, in effect from the drag coefficient of 2.0 in the first

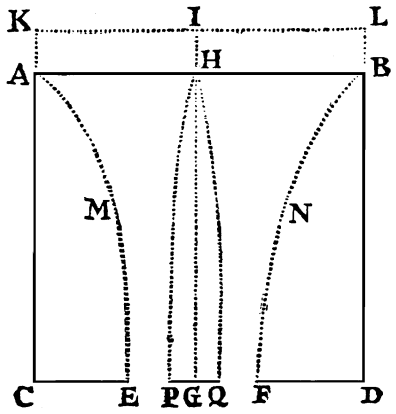

Figure 2. Newton's diagram in the second and third editions of the *Principia* for the column of "continuous" fluid borne by the moving body (PGQ).

edition to 1.0. The new efflux solution, however, also suggested that the column of fluid supported by a small disk in the middle of the efflux stream would not be the cylinder above it, for the flow incident on the disk would be at an acute angle. That is, in Figure 2 the angles HPG and HQG should be acute for the same (cataract) reason that the angles MEC and NFD are acute. Newton considered two limiting cases: (1) the column PQH must at least contain the cone PQH, and hence its weight on the circular disk PQ must be no less than one-third of that of the cylinder of fluid above this disk; and (2) the column must be contained in a half-spheroid, with angle HQP a right angle, and hence its weight on the disk must be less than two-thirds of that of the cylinder. Newton then chose the value one-half, on the grounds that "this weight is an arithmetical mean between the weights of the cone and the said half-spheroid."[48] Finally, he argued that this same column of fluid is supported by a cylinder on-end with the same circular cross-section as the disk, for the fluid on the sides of such a cylinder offers no resistance. This reasoning resulted in his new theoretical value for the purely inertial resistance on the front face of a cylinder-on-end, a value that amounts to a drag coefficient of 0.5, matching the value he had obtained experimentally for spheres in his initial vertical-fall experiments in water.

Newton's second step toward establishing a law of resistance for spheres in idealized inviscid fluids was to equate this inertial resistance on the front face of a cylinder-on-end with the total resistance on spheres in such fluids. This step involves a sequence of assertions. First, the pressure on the rear face of the cylinder remains the same when the efflux stream is passing it (or equally, when it is moving forward in the fluid). In other words, in contrast to the first edition, Newton has now decided that the question of the induced change in fluid pressure on the rear face of a body moving in an inviscid fluid can be answered from theory alone, and that answer is zero. Second, any oblique motion transferred to the fluid by the moving cylinder has no effect on the inertial resistance on the it. And third, "if a cylinder, a sphere, and

a spheroid, whose widths are equal, are placed successively in the middle of a cylindrical channel . . . , these bodies will equally impede the flow of water through the channel."[49] Newton offers qualitative reasoning in support of these three claims: for the first, appealing to the supposed instantaneous propagation of any disturbance in an incompressible fluid; and for the other two, asking the reader to think of the fluid not contributing resistance as frozen in the manner that he had earlier chosen to visualize the portions of the fluid FNBD and EMAC in Figure 2 in his cataract solution for the efflux. The net result of all of this was a new *purely theoretical* law for the resistance force on a sphere (or cylinder-on-end) in an idealized incompressible, inviscid fluid: $0.25\rho_{\mathrm{f}}A_{\mathrm{front}}v^2$.

One worrisome feature in this reasoning is that it seems to be tailored to match a known experimental result. Its real fault, however, lies in why it seems so tailored. It is supposed to be a purely theoretical result for the case of an idealized inviscid fluid. By Newton's standards, the appropriate way to arrive at such a result is to derive it from laws of motion, if need be by first deriving equations of motion for such a fluid passing a sphere or cylinder-on-end, and then determining from these equations the resulting pressures on the front and rear faces of the body. Instead of doing this, Newton simply makes assertions about what happens when an inviscid fluid passes a sphere or cylinder-on-end, and then offers us a way of visualizing parts of the fluid as frozen that allow his assertions to be true. At most, this shows only that his law for resistance on a sphere in an inviscid fluid is plausible! No grounds are given for concluding that the fluid passing around a sphere or cylinder-on-end must behave in the manner he says. As d'Alembert later remarked, Newton's reasoning here is "intended to elude rather than surmount the difficulty of the problem."[50]

The third step in Newton's new approach is to compare the theory with vertical-fall data in air and water, concluding that "almost all the resistance encountered by balls moving in air as well as in water is correctly shown by our theory, and is proportional to the density of the fluids—the velocities and the sizes of the balls being equal."[51] To the extent that both the choice of one-half for the fraction of the weight of the fluid cylinder supported by the disk in the efflux stream and the claims about the induced action on the rear face and the effects of shape were tailored to match the vertical-fall data, this conclusion should scarcely be surprising. The conclusion thus underscores the question, was his derivation of the theory sufficiently justified in the absence of the data? As already suggested, the answer is no.

Alternatively, one might view Newton's theory as a hypothesis being tested by the vertical-fall data. This, however, is unsatisfactory, and not just because of Newton's disdain for hypothetico-deductive evidence. The theory is formulated for idealized inviscid fluids, and the vertical-fall data are for air and water, neither of which is inviscid. The only thing that the vertical-fall data can show is just what Newton so carefully says they show in the above quote: his $0.25\rho_{\mathrm{f}}A_{\mathrm{front}}v^2$ law for inviscid fluids provides a close approximation to the resistance measured in the vertical-fall experiments. The fact that it does provide such a close approximation is in no way evidence that this law correctly describes resistance on spheres in an idealized inviscid fluid, much less evidence for his claims about what would happen when such

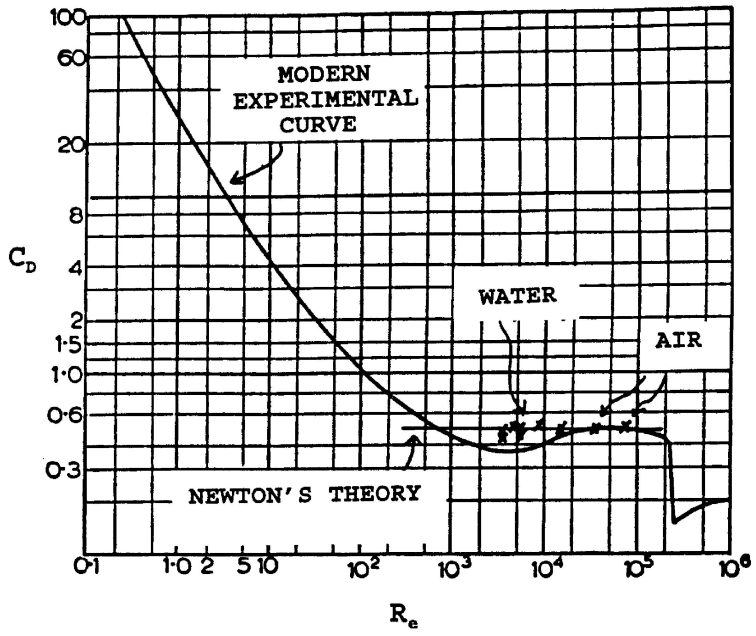

Figure 3. Drag coefficient versus Reynolds number for spheres: a modern assessment of Newton's theory and his vertical-fall data from the second and third editions of the *Principia*.

an inviscid fluid flows past such a sphere. If Newton is trying to give the reader the impression that his vertical-fall data confirm his theory, then Truesdell is correct when he speaks of Newton as engaged in "bluff."

A closer look at the data shows that the actual situation is worse than this. Figure 3 compares the drag-coefficients from Newton's vertical-fall data with the 0.5 constant value from his theory and with modern measured values, taking as independent variable the modern Reynolds number.[52] While most of the vertical-fall data lie close to his theory, some of the experimentally determined resistances were non-trivially larger than the theory implied, and some were *less*. Newton noted the cases where the experimental resistances were significantly larger, calling attention to the high velocities in these cases and arguing that when the velocity becomes high enough, there is a loss of fluid pressure on the rear face of the moving body, contrary to the assumption made in deriving his theory. He says nothing, however, about the cases in which the experimental resistances fell non-trivially below his theoretical value. (Notice that these cases represent ones where modern data indicate that the drag coefficient indeed falls significantly below 0.5; this is the strongest sign of the high quality of Newton's data.) In putting forward his theory, Newton remarked that it is supposed to give the least resistance that can occur insofar as the viscous and surface-friction effects will augment the resistance from fluid inertia. Unless the experimental resistances falling

below his theory are attributed to experimental error, therefore, they cannot help but raise questions about whether the theory is correct.

In sum, Newton's treatment of the resistance arising from the inertia of the fluid in the second edition, besides being wrong, appears to be bad science both by his standards and those of most anyone else. The theoretical part of this treatment involves several assertions – "pure guessing", to use the word Truesdell conjoins with "bluff" – about how an inviscid fluid would behave as it flows around a spherical obstruction. These assertions were not obtained by solving equations of motion for an inviscid fluid, equations that by Newton's standards ought to be derived from his laws of motion. Consequently, the theoretical part of the treatment has either been tailored to fit the data and hence is ad hoc or has been put forward as a hypothetical conjecture calling for independent evidential support. Either way, Newton's vertical-fall data for water and air provide no real evidence for his theory of resistance in inviscid fluids. To whatever extent Newton's presentation gives an impression that these data do constitute evidence for the theory, that presentation is misleading. The only element of good science here appears to be the vertical-fall data themselves.

AN ALTERNATIVE INTERPRETATION OF WHAT NEWTON WAS UP TO

Part of this critique of Newton's treatment of resistance in the second and third editions presupposes that he was offering the vertical-fall data in support of his theory of resistance in incompressible, inviscid fluids. This, however, is not what the conclusion quoted above actually says: "almost all the resistance encountered by balls moving in air as well as in water is correctly shown by our theory, and is proportional to the density of the fluids."[53] Moreover, the phrasing he uses when he first turns to the vertical-fall data, at the end of the presentation of the theory, does not imply that the data are to be taken as evidence for the theory:

> This is the resistance that arises from the inertia of matter of the fluid. And that which arises from the elasticity, tenacity, and friction of its parts can be investigated as follows.[54]

What immediately follows this statement is an exact solution for vertical fall with resistance given by $0.25\rho_{\mathrm{f}}A_{\mathrm{front}}v^2$, ending with the remark that this solution covers the case in which

> . . . the sphere encounters no other resistance than that which arises from the inertia of matter. But if it encounters another resistance in addition, the descent will be slower, and the quantity of this resistance can be found from the retardation.[55]

If we take Newton at his word, then, the vertical-fall data were not intended simply to confirm his theory.[56] Rather, the theory was supposed to be taken as established by its derivation, and the vertical-fall data were supposed to be showing two things: (1) virtually all of the resistance the spheres encounter in his vertical-fall experiments

is from the inertia of the fluid alone; and (2) the theoretical solution can be adopted as an idealized first-approximation to the true resistance, and discrepancies between it and true resistances measured in vertical-fall can be used in researching the contributions made to resistance by factors other than the inertia of the fluid.

Both parts of this conclusion were still wrong. The question is whether Newton is equally open to criticism under this interpretation of what he was up to. In considering this question, we must not loose sight of the importance Newton attached to part (1) of the proposed two-part conclusion. It provides the basis for the further conclusion with which Section 7 ends:

> And even if air, water, quicksilver, and similar fluids, by some infinite division of their parts, could be subtilized and become infinitely fluid mediums, they would not resist projected balls any the less. For the resistance which is the subject of the preceding propositions arises from the inertia of matter; and the inertia of matter is essential to bodies and is always proportional to the quantity of matter. By the division of the parts of a fluid, the resistance that arises from the tenacity and friction of the parts can indeed be diminished, but the quantity of matter is not diminished by the division of its parts; and since the quantity of matter remains the same, its force of inertia – to which the resistance discussed here is always proportional – remains the same. For the resistance to be diminished, the quantity of matter in the spaces through which bodies move must be diminished. And therefore the celestial spaces, through which the globes of the planets and comets move continually in all directions freely and without any sensible diminution of motion, are devoid of any corporeal fluid, except perhaps the very rarest of vapors and rays of light transmitted through those spaces.[57]

Crucial to this reasoning is the claim that virtually all of the resistance measured in the actual fluids is from their inertia and would remain even if these fluids were perfectly inviscid; and this claim depends first on the close agreement between the vertical-fall data and Newton's theory and second on the claim that his theory gives resistance in an inviscid fluid. When d'Alembert showed that the resistance in an idealized inviscid fluid is exactly zero, Newton's argument that celestial space is devoid of fluid lost much of its force.[58]

Thus, one reason for criticism of Newton remains under the proposed interpretation: his law for resistance on spheres in idealized inviscid fluids was not derived from his laws of motion, but instead was predicated on a number of seemingly arbitrary guesses about what would happen when such a fluid flows past a sphere. Still, this line of criticism is somewhat mitigated on two counts. First, the close agreement between theory and experiment in air as well as in water provided evidence that, at least across the range covered in the experiments, the total resistance on a sphere varies to a high approximation as the density of the fluid. Similarly, the close agreement between theory and experiment throughout provided evidence that the total resistance across this range varies more or less as the velocity squared – as it should if the underlying mechanism is one of direct momentum transfer to fluid particles. Therefore, this agreement provided evidence that the physical factor dominating the observed resistance

forces is the inertia of the fluid.[59] This, in turn, provided reason, beyond its derivation, to take Newton's theory as a *promising candidate* for resistance in idealized inviscid fluids – at least more reason than it would have if the agreement between theory and experiment had not been so good across the range of the experiments.

Second, and more important, the proposal of using the discrepancies between this theory and measurement as a source of evidence for investigating the contributions made by the compressibility, surface friction, and viscosity of the fluid opened a *potential* avenue for evidence supporting the theory. For, suppose these discrepancies were to enable laws to be obtained describing the contributions made by those other factors; this would give strong indirect evidence for the claim that the resistance on spheres in idealized inviscid fluids is given by $0.25\rho_f A_{front} v^2$, the arbitrary steps in the derivation of this relation notwithstanding. Equally, failure in the attempt to marshall the discrepancies into evidence on the contributions made by the other mechanisms would provide grounds for questioning the $0.25\rho_f A_{front} v^2$ law for spheres in inviscid fluids. So, at the very least, Newton's theory for resistance in inviscid fluids was creating a possibility for bringing more effective evidence to bear on the question of resistance forces than the pendulum-decay experiments had brought.

On this interpretation, then, Newton's treatment of resistance in the second and third editions was a not entirely unreasonable response to the limitations of the pendulum-decay experiments. The initial vertical-fall experiments and the revised efflux solution put him in a position to take a stab at a law of resistance for idealized inviscid fluids; and this law, together with his exact solution for vertical fall under resistance forces varying as v^2, offered at least a hope for obtaining more informative data. This interpretation puts the arbitrary claims Newton made about flow of inviscid fluids around bodies in a somewhat different light. For example, the claim that motion does not alter the fluid pressure on the rear face of a body can be viewed not as a final pronouncement, but as an assumption that any such effect on the rear face is at most second-order compared with the inertial force on the front face. Newton's comment that his vertical-fall data indicate a loss of pressure on the rear face at very high velocities is consistent with this weakened reading of the claim. The proposed law for inviscid fluids can then itself be treated as a mere first-approximation to the inertial contribution, with discrepancies between it and experiment serving not only to provide more effective experimental access to the other contributions, but also data that might help toward refining it. On this view, Newton was leaving it to subsequent investigators either to conduct further, perhaps more exacting, vertical-fall experiments and evaluate what the discrepancies between them and his theory showed, or to devise a properly rigorous account of the flow of an inviscid fluid around a body in order to assess and refine his assumptions about which effects are first-order and which, second-order.

The complicated variation of the modern measured values in Figure 3 for the drag coefficient on spheres over the range covered by Newton's data shows that measured deviations from his theory in further vertical-fall experiments were not going to do what he had hoped. Further experiments, however, would not have readily shown that his proposed law for resistance in idealized inviscid fluids is categorically wrong. Indeed, Newton's derivation left room to reduce the value of the leading coefficient in

the law below 0.25, so that a slightly modified version of the law would have removed the problem of measured resistances falling below it. Only with d'Alembert's result, and its subsequent confirmation with Euler's more canonical equations of motion for idealized inviscid fluids,[60] could the full extent of Newton's mistakes have become clear. Not only were his guesses wrong, but his supposed resistance law is not even so much as a first approximation, and the idea of using measured discrepancies from any theoretical law for a purely inertial contribution in order to disaggregate the different contributions is a total dead end. For, there simply is no such thing as a purely inertial contribution to resistance. Newton, in truth, had done nothing more in the second edition than fit a simple curve to his vertical-fall data.

Still, Newton's treatment of resistance in the second and third editions is not necessarily so radical a violation of his standards for good science as it may first appear to be. One might criticize him for not stating his bottom-line anti-Cartesian conclusion in a more qualified fashion, or for not calling explicit attention to the fact that the resistance forces found in some of his vertical-fall experiments were less than his theory required. Yet he scarcely hid the lack of rigor in the derivation of his theory, and he made clear that he had his initial vertical-fall results before formulating it. Anyone who reads the text closely without a pre-commitment to the view that all evidence in science must be hypothetico-deductive can see what he was up to. He was taking a stab at an idealized first-approximation that gave hope of opening the way to more tractable research into the complexities of resistance. Moreover, one conclusion Newton drew from the good agreement between his theory and his vertical-fall data was entirely correct: over the range of these experiments the resistance is predominately from the inertia of the fluid. Should he be criticized for not realizing that inertial effects of the fluid can be predominant without their arising from some independent, purely inertial mechanism?

GOOD SCIENCE OR BAD?: A PHILOSOPHICAL ASSESSMENT

We still do not have a law for resistance forces of the sort Newton wanted, even just for spheres. We have only experimentally determined relationships between the drag coefficient and the Reynolds number for specific shapes, like the one for (smooth) spheres shown in Figure 3.[61] The failure of 300 years of effort to produce a general law for a kind of force that has become of great practical importance in the aerospace age is a reminder of just how elusive success can be in theoretical science. So long as "good" science is not simply equated with successful science, no standards for good science are going to provide a guarantee of success. Sound methodology alone is never enough.

On my view, the fundamental problem in doing science is turning data into evidence. Because evidence is a relationship between data and claims that reach beyond them, some mediating element is always needed in order for data to become evidence. For example, many of the propositions of Newton's mathematical theory of centripetal force in Book 1 of the *Principia* become such mediating elements in Book 3, enabling phenomena of celestial motion to become evidence for claims

about the physical forces acting on orbiting bodies. Similarly, Newton's mathematical solution for pendulum-decay under resistance that varies as velocity to an arbitrary power had the promise of enabling pendulum-decay data to yield evidence about how resistance forces vary. In spite of its already noted limitations, even Newton's exact solution for vertical fall with resistance varying as velocity squared enabled his vertical-fall data to become strong evidence that the inertia of the fluid is the dominant factor over the (then-yet-to-be-characterized) range covered in his experiments. As these examples indicate, the mediating element that allows data to be turned into evidence is invariably theoretical in character, and hence to some extent provisional – sometimes exceedingly so.

I am prepared to call any science good if it has clear promise of bringing stronger evidence to bear in some area of research than was previously available, *provided* that it includes safeguards against being misled by the apparent high quality of the initial evidence obtained from the readily accessible data. The proviso is needed because of the risk of being led down extended garden paths in research, especially when the new evidence is being compared with a virtual absence of prior evidence. Lacking an omniscient external perspective, the only way we have to judge the quality of evidence initially is from various of its internal features. For example, are the data tight and well-behaved, and are the mediating elements involved in turning these data into evidence free from obvious objections? The trouble is that accidental or parochial factors can still make the initial evidence promising, launching research down a path that later turns out to have all along been a dead end. Newton's standards for good science can be viewed as an attempt to protect against such garden paths. Thus, for example, the requirement that force laws be "deduced" from phenomena assures that these laws hold at least to high approximation of the specific data defining the phenomena (unless something is seriously wrong in his laws of motion).[62] Similarly, the demand that theory do more than just explain these and related phenomena, that it immediately become an instrument in ongoing research, helps to expose any parochially misleading promise in the initial evidence, before being led far down a garden path.

By my proposed standards for good science, Newton's efforts on resistance forces in neither the first nor the second edition of Book 2 are especially bad science, their failure notwithstanding. His approach in the first edition required that pendulum-decay data from any one pendulum bob in any one medium yield constant values for the coefficients A_n in a mathematical framework with no more than two or three $A_n v^n$ terms; and his mathematical solution for pendulum-decay, which enabled values for A_n to be obtained from data, put no restriction on the exponents n. This placed a strong internal demand on the initial evidence, so that it was easy to see that the evidence the data were actually yielding was of limited quality. Because the approach taken in the second edition to a significant extent tailored the v^2 term to fit available vertical-fall data, it had no way of placing such an internal demand on the initial evidence. Nevertheless, the requirement that the residual discrepancies between this theoretical v^2 term and experiment provide a basis for characterizing the contributions made by such other factors as compressibility and viscosity safeguarded against making too much of the close agreement between the vertical-fall results and the theory of the

inertial contribution. Moreover, Newton's "derivation" of the supposed purely inertial contribution clearly identified places where, in the absence of equations of motion for an inviscid fluid, he had to make assumptions about how such a fluid would behave flowing past an obstacle; and each of these could be reassessed once the relevant equations of fluid motion had been formulated. In short, each of the two approaches had promise of bringing stronger evidence to bear on questions about resistance forces than was previously available, and each included safeguards against being led down an extended garden path should the initial evidence appear to be of high quality.

Still, one might argue, this defense of Newton does not absolve him. For it says nothing about the error that was fundamentally responsible for the failure of both of his approaches: his assumption that resistance forces result from a combination of independent physical mechanisms, each of which varies as velocity to some characteristic power, and one of which represents the contribution made purely by the inertia of the fluid. This assumption was indispensable to his attempt to use pendulum-decay data to establish a law of resistance forces, and thus doomed this approach from the outset. It was also crucial to the main error made in his attempt to use vertical-fall data, namely the inference from the sound conclusion that inertial effects of the fluid were predominate over the range of these data to the mistaken conclusion that the magnitude of the total resistance over this range was giving a good approximation to the magnitude of the supposed purely inertial contribution. Newton claimed that "hypotheses have no place" in his experimental philosophy, "unless as conjectures or questions proposed to be examined by experiments."[63] Yet here is an erroneous hypothesis on which all his efforts on resistance were predicated and which was the ultimate source of their failure. Why then isn't this an archetypal example of wrong science that was bad science?

The purpose to which Newton put this erroneous assumption was to license the weaker assumption that resistance forces can be characterized by means of a sum of a few terms, each of which varies as some power of velocity. This weaker assumption then warranted his derivation of the propositions that comprise most of Book 2, propositions relating motions and resistance forces that vary as a single power of velocity – propositions that were intended to enable data from motions to be turned into evidence about resistance forces. The counterpart to this assumption in the more famous part of the *Principia* is the assumption that the forces retaining the planets and their satellites in orbit vary as a function of their distance from the respective central bodies, and not also as a function of their angular position, that is, as a function of r, and not of θ and φ, in spherical coordinates. (This latter assumption, it should be noted, was in direct conflict with the Cartesian vortex theory, for the force retaining planets in their orbits in this theory depends on contiguous, surrounding vortices, and hence would be a function of both θ and φ.) Neither of these assumptions is a hypothesis in the broad sense that Newton railed against, for both are minimal from a physics standpoint, and neither is explanatory. Moreover, neither is directly testable, but only indirectly through the success of the research predicated on them. I prefer to call such assumptions *working hypotheses* in the restricted sense that they are not testable in and of themselves, yet they are indispensably presupposed in a train of

evidential reasoning. As said above, propositions with theoretical content of some sort are needed as mediating elements in order to turn data into evidence. At least in the initial stages of theory development, working hypotheses in the sense I am proposing play this role. They are crucial to getting the process of marshalling evidence off the ground.

The thought behind any such working hypothesis is clear. If it holds, at least to a suitable approximation, then there are prospects for empirically driven, sustained research; if it does not – if the empirical world does not cooperate – then there is no apparent way to get beyond mere conjecture. Obviously, any such working hypothesis involves an element of wishful thinking. Still the thought behind it is not idle.

The only way to establish such a working hypothesis is through the success of the research predicated on it. The short-term requirement is success in turning data into well-behaved evidence, and the long-term requirement is continuing improvement in the agreement between theory and observation. When these requirements are met, the hypothesis becomes part of established science, often without notice or fanfare. Newton's working hypothesis for orbital motion, and the almost total absence of comment, both then and now, about its flying in the face of Cartesian vortex theory, illustrate this. Also, a working hypothesis does not have to hold exactly in order to meet these requirements.[64] For, a suitable first approximation can open the way to sustained research that ultimately yields a refinement of it without invalidating the prior evidential reasoning predicated on it.

The long-term requirement on working hypotheses and the risk of being led down a garden path by short-term success offer philosophical reasons to prefer Newton's unusually demanding standards for good science. One objective to which these standards are responsive is to expose irretrievable shortcomings in a working hypothesis at the earliest moment. His demand that the data unequivocally determine the values of such theoretical parameters as the exponent in the centripetal force law has the effect of highlighting any vagaries in the data, preventing their being dismissed or swept under a rug. Thus, had Newton been able to measure the arc lost in the first swing of pendulum-decay, thereby avoiding the confounding effects of motion induced in the surrounding fluid, the failure to obtain a stable value for the coefficient of the v^2 term from the data would have raised immediate questions about his working hypothesis. Indeed, one can view his initial vertical-fall experiments as an attempt to determine whether the vagaries in the pendulum-decay results came from defects in the experiment or shortcomings in his working hypothesis.

Similarly, Newton's demand that theory do more than just explain the phenomena providing the initial evidence for it – that theory become an instrument in further research – has the effect of putting a working hypothesis under continuing scrutiny, probing for shortcomings. Newton's theory of resistance in idealized inviscid fluids provided a seemingly adequate explanation for why inertial effects were predominate in his vertical-fall experiments; even so, a subsequent failure in trying to use discrepancies between this theory and vertical-fall data to make progress on the non-inertial contributions to resistance would in time have raised doubts about whether resistance does arise from the additive contributions of independent mechanisms. Of course,

d'Alembert short-circuited this process when he showed that the rigorously derived value for the resistance in an incompressible, inviscid fluid is exactly zero. This theoretical result, together with Newton's strong evidence that inertial effects dominate across a range of fluids and sizes of spheres, made clear that the additive working hypothesis was not even so much as a suitable first approximation for purposes of developing evidence toward a law of resistance forces.

The precise respect in which Newton's additive working hypothesis was not a suitable first approximation is important. A moment's thought about the modern measured variation of the drag coefficient with Reynolds number in Figure 3 shows that a curve of the form, $a_1 v + a_2 v^2$, can provide a reasonably good fit to measured resistance on spheres up to Reynolds numbers of 2×10^5, where the flow becomes fully turbulent. Indeed, engineers still use a formula of this form when all they need is a rough approximation. Mere numerical approximation, however, is not enough to initiate sustained development of theory in successive approximations. The deviations from a mere numerical approximation are too likely not to be physically informative, but instead just reflections of the choice of mathematical framework. This is why a proper statement of Newton's working hypothesis is not simply that resistance forces can be characterized by a sum of a few terms in powers of v, but beyond this that these terms represent independent physical contributions. It also gives a further reason, in addition to his goal of arguing against Descartes, for why it was important to show that the v^2 term represents a purely inertial contribution. The greatest promise for initiating a sustained program of successive approximations is when the initial approximation is an idealization that, according to the theory, would hold exactly in certain identifiable circumstances – that is, the kind of idealization that Keplerian motion turned out to be on Newton's theory of gravity.

A century and a half elapsed between d'Alembert's publication of his paradox and Ludwig Prandtl's resolution of it.[65] Prandtl showed that, at least for many engineering purposes, fluid flow at Reynolds numbers above 10^3 is best regarded as involving two regimes: inviscid flow in the main stream and viscid flow in "boundary layers" in the close vicinity of solid surfaces. The resistance force on a body is governed by the flow in the boundary layer and its consequent effects on the flow in the wake of the object. (Newton was right to be worried about the induced fluid effects on the rear face of the body.) The reason for the complex variation of the drag coefficient curve over Reynolds numbers between 10^3 and 10^5 (as well as for the abrupt drop in the curve above 10^5) is the onset of vortices and turbulence within the boundary layer and in the wake.[66] We will not have a law of resistance forces of the sort Newton sought, even for spheres, until we have a theory of turbulence.

During the century and a half between the discovery of d'Alembert's paradox and Prandtl's resolution of it, the theory of idealized inviscid fluids continued to be developed, under the rubric of "hydrodynamics." Some results of this effort have come to have useful engineering applications in the post-Prandtl period, and a few of the results were lasting contributions to physics. Nevertheless, the field of hydrodynamics was much more a mathematical exercise than an attempt to find ways for turning data into evidence in fluid mechanics. This led to an odd situation summarized by a now

famous remark made by Sir Cyril Hinshelwood: "fluid dynamicists were divided into hydraulic engineers who observed what could not be explained, and mathematicians who explained things that could not be observed."[67] If one is looking for an example of bad science, then by my standards a far better candidate than Newton's failed efforts in Book 2 of the *Principia* is the century and a half of hydrodynamics between d'Alembert and Prandtl.[68]

Philosophy Department, Tufts University, USA

NOTES

[1] Clifford Truesdell, "Reactions of Late Baroque Mechanics to Success, Conjecture, Error, and Failure in Newton's *Principia*," in *Essays in the History of Mechanics* (New York: Springer-Verlag, 1968), p. 149. I have ended this quotation here, rather than including the remainder of the paragraph, because Truesdell assigns an importance to the Bernoullis and Euler in the subsequent history that seems to me to seriously understate the contributions of Clairaut and d'Alembert.

[2] Truesdell makes clear in the cited article and in other articles in the same book that he viewed science at its best as driven more by the pursuit of mathematically elegant "rational" representations of the world than by the pursuit of compelling empirical evidence. See also his *An Idiot's Fugitive Essays on Science: Methods, Criticism, Training, Circumstances* (New York: Springer-Verlag, 1984) and *The Elements of Continuum Mechanics* (New York: Springer-Verlag, 1966).

[3] In the first edition of the *Principia* Newton appeals to values of the speed of sound put forward by Mersenne and Roberval and to experiments he conducted in Trinity College to conclude that his theoretical value lies within the error bounds of experiment. See Isaac Newton, *The Principia: Mathematical Principles of Natural Philosophy*, trs. I. Bernard Cohen and Anne Whitman (Berkeley: University of California Press, 1999), p. 776f, n. aa. The reference to "fudge factors" in the text is an allusion to Richard S. Westfall's "Newton and the Fudge Factor," *Science* 179: 751–758 (1973), where pp. 752 and 753f discuss Newton's handling of the speed of sound.

[4] See S. Chandrasekhar, *Hydrodynamic and Hydromagnetic Stability* (Oxford: Clarendon Press, 1961), p. 272f, for a derivation of the correct relationship; Newton's error was noted by G. G. Stokes in his "On the Theories of the Internal Friction of Fluids in Motion, and of the Equilibrium and Motion of Elastic Solids" of 1845, p. 103, reprinted in *Mathematical and Physical Papers*, vol. 1 (Cambridge: Cambridge University Press, 1880), pp. 75–129. The error Newton makes in treating vortices generated by rotating cylinders and spheres is discussed more recently in Geoffrey J. Dobson, "Newton's Errors with the Rotational Motion of Fluids," *Archive for History of Exact Sciences*, 54: 243–254 (1999).

[5] Stokes, *ibid*. Specifically, Stokes points out,

> It may be shewn from the general equations that in this case [a rotating sphere] a permanent motion in annuli is impossible, and that, whatever may be the law of friction between the solid sphere and the fluid. Hence it appears that it is necessary to suppose that the particles move in planes passing through the axis of rotation, while at the same time they move round it. In fact, it is easy to see that from the excess of centrifugal force in the neighborhood of the equator of the revolving sphere the particles in that part will recede from the sphere, and approach it again in the neighborhood of the poles, and this circulating motion will be combined with a motion about the axis. If however we leave the centrifugal force out of consideration, as Newton has done, the motion in annuli becomes possible, but the solution is different from Newton's, as might have been expected.

Dobson gives such a Newtonian steady-state solution in his article (*ibid.*) citing I. G. Currie, *Fundamental Mechanics of Fluids* (New York: McGraw-Hill, 1974), p. 267. This, however, is a so-called "low-Reynolds-number solution" – i.e., a solution obtained by eliminating the inertial terms from the Navier–Stokes equations, leaving only the pressure and viscous terms. In other words, unlike the solution for rotating cylinders, the corrected Newtonian solution for rotating spheres given by Dobson does not represent a physically realizable condition. Nor does it represent a condition that Newton himself would have considered appropriate for arguing against Cartesian vortices insofar as his main argument against these vortices, put forward forcefully in the second and third editions of both the *Principia* and the *Opticks*, emphasized the ineliminability of the inertial effects of the fluid. (This argument will be discussed in Section 6 below.)

[6] Another place in the *Principia* where Newton's treatment of angular motion is inadequate is his derivation of the rate of the precession of the equinoxes; see Geoffrey J. Dobson, "Newton's Problems with Rigid Body Dynamics in the Light of his Treatment of the Precession of the Equinoxes," *Archive for History of Exact Sciences* 53: 125–145 (1998) and "Against Chandrasekhar's Interpretation of Newton's Treatment of the Precession of the Equinoxes," *ibid.*, 577–597; and Curtis Wilson, "D'Alembert versus Euler on the Precession of the Equinoxes and the Mechanics of Rigid Bodies," *Archive for History of Exact Sciences* 37: 233–273 (1987). In his youthful tract, "The Lawes of Motion," Newton recognizes that motion upon impact can be transferred to both translational and rotational components, but the only way he sees to determine the fraction between these two is by means of experiment; see *Unpublished Scientific Papers of Isaac Newton*, eds. A Rupert Hall and Marie Boas Hall (Cambridge: Cambridge University Press, 1962), pp. 157–164.

[7] In other words, the correct speed of sound is given by $(\gamma p/\rho)^{1/2}$, where γ is the ratio of the specific heats, p is the pressure, and ρ is density, and not by $(p/\rho)^{1/2}$, as Newton had it. This error was first corrected by Laplace; see his *Traité de Mécanique Céleste*, vol. V (Paris: Bachelier, 1825), as reprinted in *Celestial Mechanics* (New York: Chelsea, 1969), pp. 133–145.

[8] Specifically, in the text of Bk. 2 Prop. 34, *Principia*, p. 729.

[9] R. Giacomelli (with the collaboration of E. Pistolesi), "Historical Sketch," *Aerodynamic Theory: A General Review of Progress*, ed. William Frederick Durand, vol. 1 (New York: Dover Publications, 1963), p. 311.

[10] *Principia*, p. 730.

[11] M. le Chevalier de Borda, "Expériences sur la Résistance des Fluides," *Mémoires de l'Académie Royale des Sciences* (1763), pp. 358–376, and "Expériences sur la Résistance des Fluides," *ibid.* (1767), pp. 495–503. Newton did not publish the mathematical derivation of the surface of least resistance, which employed the calculus of variations decades before Euler "invented" it. See *The Mathematical Papers of Isaac Newton*, vol. VI, ed. D. T. Whiteside (Cambridge: Cambridge University Press, 1974), pp. 456–480.

[12] Much the same excessive sense of Newton's infallibility seems to have occurred in the case of his even more openly exploratory conclusion in Bk. 2 Prop. 23 that Boyle's law would hold exactly in a fluid composed of static particles which repel their immediate neighbors by means of "centrifugal" forces inversely proportional to the distances among the neighboring particles. Daniel Bernoulli subsequently demonstrated that Boyle's law would also hold in a fluid consisting of particles in motion in which the only forces stem from impact. The excessive regard accorded Newton's exploratory conjecture may explain why Bernoulli's result received so little attention, perhaps delaying for decades efforts toward a kinetic theory of gases (Truesdell, "Early Kinetic Theory of Gases," *op. cit.*, pp. 272–304). Even in 1875 Maxwell took the trouble to respond to Newton's static particle conjecture ("On the Dynamical Evidence of the Molecular Constitution of Bodies," *The Scientific Papers of James Clerk Maxwell*, vol. 2 (New York: Dover reprint, 1890), p. 422). I have not included Prop. 23 in my list of errors in Book 2 precisely because Newton is so explicit about its status: "Whether elastic fluids consist of particles that repel one another is, however, a question for physics. We have mathematically

demonstrated a property of fluids consisting of particles of this sort so as to provide natural philosophers with the means with which to treat this question." (p. 699) Indeed, I regard Prop. 23 as an example of science at its best because it calls attention to the possibility that Boyle's law might help provide experimental access to the microstructure of air.

[13] For a translation of the subsequently replaced portions of Bk. 2 Section 7 from the first edition, see "Newton on Fluid Resistance in the First Edition: English Translations of the Passages Replaced or Removed in the Second and Third Editions," the Appendix to my "The Newtonian Style in Book II of the *Principia*," *Isaac Newton's Natural Philosophy*, eds. Jed Z. Buchwald and I. Bernard Cohen (Cambridge: MIT Press, 2001), pp. 299–313. The contrast between Newton's theoretical treatments of fluid resistance in the first and second editions is discussed in detail in my "Fluid Resistance: Why Did Newton Change His Mind?," in *The Foundations of Newtonian Scholarship*, eds. Richard H. Dalitz and Michael Nauenberg (Singapore: World Scientific, 2000), pp. 105–136.

[14] The modern drag coefficient is given by

$$C_{\mathrm{D}} = \frac{2F_{\mathrm{resist}}}{\rho_{\mathrm{f}} A_{\mathrm{front}} v^2}$$

where ρ_{f} is the fluid density, A_{front} is the frontal area of the body, and v is the (relative) velocity between the fluid and the body. All but the factor of 2 in this non-dimensionalization is from Newton's *Principia*. Newton uses a different non-dimensionalization: the resistance to a sphere is to the force by which its whole motion could be either destroyed or generated, in the time in which it describes $x/3$ of its diameter, as the density of the fluid to the density of the sphere. Here, $x = 4/C_D$.

[15] *Principia*, pp. 733–761.

[16] D'Alembert's "paradox" asserts that the resistance on a body immersed in an inviscid, incompressible fluid is exactly zero, regardless of shape. He published this result for a range of shapes in his *Essai d'une nouvelle théorie de la résistance des fluides* (Paris: David, 1752), and he revisited the topic in his "Paradoxe proposé aux Géometres sur Résistances des Fluides," *Opuscules mathématiques*, vol. V (Paris: Jombert, 1768), pp. 132–138. For an especially clear modern analysis, see L. Prandtl and O. G. Tietjens, *Applied Hydro- and Aerodynamics*, tr. J. P. Den Hartog (New York: Dover, 1934), pp. 104–107. Truesdell discusses the history of d'Alembert's paradox in his *Rational Fluid Mechanics, 1687–1765*, Editor's introduction to *Leonhardi Euleri Opera Omnia*, ser. 2, vol. 12 (Lausanne: Societatis Scientiarum Naturalium Helveticae, 1954), pp. LII–LVIII and XL (where he indicates that Euler had already come upon the paradoxical result before d'Alembert published it).

[17] From a letter of 6 February 1672 to Henry Oldenburg, in *Correspondence of Isaac Newton*, vol. 1, ed. H. W. Turnbull (Cambridge: Cambridge University Press, 1959), p. 96.

[18] In Newton's reply to the second letter of P. Pardies, in *Philosophical Transactions of the Royal Society* (1672), p. 740, reprinted in *Isaac Newton's Papers and Letters on Natural Philosophy*, ed. I. Bernard Cohen (Cambridge: Harvard University Press, 1958), p. 106.

[19] *Isaac Newton's Papers and Letters on Natural Philosophy*, p. 119.

[20] "De Motu Corporum in Gyrum", *The Mathematical Papers of Isaac Newton*, vol. VI. p. 78. The translation of the excerpt from the "Copernican Scholium" of the augmented version is by Curtis Wilson in "The Newtonian Achievement in Astronomy," in *Planetary Astronomy from the Renaissance to the Rise of Astrophysics*, *Part A: Tycho Brahe to Newton*, *The General History of Astronomy*, vol. 2A, eds. R. Taton and C. Wilson (Cambridge: Cambridge University Press, 1989), p. 253. The material added to the original version of "De Motu" to form the so-called "Augmented Version" was first published in 1893 in W. W. Rouse Ball's *An Essay on Newton's "Principia"* (London: Macmillan, 1893), pp. 51–56; hence this remarkable passage was for all practical purposes unknown until then.

[21] Newton was aware of (at least) the alternative approaches by Ismaël Boulliau, Thomas Streete, and Vincent Wing. In the *Principia* Newton expressly cites Boulliau's alternative

(p. 800), which uses a geometric construction to replace the area rule. Wing originally used an oscillating equant in place of the area rule, later changing to still another geometric construction. Streete employs Boulliau's construction, and he follows Jeremiah Horrocks in using Kepler's 3/2 power rule to infer the length of the semi-major axis from the better known period. For details, see Curtis Wilson, "Predictive Astronomy in the Century after Kepler," Taton and Wilson, *op. cit.*, pp. 161–206).

[22] The quotation is from Newton's fourth Rule of Reasoning, added to the third edition of the *Principia*, p. 796.

[23] *Principia*, p. 588f.

[24] Thus, for example, Newton never infers the inverse-square from the Keplerian ellipse, for Bk. 1 Prop. 11 – if a Keplerian ellipse, then inverse-square – ceases to hold in *quam proxime* form. See my "From the Phenomenon of the Ellipse to an Inverse-Square Force: Why Not?," in *Reading Natural Philosophy: Essays in the History and Philosophy of Science and Mathematics to Honor Howard Stein on His 70th Birthday*, ed. D. Malament (La Salle: Open Court, 2001).

[25] Specifically, Newton remarks in the *Principia* that he was "still not certain whether that proportion of the inverse-square obtained exactly in a total force compounded of a number of forces, or only nearly so" until he had proved the propositions for spheres of uniform and spherically symmetric density (p. 811). In Section 7 of Book 2 he shows similar concern for whether the $\rho_f A_{\text{front}} v^2$ proportion holds exactly by deriving it from microphysical models of fluid action. One element of his concern in both instances is inexactitude in the force law introduced by appealing exclusively to macroscopic phenomena as evidence for it.

[26] As Newton knew, Descartes had called attention to the likelihood that the planetary trajectories are constantly changing over time, and hence the observations of them available in the 17th century were likely to be epochally parochial (*Principles of Philosophy*, trs. V. R. Miller and R. P. Miller (Dordrecht: D. Reidel Publishing, 1983), p. 98). The key parochialism that in fact emerged with Einstein's relativity theory is that gravitational fields in our planetary system are very weak.

[27] For a more extended discussion of the importance of the approximate-exact distinction in Newton's *Principia*, see my "The Methodology of the *Principia*," in *The Cambridge Companion to Newton*, eds. I. B. Cohen and G. E. Smith (Cambridge: Cambridge University Press, 2001).

[28] *Principia*, pp. 832–874.

[29] Specifically, Descartes had remarked in a letter of 13 November 1629 to Mersenne,

> As for the cause of the air resistance which you asked me about, in my view it is impossible to answer this question since it does not come under the heading of knowledge. (*The Philosophical Writings of Descartes*, vol. 3 (Correspondence), trs. John Cottingham, Robert Stoothoff, Dugald Murdoch, and Anthony Kenny (Cambridge: Cambridge University Press, 1991), p. 9.)

This letter was not included in Clerselier's 17th century edition of Descartes's correspondence, but was instead part of a collection la Hire put together, but then did not publish. Hence, Newton was most likely unaware of it.

[30] In *Two New Sciences*, Galileo had remarked:

> As to the perturbation arising from the resistance of the medium this is more considerable and does not, on account of its manifold forms, submit to fixed laws and exact description. . . . Of these *accidenti* of weight, of velocity, and also of shape, infinite in number, it is not possible to give any exact description; hence, in order to handle this matter in a scientific way, it is necessary to cut loose from these difficulties; and having discovered and demonstrated the theorems, in the case of no resistance, to use them and apply them with such limitations as experience will teach. (*Dialogues Concerning Two New Sciences*, trs. Henry Crew and Alfonso de Salvio (Buffalo: Prometheus Books, 1991), p. 252.)

Newton almost certainly had not read Galileo's book, but he had learned a good deal about it from secondary sources.

[31] Newton's thinking about the different mechanisms contributing to fluid resistance evolved over time. While working on the first edition of the *Principia* he came to think that only one mechanism besides the inertia of the fluid contributes to resistance forces, a mechanism that he referred to as the "tenacity" of the fluid. In the first edition of the *Principia*, however, the word '*tenacitas*' never occurs, and Newton speaks instead of a "defect of lubricity." The pair of mechanisms, inertia and tenacity, reappear with the first Latin edition of the *Opticks* in 1706 (p. 310), where Newton leaves open the question whether the resistance force arising from the tenacity of the fluid varies with v^0 or v^1. The three-mechanism view I have given in the text is first put forward in the third (1726) edition of the *Principia*, in a scholium added at the end of Book 2, Section 3. I have chosen to use it here for ease of exposition. For more details of the evolution of Newton's views on the mechanisms contributing to fluid resistance, see my "On the Origins of Book 2 of Newton's *Principia*," forthcoming.

[32] Continuing with the point made in the preceding note, Newton had concluded from his preliminary pendulum-decay experiments that the resistance force can be expressed to high accuracy by

$$f_{\text{resist}} = a_1 v + a_2 v^2$$

where the first term represents the effects of the tenacity and the second, the inertia, of the fluid. As described in my "On the Origins of Book 2 of Newton's *Principia*," Newton had realized from further pendulum-decay experiments that the results from these preliminary experiments were highly misleading. The preliminary experiments are never mentioned in any edition of the *Principia*. As discussed in Section 3 of the present paper, the results from the pendulum-decay experiments that were reported in the first, and subsequent, editions left Newton uncertain about whether one or two terms with velocity to a power less than 2 are needed and what the exponents of these terms should be. As we shall see in subsequent sections, Newton never resolved this issue. Of course, the logic of the problem of disaggregating the contributions made by different mechanisms remains essentially the same regardless of whether two or three distinct mechanisms contribute and regardless of the specific exponents.

[33] *Principia*, Bk. 2 Prop. 30, pp. 708–711. For a discussion of this and the other propositions on resisted pendulum motion in Section 6 of the *Principia*, see D. T. Whiteside's notes in Volume VI of *The Mathematical Papers of Isaac Newton* (Cambridge: Cambridge University Press, 1974), pp. 439–450.

[34] *Principia*, Bk. 2 Props. 30 and 31, pp. 708–712. For an analysis of the derivation of this result, see Whiteside's notes in *The Mathematical Papers of Isaac Newton*, vol. VI, pp. 446–449, or Michael Nauenberg, "Addendum to G. Smith's 'Fluid Resistance: Why Did Newton Change His Mind?'," in Dalitz and Nauenberg (ed.), *op. cit.*, p. 137f; and, for the use to which it is put, see my "The Newtonian Style in Book II of the *Principia*," in Buchwald and Cohen (ed.), *op. cit.*, pp. 259–262.

[35] Newton had used a 12 foot long pendulum in a preliminary pendulum-decay experiment, as described in I. Bernard Cohen's *Introduction to Newton's "Principia"* (Cambridge: Harvard University Press, 1978), pp. 101–103. These preliminary experiments and their relationship to those presented in the *Principia* are discussed in my "On the Origins of Book 2 of Newton's *Principia*," forthcoming.

[36] The quoted phrase is taken (as typical) from Christiaan Huygens, "De Vi Centrifuga," in *Oeuvres Complètes de Christiaan Huygens*, vol. 16 (The Hague: Martinus Nijhoff, 1929), p. 298.

[37] See "Newton on Fluid Resistance in the First Edition," cited in note 13, p. 312f.

[38] *Ibid.*, pp. 308–312.

[39] *Ibid.*, p. 313; this complementary measurement of resistance using a conical pendulum was dropped in the second and third editions.

[40] The fact that Newton proposed this solution is evidence that he (unlike Huygens) was unaware of Evangelista Torricelli's efforts on the efflux problem – as described in Cesare Maffioli's *Out of Galileo: The Science of Waters, 1628–1718* (Rotterdam: Erasmus Publishing, 1994), pp. 71–89 (and also pp. 403ff). I am grateful to Domenico Bertoloni Meli for first calling my attention to this book.
[41] Fatio claims in a letter, "I could scarcely free our friend Newton from this mistake, and that only after making the experiment with the help of a vessel which I took care to have provided." (*The Correspondence of Isaac Newton*, vol. 3, ed. H. W. Turnbull (Cambridge: Cambridge University Press, 1961), p. 169.)
[42] *Principia*, p. 735f.
[43] *Keynes MS*, 130.15, quoted in Richard S. Westfall, *Never At Rest: A Biography of Isaac Newton* (Cambridge: Cambridge University Press, 1980), p. 455.
[44] *Principia*, p. 752.
[45] See R. G. Lunnon, "Fluid Resistance to Moving Spheres," *Proceedings of the Royal Society of London*, Series A, 10: 302–326 (1926).
[46] Newton makes this intent clear in his introductory explanation of the vertical-fall experiments, *Principia*, p. 749.
[47] The two other principal candidates are Newton's highly contrived derivation of the rate of the precession of the equinoxes in the second and third editions (see Geoffrey Dobson's and Curtis Wilson's articles cited in note 6 above); and his lack of candor about the calculated mean precession of the lunar perigee falling a factor of 2 below the value from observation (see my "The Motion of the Lunar Apsis" in "A Guide to Newton's *Principia*" by I. Bernard Cohen in *Principia*, pp. 257–264).
[48] *Principia*, p. 741.
[49] *Principia*, p. 746.
[50] D'Alembert, *op. cit.*, p. xx.
[51] *Principia*, p. 759.
[52] The Reynolds number for a sphere moving with velocity v in a viscid medium is given by $\rho_f dv/\mu$, where d is its diameter and μ is the fluid viscosity.
[53] *Ibid.*, p. 759.
[54] *Principia*, p. 749.
[55] *Ibid.*
[56] The only remark in the text strongly suggesting that the data were to be taken as confirming the theory is made at the end of the discussion of the experiments in water, just before turning to those in air: "The theory therefore agrees with the phenomena of bodies falling in water; it remains for us to examine the phenomena of bodies falling in air." (p. 756) Even this remark, however, can be taken as pointing to nothing more than the extent to which the theoretical contribution made by the inertia of the fluid accounts for the total measured resistance.
[57] *Principia*, p. 761.
[58] Of course, Cartesians would still have faced the problem of explaining how the planets can be carried around in their orbits by vortices in an *inviscid* fluid.
[59] The claim that the inertia of the fluid is the dominant factor in resistance (at high Reynolds numbers) should not be confused with the claim that there is a separate, independent inertial mechanism contributing to resistance. The Reynolds number – for spheres, $\rho_f dv/\mu$, where d is the diameter and μ is the fluid viscosity – represents the ratio between inertial and viscous effects in a fluid regime. In real fluids, the viscosity is never zero, and hence there is always some interaction between inertial and viscous effects, an interaction that becomes especially significant near any solid boundary. For more details, see D. J. Tritton, *Physical Fluid Dynamics* (Oxford: Oxford University Press, 1988), pp. 100–105.
[60] In modern vector form, Euler's equations of motion for an incompressible, inviscid fluid in the absence of external body forces are given by

$$\rho(\vec{u} \cdot \nabla \vec{u}) = -\nabla p$$

for the steady state case, and

$$\rho\left(\frac{\partial\vec{u}}{\partial t}+\vec{u}\cdot\nabla\vec{u}\right)=-\nabla p$$

for the unsteady case. Supplementing these equations is an equation for continuity of mass. Euler first published these equations in 1755 after formulating them three years earlier in a tract that was not published until later. For historical details, see Truesdell, *Rational Fluid Mechanics, 1687–1765*, Editor's introduction to *Leonhardi Euleri Opera Omnia*, ser. 2, vol. 12 (Lausanne: Societatis Scientiarum Naturalium Helveticae, 1954), pp. LXII-C.

[61] The general equations of motion – i.e., the Navier–Stokes equations – for an incompressible viscid fluid in the absence of external body forces, in modern vector form, add another (viscous) term to Euler's equations:

$$\rho\left(\frac{\partial\vec{u}}{\partial t}+\vec{u}\cdot\nabla\vec{u}\right)=-\nabla p+\mu\nabla^2\vec{u}$$

Notice that these are second-order equations, requiring two boundary conditions at any solid surface; by contrast, Euler's equations are first-order, requiring (and allowing) only one boundary condition at any solid surface. That one boundary condition is, zero flow through the solid surface. In the case of the Navier–Stokes equations, the second boundary condition allows a specification of flow conditions tangential to the solid surface. The inability to specify a boundary condition tangential to a solid surface is one way of explaining why d'Alembert's paradox holds for a perfect fluid (i.e., one governed by Euler equations), but not for a fluid with any viscosity at all, no matter how small. For, this is why the flow governed by the Navier–Stokes equations does not converge in the limit as the viscosity approaches zero to the flow governed by the Euler equations. Ultimately, of course, this failure of convergence was what made the d'Alembert paradox a paradox.

The Navier–Stokes equations do not admit of closed form solution for virtually any interesting cases (Tritton, *op. cit.*), and even their numerical solution on current computers requires problematic simplifying assumptions in the vicinity of any solid surface (P. Bradshaw, T. Cebeci, and J. H. Whitelaw, *Engineering Calculation Methods for Turbulent Flow* (London: Academic Press, 1981)). For a history of the Navier–Stokes equations, see Olivier Darrigol, "Between Hydrodynamics and Elasticity Theory: The First Five Births of the Navier–Stokes Equations," *Archive for History of Exact Sciences* 56: 95–150 (2002).

[62] A philosophical point: even though the deduced laws describe the specific data to high approximation, they still may not be suitable for inductive generalization, for the formulation of the laws may incorporate accidental, parochial features. See Nelson Goodman, "The New Riddle of Induction," *Fact, Fiction, and Forecast* (Indianapolis: Bobbs-Merrill, 1973), pp. 59–83.

[63] Isaac Newton, "An Account of the Book entituled *Commercium Epistolicum Collini & aliorum*," *Philosophical Transactions of the Royal Society* 29: 173–224 (1715), p. 222; reprinted in A. Rupert Hall, *Philosophers at War: The quarrel between Newton and Leibniz* (Cambridge: Cambridge University Press, 1980), pp. 263–314.

[64] In fact, because the various celestial bodies are generally oblate spheroids, the working hypothesis that the gravitational fields around them do not vary with θ and φ is an example of one that is false, yet nevertheless fully useful; of course, the variation in question is usually negligible at celestial distances, though not, for example, in the case of the motion of our Moon.

[65] Ludwig Prandtl, "Über Flüssigkeitsbewegung bei sehr kleiner Reibung," *Proceedings of the Third International Mathematical Congress*, Heidelberg, 1904, pp. 484–491. An English translation of this paper is available under the title "On the motion of fluids with very little friction" in J. A. D. Ackroyd, B. P. Axcell, and A. I. Ruban, *Early Developments of Modern Aerodynamics* (Reston, VA: American Institute of Aeronautics and Astronautics, 2001),

pp. 77–84. For a discussion of the significance of this paper, see Hermann Schlichting, *Boundary Layer Theory*, 7th ed., tr. J. Kestin (New York: McGraw-Hill, 1987), pp. 1–4.
[66] See Richard P. Feynman, Robert B. Leighton, and Matthew Sands, *The Feynman Lectures on Physics* (Reading, MA: Addison-Wesley, 1964), Lect. 41, pp. 7–9.
[67] Quoted from Garrett Birkhoff, *Hydrodynamics: A Study in Logic, Fact, and Similitude*, rev. ed. (Princeton: Princeton University Press, 1960), p. 4. Birkhoff obtained the quote from a summary article covering a conference on very high speed flow, where M. J. Lighthill attributed it to Hinshelwood. See *Nature* 178: 343 (1956).
[68] I express my gratitude to Allan Franklin for proposing the topic of this paper.

XIANG CHEN

VISUAL PHOTOMETRY IN THE EARLY 19TH CENTURY: A "GOOD" SCIENCE WITH "WRONG" MEASUREMENTS

INTRODUCTION

Although the law of illumination, that is, the intensity of light on a surface being in inverse proportion to the square of the distance between the surface and the light source, had been suggested by Kepler in 1604, photometry as a science did not appear until the mid-18th century.[1] Pierre Bouguer's *Optical Treatise on the Gradation of Light* printed in 1760 was the first major publication in which the methodological principles and instruments for photometric measurements were systematically discussed. Despite strong skepticism from some critics who believed that the brightness of light was not a stable physical phenomenon, photometry developed rapidly in the last few decades of the 18th century. In the works of Pierre Bouguer, Johann Lambert, Court Rumford, William Herschel and John Leslie, a variety of photometric phenomena were investigated and measured, including the intensity of various natural lights, the intensity of reflected light from different media and the transparency of different materials.

By the beginning of the 19th century, there had emerged two rival approaches to photometric research: a visual approach and a physical approach. The visual approach was based on the belief that the eye was an ideal photometric instrument and should play an essential role in all photometric measurements, and many meticulous experimental protocols and measuring procedures were developed to ensure that the eye was in optimal conditions during photometric experiments. On the other hand, the physical approach was rooted in doubts about the reliability of the eye in comparing and judging the brightness of light. To reduce the impact of the eye, the physical approach employed a variety of instruments such as differential thermometers to compare and measure the intensity of light. Despite that they had been working hard in developing appropriate measuring methods and that some of their efforts were successful, neither the visual nor the physical approach in the early age of photometry could offer accurate photometric measurements. Most of the measuring results available in this period carried double-digit relative error, compared to currently accepted values offered either by contemporary experiments or by theoretical calculations.[2] But it seems that the scientific community still made a choice between these two immature approaches. By the mid-19th century, the visual approach had been accepted as the "good" photometry by the majority in the community. The victory of the visual approach raises some interesting questions. How could the visual approach be justified

J.Z. Buchwald and A. Franklin (eds.), Wrong for the Right Reasons, 161–183.

by its "wrong" results? Or, more generally, how could a "good" science result in "wrong" measurements?

To answer these questions, we need to examine the evaluation standards for photometric methods and photometric measurements in their historical contexts. In the following sections, I will review the development of photometric methods during the first half of the 19th century, and particularly examine two debates between the visual and the physical approaches. The historical analysis will show that whether a particular photometric measurement was perceived as wrong depended both on the acceptable range of relative error required by the practice, and on the system errors determined by the available instruments. Furthermore, the evaluations of photometric methods in the early 19th century did not rely solely on the degree of accuracy, and the practitioners in the field also used the scope of application and the level of efficiency as evaluation standards. The irony of a "good" science with "wrong" measurements will disappear when we put the meanings of "good science" and "correct measurement" in their historical contexts.

EARLY VISUAL PHOTOMETRY

Pierre Bouguer (1698–1758), Professor of Hydrography at Havre, laid the foundation of photometric measurements in the early 18th century. Bouguer was originally intrigued by the problem of determining the intensity of light from various celestial subjects, and in 1725 he succeeded in determining the intensity of moonlight by comparing its light to that of a candle. Many later believed that this was the birth of visual photometry (Walsh, 1958, p. 1). Bouguer's contribution to photometry was essential. Although all of his photometric measurements turned out to be inaccurate by today's standards, the methodological principles that he proposed were sound and dominated the practice in the next 200 years.

The most important methodological principle proposed by Bouguer was the condition of simultaneous comparison. One of the earliest attempts to determine the intensity of light was from Huygens, who in the 17th century had tried to compare the brightness of the sun and that of the star Sirius. Huygens used a long tube with an adjustable aperture at the far end as the instrument. To make the comparison, he first pointed the tube to the sun and observed a diminished sunlight by reducing the size of the aperture. He then waited until the evening and pointed the tube to the star. Through adjusting the size of the aperture, he obtained a starlight with the same brightness as that of the sunlight. The intensity ratio of the sun and that of the star was in inverse proportion to the area ratio of the two openings (Huygens, 1698). According to Bouguer, Huygens's measuring procedure was fundamentally wrong. The main problem was that the eye, an extremely delicate organ, would have different levels of sensibility over a day and could respond differently to the same level of intensity. Even worse, no one could remember the intensity of light even after a very small period of time, so that comparing the intensities of sunlight and starlight observed at very different times was extremely unreliable. Bouguer insisted that "we can only judge directly the strength of the two sensations when they affect us at the same instant" (Bouguer, 1760, p. 46). Thus, when the eye was used, photometric comparison must be made

simultaneously. The comparison of the intensities of sunlight and starlight, for example, could only be achieved in two steps – first to compare sunlight (or starlight) with an auxiliary light, and then to compare the auxiliary light with starlight (or sunlight).

Another aspect of the simultaneous comparison condition concerned the subjects to be compared. Before Bouguer, a variety of parameters had been used in the comparison of illuminations. Celsius in the early 18th century, for example, had used the degree of distinctness as the subject of comparison in his photometric measurements. He put two small objects at different distances and adjusted the intensity of light falling on each object until he saw both of them with the same degree of distinctness. The intensity ratio was then calculated according to the distance ratio.[3] According to Bouguer, Celsius's measuring procedure was also problematic, because the estimation of distinctness was highly subjective. People with different visual problems, such as long-sighted or short-sighted, could have different judgments regarding the distinctness of an object. To avoid arbitrary decisions, Bouguer suggested that brightness should be the only parameter used in photometric comparison. "As we are considering only the amount of light or its brightness, it does not matter whether the observer has long or short sight, good sight or bad. If the rays cross before having reached the retina or if they come together farther back, nevertheless they act on the back of the eye. There is nothing lost, and the total effect is always the same as regards the intensity of the impression" (Bouguer, 1760, p. 48).

The simultaneous comparison condition thus implicitly defined the role of the eye in photometric measurements. According to Bouguer, the eye should only be used as a null indicator to mark the equality of brightness of two luminous bodies, rather than being used to estimate the degree of illuminations directly. Using the eye as a null indicator reflected the fact that the eye was extremely sensitive and reliable in judging the equality of brightness. Bouguer designed an experiment to demonstrate this capacity of the eye. The apparatus of the experiment included a white screen, two identical candles, and a rod placed between the screen and the candles, casting two shadows on the screen (Figure 1). Bouguer set one of the candles at a distance of

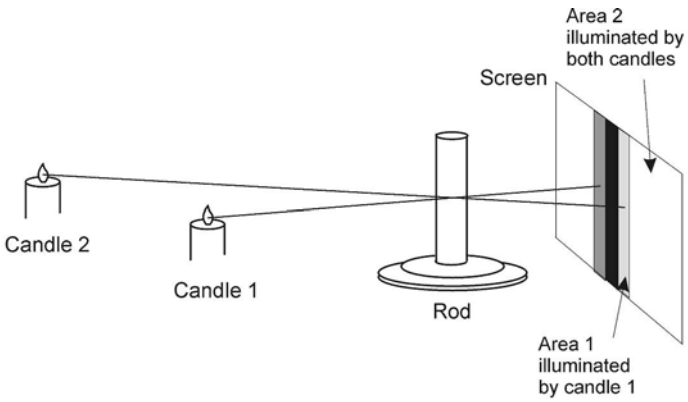

Figure 1. Bouguer's apparatus for determining the sensitivity of the eye.

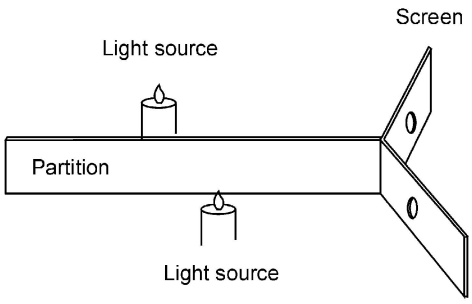

Figure 2. Bouguer's photometer.

one foot from the screen, and slowly increased the distance of the other candle until its shadow disappeared, that is, until the difference of brightness between Area 1 and Area 2 was invisible. After repeated experiments, he found that the shadow of the second candle always remained visible when its distance from the screen was less than 7 feet, but it disappeared when the distance was about 8 feet. These results suggested that two illuminated areas became visually indistinguishable when their brightness difference was smaller than $1/64$ ($1/8^2$), or about 1.5%. In other words, the eye was able to distinguish differences in brightness as small as 1.5%.[4]

Using the principle of simultaneous comparison, Bouguer constructed several photometers to compare the brightness of different light sources. One of these photometers was simply made of two pieces of cardboard, one of which served as a screen and the other as a partition, preventing the light from the sources from interfering with one another (Figure 2). Observations were made through two holes in the screen, which had the same size, about a quarter of an inch in diameter, and were covered by oiled paper. By moving one of the light sources to a greater or smaller distance form the screen, one could obtain equal brightness between the two holes. The intensity ratio between the sources could then be calculated in terms of their distances according to the inverse square law.

Armed these photometers, Bouguer conducted many photometric measurements. He measured the optical properties of various materials, including the reflective power of glass, metals, water and rough surfaces, as well as the transparency of glass, sea water and air. He also measured the intensity of light from many astronomical objects, such as the sun and the sky. Although Bouguer's measurements were built upon a sound methodological principle, his measuring results were not satisfactory. Most of his measurements of reflective power, for example, carried double-digit relative error. He reported that the reflective power of plate glass at 0° was 2.5%, about 40% lower than the value (0.04193) given by the wave theory. His measurements of the reflective power of water, 1.8% at 0° and 72.1% at 89.5°, were 10–20% lower than the values given by the same theory. And his measurement of the reflective power of mercury, 63.7% at 69°, was also more than 10% lower than the value given by the electromagnetic theory (Bouguer, 1760, pp. 93–99).[5] Some of Bouguer's measurement error may have been

caused by negligence. For example, when he measured the reflective power of glass, Bouguer did not consider the effect of tarnished surfaces, which could significantly reduce the intensity of the reflected light (Rayleigh, 1886). Some other error might result from the limits of experimental conditions. In the measurements of water and mercury, for example, Bouguer found that the surface of liquid was not perfectly plane but convex, especially when the containing vessel was not sufficiently large, and that the convexity weakened the intensity of the reflected light. All of these offer a possible explanation of why most of Bouguer's measurements of reflective power were lower than the currently accepted values.

Photometry developed rapidly in the second half of the 18th century on the foundation laid out by Bouguer. The first to appear was that of the French mathematician, Johann Lambert (1728–1777). In a book published in 1760, Lambert outlined several fundamental laws of photometry, including the cosine law of illumination (the illumination of a surface being proportional to the cosine of the incident angle), the cosine law of emission (the radiation from a single-unit surface being proportional to the cosine of the emission angle), and the law of addition of illumination (the illumination produced by multiple light sources being equal to the sum of the illumination produced by each source) (Lambert, 1760). Lambert also designed a photometer, but it is unclear if he actually constructed it (Walsh, 1958, p. 12). The major improvement of photometers in the late 18th century was achieved in the hands of Court Rumford (1753–1814), the American philanthropist and statesman.

Rumford's purpose in studying photometry was practical: to search for an economical method of lighting. To compare the intensity of light from different sources, he constructed a shadow photometer. The key component of this instrument was a photometric head, where two identical rods were placed a few inches from a screen (Figure 3). The two rods cast shadows on the screen, corresponding to two light sources. A wing was attached to each rod so that the observer could bring the two shadows in contact by rotating the rods around their axes. An aperture with a square opening was placed immediately in front of the screen to define the boundaries of a matching field. Unlike Bouguer's photometer where the observer compared the

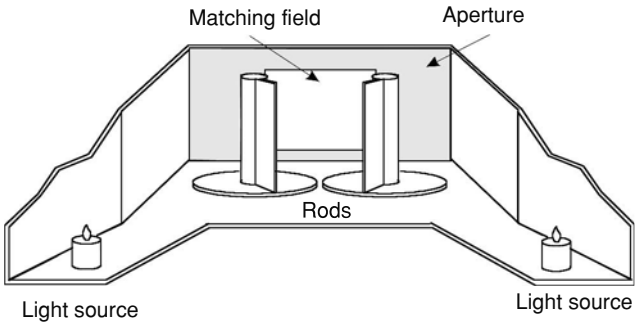

Figure 3. Rumford's photometric head.

surfaces directly illuminated by the light sources, Rumford's photometer compared the shadows of the rods. This new design significantly reduced the intensity of light in the matching field and thereby protected the eye from developing fatigue (Rumford, 1794, pp. 71–78).

Rumford also realized that a constant observation was necessary in order to compare brightness reliably. Questions frequently appeared as to whether the matching of brightness was due to adjustment of the light source or to fatigue of the eye. To avoid confusion, the observer must not turn his eyes away from the matching field while adjusting the distance of the light sources. To make constant observations possible, Rumford introduced a "remote control" device. He placed each light source upon a sliding carriage drawn along a straight groove by means of a cord, and he passed the cord over pulleys at the ends of the groove. By pulling the cord, he could adjust the distances of the light sources without turning his eyes away from the matching field (Rumford, 1794, pp. 80–81).

Using the shadow photometer, Rumford compared the intensities of various lamps and candles. He also measured the optical properties of glass, both the reflective power and the transparency. Because he did not use a standard unit of illumination (such a unit did not appear until the end of the 19th century), it is difficult to determine the accuracy of many Rumford's measurements. But evidently, with the help of the more sophisticated shadow photometer, Rumford was able to present more accurate results in some of his measurements. For example, he reported that the transparency of a very thin plate of window glass was 87.37%, which is only 5% lower than theoretical value (91.74%).[6] Considering the threshold of identifying brightness difference set by the eye (1.5%) and the uncertainty in distance measurements, which would be magnified due to the inverse square relation between illumination and distance (a 1% error in distance measurement would become a 2% error in illumination measurement), Rumford's measurements may very well have reached the limit allowed by the instrument.

EARLY PHYSICAL PHOTOMETRY

Around the end of the 18th century when visual photometry was still in its cradle, a competing photometric method emerged. John Leslie (1760–1832), Professor of Natural Philosophy at Edinburgh, invented a novel photometer in 1797.[7] Leslie's photometer was in fact a differential thermometer, with one of its balls coated with black ink (Figure 4). The thermometer was enclosed within a glass case to prevent the irregular effect of winds. When the thermometer was exposed to the light to be measured, the blackened ball absorbed the incident light, while the other one transmitted it freely. According to Leslie, the absorbed light would assume a latent form and act as heat. "Light, in proportion to its absorption, causes heat, whether uniting with bodies it really constitutes the matter of heat, or only excites heat in the act of combination," he reasoned (Leslie, 1800, p. 466). The heat caused by the incident light would continue to accumulate, until it was offset by the emission of heat

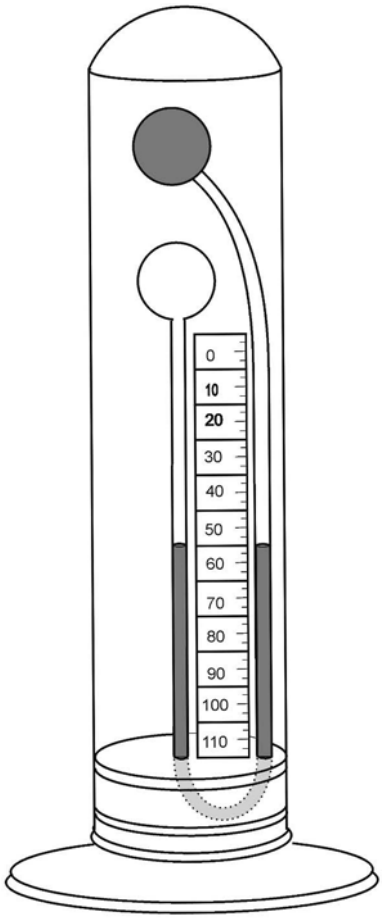

Figure 4. Leslie's thermometric photometer.

to the surrounding air. At the point of equilibrium, the accumulation of heat derived from the action of the incident light was exactly the same as the amount of heat radiated in the cooling process. "The depression of the liquor, therefore, will measure the momentary afflux of light" (*Ibid.*). In other words, this differential thermometer was able to measure the momentary impression of light, that is, its actual intensity.

Armed with this thermometric photometer, Leslie conducted a series of experiments to determine the intensities of various luminous sources, including sunlight, moonlight, skylight and many artificial lights, by using an arbitrary unit that measured the depression of the liquor of the thermometer. He found that the direct impression of the sun at noon in Edinburgh during the summer solstice amounted to 90°, and the impression of the skylight, 30°. Leslie thus concluded that the intensity of the sunlight was about three times more powerful than that of the light reflected from the

sky. Following the same procedure, he found that the intensity of the sunlight was 150,000 times stronger than that of the moonlight, and 12,000 times of that of a wax candle (Leslie, 1813, p. 46; Leslie, 1824, p. 397). Using the thermometer to measure the temperatures caused by the lights passed through various transparent materials, Leslie determined their transparencies. He reported that the transparency of oiled paper was 80% and fine white paper, 49% (Leslie, 1813, p. 46). Leslie even tried to measure the relative intensities of different colors in the prismatic spectrum, a task that had perplexed Bouguer because the eye was inefficient in matching lights with different colors. Dividing the spectrum into four parts that coincided with the blue, the green, the yellow and the red, Leslie found that their relative intensities were in the ratios of 1, 4, 9, and 16 (Leslie, 1801, p. 347).[8]

Leslie was confident of the effectiveness of his thermometric photometer, because it was able to "perform with the utmost facility all those ingenious experiments which have exercised the sagacity of Bouguer and Lambert" (Leslie, 1800, p. 467). However, Leslie's instrument did not receive a warm response from the photometric community. According Leslie, "owing to a combination of circumstances, this elegant instrument [the thermometric photometer] has only been partially and reluctantly admitted; and the philosophic world has still to discharge an act of justice, by receiving it into the favor and distinction which it so well deserves" (Leslie, 1838, p. 313). The main concern of the community was that Leslie's instrument was not really a photometer but merely a thermometer, because it measured heat instead of light. Responding to the criticisms, Leslie first insisted that measuring the intensity of light indirectly through its heat effect was reliable. "What does the thermometer itself indicate, except *expansion*? As heat is measured by the expansion it occasions, so light is determined by the intensity of the heat which, in every supposition, invariably accompanies it" (Leslie, 1838, p. 313; original emphasis). Furthermore, Leslie claimed that measuring light through its heat effect was, in fact, the only legitimate way. According to Leslie, a measuring unit was needed for accurate photometric measurements. "What other mode, after all, could be imagined for detecting the presence of light? How can an unknown quantity be expounded, but in terms of one already known?", he asked (*Ibid.*). Using a thermometer, the intensity of light could be measured accurately in terms of degrees of temperature, a known parameter. On the contrary, the visual approach was problematic and unreliable simply because it attempted to determine the intensity of light without using any known parameter as the measuring unit.

Leslie did have many reasons to be confident of his thermometric approach. In terms of scope of application, his thermometric photometer was able to measure phenomena that neither Bouguer's nor Rumford's visual photometer was able to handle, such as direct sunlight (too intense for the visual approach) and colored light (too complicated for the eye to handle). In terms of application procedure, Leslie's approach was relatively simple. His photometer could produce quantitative readings without demanding considerable skill and complex preparation, while all visual photometers required extreme precautions in preparation and operation. Even in terms of accuracy, Leslie's photometer also demonstrated its advantage. Although the accuracy of most of Leslie's measurements cannot be determined due to lacking of a standard unit, it is

evident that his measurement of the transparency of sea water was indeed better than the one offered by Bouguer.

Bouguer had used the visual method to determine the transparency of sea water. He constructed a 115-inch-long canal with wood boards, and placed two pieces of glass at its ends. Using two light sources, a torch and a candle, he let the light of the former pass through the canal and the light of the latter fall directly to the matching field of a photometer. He found that, when the canal was empty, equal brightness appeared in the matching field when the candle was 9 feet away, but when the canal was filled with sea water, the candle must be moved to 16 feet away (Bouguer, 1760, pp. 55–57). It followed that about $(9/16)^2$, or 31%, of the incident light passes through 115 inches of sea water, a measurement that is only one third of the theoretical value (0.9712).[9] In Leslie's measurement, he first used the thermometric photometer to determine the intensity of light at the surface of the sea and, by submerging the photometer in the sea, he then determined the intensity of the transmitted light. He reported that, when the photometer was 15 feet below the surface, the reading of the thermometer was only a half of that at the surface of the sea, that is, 50% of the incident light can pass through 15 feet of sea water (Leslie, 1838, p. 491). Leslie's measurement was still too low, about 40% from the theoretical value (0.936), but still an improvement compared to Bouguer's.[10]

Despite these advantages, the photometric community did not accept Leslie's thermometric approach. The first comparison between the visual and the physical approaches occurred in the mid-1820s when photometry was called on to perform measurements for the gas industry. The practical questions by the times were how to determine the illuminating powers of different gas burners and how to compare the illuminating efficiencies of coal gas and oil gas. Robert Christison, Professor of Medical Jurisprudence at Edinburgh, and Edward Turner, Lecturer of Chemistry at Edinburgh, conducted a series of experiments in the mid-1820s to answer these questions.

When Christison and Turner started their investigations, they did not have a preference for either the visual or the thermometric method. They acquired two photometers, a thermometric one built by Leslie himself and a shadow photometer identical to the one designed by Rumford. But very soon they realized that Leslie's thermometric photometer was unfitted for the tasks.

The first defect of Leslie's photometer, Christison and Turner reported, was its low sensibility. To compare the illuminating power of gases, they sometimes needed to measure light with low intensity, about a quarter of that of a candle. Leslie's photometer was not able to perform in these cases. Furthermore, Leslie's photometer was inefficient – it took about 40 minutes for it to reach the equilibrium point and then to reset. Worst of all, Christison and Turner found that Leslie's photometer was unable to distinguish the impact of light from that of nonluminous heat. To prove this, Christison and Turner designed a simple experiment. They set up two light sources, a burning chamber fire that was hot but emitted little light and an Angand oil lamp. They estimated that the illuminating power of the lamp was about 16 times of that of the fire, according to the distances at which an object was distinct while illuminated by these

two sources. They placed Leslie's photometer 16 inches from the chamber fire and it yielded a reading of 25. They then put the photometer 6.5 inches from the lamp and the reading was 3. Thus, according to Leslie's photometer, the illuminating power of the fire was about 40 times of that of the lamp, an obviously faulty measurement with respect to visible light. Christison and Turner reasoned that Leslie's photometer must have inflated the illuminating power of the fire by confusing light and nonluminous heat (Christison and Turner, 1825, pp. 4–6). Because of these problems, Christison and Turner declared that Leslie's photometer was unfitted for their tasks, and only used the shadow photometer to measure the illuminating power of gases and the efficiency of burners.

POTTER'S VISUAL PHOTOMETER

Another debate between the visual and the physical approaches occurred in the 1830s when photometric measurements were introduced to physical optics. One of the major players in this debate was Richard Potter (1799–1868), Professor of Natural Philosophy and Astronomy at University College, London. In his earlier years when he was still an amateur scientist, Potter had built a reflecting telescope. The need for evaluating the telescope, specifically the reflective power of its mirrors, triggered Potter's photometric research. To measure the reflective power of mirrors at various angles of reflection, Potter constructed a reflective photometer in 1830 (Figure 5).

The main components of this photometer were an upright screen with an aperture and a horizontal board divided by a blackened partition. Two identical lamps were used, each of which was put on the end of a slide and placed on either side of the partition. To determine the reflective power at various angles, Potter added a couple special devices

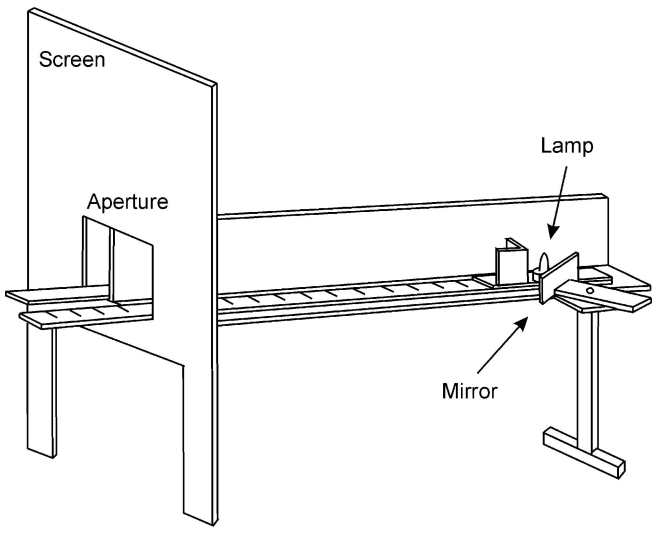

Figure 5. Potter's reflecting photometer.

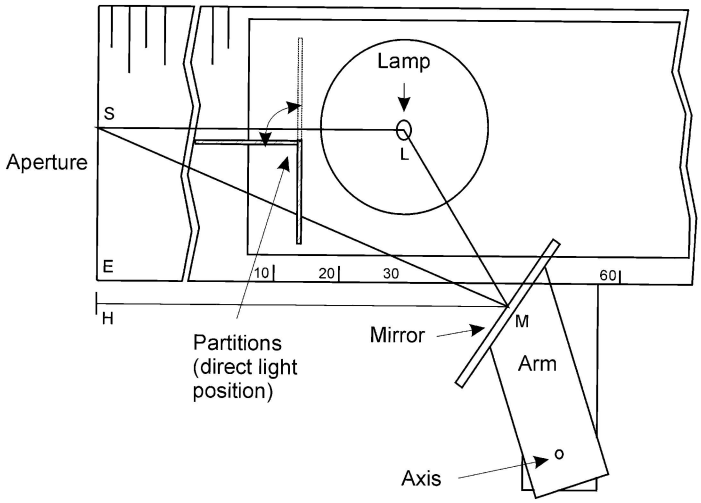

Figure 6. Potter's reflecting photometer (details).

to the photometer. He fixed the mirror to be measured to an arm, which could be turned around an axis attached to the right-hand slide (Figure 6). To intercept alternately the direct and reflected light, Potter installed two upright partitions perpendicular to each other. When the partitions were in the direct light position as shown in Figure 6, they intercepted the reflected light; when the partitions were turned 90° clockwise to the reflected light position, they stopped the direct light.

The major difficulty that Potter experienced in the experiments arose from "the fatigue of the eye experienced by looking long and intently at bright objects surrounded by darkness, which prevents it after some time judging accurately of very small differences [in brightness]" (Potter, 1830, p. 279). To reduce the fatigue of the eye, Potter covered the aperture with semi-translucent paper, which reduced the contrast between the light sources and the background. Potter also invented several "remote-control" devices, which allowed him to conduct the experiments without exposing himself to the direct light from the lamps. He put the lamps on moveable slides, and because the ends of these slides extended over the screen, he could adjust the distances to the lamps by simply pulling or pushing the slides while staying behind the screen. He marked the right-hand slide with divisions, in 0.25 inch intervals so that he could determine the distance between the lamp and the screen by simply reading off the divisions. By attaching strings to the corners of the perpendicular partitions, he could turn them in either direction without leaving his seat behind the screen.

Potter first measured the reflective power of several metallic mirrors, one composed of cast steel and the rest of tin–copper alloy. To begin with, Potter put the right-hand lamp and the mirror in preset positions and turned the perpendicular partitions to stop the reflected light. He then made the first brightness match, adjusting the left-hand lamp until equal brightness appeared in the aperture, and measuring the distance between the right-hand lamp and the screen (the distance of the direct light). Next,

he turned the partitions to stop the direct light, and made the second brightness match by pulling the right-hand slide together with the lamp and the mirror closer to the screen until equal brightness appeared in the aperture. He again measured the distance between the right-hand lamp and the screen (the distance of the reflected light). Finally, with the distances of the direct and the reflected light, he calculated the reflective power according to the inverse square law.

Among these operations, the measurements of distances deserve our attention. Potter's measurement of the distance of the direct light was straightforward. He obtained this parameter by simply reading off the divisions on the slide. But Potter's method of measuring the distance of the reflected light was peculiar. This parameter is the sum of the distance from the lamp to the mirror (LM in Figure 6) and the distance from the mirror to the center of the aperture (MS). The value of LM was available before the experiment from the preset positions of the lamp and the mirror, but the value of MS was not because, after the second brightness matching, the reflected light no longer fell into the center of the aperture. Potter made it clear that he did not actually measure MS. "It will be seen that the divisions commencing only at the thicker piece of wood, the distance of the lamp in the direct measurements, and the sum of the distances of the lamp to the mirror, and the mirror to the commencement of the divisions, must be added afterwards in the reflected ones," he said (Potter, 1830, p. 286). In other words, Potter made an approximation by substituting for MS the horizontal span between the mirror and the screen (MH), which was available by reading off the slide. The reason for making this approximation was probably to protect the eye. If Potter measured MS directly, he would have exposed himself to the direct light from the lamps and quickly developed eye fatigue. In hindsight, we know that this approximation had little notable effect on the measurements of metallic mirrors. Because metals have relatively high reflective power, the right-hand lamp in Potter's setting was still quite far away from the screen after the second brightness matching, usually more than 30 inches. Potter's approximation of distance caused only about 0.1% deviation in the final measurements.

Although his photometer was not in any way sophisticated, Potter's metallic measurements were surprisingly accurate. He reported that the reflective power of his alloy mirror (69% copper and 31% tin) was 67.5% at 10°, 66.05% at 30°, 65.07% at 50°, 64.91% at 60°, and 65.16% at 70°.[11] The discrepancies between Potter's measurements and the calculated values derived from the electromagnetic theory are very small, most of which are less than 5%.[12]

Potter's measurements immediately drew the attention of many in the optical community. David Brewster first heard of Potter's measurements in 1830, and immediately invited Potter to publish the results in *Edinburgh Journal of Science*. Apparently, Brewster believed that Potter's measurements were useful for constructing reflecting telescopes. In his *Treatise on Optics* printed in 1831, Brewster cited Potter's results in the section on reflecting telescopes, and proposed to use an achromatic prism to replace the plane metallic mirror in the traditional reflecting telescopes (Brewster, 1831).

Potter soon turned his attention to the reflective power of glass. In his glass experiments, he used the same photometer and followed essentially the same procedures

as those adopted in the metallic experiments. In the glass experiments, however, he had to place the right-hand lamp very close to the screen during the second brightness matching because of the low reflective power of glass. A significant amount of light scattered by the parts surrounding the lamp reached the aperture and inflated the measurements. Thus, Potter added a new procedure to estimate the amount of the scattered light and then subtract it from the gross readings including both the reflected and the scattered light. He started with a setting in which light reflected by the glass mirror and scattered by the surrounding parts all reached the aperture. He attached a roughly ground glass plate in front of the left-hand lamp, and adjusted the luminous area of the plate (by covering it with black paper) until equal brightness appeared in the aperture. Next, he removed the glass mirror from the photometer so that only scattered light reached the aperture, and reduced the luminous area of the glass plate in the left-hand side until equal brightness appeared. Finally, he used the ratios between the two luminous areas to determine the amount of the scattered light.

In 1831, Potter published his measurements of the reflective power of plate, crown and flint glass at various reflection angles, from 10 to 80°. He reported that the reflective power of plate glass was 3.66% at 10°, 4.09% at 30°, 5.57% at 50°, and 14.06% at 70°, and for crown glass, the numbers were 3.66, 4.17, 5.25, and 13.7%, respectively. The accuracy of these measurements was much lower than that of his metallic measurements. According to the wave theory of light, the values of the reflective power of plate glass, for example, should be 4.23% at 10°, 4.37% at 30°, 6.02% at 50°, and 17.41% at 70°, and for crown glass, 4.31, 4.46, 6.12, and 17.55%, respectively. In most cases, the discrepancies between Potter's measurements and the theoretical values are more than 10%.[13]

In his glass experiments, Potter used the same instrument and followed virtually the same procedures that had produced relatively accurate results in his metallic experiments, but why the accuracy of his measurements was so poor? A possible answer could be the approximation of the reflected distance that he made in the measuring process. In the metallic experiments, the impact of the approximation was negligible, but its consequences became significant in the glass experiments. Because the reflective power of glass was low at small reflection angles, Potter had to pull the glass mirror to be measured very close to the aperture in the second brightness matching. When the reflection angle was 10°, for example, the glass mirror was placed only about 6 inches away from the aperture. In this setting, the approximation of the reflected distance was about 2% lower than the true value, which would generate a 4% error after the calculation according to the inverse square law.

FORBES' THERMOMETRIC PHOTOMETER

Potter's photometric measurements received many criticisms. The first response occurred in 1834 when Humphrey Lloyd presented his "Report on Physical Optics" to the British Association. In the report, Lloyd briefly mentioned Potter's photometric measurements and cast doubts on their accuracy. Without replicating Potter's

experiments, Lloyd did not have solid evidence, but he raised reasonable doubt by questioning the reliability of the eye at matching brightness, which was the crucial procedure of the visual approach (Lloyd, 1834, pp. 74–75).

In a paper presented to the 1838 meeting of the British Association, Baden Powell picked up the issue raised by Lloyd and continued questioning the accuracy of Potter's photometric measurements. Powell used a "thought experiment" to reveal the problems of the visual approach. He asked the audience to imagine the result of a simple experiment in which the light from a candle first fell onto a screen directly. Then a thin and clear glass plate was inserted between the candle and the screen. Because reflection took place at both surfaces of the plate, more than one half of the incident light was reflected. If the eye was reliable, Powell reasoned, we should have seen a near two-to-one difference caused by the glass plate. But Powell noted that, in our daily experience, we did not perceive such a dramatic difference. Thus, he concluded that, because the eye could not accurately judge the intensity of light, photometric measurements using the visual method could not be unreliable (Powell, 1838, p. 7). But apparently Powell did not fully comprehend the procedures of the visual approach. In the "thought experiment," he "compared" the brightness consecutively – he first "observed" the illumination of the direct light and then the illumination after the reflection. This procedure violated an essential requirement of the visual method, namely that illuminations must be compared simultaneously.[14]

The major challenge to Potter came from James Forbes (1809–1868), Professor of Natural Philosophy at the University of Edinburgh. In a paper presented to the Royal Society of Edinburgh in 1838, Forbes questioned the reliability of Potter's photometric measurements with experimental evidence. Again, he did not replicate Potter's experiments. Instead, Forbes built his criticisms on experiments in which he used a thermometric photometer to measure the reflection of heat, assuming that the laws of reflection for heat and those for light, if not identical, would at least be analogous. Forbes's photometer was in principle similar to the one designed by Leslie, but Forbes employed an electric thermometer instead of a differential liquor thermometer. Forbes' photometer consisted of a thermoelectric pile and a galvanometer (Figure 7). The pile contained 30 pairs of bismuth–antimony bars that generated electricity when they were heated. The galvanometer consisted of a magnetic needle hung over a flattened coil of silver-wire, and it measured the electric current in terms of the angular deviation of the needle. The extent of the angular deviation was read off in reference to the attached divided circle. With the help of a small telescope that focused upon the divided circle, Forbes was able to observe angular deviations of the needle as small as 6 arc-minutes, which amounted to a sensitivity of 0.005 centigrade degrees (Forbes, 1835, pp. 134–140).[15]

Forbes used this thermometric photometer in 1837 to measure the intensity of heat reflected by glass and found that about 8% of the heat was reflected at 55°, a result close to the prediction given by Fresnel's formula (7%), assuming that Fresnel's formula could be applied to the reflection of heat. But Forbes soon realized that his measurement was invalid because he had not excluded the reflection from the second surface of the glass. Forbes improved his experiment in 1838, in which he used wedges

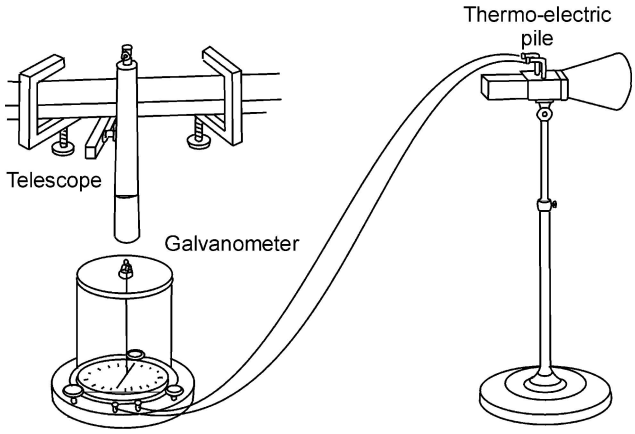

Figure 7. Forbes's thermometric photometer.

of plate glass to exclude the reflection from the second surface. He also constructed square tubes to guide the heat rays and to reduce the impact of scattered heat from the background. Using the thermometric photometer to measure the intensity of the source and that of the reflected heat directly, Forbes determined the reflective power of glass. He reported that the reflective power of plate glass was 4% at 10°, 5.1% at 30°, 7.6% at 50°, and 18.5% at 70° (Forbes, 1851). Except for the one at 10°, all of these measurements were significantly higher than Fresnel's predictions. Forbes could not say that his measuring results verified Fresnel's formula, but he compared his findings with Potter's measurements and claimed that Potter must have underestimated the reflective power of glass.

Forbes also measured the intensity of heat reflected by metallic mirrors at various angles and compared his thermal measurements with Potter's visual ones, again under the assumption that reflections of light and heat were analogous. Forbes found that his measurements verified Potter's observations that metallic reflection was less intense when the angle of reflection increased. However, Forbes also reported that the amounts of heat reflected from metallic surfaces were significantly higher than those reported by Potter. "The quantity of heat reflected by the metals is so much greater than Mr. Potter's estimate for light, as to lead me to suspect that his photometric ratios are all too small, which would nearly account for their deviation from Fresnel's law," he claimed (Forbes, 1839, p. 480).

To explain the discrepancies between his measurements and Fresnel's predictions, Forbes blamed the impact of scattered heat from the background. Because scattered heat was distributed unevenly in the background, the directed heat rays from the source and the reflected heat rays from the glass could have mixed with different levels of scattered heat once they took different paths. To control the scattered heat, Forbes designed a new experiment, in which he transmitted the direct and the reflected heat rays along the same path.

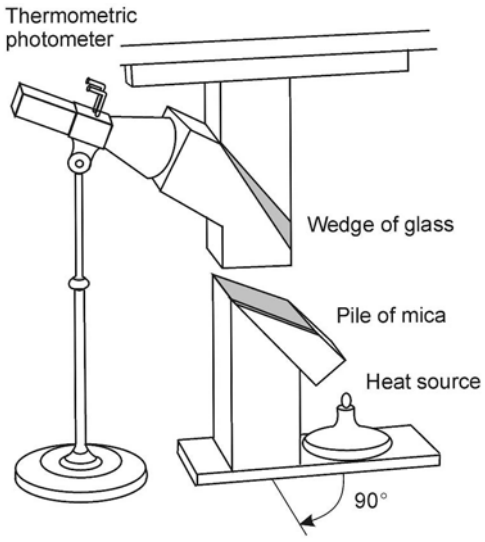

Figure 8. Forbes's proposed experiment.

The key to Forbes's proposed design was measuring the intensity of polarized heat by reflection. Partially polarized heat, or, more precisely, elliptically polarized heat, could be mathematically decomposed into two components with their planes of polarization perpendicular to one another. The two fractions in Fresnel's formula corresponded to the intensities of these two components. Fresnel's formula could then be tested by measuring the difference in intensity of polarized heat between the two components after reflection. Forbes proposed the following experiment. A beam of heat was first passed through a pile of mica sheets, which rendered the heat polarized by successive refraction. The heat rays then reached a wedge of plate glass, which reflected the incident rays to the photometer (Figure 8). According to Fresnel's formula, the intensity of the reflected heat should be:

$$I_1 = \frac{m}{2}\frac{\sin^2(\theta - \theta')}{\sin^2(\theta + \theta')} + \frac{n}{2}\frac{\tan^2(\theta - \theta')}{\tan^2(\theta + \theta')}$$

Here, m and n were the relative intensities of the two perpendicular components in the polarized incident rays (m was the one with the plane of polarization parallel to the plane of reflection). In his previous studies, Forbes had determined that the ratio of m to n was 100 to 27 in the polarized rays that passed through the mica pile (Forbes, 1838, p. 551).[16] After the intensity of the reflected heat was recorded, Forbes rotated the pile of mica 90° and made a new measurement. Because turning the mica pile did not alter the path of the heat rays, the impact of scattered heat was effectively controlled. Now, according to Fresnel's formula, the intensity of the reflected heat

should become:

$$I_2 = \frac{n}{2}\frac{\sin^2(\theta-\theta')}{\sin^2(\theta+\theta')} + \frac{m}{2}\frac{\tan^2(\theta-\theta')}{\tan^2(\theta+\theta')}$$

The difference between these two measurements was:

$$I_1 - I_2 = \left(\frac{m-n}{2}\right)\left(\frac{\sin^2(\theta-\theta')}{\sin^2(\theta+\theta')} - \frac{\tan^2(\theta-\theta')}{\tan^2(\theta+\theta')}\right)$$

Since the values of m and n were already known, Fresnel's formula could then be tested by comparing the difference between the two fractions with the difference between the two measurements.

Forbes' design was beautiful, but he could not carry out the experiment. The obstacle was the intensity level of the reflected heat, which was too weak to be measured after both refraction and reflection. "I fear we must wait for yet more delicate instruments to measure it," he conceded (Forbes, 1839, p. 480). Nevertheless Forbes insisted that his approach was better than Potter's visual method. Although his photometer only measured the reflection of heat, and his verification of Fresnel's formula could only be analogical, he believed that his approach was reliable; on the contrary, "photometric methods are so very imperfect as I still consider them to be, however dexterously employed" (Forbes, 1840, p. 103). According to Forbes, the reliability of his thermal/physical approach came from the measuring procedure, which converted the thermal effect to angular deviation and thus reduced the dependence upon the eye to a minimum. In contract, although Potter's visual photometer measured the reflection of light directly, it relied upon the eye to estimate the intensity of light and thus was in essence unreliable no matter how it was carefully operated.

These criticisms prompted a quick response from Potter. In 1840, he published a paper in *Philosophical Magazine* defending his photometric research. Potter apparently did not understand why his critics questioned the reliability of the eye, and did not offer any argument nor evidence to justify the extensive use of the eye in his photometer. He instead accused his critics, particularly Lloyd and Powell, of ignorance. "As Lloyd and Powell did not think it necessary to make themselves acquainted with the subject they undertook to discuss," he claimed, "their observations do not call for any further notice in this place" (Potter, 1840, pp. 17–18).

Potter devoted most of his paper to answering Forbes' criticisms. He first questioned the reliability of Forbes' measurements of reflected heat. Without experience in dealing with heat phenomena, nor the necessary skills of operating the thermometric photometer, Potter was unable to replicate Forbes' experiments and could only play with rhetoric. Because Forbes admitted that he had experienced many "unforseen difficulties" in his experiments, Potter seized this chance and insisted that because of these "unforseen difficulties" Forbes's methods "are not likely to furnish results accurate enough for testing important laws of nature" (Potter, 1840, p. 19). Responding to Forbes's suspicion that his photometric measurements were all too

small, Potter offered some empirical evidence by citing the work of Michael Faraday. He gave details of Faraday's photometric measurements presented in the 1830 Bakerian Lecture on the manufacture of optical glass, in which Faraday adopted the visual method to measure the reflective power of plate, crown and flint glass at 45°. Faraday's measurements were also at odds with Fresnel's formula, and more importantly, Faraday's measurements were even smaller than Potter's. For example, Faraday reported that the reflective power of his No. 6 crown glass at 45° was 4.52%, much smaller than the prediction from Fresnel's formula (5.366%). By pointing out the consistency between Faraday's and his own measurements, Potter claimed that the discrepancies between Fresnel's predictions and photometric measurements were substantial. Furthermore, Potter noted that, in effect, Faraday's photometric measurements could be used as *experimenta crucis* to test the wave theory, because "in high refracting bodies the discordance of Fresnel's formula with experiments is palpable, for it gives results frequently one-half more, to twice as much as experiment" (Potter, 1840, p. 20).

CONCLUSION

In the two debates between the rival photometric methods, the visual approach demonstrated its advantages. Neither Leslie's nor Forbes' thermometric photometer was adopted by the practitioners. By the mid-19th century, the visual approach had won the competition and become dominant in photometry. There were several factors responsible for the victory of the visual approach.

A significant technical improvement in the design of visual photometers emerged in 1843 when Robert Bunsen, the German chemist, introduced a grease-spot photometer. The main component of this photometer was a piece of white paper with a grease-spot at the center, placed in between the two light sources to be compared. When an equal quantity of light fell into each side of the paper, the grease-spot would be equally bright as the other part of the paper and thus become invisible. This grease-spot photometer was soon accepted widely, but the reason was not that it could in any way significantly improve the accuracy of photometric measurements. Bouguer had demonstrated that a shadow photometer could reach an accuracy level of 1.5%, and the Bunsen photometer could not make any substantially improvement. Breakthrough in accuracy did not occur until the end of the century when the Lummer–Brodhun photometer became available.[17] What made the Bunsen photometer popular was its simplicity in both construction and operation. Compared to thermometric photometers, a piece of paper with a grease-spot was strikingly simple and inexpensive. Furthermore, thermometric photometers required complicated procedures to calibrate measuring results and to translate thermal readings to photometric measurements, but the measuring process of the Bunsen photometer was plain and straightforward. Many practitioners preferred the Bunsen photometer simply because "it speaks more directly to the senses" (Wright, 1850, p. 326). Compared to the Bunsen photometer, thermometric photometers no longer had any competitive edge.

The significance of the technical breakthrough was further amplified by certain contextual factors. Around the mid-19th century, the major photometric practitioners in Britain were gas examiners from the lighting industry. The practical needs for measuring the illuminating power of gases and the performance of burners demanded efficiency and portability. Compared to thermometric photometers that required a long responding time, Bunsen's grease-spot photometer was clearly a winner. To gas examiners, accuracy was merely secondary – a measurement accurate to 5% "would do exceedingly well" (Wright, 1850, p. 326). In stellar photometry, another major application field of photometry, the key concern was to categorize more stars, many of which were very dim. Stellar measurements thus demanded high sensitivity, and thermometric photometers, poor at measuring low illumination, were inappropriate.

Around the mid-19th century, the visual approach became dominant. Rumford's and Bunsen's photometers were the standard equipments for gas examiners. In the eye of William Dibdin, the superintending gas examiner to the London County Council, the Bunsen photometer was simply the best. "Although confronted with innumerable processes, chemical, optical and photographical, this [Bunsen's photometer] has calmly held its own; and not all the ingenuity and profound research of modern laboratories has been able to shake it on the throne to which its own simplicity and admirable adaptability to practical photometrical requirements have raised it" (Dibdin, 1889, p. 3). Similarly, the majority of photometers used by the practitioners in stellar photometry were visual. The dominance of the visual approach continued more than a half century until photoelectric photometry became available in the 1920s.

When we consider how photometric measurements and photometric methods were evaluated by the practitioners, we can see why the visual approach was accepted as "good" science even though it could not offer accurate measurements. For when we put the meanings of "good science" and "correct measurement" in historical context, the irony of a "good" science associated with "wrong" measurements in early 19th century visual photometry disappears.

First, whether a photometric measurement was wrong depended on the acceptable range of error defined by the practice. This acceptable range was partly determined by the level of systematic error due to the instruments and the related procedures. Using the eye as the essential part to match brightness, the visual approach inherited a systematic error about 1.5%, caused by the limited sensibility of the eye in comparing brightness difference. When there were two brightness-matching operations, such as in the measurements of transparency and reflective power, the systematic error caused by the eye can be doubled. Acceptable range was also determined by the inevitable error in distance measurements since a 1% error in distance would result in a 2% error in photometric measurements due to the inverse square relation. Thus, a 3–5% range of error was unavoidable for the visual approach. In addition, the acceptable range of error also reflected practical needs. In the 19th century, the gas industry required an accuracy of 5%, and stellar photometry demanded an accuracy of 0.1 magnitudes, that is, about 25% in terms of brightness (Walsh, 1958, p. 478). Thus, although measurements offered by the visual approach were wrong according to contemporary values, many of them, such as Rumford's measurements of transparency and Potter's

of reflective power of metals, were "good enough," within the acceptable range defined by prevailing standards.

Furthermore, accuracy was not the only standard that the practitioners used to evaluate photometric methods in the 19th century. Although photometry was an experimental science with measuring illumination as its main purpose, its soundness did not depend solely upon the accuracy of its measurements. Similar to theory evaluation where the question of truth or falsity is not the only concern (for example, simplicity, consistency, and scope of application are frequently used in theory appraisal), the evaluation of experimental methods frequently involves multiple standards in addition to accuracy. In 19th century photometry, such issues as simplicity (instrument design and experiment operation), efficiency (time and cost), and scope of application (sensibility) were important to the evaluation of the rival methods. Because gas inspection and star categorization were the major applications of photometry, accuracy became a secondary issue. The physical approach that utilized thermometers to measure illumination was defeated in the competition mainly because of its low efficiency (long response time) and limited application (low sensibility). Thus, one should not be surprised that the visual approach was accepted as "good (enough)" in the 19th century, although most of its measurements were inaccurate according to contemporary values.

Department of Philosophy, California Lutheran University, USA

NOTES

[1] Kepler did not explicitly indicate that distance was the parameter affecting the intensity of light (Kepler, 1604, p. 10). The relation between illumination and distance was specified by Castelli in 1634, and empirically verified by Monstanari in 1676. For more about the law of illumination before Bouguer, see Ariotti and Marcolongo (1976).

[2] Since the mid-19th century theoretical values from well established theories such as the wave theory of light and the electromagnetic theory had been frequently used to calibrate photometric measurements. For an example of how theoretical values affected photometric measurements, see Rayleigh (1886).

[3] Celsius claimed that, to see an object placed twice as far as another with the same degree of distinctness, it should be illuminated 256 times as much, or eight power of the distance. He however did not justify this quantitative relation; see Bouguer (1760, p. 48).

[4] This quantity was called the brightness difference threshold. Bouguer admitted that this threshold varied between individuals, but he insisted that 1.66% should be the highest under normal circumstances. Later in the mid-19th century Fechner reported that he could distinguish illumination difference as small as 1%, and Helmholtz pushed the threshold to the level of 0.6% (Palaz, 1896, p. 11).

[5] The wave theory of light calculates the reflective power of transparent materials by the Fresnel formula:

$$R = \frac{1}{2}\left(\frac{\sin^2(\theta - \theta')}{\sin^2(\theta + \theta')} + \frac{\tan^2(\theta - \theta')}{\tan^2(\theta + \theta')}\right)$$

where θ is the angle of reflected light and θ', the angle of refracted light ($\sin\ \theta' = \sin\ \theta/n$, where n is refractive index). For plate glass $n = 1.515$; its reflective power at 0° should be 0.04193; for

water $n = 1.33$; its reflective power at zero and 89.5° should be 0.02006 and 0.9465, respectively. The reflective power of metals is calculated according to the electromagnetic theory in terms of the following formula:

$$R = \frac{1}{2}\frac{n^2(1+k^2)\cos^2\theta - 2n\cos\theta + 1}{n^2(1+k^2)\cos^2\theta + 2n\cos\theta + 1} + \frac{1}{2}\frac{n^2(1+k^2) - 2n\cos\theta + \cos^2\theta}{n^2(1+k^2) + 2n\cos\theta + \cos^2\theta}$$

where n is refractive index, k, absorptive index, and θ, angle of reflection (Ditchburn, 1991, p. 444). For mercury, $n = 1.62$, $k = 2.71$; its reflective power at 69° should be 0.7165.

[6] Rumford did not describe the thickness of the glass plate, except by saying that it was very thin. To estimate the theoretical value, it is assumed that the absorption is minimum and the lost of light is caused only by the two reflections from both surfaces. Given that the refractive index of plate glass is 1.515, its reflective power R at 0° should be 0.04219, and its transparency T is calculated according to this formula: $T = (1 - R)^2$.

[7] Before he won the professorship in 1805, Leslie earned his living as a private tutor, a translator, and a reviewer (Morrell, 1975).

[8] The peak of solar radiation at sea level is in the region between the blue and the green (Moon, 1940). Leslie's photometer must have been influenced by the infrared and thus inflated the intensity of the red region.

[9] The absorptive index of sea water to visible light is 10^{-4} cm^{-1}. The transparency (T) of 115-inch (292 cm) sea water should be: $T = e^{-(292\times0.0001)} = 0.9712$. Bouguer' measurement error may have been caused by the low brightness in the matching field. In Bouguer's experimental setting, the brightness of the matching field was at the level of 10^{-2} candle/m^2 after the incident light had traveled 16 feet from the source of a single candle. The brightness difference threshold would increase from 1.5 to 7.5% under this circumstance (Walsh, 1958, p. 62).

[10] The transparency in Leslie's setting is calculated by the following formula:

$$T = \left(\frac{1-n}{1+n}\right)^2 \times e^{-kl}$$

where n is the refractive index of water (1.33), k is the absorptive index (10^{-4} cm^{-1}), and l is the depth of the seawater (457.2 cm).

[11] Means are used when more than two measurements of a specific reflection angle were made. Potter's measurements invalidated the received view that, similar to other substances, metals reflected more light when the reflection angle increased. Potter's discovery stimulated James MacCullagh to study metallic reflection and to discover in 1836 an empirical law to describe the reflective power of metals (MacCullagh, 1880, p. 61).

[12] For copper–tin alloy, $n = 1.22$ and $k = 2.7$ when $\lambda = 6000$Å (National Research Council, 1926–1930, vol. 5, pp. 248–256). Its reflective power should be 69.07% at 10°, 69.01% at 30°, 68.59% at 50°, 68.17% at 60°, and 68.26% at 70°.

[13] Potter believed that the discrepancies discredited the wave theory of light; see Potter (1831, p. 322).

[14] No one at the time detected the fault in Powell's "thought experiment," which may indicate that the methodology of visual photometry was not well understood outside the photometric community.

[15] Forbes calibrated his electric thermometer by using a Leslie photometer. He found that one centigrade degree measured by the Leslie photometer corresponded to 42° deviation of the needle in his thermometer.

[16] To determine the ratio, Forbes passed a beam of heat through two mica piles, and used the "thermal photometer" to measure the intensities of the transmitted heat when the axes of the two mica piles were parallel and perpendicular. He found that the ratio depended upon many factors, including the angle of refraction, the refracting medium, and the heat source.

[17] Through a large number of readings, a Lummer–Brodhun photometer could produce measurements accurate to about 0.2% (Walsh, 1958, p. 189, 204).

REFERENCES

Ariotti, P. E. and F. J. Marcolongo (1976). "The law of illumination before Bouguer (1729): Statement, restatements and demonstration." *Annals of Science* **33**: 331–340.

Bouguer, P. (1961 [1760]). *Optical Treatise on the Gradation of Light*, trans. W. Middleton. Toronto: University of Toronto Press.

Brewster, D. (1831). *A treatise on Optics*. London: Longman.

Christison, R. and E. Turner (1825). "On the construction of oil and coal gas burners, and the circumstances that influence the light emitted by the gas during their combustion; with some observations on their relative illuminating power, and on the different modes of ascertaining it." *Edinburgh Philosophical Journal* **13**: 1–39.

Dibdin, W. (1889). *Practical Photometry: A Guide to the Study of the Measurement of Light*. London: Walter King.

Ditchburn, R. W. (1991). *Light*. New York: Dover.

Forbes, J. (1835). "On the refraction and polarization of heat." *Philosophical Magazine* **6**: 134–142, 205–214, 284–291, 366–371.

Forbes, J. (1838). "Research on heat, second series." *Philosophical Magazine* **12**: 545–559.

Forbes, J. (1839). "Memorandum on the intensity of reflected light and heat." *Philosophical Magazine* **15**: 479–481.

Forbes, J. (1840). "Letter to Richard Taylor, Esq., with reference to two papers in the *Philosophical Magazine* for January, 1840." *Philosophical Magazine* **16**: 102–103.

Forbes, J. (1851). "On the intensity of heat reflected from glass." *Proceedings of the Royal Society of Edinburgh* **2**: 256–257.

Huygens, C. (1698). *Cosmotheoros, sive De Terris Coelestibus, earumque ornatu, conjecturae*. The Hague: Hagae-comitum, apud A. Moetjens, 1698.

Kepler, J. (1604). *Ad Vitellionem paralipomena quibus Astronomiae pars optica traditur*; *potissimum de artificiosa observatione et aestimatione diametrorum deliquiorumq́; solis & lunae, cum exemplis insignium eclipsium . . . Tractatum luculentum de modo visionis, & humorum oculi usu, contra opticos & anatomicos.*, Francofurti: Francofurti, apud Marnium & Heredes Aubrii, 1604.

Lambert, J. (1760). *Photometria sive de mensura de gradibus luminis, colorum et umbrae*. Augustae Vindelicorum, Sumptibus Vidvae Eberhardi Klett, typis Christophori Petri Detleffsen, 1760.

Leslie, J. (1800). "Description of an hygrometer and photometer." *Journal of Natural Philosophy* **3**: 461–467.

Leslie, J. (1801). "Observations and experiments on light and heat, with some remarks on the inquiries of Herschel." *Nicholson's Journal* **4**: 344.

Leslie, J. (1813). "Description of an atmometer, and an account of some photometric, hygrometric, and hygroscopical experiments." *Philosophical Magazine* **42**: 44–52.

Leslie, J. (1824). "Letter from Professor Leslie to the Editor on Mr. Ritchie's experiments on heat, and new photometer." *The Edinburgh New Philosophical Journal* **4**: 170–172.

Leslie, J. (1838). *Treatises on Various Subjects of Natural and Chemical Philosophy*. Edinburgh: Adam and Charles Black.

Lloyd, H. (1834). "Report on the progress and present state of physical optics." In: *Miscellaneous Papers Connected with Physical Science (1877)*, ed. H. Lloyd. London: Longman, pp. 19–146.

MacCullagh, J. (1880). *A Collected Works of James MacCullagh*. Dublin: Hodges.

Moon, P. (1940). "Proposed standard solar-radiation curves for engineering use." *Journal of the Franklin Institute* **230**: 583–617.
Morrell, J. B. (1975). "The Leslie affair: Careers, kirk, and politics in Edinburgh in 1805." *Scottish Historical Review* **54**: 62–82.
National Research Council (1926–1930). *International Critical Tables of Numerical Data, Physics, Chemistry and Technology*. New York: McGraw-Hill.
Palaz, A. (1896). *A Treatise on Industrial Photometry*. New York: Van Nostrand.
Potter, R. (1830). "An account of experiments to determine the quantity of light reflected by plane metallic specula under different angles of incidence." *The Edinburgh Journal of Science* **3**: 278–288.
Potter, R. (1831). "Experiments relating to the reflective powers of crown, plate, and flint glass, with theoretical considerations." *The Edinburgh Journal of Science* **4**: 320–328.
Potter, R. (1840). "On photometry in connexion with physical optics." *Philosophical Magazine* **16**: 16–23.
Powell, B. (1838). "On some points connected with the theory of light." *Annual Reports of the British Association* **8**: 6–7.
Rayleigh, J. (1886). "On the intensity of light reflected from certain surfaces at nearly perpendicular incidence." In: *Scientific Papers, Vol. 2 (1964)*, ed. J. Rayleigh. New York: Dover, pp. 522–542.
Rumford, C. (1794). "Experiments on the relative intensities of the light emitted by luminous bodies." *Philosophical Transactions of the Royal Society of London* **84**: 67–106.
Walsh, J. (1958). *Photometry*. London: Constable.
Wright, A. (1850). "Lecture by Alexander Wright." *The Journal of Gas Lighting* **1**: 325–326.

JED Z. BUCHWALD

AN ERROR WITHIN A MISTAKE?

AN ERROR WITHIN HELMHOLTZ'S ELECTRODYNAMICS?

In recent years much work has been done on the system of electrodynamics that was developed by the German polymath Hermann Helmholtz in the late 1860s and throughout the 1870s.[1] In 1874 Helmholtz used this new electrodynamics briefly to examine a situation that seems to be quite similar to one analyzed half a decade later by his student Heinrich Hertz in a manuscript written for Helmholtz's eyes. Though Hertz did not refer explicitly to Helmholtz's previous considerations of 1874, his analysis was based unequivocally on equations that were unique to Helmholtz. Yet in this particular application of the master's system, its creator in 1874 and his student in 1879 seem to have arrived at markedly different, indeed at conflicting, results.

Here, it seems, we have a situation in which one of the two must have erred either in calculation or else in setting out the problem's conditions. Both were using the very same system of electrodynamics, a system that was abandoned in Germany shortly after Hertz himself discovered electric waves late in 1887. We have accordingly found a most interesting circumstance, in which one of two practitioners apparently made some sort of mistake within the confines of a system that became altogether defunct about a decade later. It is as though we had come across a disagreement between, say, two proponents of 18th century affinity chemistry concerning a process that has no significance at all in the post-Lavoisieran world. We have stumbled across an error within a mistake: something done wrong within a now-rejected system.

Since the details of Helmholtz's electrodynamics have been given several times (see note 1), we may begin our excavation of error directly with the system's fundamental assumption, which is the existence of a 'potential' function from which the electrodynamic interaction between paired differential volumes of (e.g.) conducting bodies may be deduced. This function, P, can be interpreted, and used, as an energy of the system; it depends on the electric current within each volume element, as well as upon the distances between the elements. In its most general form P is:

$$P = \underbrace{-\frac{A^2}{2}\int_r\int_{r'}\frac{\mathbf{C}\cdot\mathbf{C}'}{|\mathbf{r}-\mathbf{r}'|}\,\mathrm{d}^3r'\mathrm{d}^3r}_{\text{'Neumann' function for extended objects}} \underbrace{-\frac{A^2}{2}(1-k)\int_r\mathbf{C}\cdot\left[\nabla_r\int_{r'}\mathbf{C}'\cdot\nabla_{r'}\,|\mathbf{r}-\mathbf{r}'|\,\mathrm{d}^3r'\right]\mathrm{d}^3r}_{\text{generalized term that vanishes if either circuit is closed}}$$

J.Z. Buchwald and A. Franklin (eds.), Wrong for the Right Reasons, 185–208.

Here A is a universal electrodynamic constant, $\mathbf{C}$ and $\mathbf{C}'$ are interacting currents, $\mathbf{r}$ runs from the origin to $\mathbf{C}$, $\mathbf{r}'$ runs from the origin to $\mathbf{C}'$, and k distinguishes among different permissible expressions for the interaction.

The function P was designed by Helmholtz to yield the by-then standard Ampère force between current-bearing circuits when the circuits are closed. Indeed, if either of the interacting systems to which the volume elements d^3r and d^3r' respectively belong is closed, then this most general expression for the potential reduces to its first term. That expression (for closed, linear circuits) was first obtained by Franz Neumann at Königsberg in his successful attempt to find a single function from which both electrodynamic force and electromagnetic induction could be obtained by, respectively, space and time differentiation.[2] In the system considered by Hertz in 1879, as well as in the actual experimental systems that were examined by Helmholtz and his collaborators earlier in the 1870s, one of the current-bearing objects always forms an effectively closed circuit, so that, with them, we may limit our considerations here to the implications of this first term in P.

The forces that act on the current bearing objects are calculated by varying the function P. The volume elements may experience two kinds of effect. Each may be acted on by a force that tends to move the element physically from one location to another in the usual way (i.e., by producing an acceleration equal to the force divided by the mass of the element) – contemporary language referred to this kind of force as *ponderomotive*. Each may also experience an action that tends to change the current that exists within it – or, in common parlance, an *electromotive force* (or *emf*). The important point for our purposes is this: Helmholtz's potential function yields the same *emf* as other theories of the day (and as modern electrodynamics) *only* when the element that is being acted on itself forms part of a closed system; otherwise Helmholtz's scheme entails an altogether novel force. The second major task that Helmholtz assigned his apprentice, the talented young Heinrich Hertz, was to see whether one could polarize dielectrics by means of electromagnetic induction. To do so Hertz thought to use this new *emf* that his mentor's system entailed; to that end he produced a feasibility study for Helmholtz's own eyes, a *prospectus* as it were of what might be done (Hertz, 1879).

Yet in this work of 1879 that Hertz drew up for Helmholtz himself, and wherein he used his mentor's novel electromotive force, he arrived by computation at results that seem to conflict with ones that Helmholtz had explicitly set out in print five years before. Hertz had surely read Helmholtz's paper, though he might have overlooked the remark. But even if he had missed it, or if he had not read the paper at all, how, using Helmholtz's own electrodynamics, could he have reached a different result? Is it a simple case of a mistaken calculation by a young apprentice? Or had Helmholtz himself erred? And, if so, why did neither Helmholtz (who read Hertz's MS) nor Hertz apparently ever notice it? Have we just found an error within a mistaken theory? Or is there something more to it?

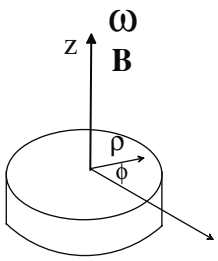

Figure 1. Helmholtz's spinning disk.

HERTZ AND HELMHOLTZ SEEM TO DISAGREE

The problem first appeared in 1874. In April of that year Helmholtz ended an article entitled "Kritisches zur Elektrodynamik", which considered objections to his formulation of electrodynamics, by pointing to a specific case in which his system yielded results that are different from those that are implied by all of the others. Helmholtz described the situation in the following words:

> Imagine a metal disk [see Figure 1] that is spinning rapidly about its axis and that is crossed by magnetic lines of forces that are parallel to the axis and symmetrically distributed about it, then the edge of the disk will be electrified according to the Ampère law,[3] but it will not be according to the potential law. (Helmholtz, 1874, p. 762)[4]

To understand what was at stake here, let's begin with modern electrodynamics and generalize the problem to an object that moves with a velocity **v** through a magnetic field **B**; a point in the object is specified by the vector **r** that is drawn to the point from the origin of coordinates, and the object's center of mass is itself located by the vector $\mathbf{r}_{cm}$ (see Figure 2). The object will experience an electromotive force (*emf*) **F** at a given point that is given by the cross-product there of the velocity with the magnetic field:

$$\mathbf{F_{MAX}} = \mathbf{v} \times \mathbf{B} \tag{1}$$

(*The "Ampère" expression for the electromotive force*)

If our object spins with angular velocity ω about its center of mass then the linear velocity **v** at the point in the object that is specified by the vector **r** will be:

$$\mathbf{v} = \boldsymbol{\omega} \times (\mathbf{r} - \mathbf{r}_{cm}) \tag{2}$$

(*The velocity at a point of a spinning object*)

Consequently, according to the Ampère expression the electromotive force at **r** will be:

$$\mathbf{F_{MAX}} = [\boldsymbol{\omega} \times (\mathbf{r} - \mathbf{r}_{cm})] \times \mathbf{B} \tag{3}$$

(*The Ampère emf for a spinning object*)

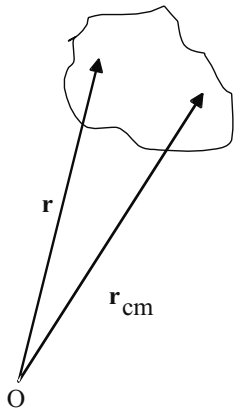

Figure 2. Object vectors.

Suppose now that the angular velocity is parallel to the magnetic field. According to this equation, the *emf* will not vanish.

Yet according to Helmholtz in 1874 the *emf will* disappear in these circumstances. Why? The answer is, in one sense, quite simple: Helmholtz's expression for the *emf* contains an additional term that can in the right circumstances annul the Ampère expression. According to Helmholtz, the electromotive force that acts on an object which moves with velocity **v** in the presence of what we shall call a vector potential **A** is:

$$\mathbf{F}_{\mathrm{HELM}} = \underbrace{\mathbf{v} \times (\nabla \times \mathbf{A})}_{\text{Ampère term}} - \underbrace{\nabla(\mathbf{v} \cdot \mathbf{A})}_{\text{Helmholtz term}} \qquad (4)$$

(*emf according to Helmholtz*)

As the Appendix below shows, this auxiliary vector is defined by Helmholtz (and so by Hertz) primitively in terms of currents (or derivatively in terms of magnetization). For reasons that will become clear below, it's important to note that this potential gains significance altogether from its role as the vector that a current multiplies in calculating the energy of the system comprised of the current in question and the currents or magnetization with which it is interacting (*vide* equation (21)).[5] In Helmholtz's energy-based electrodynamics the vector potential has no other function or meaning than this, but, just because of its immediate presence in the energy, the potential was more fundamental than the forces to which the energy gave rise.

The new term in Helmholtz's expression for the force can cancel out Ampère *emf*s – and, according to Helmholtz, it does so when an object spins about an axis that is parallel to a magnetic field which is symmetric about the axis. Despite the fact that the young Hertz used exactly Helmholtz's formula for this very force, he – unlike Helmholtz himself in 1874 – did not find that the force must vanish when the angular velocity and the magnetic field are parallel to one another, as we shall see in detail.

Yet in both cases essentially the same magnetic field and velocity are involved. One of the two, it seems, must have erred in computation.

HERTZ'S SPHERE

Hertz did not consider precisely the same configuration that his mentor Helmholtz had, but the one which he did examine provides a more general case that embraces Helmholtz's, as we shall see. Hertz's particular goal was to find a way of experimentally testing whether or not the electromotive force that is generated by motion through a magnetic field can polarize dielectrics just as it can generate currents in conductors. To do so he thought to use the force that would be generated in a small object by spinning it in the earth's magnetic field. To that end he had first to calculate the magnetic force at the earth's surface, for which he used auxiliary functions that were in reasonably standard German employ at the time.

Hertz began with a quantity $\boldsymbol{\lambda}$ which he used to represent the potential of the earth's magnetization $\mathbf{M}$ (i.e., its magnetic moment per unit volume). With $\boldsymbol{\lambda}$ given by $\int (\mathbf{M}(r')/|\mathbf{r} - \mathbf{r}'|)\, \mathrm{d}^3 r'$, the corresponding vector $\mathbf{A}$ that is to be used in Helmholtz's equation (4) has the form $-\nabla \times \boldsymbol{\lambda}$:[6]

$$\mathbf{A} = -\nabla \times \boldsymbol{\lambda}$$

where

$$\boldsymbol{\lambda} \equiv \int (\mathbf{M}(r')/|\mathbf{r} - \mathbf{r}'|)\, \mathrm{d}^3 r' \tag{5}$$

(*The vector potential A for magnetization M*[7])

We can substitute $\mathbf{A}$ into Helmholtz's basic expression for the force to obtain:[8]

$$\mathbf{F}^{\mathrm{MAG}}_{\mathrm{HELM}} = \mathbf{v} \times \nabla(\nabla \cdot \boldsymbol{\lambda}) - \nabla(\mathbf{v} \cdot (\nabla \times \boldsymbol{\lambda})) \tag{6}$$

(*The force expressed in terms of the magnetization potential*)

In equation (6) the force is labeled $\mathbf{F}^{\mathrm{MAG}}_{\mathrm{HELM}}$ to emphasize that it is not as yet in a form appropriate to Hertz's specific application to the earth, in that $\boldsymbol{\lambda}$ may to this point derive from any source of magnetization whatsoever.

Let's now follow Hertz's application of the formula to the case of an object spinning in the earth's field, adding in a few details that he omitted in order to facilitate our comparison of his results with those of Helmholtz. Take the earth's magnetization m to be directed along the z axis, and assume that the magnetic effects which are responsible for the earth's field are localized near the earth's center, so that we can take the distance from the earth's center to our spinning object also to be the effective distance between the object and the earth's magnetization (see Figure 3). As a result, the vector λ will be:

$$\boldsymbol{\lambda}_{\mathrm{HTZ}} = \frac{m}{r}\mathbf{e}_z \tag{7}$$

(*Hertz's magnetization potential* $\boldsymbol{\lambda}_{HTZ}$ *for the earth's field*)

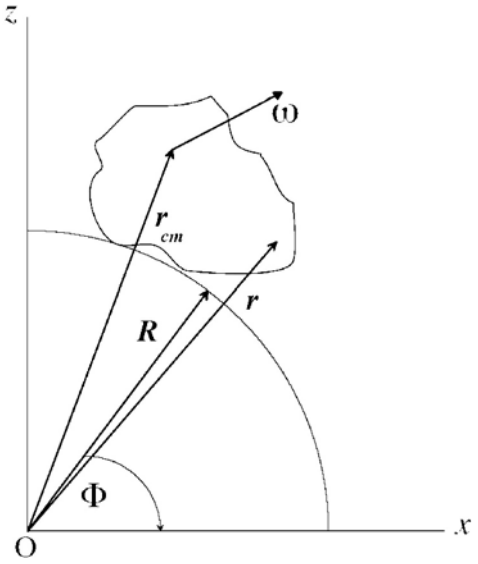

Figure 3. Spinning object near earth's surface.

Hertz did not provide the vector **A** for calculating the magnetic force since he had no specific need for it given expression (6), but for future reference we note that **A** has the following exact and approximate forms:

$$\mathbf{A} \text{ is exactly } -\nabla \times \left(\frac{m}{r}\mathbf{e}_z\right) \text{ or approximately } \frac{m}{R^2}\cos(\Phi)\mathbf{e}_y \tag{8}$$

Here $\mathbf{e}_z$ lies along the polar axis, R is the earth's radius, Φ is the latitude, and $\mathbf{e}_y$ is orthogonal to a meridian plane.

Hertz next moved almost directly to give expressions for the *emf*s that result according to equations (6) and (7) when a small object spins near the surface of the earth. First of all, let's assume, as he did, that our object's angular velocity lies entirely in the plane formed by the earth's polar axis (to which the magnetization is assumed to be parallel) and the line from the earth's center to the object. That is, our object spins only about an axis that lies in the plane of the local meridian – we will not examine the effect of an east–west component. Since the earth's field is axially symmetric about the polar (say z) axis, we can in full generality consider the forces that act in any plane section that contains the axis. For simplicity we will take the xz plane as the one in which we calculate forces.[9] The object's angular velocity accordingly has (by assumption) no component along the y axis, but it may have components along the z (polar) and x (equatorial) axes. Denote the latitude at our object by Φ, and assume as well that the object's dimensions are small to first order in respect to the radius R of the earth. In the final result we can accordingly replace r (the distance to the object point) with R (the radius of the earth – see Figure 3).

We can now proceed to substitute Hertz's magnetization potential (equation (7)) into Helmholtz's force (equation (6)), after which we replace both of the distances r and r_{cm} with the earth's radius R (Figure 3). We then drop all expressions in which the third or higher power of the earth's radius appears in the denominator, on the grounds that other terms remain that contain a factor of only $1/R^2$, as we shall see in a moment. This last assumption completely removes the expression that corresponds to the Ampère *emf* (*viz.* $\mathbf{v} \times \nabla(\nabla \cdot \boldsymbol{\lambda})$).[10] Limiting our consideration to the *xz* plane, we find with Hertz that the spinning body will experience the following *emf*:

$$\begin{aligned} F^x_{\mathrm{HTZ}} &= -\frac{m}{2R^2}\omega_z \cos(\Phi) \\ F^y_{\mathrm{HTZ}} &= 0 \\ F^z_{\mathrm{HTZ}} &= -\frac{m}{2R^2}\omega_x \cos(\Phi) \end{aligned} \tag{9}$$

(*The emf on the spinning object according to Hertz*)

These *emf*s vanish altogether at the poles and are a maximum at the equator, for a given angular velocity.

We ask next what direction the magnetic force itself has at the latitude Φ. For consistency we must use Hertz's expression for the magnetization potential in our computation (equation (7)). Since the corresponding magnetic force must be $-\nabla \times (\nabla \times \boldsymbol{\lambda})$, we find (again under the approximation that in the end we replace both r and r_{cm} with R):

$$\begin{aligned} B^x_{\mathrm{HTZ}} &= \frac{m}{R^3}(3\ \sin(\Phi)\cos(\Phi)) \\ B^y_{\mathrm{HTZ}} &= 0 \\ B^z_{\mathrm{HTZ}} &= \frac{m}{R^3}(2 - 3\cos^2(\Phi)) \end{aligned} \tag{10}$$

(*The magnetic force corresponding to Hertz's magnetization potential*)

We can obviously adjust the angular velocity so that it parallels the magnetic force at a given latitude.[11] In fact, we can rewrite Hertz's expressions for the *emf* in terms of the local components of the magnetic force in the following way:

$$\begin{aligned} F^x_{\mathrm{HTZ}} &= -\frac{R}{2}\omega_z(B^z_{\mathrm{HTZ}}\cos(\Phi) - B^x_{\mathrm{HTZ}}\sin(\Phi)) \\ F^y_{\mathrm{HTZ}} &= 0 \\ F^z_{\mathrm{HTZ}} &= -\frac{R}{2}\omega_x(B^z_{\mathrm{HTZ}}\cos(\Phi) - B^x_{\mathrm{HTZ}}\sin(\Phi)) \end{aligned}$$

To take a simple example, we can locate ourselves at the equator, where the magnetic force runs along a north–south axis (c.f. equation (10), with Φ set to zero), and where the *emf* reaches a maximum. We can set a sphere of radius r_{s}, say, spinning about its center around this same axis (tangent to the local meridian), in which case the *emf* will point directly downwards (see Figure 4). In this same situation, the

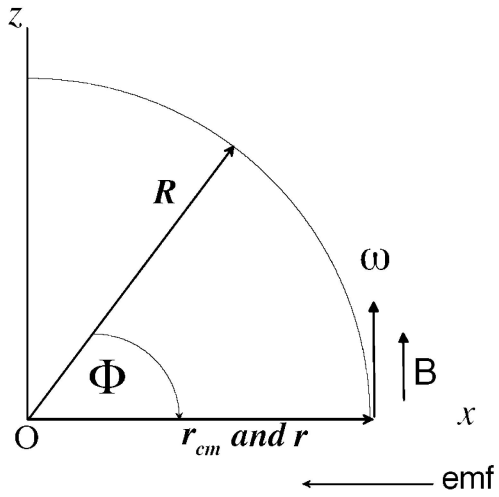

Figure 4. The *emf* at the equator for an object spinning parallel to the polar axis, according to Hertz's equations.

Ampère expression ($\mathbf{v} \times \mathbf{B}$) yields an *emf* directed at each point along a line that is perpendicular to the object's axis of spin, aiming directly away from the axis and towards the surface of the sphere. However, the Ampère *emf* will be incomparably smaller than this new one that Hertz has calculated, since it will contain a factor r_s, whereas the Hertz force contains a corresponding factor R.[12] That is, the new force is larger than the Ampère *emf* by the immense ratio R/r_s. Of course, the Hertz force should not exist at all according to Helmholtz's remarks in 1874, who had at the time used precisely the same expression for calculating the *emf* that his student Hertz used later (*viz.* equation (4)). Turn now to Helmholtz's claim.

HELMHOLTZ'S DISK

Helmholtz had not specifically discussed an object of any shape whatsoever spinning in an arbitrary direction in the earth's magnetic field. His comment referred to a disk that turns about its axis of symmetry in a field of magnetic force that is parallel to, and symmetric about, the axis. Under these conditions, Helmholtz had asserted in 1874, the Ampère expression requires the existence of an *emf* that is directed from the central axis towards the disk's perimeter. But his own force law, he continued, implies that there will be no *emf* at all. We will turn below to the reasoning that may lie behind Helmholtz's claim. Let's first consider whether, and if so in what manner, it applies to the situation that Hertz envisioned half a decade later.[13]

One might argue that the two situations (Helmholtz's and Hertz's) differ from one another because Helmholtz specified a magnetic field that is symmetric about and parallel to a disk's axis of spin, whereas Hertz considered the earth's field, which

certainly seems not to satisfy the requirement, *vide* equation (10). However, Hertz's spinning object is vastly smaller than the earth, and in its vicinity the earth's field should certainly be effectively uniform, thereby trivially fulfilling Helmholtz's symmetry requirement. Nevertheless, Helmholtz's conclusion implicated the symmetry of a field of magnetic force, whereas Hertz's calculation was based upon a specific expression for the magnetization potential, from which the force was computed. In order to clarify the plausible assertion that the (locally insignificant) inhomogeneity of Hertz's magnetic force cannot be the source of the difference between his and Helmholtz's claims, we will first connect Hertz's calculation to a vector **A** that does yield a strictly uniform force.

We need to find a vector whose curl will be equal to a homogeneous magnetic field **B**.[14] One such is $\mathbf{A}_r$:

$$\mathbf{A}_r = \tfrac{1}{2}(\mathbf{B} \times \mathbf{r}) \tag{11}$$

(*A vector potential that produces a uniform magnetic field*)

Inserting $\mathbf{A}_r$ into the general Helmholtz expression for the force ($\mathbf{F}_{HELM}$, equation (4)) we obtain (naming the result $\mathbf{F}_r$):

$$\mathbf{F}_r = \tfrac{1}{2}\,\mathbf{v} \times \mathbf{B} + \tfrac{1}{2}\boldsymbol{\omega} \times (\mathbf{B} \times \mathbf{r}) \tag{12}$$

(*First expression for the emf corresponding to* A_r)

This new *emf* can be written in strict mathematical equivalence as:

$$\mathbf{F}_r = \tfrac{1}{2}(\mathbf{r} - \mathbf{r}_{cm}) \times (\mathbf{B} \times \boldsymbol{\omega}) + \tfrac{1}{2}\boldsymbol{\omega} \times (\mathbf{B} \times \mathbf{r}_{cm}) \tag{13}$$

(*Second expression for the emf corresponding to* A_r)

We next insert into equation (13) the very same expressions for the magnetic field that result from Hertz's magnetization (equation (10)). In addition, we also approximate the center-of-mass distance (r_{cm}) by the earth's radius (R). Doing so yields, as expected, precisely the same expression for the *emf* on the spinning object that Hertz himself had obtained (equation (9)). In other words, the vector $\mathbf{A}_r$ produces the very same *emf* as the vector **A** (equation (8)) that corresponds to Hertz's magnetization potential, $\boldsymbol{\lambda}_{HTZ}$ (equation (7)), when the same approximations are used.

Since we now see that Hertz's expressions for the *emf* follow perfectly well from a calculation based on the assumption that the local magnetic force is uniform in direction and magnitude, it follows that the difference between his and Helmholtz's assertions can have nothing to do with any slight local inhomogeneity. If Hertz's claim is correct, then it seems that Helmholtz's simply cannot be, and *vice versa*.

Or have we missed something essential here? To see whether or not we have, turn first to Helmholtz's original statement. Helmholtz had there referred explicitly to "magnetic lines of force . . . that are parallel to the axis and symmetrically distributed about it". Although he used the phrase "magnetic lines of force", Helmholtz just might have been thinking of a field of vector potential, since his entire discussion of the *emf*s involved in motion proceeds from his fundamental interaction energy, which

is formulated in terms of the vector potential and not (by necessity) the corresponding magnetic force. If that were so, then Helmholtz's conclusion would be almost obvious, given the foundation of his electrodynamics in variational calculations based on interaction energy: for if the vector potential is itself symmetric about the disk's axis of rotation, then the potential that will be seen by any point of the rotating sphere or disk must always be the same – in which case the energy-variation that underpins Helmholtz's calculations can yield no resultant force at all, just as he asserted.[15] Under this interpretation there is no conflict between Hertz's and Helmholtz's claims; we are instead left with a sloppy statement on the part of Helmholtz – and worse, one that would not correspond to any reasonable experimental situation, since the originating currents follow the vector potential in direction.[16]

There is another possibility. What if Hertz's *emfs* are not the only ones that are consistent with the assumption that the field of magnetic force is uniform (or symmetric about the axis of spin)? This seems unlikely, since we have already found that there is nothing at all wrong with his computation, and, moreover, that it is entirely compatible with the local uniformity of the magnetic force. But to imply a proposition is not necessarily to be implied by it.

Let's return to the vector potential that corresponds to a uniform magnetic force. We considered $\mathbf{A}_{\mathrm{r}}$, which, we saw above, produces Hertz's *emf* when we require (as we may) that the expressions for the **B** field in the resultant force (equation (12)) be the same as those that are implied by Hertz's magnetization. But this is not the only vector potential that can produce the requisite magnetic force. In fact, we can clearly add any constant, or the gradient of any function, to $\mathbf{A}_{\mathrm{r}}$ and still obtain what we need if we are concerned only with the resultant magnetic force. For example, we could if we like replace the distance **r** to the point in the object at which the *emf* is calculated with the distance from the object's center of mass to that point, because the additional term that results (namely $-\frac{1}{2}(\mathbf{B} \times \mathbf{r}_{\mathrm{cm}})$) is itself a constant. The potential $\mathbf{A}_{\mathrm{cm}}$ would then be:

$$\mathbf{A}_{\mathrm{cm}} = \tfrac{1}{2}[\mathbf{B} \times (\mathbf{r} - \mathbf{r}_{\mathrm{cm}})] \tag{14}$$

(*A magnetically-equivalent vector potential*)

Note that if the magnetic field **B** is parallel to the angular velocity $\boldsymbol{\omega}$ then this vector potential will itself parallel the linear velocity **v**. Note also that our new vector potential is axially symmetric since its direction and magnitude always have the same values with respect to the disk's radius. This is not true for $\mathbf{A}_{\mathrm{r}}$ because the position vector **r** is not perpendicular to the disk's axis (*vide* note 15).

If we now insert $\mathbf{A}_{\mathrm{cm}}$ into Helmholtz's formula then we obtain, after considerable but standard manipulation (recalling that the location of the center of mass and the angular velocity of the spinning body are both to be considered constant):

$$\mathbf{F}_{\mathrm{cm}} = \mathbf{v} \times (\nabla \times \mathbf{A}_{\mathrm{cm}}) - \nabla(\mathbf{v} \cdot \mathbf{A}_{\mathrm{cm}}) = \tfrac{1}{2}\,[\mathbf{v} \times \mathbf{B} - \boldsymbol{\omega} \times (\mathbf{B} \times (\mathbf{r} - \mathbf{r}_{\mathrm{cm}}))] \tag{15}$$

(*First expression for the emf corresponding to A_{cm} according to Helmholtz's formula*)

We can immediately see that this new force differs by the term $-\frac{1}{2}[\mathbf{v} \times \mathbf{B} + \boldsymbol{\omega} \times (\mathbf{B} \times (\mathbf{r} - \mathbf{r}_{\mathrm{cm}}))]$ from the expression (equation (3)) for the Ampère *emf*. Of course, the force that derives from $\mathbf{A}_{\mathrm{r}}$ (equation (12)) also differs from the Ampère expression. The question is whether our new force, which derives from $\mathbf{A}_{\mathrm{cm}}$, differs in an appropriate manner from the one that is implied by $\mathbf{A}_{\mathrm{r}}$.

Indeed it does. The new expression can be manipulated to yield, in strict equivalence:

$$\mathbf{F}_{\mathrm{cm}} = \tfrac{1}{2}\,[(\mathbf{r} - \mathbf{r}_{\mathrm{cm}}) \times (\mathbf{B} \times \boldsymbol{\omega})] \tag{16}$$

(*Second expression for the force corresponding to A_{cm} according to Helmholtz's formula*)

According to this equivalent second expression, the *emf* will indeed vanish altogether whenever the angular velocity parallels the magnetic field. We have therefore found a vector potential that yields a uniform field of magnetic force and that nevertheless produces the very effect that Helmholtz had claimed.

How can this be so? The answer is deceptively simple: although a constant addition to the vector potential has no affect at all on the magnetic force, it certainly may have one on the electromotive force according to Helmholtz's electrodynamics, because Helmholtz's general expression for *emf* contains the additional term (in comparison to Ampère) $-\nabla(\mathbf{v} \cdot \mathbf{A})$. Even if the addition (call it $\mathbf{A}'$) to the vector potential is constant, this extra term in the force will yield two novel contributions: namely, $-(\mathbf{A}' \cdot \nabla)\mathbf{v}$ and $-\mathbf{A}' \times (\nabla \times \mathbf{v})$. Neither of these necessarily vanishes, because $\mathbf{v}$ may depend upon r (*vide* equation (2)). As a result, $\mathbf{F}_{\mathrm{cm}}$, but not $\mathbf{F}_{\mathrm{r}}$, does indeed disappear when the angular velocity is parallel to the magnetic force. It's instructive to rewrite the force that arises from the Hertz potential ($\mathbf{A}_{\mathrm{r}}$) in the following manner, since we can then easily see how it differs from the one that arises from $\mathbf{A}_{\mathrm{cm}}$:

$$\mathbf{F}_{\mathrm{r}} = \overbrace{\underbrace{\tfrac{1}{2}\,[(\mathbf{r} - \mathbf{r}_{\mathrm{cm}}) \times (\mathbf{B} \times \boldsymbol{\omega})]}_{\mathbf{F}_{\mathrm{cm}},\ \text{the Helmholtz force from } \mathbf{A}_{\mathrm{cm}}} + \underbrace{\tfrac{1}{2}\,[\boldsymbol{\omega} \times (\mathbf{B} \times \mathbf{r}_{\mathrm{cm}})]}_{\text{addition from } \mathbf{A}_{\mathrm{r}}}}^{\text{The Hertz force}} \tag{17}$$

(*Comparison of the Hertz and Helmholtz forces*)

Here we see clearly that the Hertz force can yield a result even when the Helmholtz *emf* vanishes altogether. Unlike field theory, Helmholtz's system is manifestly not gauge-invariant, and in this case of the spinning disk or sphere we have found a situation in which the lack of invariance has a testable consequence.

We can naturally ask whether Helmholtz might have envisioned such an expression as $\mathbf{A}_{\mathrm{cm}}$. If we recognize that he, unlike Hertz (who started from the earth's magnetization), began with a field of magnetic force and a spinning object, then it seems plausible that Helmholtz would have thought of this expression for the vector potential, had he produced any at all, and not the one that Hertz's lengthy computation entailed. Unlike Hertz, who naturally reckoned from the earth's center, Helmholtz (thinking of a locally-produced magnetic field) would no doubt have worked in terms

of local cylindrical coordinates, placing the origin at the center of his spinning disk. The potential $\mathbf{A}_{\rm cm}$, unlike $\mathbf{A}_{\rm r}$, contains the vector $\mathbf{r} - \mathbf{r}_{\rm cm}$, or $\boldsymbol{\rho}$, which represents the distance from the center of mass of the spinning object to the point on it at which we wish to calculate the *emf*. This same distance appears in the velocity $\mathbf{v}$ (equation (2)) of such a point. Accordingly, if Helmholtz had wondered at all about an appropriate vector potential to correspond to his magnetic field, then he would likely have used the very same vector that appears in the velocity, thereby ensuring the absence of *emf*. We will turn in a moment to the possible course of Helmholtz's reasoning during the year following the publication of his remark concerning the *emf* in a spinning disk, but let's first consider the difference between Hertz's and Helmholtz's attitudes in respect to this sort of problem.

Hertz was in an altogether different frame of mind from Helmholtz when he considered the spinning sphere. Helmholtz in 1874 was looking for a situation that contrasted strikingly with the claims of the Ampère *emf*. Hertz was looking for a way to test whether electromagnetic induction can polarize dielectrics. Where Helmholtz was looking to provide evidence for a new force law, Hertz was looking to use this same force law as a tool in order to see whether a particular kind of novel effect that would otherwise be difficult to produce could actually be elicited. Hertz therefore began directly with the specific physical situation that he had in mind, and he then proceeded in a straightforward way to calculate the force from it. He started with what he took to be the most fundamental assumption possible, namely that the earth's field results from a magnetic dipole located near its center. Hertz had to work with the dipole's potential, and not the force to which it gives rise, because Helmholtz's law was expressed in terms of potentials.

Do we have any evidence concerning Helmholtz's own thoughts in respect to the requirements of his new force law, based as it was on the vector potential and not on magnetic force? To answer that question, let's first return to Helmholtz's original statement of 1874. There Helmholtz specified a magnetic field that is symmetric about the disk's axis of rotation; he said nothing about the vector potential *per se*, or even about the sources of the field. We saw above that we can produce a trivially symmetric magnetic field – i.e., a constant one – using either of the following two vector potentials (with $\mathbf{B}$ constant of course):

$$\mathbf{A}_{\rm r} = \tfrac{1}{2}(\mathbf{B} \times \mathbf{r}) \quad \text{or} \quad \mathbf{A}_{\rm cm} = \tfrac{1}{2}[\mathbf{B} \times (\mathbf{r} - \mathbf{r}_{\rm cm})]$$

Neither of these two potentials corresponds to a physically-realizable distribution of (closed) currents, simply because the curl of their curl – which represents current – vanishes.[17] Nevertheless, the difference between these two expressions contains a hint that may be historically significant.

We have seen that $\mathbf{A}_{\rm r}$ yields a force on a spinning disk or sphere when the magnetic field parallels the rotation, whereas $\mathbf{A}_{\rm cm}$ does not. If the field is parallel to the angular velocity, then we can rewrite $\mathbf{A}_{\rm cm}$ as $\frac{1}{2}[(\frac{B}{\omega})\boldsymbol{\omega} \times (\mathbf{r} - \mathbf{r}_{\rm cm})]$. The expression $\boldsymbol{\omega} \times (\mathbf{r} - \mathbf{r}_{\rm cm})$ is just the linear velocity $\mathbf{v}$ at the circumference of our rotating sphere or disk. Here, then, the vector potential circulates symmetrically about the disk's axis, while the corresponding magnetic field parallels the axis.[18]

Consider any given radius of the rotating disk. No matter what the position of the radius may be at any given moment, it always sees precisely the same value $\mathbf{A}_{\text{cm}}$ because it has always the same velocity $\mathbf{v}$. And here we perhaps spy a clue to Helmholtz's reasoning during the year after his remark was printed. Suppose we assume that the vector potential is produced by currents that are concentric to, and symmetric about, the disk's axis. In such a case as well, the rotating radius will always see the same potential. It is not a difficult leap from the symmetry of an $\mathbf{A}_{\text{cm}}$ that produces a constant magnetic field to the potential (call it $\mathbf{A}_{\text{symcurr}}$) that is produced by axially-symmetric currents proper. Neither $\mathbf{A}_{\text{cm}}$ nor $\mathbf{A}_{\text{symcurr}}$ will produce any *emf* in the rotating disk or sphere, and for precisely the same reason.

We may now fruitfully examine the consequences of the fact that the *emf* will vanish only when the vector potential is axially symmetric. Specifically, suppose that $\mathbf{A}$ has the general, axially-symmetric form $h(\rho)\mathbf{e}_\varphi$ where $h(\rho)$ depends solely on the distance from the central axis. Here the cylindrical-coordinate ρ specifies the distance to a given point from the z axis, while φ specifies the angle of ρ in a plane orthogonal to z. Then the magnetic field $\mathbf{B}$ becomes $[(h + \rho h')/\rho]\mathbf{e}_z$, and $\mathbf{F}_{\text{emf}}$ vanishes.[19] This magnetic field is itself axially symmetric (although orthogonal to its vector potential), so we have now found a situation that corresponds directly to Helmholtz's requirement and claim in 1874. The point that Helmholtz seems to have missed is this: namely, that $\mathbf{A}$ fields which are not themselves axially-symmetric can nevertheless generate $\mathbf{B}$ fields that are, with non-zero *emf*s resulting thereby. One such $\mathbf{A}$ field, for example, is $h(\rho)\mathbf{e}_\varphi + \varphi\mathbf{e}_\rho$. The corresponding magnetic field is then $[(h + \rho h')/\rho - 1]\mathbf{e}_z$, which is itself axially-symmetric, but the *emf* no longer vanishes, becoming in fact $-\omega\mathbf{e}_\rho$.

Hertz's magnetization potential for the earth is just another example, albeit one in which the magnetic field is symmetric about the spin axis by virtue of its near uniformity in the neighborhood. This is most simply understood by considering the potential's approximate form, in which we replace the vectors to the object point and to the object's center of mass with the earth's radius. For then we can at once see that the approximate potential (see equation (8)) has the form $\mathbf{B} \times \mathbf{r}$ (see equation (10)), and this, as we have seen, does not abolish the Helmholtz *emf*.

DISAGREEMENT AVOIDED, WITH REMARKS ON MISTAKES, NOVELTY AND PRACTICAL WORK

We began our discussion by pointing to a conflict between Helmholtz and his student Hertz concerning the *emf* that is generated in a spinning object subject to a magnetic field. It's certainly possible that the difference remained unresolved, and that it was perhaps never even recognized at all by either of them. But Helmholtz did not cease working on electrodynamics after the paper containing his claim about the spinning disk was printed. Not at all – he continued to write articles on the subject, and a good deal of related experimental work occurred in his Berlin laboratory. Is there any evidence in this subsequent activity that Helmholtz ever recognized, if only implicitly, that his remarks concerning the spinning disk were problematic?

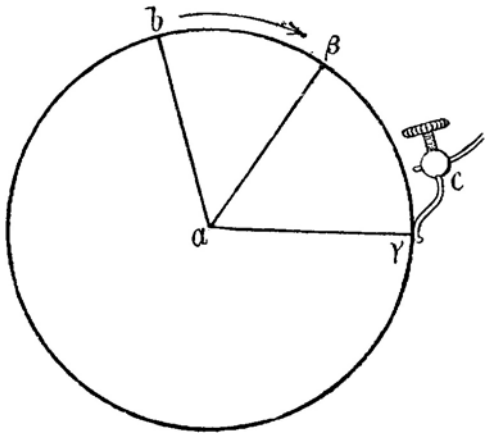

Figure 5. Helmholtz's test for *emf*.

Indeed there is. The very next year Helmholtz made the following remark in a paper concerning experiments done on induction produced by motion in open circuits. Helmholtz wrote:

> Let the endpoint *a* of the conductor (*ab* Fig. 1 [Figure 5 here]) be fixed, *b* however being able to rotate in a circle about *a*, further let the acting magnets and current elements be so arranged that the first of these constitute rotationally-symmetric bodies, whose magnetic axes, as well as whose axes of rotational symmetry, coincide with the normal erected at the midpoint of the circle, while the circuits build concentric circles about this axis. With such an arrangement, the relative position of the radius $a\beta$ with respect to the magnets or the currents is precisely the same as {that of} *ab*; the electrodynamic potential has the same value in both cases, namely zero, and the potential law would as a result have the consequence that in this case no electromotive force will act along it during the course of the rotation of the radius *ab* into position $a\beta$. (Helmholtz, 1875, p. 782)

Here we see that Helmholtz has now recognized the conditions that must be satisfied in order to guarantee the absence of *emf*. The currents that act upon the moving radius all lie in concentric circles having as axis the line about which the arm *ab* in the figure rotates. Further, any magnetic bodies have their axes of magnetization and rotational symmetry along this same axis, and so here too the rotating radius can never see any change in the vector potential.[20] Not only are all magnetic fields axially-symmetric, so too are the corresponding vector potentials.

Clearly, during the time between his remark the year before and this one Helmholtz had understood the need to specify conditions on the symmetry of the vector potential rather than the magnetic field. He was undoubtedly pressed to do so by the demands of an experiment to test the *emf* produced by motion, for that required producing an appropriate physical configuration of currents and magnets. The situation described

here is what Helmholtz had had in mind the previous year, but with a notable difference: the magnetic field in this new situation is not necessarily parallel to the arm's axis of rotation. It is however always axially-symmetric, as are any currents. For the latter reason alone there can be no resultant *emf*.

And so we have solved our apparent conundrum. There is in the end no persistent disagreement between Helmholtz and his student Hertz, because Helmholtz in 1875 altered his inadequate remark of 1874. Hertz's quick and easy use of Helmholtz's equation for *emf* needs no explanation at all, at least insofar as a putative conflict with Helmholtz is concerned. Unlike Helmholtz himself, who had deduced the expression, Hertz had learned it. For him applying the formula to an object spinning in the earth's field was just an exercise in using what he had learned from Helmholtz at a comparatively early stage in his career. For Helmholtz, on the other hand, the new expression for *emf* had come as the result of considerable work trying to build a general foundation for electrodynamics. He had not learned it as a student, and for Helmholtz, its creator, the new formula undoubtedly did not have the character of intuitive directness that, for some time in the early 1880s, it had for the young Hertz.

Four different moments in the production of a novel physical system are nicely illustrated here. Within five years of the system's initial production by Helmholtz we find him applying it incorrectly on paper. Helmholtz had indeed erred within the confines of his own new system. But not for long: the very next year, faced with the concrete demands of a real experimental structure, Helmholtz corrected his error. Then, four years later, the neophyte Hertz, who had learned the new system without having been thoroughly immersed in alternatives to it, applied the scheme almost mechanically, without questioning the elements in it that experiments had begun to make problematic even to its originator.

Novelty, error, error rectified, and finally rote application – these are issues that raise questions for understanding how systems that live both on paper and in the world of material devices evolve. Initially, the specific novelties of Helmholtz's system had little relevance for the contemporary electrodynamics laboratory; there simply weren't any devices that worked with the new forces that Helmholtz had created on paper. Neither were any experimental oddities clarified thereby. More to the point, the world of electrodynamic devices and objects had long been designed and understood on the basis of symmetries that were scarcely compatible with the new system.

Symmetries often constitute critical elements in apparatus design, usually for intensely practical reasons. In lens design, for example, it had always been important to avoid astigmatism, which meant that the lens had to have the same form in any plane section through its center and including its lenticular axis. Moreover, the very manner in which lenses were ground meant that any asymmetries that did occur had to be the result of undesirable and hard to control factors. Lenticular symmetry accordingly represented both abstract desiderata (the avoidance of astigmatism) and practical necessity (lens grinding methods). The motors and induction coils of electrodynamic devices – the existing world that Helmholtz's system had to accommodate – had similar design and practical symmetries built in. Apparatus builders and paper analysts had long concentrated on the magnetic forces that push motors around and that induce

electric currents. It wasn't practical to make, or to calculate, complicated force patterns, and so devices were constructed with extremely simple symmetries. Usually the goal was to keep the magnetic field as uniform as possible within the motor or the induction coil, and to avoid complexity in the winding of coils or armatures. Symmetry of calculation connected to symmetry of design.

Uniform magnetic forces, or forces that are nicely and simply distributed about well-chosen axes, were all that were needed to build working apparatus until Helmholtz's intervention in 1870. Intuitions had been developed over decades for designing and building apparatus that had the right kinds of symmetries to produce the desired actions. Helmholtz himself undoubtedly possessed just this kind of intuitive sense, and it was precisely this that led him into error in 1874, because his new system broke apart the prevailing concordance of symmetries.

It's not generally wise to attribute 'error' to work done long ago, because it is entirely too easy to ignore contemporary factors that make reasonable what was done or said at the time, or to import into past work irrelevant present views. But error *per se* certainly can and does exist. It can be recognized at the time, and, even if it isn't, the historian who has mastered the tools and techniques of the era has license to point out mistaken calculations or claims that shed revealing light on what took place. It is not easy to specify a precise set of rules that might govern the act of error-excavation (and in itself error is not perhaps intrinsically interesting), but at least this much should hold true: the error-excavator must be reasonably certain that the error-maker could have been persuaded to acknowledge and to correct his mistakes had they ever been pointed out.

In Helmholtz's case there is no doubt that he would have acknowledged error, because he in fact did so (albeit implicitly) the very next year by altering his specification for a symmetry that would abolish the electromotive forces. During the year between 1874 and 1875 Helmholtz recognized that intuitions based on force symmetries did not work for his scheme. Symmetries had rather to apply to the vector potential than to the magnetic force that arises by taking its curl, which meant that apparatus had to be designed by arranging the wires themselves according to the desired patterns. Intuitions about symmetric forces had to be replaced by intuitions about symmetrically-placed wires.

This is particularly significant when we recognize that Helmholtz's system worked entirely and directly with entities that formed the tangible electrodynamic workplace. His scheme did not base itself upon electric particles, as did his rival Wilhelm Weber's (and many others in Germany at the time), nor did it work at a fundamental level with force fields, as the British did. It spoke instead of wires that carried currents, or of electrically-polarized dielectrics, or of magnetic bodies. These were its primordial elements, at least in the 1870s, and despite the certain fact that Helmholtz did think that something more fundamental might lie hidden beneath the tangible world, he did not for the most part build his theories at a deep level on conjectures about the invisible realm.

Although our subject has not been a large one since it does not involve many people over a long period of time, it does have broader implications than might seem

to be the case because its argument runs counter to contemporary historical trends that resolutely deny the existence of anything beyond the purely local – of beliefs and behaviors that transcend immediate circumstances and that may hold across national, cultural, and economic boundaries. Much history of science today sees all events as irremediably local, as having no counterparts among other people, at other times, and in different places. Innumerable articles have been written in recent times with the adjective 'local' prominently displayed for admiration in title or body to show that the author does not adhere to the disreputable notions of unity or generality. Heterogeneity in society is, no doubt, morally and socially salutary. The last century provides too many examples of what happens when passions for homogeneity govern life and desire. But, to state what should be trivially obvious, attempts to achieve internal consistency and general applicability in technical systems do not necessarily have much in common with attempts to impose social uniformity by tyrants or fanatics.

The events that we have examined are certainly 'local' in the sense that they took place in particular places, at specific times, and among certain people. And they are local even in their express content, since the peculiarities of Helmholtz's electrodynamics were pursued mostly in Berlin. But in a broader sense much is entirely general here. Helmholtz erred in thinking that it was sufficient to specify a symmetry for the magnetic force, and he later knew as much. Hertz correctly and even mechanically carried out an internally-consistent computation based on Helmholtz's system. It is entirely reasonable to assert that Hertz in 1879, but not Helmholtz in 1874, worked correctly and without error. Locality in our case pertains rather to the specifics of Helmholtz's system, which were certainly not shared by many of his German or British contemporaries, than to the pragmatics of calculation or even of instrumentation. It would have been quite possible for the Maxwellian J.J. Thomson, e.g., to uncover Helmholtz's 1874 error, to follow Hertz's 1879 calculation, and to see exactly how Helmholtz's system had correctly to be applied.

Moreover, Helmholtz fully intended that his electrodynamics should be uniquely correct – that all others would have to fall in some way or other under its sway or else be abandoned altogether. In this he was no different from any of his contemporaries, who however held different views as to the best system to adapt, or for that matter from any mathematically or experimentally oriented investigator since antiquity. Views can be nuanced, and often have been, concerning such things as whether a particular scheme is physically as well as mathematically significant, or even whether mathematics can be used at all. But I do not know of anyone who has ever maintained that two systems or computations, each of which claims to treat essentially the same physical domain in similar ways, and which have conflicting empirical consequences, can both be correct.

APPENDIX: HELMHOLTZ'S POTENTIAL AND CORRESPONDING FORCES

The close links between Helmholtz's electrodynamics and energy considerations have been discussed several times by historians, as have his deductions of the corresponding electromagnetic forces.[21] Nevertheless, it is worthwhile reproducing as closely as

possible Helmholtz's own analyses in order to capture the full flavor of his theory in the manner that he intended. We will rely on previous historical work, including my own, but will diverge from it in presentation and detail in order to adhere closely to Helmholtz.

Helmholtz's own theory of electrodynamics was presented in a series of 11 papers published from 1870 through 1881. Two among these developed the system in elaborate detail, specifically 1870b and 1874. The 1870 paper developed the consequences of Helmholtz's generalized electrodynamic potential, in particular (as its title suggests) for currents in conductors at rest, but also (and importantly) for dielectrics. Here Helmholtz was not concerned with either the mechanical force that acts to move current-bearing bodies, or the electromotive force engendered by changes in the configuration of systems in which they exist. In response to a series of intense criticisms by, among others, Wilhelm Weber, Eduard Riecke and Carl Neumann in Germany, and Joseph Bertrand in France, Helmholtz carefully worked out the forces implied by his theory.

Although part of Helmholtz's purpose was to consider the most general possible form for a potential function that would be compatible with the generally accepted laws that govern closed circuits, we will here limit our considerations to that part of the potential which is given by a generalization of the expression developed by Franz Neumann. This expression was originally developed solely for linear, closed circuits. One of Helmholtz's major assumptions was that the elements in the Neumann integral could be considered independently, thereby extending the expression to open circuits. In addition, Helmholtz examined three-dimensional currents, to which we will here limit our own considerations.[22]

We begin with the electrodynamic 'potential' that two three-dimensional current distributions establish when one at least of them forms a closed system:[23]

$$P = -\frac{A^2}{2}\int_r\int_{r'}\frac{\mathbf{C}\cdot\mathbf{C}'}{|\mathbf{r}-\mathbf{r}'|}\,\mathrm{d}^3r'\mathrm{d}^3r \tag{18}$$

In this expression for P, the integrations both occur over all space, which counts each pair of volume elements $\mathrm{d}^3r'\mathrm{d}^3r$ twice. If, as Helmholtz remarks (Helmholtz, 1874, p. 732) the currents occur in physically separated conductors, and the integrations each occur over only one set, then the factor of $\frac{1}{2}$ may be dropped. In addition, the currents $\mathbf{C}$, $\mathbf{C}'$ are *fluxes*: that is, they represent the quantity of charge per unit time per unit area that flows in a given direction. Helmholtz's generalization to three-dimensions of the procedure established originally by Franz Neumann then yields forces according to the following rules.

The ponderomotive force – the force that moves a body physically – is (following Helmholtz's sign convention) to be found from the negative gradient of this function, with the operator affecting only the locus of the body on which the force acts. Helmholtz accordingly set the negative variation of the potential function equal to the product of the force sought by the variation in position of the object. If the loci of points in the body carrying current $\mathbf{C}$ are specified by vectors $\mathbf{r}$, then the force $\mathbf{F}_{\mathrm{pmf}}$ that would act on an element d^3r as a result of its change in position from $\mathbf{r}$ to $\mathbf{r}+\boldsymbol{\delta}\mathbf{r}$

must accordingly satisfy variational equation (19):

$$\int (\mathbf{F}_{\text{pmf}} \cdot \boldsymbol{\delta}\mathbf{r})\, \mathrm{d}^3 r + \delta_{\text{r}} P = 0 \tag{19}$$

The subscript 'r' in δ_{r} indicates that the displacement of the object is completely arbitrary. Note that $\mathbf{F}_{\text{pmf}}$ depends only upon the configuration of the system and the magnitudes of the currents.

Other forces, called *electromotive* (or *emf*), may also exist that act to change the magnitudes of the currents themselves. With Helmholtz we consider the *emf* that would act on a unit current in a given direction. Such a force is given by the positive rate of change with time of the potential, with the proviso that the *emf* must not depend upon the amount of charge per unit time (that is, the *linear* current) which flows through the object being acted upon. In the case of three-dimensional currents, we construct an appropriate variational equation purely formally by taking the scalar product of the *emf* with whatever current flux $\mathbf{C}$ exists at its locus, and then setting the result equal to the time-rate of change of the potential function. We will subsequently impose the condition that the resulting *emf* must be independent of linear current. This gives equation (20):[24]

$$\left(\int_r \mathbf{F}_{\text{emf}} \cdot \mathbf{C}\, \mathrm{d}^3 r \right) \delta t - \delta_t P = 0 \tag{20}$$

The subscript 't' in δ_t indicates that the change in the potential is calculated over an arbitrary increment of time. As we will see, the variation in the position of the object during this time interval is not arbitrary: it is determined by the object's velocity, and the corresponding variation represents the change as seen by the moving object.

To facilitate computation Helmholtz in 1870 had introduced a vector $\mathbf{U}$, which allowed him to express the potential in a manner that provided in the end a compact representation of the forces:[25]

$$\mathbf{U}(\mathbf{r}) = \int_{r'} \frac{\mathbf{C}'(r')}{|\mathbf{r} - \mathbf{r}'|}\, \mathrm{d}^3 r' \rightarrow P = -\frac{A^2}{2} \int_r \mathbf{C} \cdot \mathbf{U}\, \mathrm{d}^3 r \tag{21}$$

The constant A in (18) and (21) is fundamental in Helmholtz's electrodynamics, but our discussion here does not depend upon it, and so it has been suppressed below for notational simplicity. Note that in this form the energy P depends directly on the properties of what we shall now call the vector potential $\mathbf{U}$. Since everything in Helmholtz's electrodynamics follows from the basic energy expressions, intuitions about how to set up exemplary problems must be developed about current and potential, and not about the resulting forces, since the forces are derivative, not fundamental, quantities.

Helmholtz worked as follows.[26] Begin with the general expression for $\mathrm{d}P$, which contains both $\mathbf{C}$ and $\mathbf{C}'$. Consider an infinitesimal portion of current-bearing material, the volume $\mathrm{d}^3 r$ of the element being $(\mathrm{d}\sigma)(\mathrm{d}r)$. Choose $\mathrm{d}\mathbf{r}$ such that it is parallel to the current flux $\mathbf{C}$ in our element. We may then write the product $\mathbf{C}\, \mathrm{d}^3 r$ in the equivalent form $(C\, \mathrm{d}\boldsymbol{\sigma})\, \mathrm{d}\mathbf{r}$. As a result, the contribution $\mathrm{d}P$ that this element will make to the

entire potential will be:[27]

$$\mathrm{d}P = -(C\,\mathrm{d}\sigma)\,\mathrm{d}\mathbf{r}\cdot\mathbf{U} \tag{22}$$

Since the variation is done without any consideration of the linear current $C(\mathrm{d}\sigma)$, we can now set this product to one and ignore it altogether.

Return to equation (20), and consider the contribution to the entire variation $\delta_t P$ that comes from the circuit element $\mathbf{C}\,\mathrm{d}^3r$, which must now be set to $\mathrm{d}\mathbf{r}$ in the variation for the *emf* as well:

$$\boldsymbol{\delta}_t(\mathrm{d}P) = (\boldsymbol{\delta} t)\mathbf{F}_{\mathrm{emf}}\cdot\mathrm{d}\mathbf{r} \tag{23}$$

From equation (22) we can calculate the variation of the element $\mathrm{d}P$ in terms of $\mathbf{U}$:

$$\delta_t(\mathrm{d}P) = -\delta_t(\mathbf{U}\cdot\mathrm{d}\mathbf{r}) \tag{24}$$

Consequently we have:

$$(\delta t)\mathbf{F}_{\mathrm{emf}}\cdot\mathrm{d}\mathbf{r} = -\delta_t(\mathbf{U}\cdot\mathrm{d}\mathbf{r}) \tag{25}$$

Helmholtz had now to compute the change that arises when the affected object moves in relation to the external currents, and when the external currents are themselves allowed to change *in situ*, with the virtual displacement of the object occurring as a result solely of its motion with a velocity $\mathbf{v}$ during an infinitesimal time. A modern procedure can be used greatly to simplify the computation, but it is historically instructive explicitly to follow Helmholtz's own route.[28]

Let's consider separately the two parts into which the variation divides. The first part represents the change in $\mathbf{U}$ that is seen by a point fixed in the element when the element moves from a place where $\mathbf{U}$ has one value to a place where its value is different, together with the temporal change in $\mathbf{U}$. The second part of the variation represents the change in the value of $\mathbf{U}\cdot\mathrm{d}\mathbf{r}$ that occurs as a result of the alteration in the element's length. Hereafter italic boldface ($\boldsymbol{U}$) represents a vector as seen by a point that is fixed in the element:

$$\delta_t(\mathbf{U}\cdot\mathrm{d}\mathbf{r}) = \underbrace{(\delta_t\boldsymbol{U})\cdot\mathrm{d}\mathbf{r}}_{1} + \underbrace{\mathbf{U}\cdot\delta_t(\mathrm{d}\mathbf{r})}_{2} \tag{26}$$

Here $\boldsymbol{\delta}_t(\mathrm{d}\mathbf{r})$ is the change in the length of the element that occurs as a result of its motion. Since δ_t and d commute,[29] we may replace $\delta_t(\mathrm{d}\mathbf{r})$ with $\mathrm{d}(\delta_t\mathbf{r})$. And since the differential operator d is itself $(\mathrm{d}\mathbf{r})\cdot\nabla$, the second part of the variation may be written:

$$\underbrace{\mathbf{U}\cdot\delta_t(\mathrm{d}\mathbf{r})}_{2} = \mathbf{U}\cdot[(\mathrm{d}\mathbf{r}\cdot\nabla)\delta_t\mathbf{r}]$$

Furthermore, $\delta_t\mathbf{r}$ itself is just the virtual change in $\mathbf{r}$ produced by motion with velocity $\mathbf{v}$ during the time interval δt, i.e., $\mathbf{v}\delta t$:

$$\underbrace{\mathbf{U}\cdot\delta_t(\mathrm{d}\mathbf{r})}_{2} = \mathbf{U}\cdot[(\mathrm{d}\mathbf{r}\cdot\nabla)\mathbf{v}]\,\delta t \tag{27}$$

As for the first part of the variation in (26), we want to express our result in terms of the value of **U** at a fixed point in space – not at a fixed point of the displaced element. However, the $\boldsymbol{U}$ that appears in (26) refers to a specific point in the moved element. $\delta_t \boldsymbol{U}$ must therefore be calculated using the material derivative (following the point) in order to express our results in terms of **U** at a fixed spatial point:

$$\delta_t \boldsymbol{U} = \left[\left(\frac{\partial \mathbf{U}}{\partial t} + \{(\mathbf{v} \cdot \nabla)\mathbf{U}\} \right) \delta t \right] \tag{28}$$

Combining equations (25) through (28) yields (after dropping the common scalar factor δt):

$$\mathbf{F}_{\text{emf}} \cdot \mathrm{d}\mathbf{r} = -\left[\left(\frac{\partial \mathbf{U}}{\partial t} + \{(\mathbf{v} \cdot \nabla)\mathbf{U}\} \right) \right] \cdot \mathrm{d}\mathbf{r} - \mathbf{U} \cdot [(\mathrm{d}\mathbf{r} \cdot \nabla)\mathbf{v}] \tag{29}$$

After manipulation, the right-hand side of equation (29) can be put in a form that contains the scalar product of a vector with d**r**. Equating that vector to $\mathbf{F}_{\text{emf}}$ yields Helmholtz's expression for the electromotive force:

$$\mathbf{F}_{\text{emf}} = -\frac{\partial \mathbf{U}}{\partial t} + \mathbf{v} \times (\nabla \times \mathbf{U}) - \nabla(\mathbf{v} \cdot \mathbf{U}) \tag{30}$$

California Institute of Technology, USA

NOTES

[1] Buchwald (1985, 1993a, 1993b, 1994), Darrigol (1993a, 1993b, 2000), Kaiser (1993).

[2] Archibald (1989), and Olesko (1991, chap. 5). See also Darrigol (2000).

[3] This is the electromotive force due to motion that follows from an electrodynamics that also yields Ampère's original bodily force between circuit-elements carrying electric currents. In what follows we will for the sake of brevity refer to this as the "Ampère" *emf*, although Ampère himself certainly never obtained any such thing since he did not of course discover electromagnetic induction (though he probably did observe it: see, e.g., Hofmann, 1995, chap. 8).

[4] The original reads: "Denken wir uns eine drehende Metalscheibe, schnell um ihre Axe rotirend, und von magnetischen Kraftlinien durchzogen, die der Axe parallel, und rings um die Axe symmetrisch vertheilt sind, so wird der Rand der Scheibe nach dem Ampère'schen Gesetze elektrisch werden, nach dem Potentialgesetze nicht."

[5] In the case of currents $\mathbf{A}(\mathbf{r})$ has the form $\int (\mathbf{C}(\mathbf{r}')/|\mathbf{r} - \mathbf{r}'|)\, \mathrm{d}^3 r'$; in the case of a magnetization **M** the vector **A** becomes $-\nabla \times \int (\mathbf{M}(r')/|\mathbf{r} - \mathbf{r}'|)\, \mathrm{d}^3 r'$. See below, note 7.

[6] The vector $\boldsymbol{\lambda}$ functions in a way that is analogous to that of the scalar potential for electric charge since $\nabla^2 \boldsymbol{\lambda} = -4\pi \mathbf{M}$. By the time that Hertz arrived in Berlin, methods for calculating the force exerted by magnetic distributions were well known, although specific details might differ from author to author. The route from Helmholtz's definition of the auxiliary vector **A** in terms of currents (see equation (21) in the Appendix, where the vector **U** stands for **A**) to the specification of **A** for magnetization was also well known, though again details would differ from author to author. Helmholtz in any case provided the details that Hertz would have needed in this respect, if he did not already know them, in Helmholtz (1870, pp. 617–119).

[7] Expression (5) for the vector potential due to magnetization is nowadays rather unfamiliar. Using the Coulomb gauge we today write (ignoring a sign difference due to Helmholtz's convention) $\int ((\nabla_{r'} \times \mathbf{M}(r'))/|\mathbf{r} - \mathbf{r}'|)\, \mathrm{d}^3 r'$. The two forms are however equivalent: they differ

by a term $\int (\nabla_{r'} \times (\mathbf{M}(r')/|\mathbf{r} - \mathbf{r}'|))\, d^3r'$, and this vanishes on integration over all space if the magnetization is localized. See Jackson (1975), sec. 5.8.

[8] Hertz actually worked from Helmholtz's expression for the force (equation (4)) modified by the introduction of an auxiliary scalar function χ equal to $-\nabla \cdot \boldsymbol{\lambda}$, which facilitated the comparison of Helmholtz's expression for the *emf* with one that had been derived in 1864 on the basis of Weber's electrodynamics by Emil Jochmann (Jochmann, 1864). Assuming that $\nabla^2\boldsymbol{\lambda}$ vanishes – which simply means that the force calculation holds for points that are located outside the magnetization proper – then Hertz could replace $\nabla \times (\nabla \times \boldsymbol{\lambda})$ with $-\nabla\chi$ in the expression $-\mathbf{v} \times (\nabla \times (\nabla \times \boldsymbol{\lambda}))$ for the Ampère term.

[9] Precisely because Hertz computed the force assuming a magnetic dipole located at the earth's center his coordinate system had its origin there as well. We will see in what follows that the choice of coordinate systems is closely connected to the apparent difference between Hertz and Helmholtz.

[10] It removes as well an extremely small term that is linear in the distance from the object's center of mass to the point in it at which the *emf* is to be computed.

[11] The angular velocity will parallel the magnetic force if its equatorial and polar components are in the ratio $3\sin(\Phi)\cos(\Phi)/(2 - 3\cos^2(\Phi))$.

[12] For this particular example, the Ampère *emf* would be $r_s\omega_z B_z$, whereas the Hertz *emf* would be $-(R/2)\omega_z B_z$.

[13] Helmholtz's spinning disk corresponds to a slice of Hertz's sphere taken orthogonally to the sphere's axis of rotation. Whatever consequences correctly hold for Helmholtz's disk will *ipso facto* hold as well for Hertz's sphere by treating the sphere as the limit of a series of stacked disks.

[14] Note again that a field uniform in direction and magnitude is trivially symmetric about its direction and so is clearly a special case of Helmholtz's requirement.

[15] We can easily understand this by remarking that $\mathbf{A}_r$ (equation (11)) implicates the distance $\mathbf{r}$, which implies that the potential seen by a point on the rotating arm must depend upon its angular position since $\mathbf{r}$ does not remain the same during the rotation.

[16] In such a situation the curl of the vector potential (i.e., the magnetic force) would always be tangent to concentric circles having a central axis as their common normal, and it could vary with distance from the origin along, and in the plane normal to, this common central axis. The magnetic field would accordingly circulate about the disk's axis, and to produce this would require something like a closed solenoid that coils around the disk's perimeter.

[17] If, that is, we consider them to be exact and not just approximations that are useful for nearly homogeneous magnetic fields.

[18] Certainly the magnetic field is also (trivially, because constant) symmetric about the axis, but we have already seen that this alone will not guarantee the absence of *emf* (since $\mathbf{A}_r$ also produces $\mathbf{B}$): in addition, the originating vector potential must circulate symmetrically.

[19] For h equal to $B\rho$, with B constant, this reduces to $\mathbf{A}_{cm}$(equation (14)).

[20] This claim in respect to axes of magnetization is not altogether obvious, though Helmholtz certainly recognized it (perhaps as an implication of the possibility of replacing magnetization with closed, bounding currents). It can be demonstrated, as follows. Consider an object spinning with angular velocity ω about the z axis and with its center of rotation located along that axis at a distance h from the origin. At the origin place a magnetic dipole whose axis also lies along z. The velocity of an arbitrary point r in the object, and the vector potential at that point will be:

$$\mathbf{v} = \omega\mathbf{e}_z \times (x\mathbf{e}_x + y\mathbf{e}_y)$$
$$\mathbf{A} = \nabla \times \frac{\mathbf{e}_z}{r}$$

From these we easily discover:

$$\mathbf{A} = \frac{\mathbf{v}}{r^3}$$

And the corresponding 'Helmholtz' force becomes:

$$\mathbf{F} = \mathbf{v} \times (\nabla \times \mathbf{A}) - \nabla(\mathbf{v} \cdot \mathbf{A}) = \mathbf{v} \times \left(\nabla \times \frac{\mathbf{v}}{r^3}\right) - \nabla\left(\frac{v^2}{r^3}\right)$$

We thereby find that $\mathbf{v} \times (\nabla \times \mathbf{A})$ and $\nabla(\mathbf{v} \cdot \mathbf{A})$ are equal to one another, which reduces the force to zero.

[21] Buchwald (1985, 1994) and Darrigol (1993a, 2000).

[22] Buchwald (1994, pp. 25–27) for the forces that arise among linear circuits.

[23] Helmholtz (1874, p. 717) gives the potential for linear circuits, and extends it to three-dimensional ones on pp. 730–731. Helmholtz (1870, p. 568) gives the general expression for the first time.

[24] Helmholtz (1874, p. 744).

[25] Helmholtz (1870, p. 568).

[26] Helmholtz (1874, pp. 742–745).

[27] Note that the factor of $1/2$ disappears on taking the differential. The factor emerges in the first place because otherwise the contribution to the potential from $\mathbf{C} \cdot \mathbf{C}' \mathrm{d}^3 r \, \mathrm{d}^3 r'$ would be counted twice, assuming both integrals to extend over all space. In taking the differential, however, the integration over r is dropped, and the factor of $1/2$ consequently vanishes. Formally, the factor disappears on taking the differential because the total potential, P, is symmetric in the product $\mathbf{C} \cdot \mathbf{C}'$.

[28] See Darrigol (2000, Appendix 5), which indicates that Helmholtz final's result can be obtained by calculating the convective derivative of the vector potential $\mathbf{U}$ under the requirement that the integral of the potential around a curve remains constant under a virtual displacement, i.e., that $\delta_t \oint \mathbf{U} \cdot \mathrm{d}\mathbf{r}$ must vanish. Helmholtz reasoned entirely in terms of a differential element by considering explicitly both the change in the value of a vector that is seen by a point fixed in the element, and the change in the element's length. He did not examine the value of a curve-integral during a deformation.

[29] Because the variation of a differential element of length is equal to the difference between the variations of its endpoints.

REFERENCES

Archibald, T. (1989). "Physics as a constraint on mathematical research: The case of potential theory and electrodynamics." In: *The History of Modern Mathematics*, eds. D. E. Rowe and J. McCleary. Boston: Academic Press.

Buchwald, J. Z. (1985). *From Maxwell to Microphysics: Aspects of Electromagnetic Theory in the Last Quarter of the Nineteenth Century*. Chicago: The University of Chicago Press.

Buchwald, J. Z. (1993a). "Design for experimenting." In: *World Changes: Thomas Kuhn and the Nature of Science*, ed. P. Horwich. Cambridge: MIT Press, pp. 169–206.

Buchwald, J. Z. (1993b). "Electrodynamics in context: object states, laboratory practice, and anti-Romanticism." In: *Hermann von Helmholtz and the Foundations of Nineteenth-Century Science*, ed. D. Cahan. Berkeley: University of California Press.

Buchwald, J. Z. (1994). *The Creation of Scientific Effects: Heinrich Hertz and Electric Waves*. Chicago: The University of Chicago Press.

Darrigol, O. (1993a). "The electrodynamic revolution in Germany as documented by early German expositions of "Maxwell's Theory"." *Archive for History of Exact Sciences* **45**: 189–280.

Darrigol, O. (1993b). "The electrodynamics of moving bodies from Faraday to Hertz." *Centaurus* **36**: 245–260.

Darrigol, O. (2000). *Electrodynamics from Ampère to Einstein*. Oxford: Oxford University Press.

Helmholtz, H. V. (1869). "Ueber elektrische Oscillationen." In: Helmholtz (1882), **1**: 531–536.

Helmholtz, H. V. (1870). "Ueber die Bewegungsgleichungen der Elektricität für ruhende leitende Körper." *J. Reine Angwndte. Math* **72**: 57–129. In: Helmholtz (1882), **1**: 545–628.
Helmholtz, H. V. (1874). "Die elektrodynamischen Kräfte in bewegten Leitern." *Journal für die Reine und Angewandte Mathematik* **78**: 273–324. In: Helmholtz (1882), **1**: 702–762.
Helmholtz, H. V. (1875). "Versuche über die im ungeschlossenen Kreise durch Bewegung inducirten elektromotorischen Kräfte." *Annalen der Physik und Chemie* **98**: 87–105. In: Helmholtz (1882), **1**: 774–790.
Helmholtz, H. V. (1882). *Wissenschaftliche Abhandlungen*. Leipzig: J. A. Barth.
Hertz, H. (1879). *Nachweis electr. Wirkung in Dielectricität*. Hertz MS 245. Science Museum, London.
Hertz, H. (1999). *Die Constitution der Materie. Eine Vorlesung über die Grundlagen der Physik aus dem Jahre 1884*. Berlin: Springer-Verlag.
Hofmann, J. R. (1995). *André-Marie Ampère*. Oxford, UK; Cambridge, MA, USA: Blackwell.
Jackson, J. D. (1975). *Classical Electrodynamics*. New York: Wiley.
Jochmann, E. (1864). "On the electric currents induced by a magnet in a rotating conductor." *Philosophical Magazine* **27**: 506–528 (published first in 1863 in the *Annalen der Physik* and in the *Journal für die Reine und Angewandte Mathematik*).
Kaiser, W. (1993). Helmholtz's instrumental role in the foundation of classical electrodynamics. In: *Hermann von Helmholtz and the Foundations of Nineteenth-Century Science*, ed. D. Cahan. Berkeley: University of California Press, pp. 374–402.
Olesko, K. M. (1991). *Physics as a Calling: Discipline and Practice in the Königsberg Seminar for Physics*. Ithaca: Cornell University Press.

ALLAN FRANKLIN

THE KONOPINSKI–UHLENBECK THEORY OF β DECAY: ITS PROPOSAL AND REFUTATION

When experiment and theory disagree, two possible reasons for the discrepancy are: (1) the theory may be wrong, in the sense that it does not accurately describe the phenomenon observed, whereas the experimental result is correct; and (2) the experimental result may be wrong because the apparatus has not measured the phenomenon accurately,[1] and the theory correct. In the episode I will discuss, the proposal and eventual rejection of the Konopinski–Uhlenbeck (K–U) theory of β decay, both of these difficulties occurred. The situation was further complicated by an incorrect experiment–theory comparison. A discrepancy resulted when correct experimental results were compared to a correct theory, but one which did not, in fact, apply to the phenomena observed.[2]

We begin our story with β decay, the process in which an atomic nucleus emits an electron, simultaneously transforming itself into a different kind of nucleus. In the early 20th century it was thought that β decay was a two-body process (e.g., neutron decays into an electron and a proton). Considerable experimental work on β decay in this period culminated in the demonstration by Ellis and Wooster (1927) that the energy spectrum of the emitted electrons was continuous. The electrons were emitted with all energies from zero up to a maximum, which depended on the particular element.[3] Such a continuous spectrum was incompatible with a two-body process, because applying the laws of conservation of energy and momentum to such a process required the electron to be monoenergetic. In 1930, Wolfgang Pauli proposed that a third particle, one which was electrically neutral, had a very small mass, and had spin $\frac{1}{2}$, was also emitted in β decay. This solved the problem because in a three-body process (i.e., $n \rightarrow p + e + \nu$), the electron is not required to be monoenergetic, but can have a continuous energy spectrum.[4]

FERMI'S THEORY OF β DECAY

Enrico Fermi named the proposed new particle the neutrino, little neutral one, and immediately incorporated it into a quantitative theory of β decay (1934a, 1934b). Fermi assumed the existence of the neutrino, that the atomic nucleus contained only protons and neutrons, and that the electron and the neutrino were created at the moment of decay. He added a perturbing energy due to the decay interaction to the Hamiltonian describing the nuclear system. In modern notation this perturbation is

J.Z. Buchwald and A. Franklin (eds.), Wrong for the Right Reasons, 209–228.

of the form

$$H_{\mathrm{if}} = G[U_{\mathrm{f}}^{*}\Phi_{\mathrm{e}}(r)\Phi_{\mathrm{v}}(r)]O_{\mathrm{x}}U_{\mathrm{i}}$$

where U_{i} and U_{f} describe the initial and final states of the nucleus, Φ_{e} and Φ_{v} are the electron and neutrino wavefunctions, respectively, and O_{x} is a mathematical operator.

Pauli (1933) had previously shown that O_{x} could take on only five different mathematical forms if the Hamiltonian was to be relativistically invariant.[5] These were identified as S, the scalar interaction; P, pseudoscalar; V, polar vector; A, axial vector; and T, tensor.[6] Fermi knew this, but, in analogy with electromagnetic theory, and because his calculations were in agreement with existing experimental results, he chose to use only the vector form of the interaction. He also considered what he called "allowed" transitions, those for which the electron and neutrino wavefunctions could be considered constant over nuclear dimensions. He recognized that "forbidden" transitions would also exist. Konopinski and Uhlenbeck would later note that, "Fermi's theory attributes these differences between equally energetic transitions to differences in the change of angular momentum and parity which occur during each, in analogy to the emission of dipole, quadrupole, etc. radiation in atomic spectra. (Konopinski and Uhlenbeck, 1941, p. 308)."[7] The decay rate of such forbidden transitions would be much reduced from that of allowed transitions.

Fermi calculated that the energy spectrum of the electron for allowed transitions would be

$$P(E)\mathrm{d}E = G^2\,|M|^2\,f(Z, E)(E_{\circ} - E)^2(E^2 - 1)^{1/2}E\,\mathrm{d}E \qquad (1)$$

where E is the energy of the decay electron (in units of $m_{\mathrm{e}}c^2$), E_{o} is the maximum energy allowed, $P(E)$ is the probability of emission of an electron with energy E, and $f(\mathrm{Z}, E)$ is a function giving the effect of the Coulomb field of the nucleus on the emission of electrons. It was later shown that for allowed transitions the energy spectrum in β decay was independent of the choice of decay interaction (Konopinski and Uhlenbeck, 1941).

Fermi also showed that the value of $F(Z,E_{\mathrm{o}})\tau_{\mathrm{o}}$ should be approximately constant for each type of transition, that is allowed, first forbidden, second forbidden, and so forth. $F(Z,E_{\mathrm{o}})$ is the integral of the energy distribution and τ_{o} is the lifetime of the transition. Fermi cited already published experimental results in support of his theory, in particular the work of Sargent (1932, 1933). Sargent had found that if he plotted the logarithm of the disintegration constants (inversely proportional to the lifetime) against the logarithm of the maximum electron energy, the results for all measured decays fell into two distinct groups, known in the later literature as Sargent curves (Figure 1). Although Sargent had originally remarked that "At present the significance of this general relation is not apparent (1933, p. 671)," this was what Fermi's theory required, namely, that $F\tau_{\mathrm{o}}$ is approximately constant for each type of decay. "Fermi connects the two curves of Sargent's well-known graph relating the lifetime and maximum energy with allowed and forbidden transitions (Konopinski

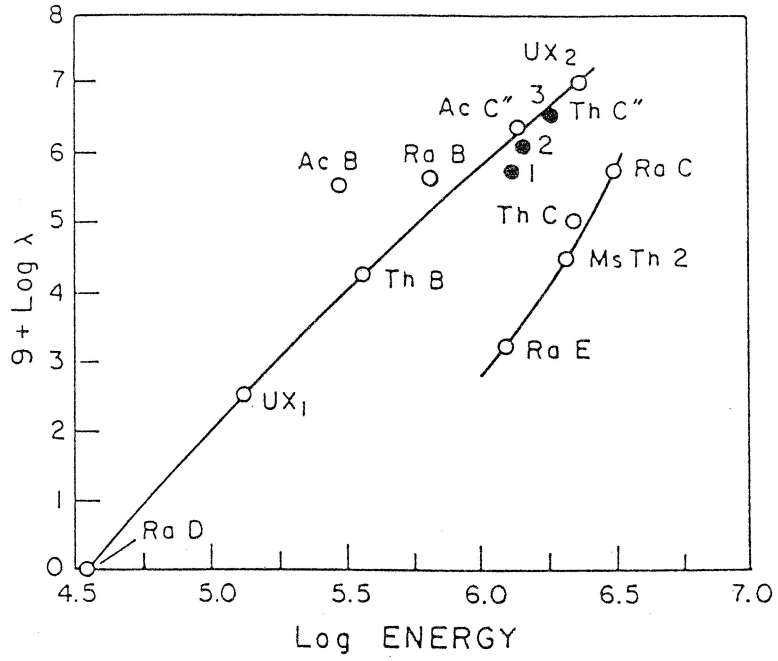

Figure 1. Logarithm of the decay constant (inversely proportional to the lifetime) plotted against the logarithm of the maximum decay energy (Sargent, 1933).

and Uhlenbeck, 1935, p. 7)." The general shape of the observed decay energy spectra also agreed with Fermi's theory.

Although as Konopinski and Uhlenbeck pointed out, Fermi's theory was "in general agreement with the experimental facts," more detailed examination of the decay energy spectra showed that there were discrepancies. Fermi's theory predicted too few low-energy electrons and an average decay energy that was too high. Konopinski and Uhlenbeck cited as evidence the energy spectrum of ^{30}P obtained by Ellis and Henderson (1934) and that of RaE (^{210}Bi) measured by Sargent (1933) (Figure 2). It is clear that the curves labeled FS (Fermi theory) are not a good fit to the observed spectra.

THE KONOPINSKI–UHLENBECK THEORY AND ITS EXPERIMENTAL SUPPORT

Konopinski and Uhlenbeck proposed a modification of Fermi's theory that would eliminate the discrepancy and predict more low-energy electrons. "This requirement could be fulfilled in a simple, empirical way by multiplying [the energy spectrum] by a power of the neutrino energy $E_0 - E$. In terms of the formalism of the Fermi theory, this is accomplished by introducing derivatives of the neutrino wave function in the Ansatz for the interaction energy (Konopinski and Uhlenbeck, 1935, p. 11)."

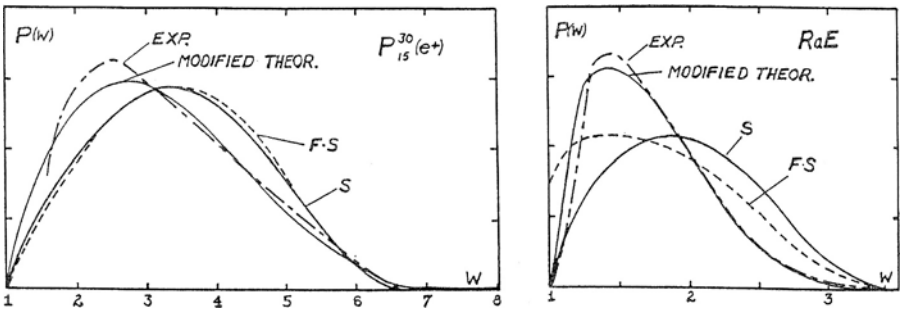

Figure 2. The energy spectra for the β decay of RaE and ^{30}P, respectively. The curve labeled EXP is the experimental result, FS is the Fermi theory, and Modified Theory is the Konopinski–Uhlenbeck theory (Konopinski and Uhlenbeck, 1935).

The energy spectrum they calculated for allowed transitions was

$$P(E)\mathrm{d}E = G^2\,|M|^2\,f(Z,E)(E_\mathrm{o}-E)^4(E^2-1)^{1/2}E\,\mathrm{d}E \qquad (2)$$

This differs from the Fermi prediction by an extra factor of $(E_\mathrm{o}-E)^2$ and thus predicts more low-energy electrons than does Fermi's theory (equation 1). As one can see from Figure 2, the curves labeled K–U (Konopinski–Uhlenbeck theory) fit the observed spectra far better than does the Fermi theory. Konopinski and Uhlenbeck also noted that, "Our modification of the form of the interaction does not of course affect Fermi's explanation of Sargent's law . . . (p. 12)." Their theory also predicted that $F\tau_0$ would be approximately constant for each type of decay.[8]

The Konopinski–Uhlenbeck theory was almost immediately accepted by the physics community as superior to Fermi's theory, and as the preferred theory of β decay. In a 1936 review article on nuclear physics, which remained a standard reference and was used as a student text into the 1950s, Bethe and Bacher, after surveying the experimental evidence, remarked, "We shall therefore accept the K–U theory as the basis for future discussions (Bethe and Bacher, 1936, p. 192)."[9]

The K–U theory received substantial additional support from the results of the cloud-chamber experiment of Kurie et al. (1936). They found that the observed β-decay spectra of ^{13}N, ^{17}F, ^{24}Na, ^{31}Si, and ^{32}P all fit the K–U theory better than did the original Fermi theory. It was in this paper that the Kurie plot, which made comparison between the two theories far easier, made its first appearance. The Kurie plot was a graph of a particular mathematical function involving the electron energy spectrum that gave different results for the K–U theory and for the Fermi theory. It had the nice visual property that the Kurie plot for whichever theory was correct would be a straight line.[10] If the theory did not fit the observed spectrum then the Kurie plot for that theory would be a curve. The Kurie plot obtained by Kurie et al. for ^{32}P is shown in Figure 3. "The (black) points marked 'K–U' modification should fall as they do on a straight line. If the Fermi theory is being followed the (white) points should follow a straight line as they clearly do not (Kurie et al., 1936, p. 377)."

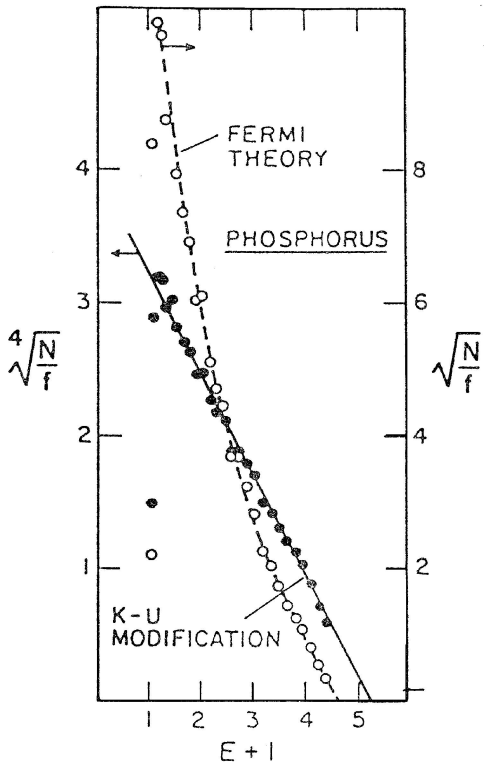

Figure 3. The Kurie plot for the decay of ^{32}P. "The (black) points marked 'K–U' modification should fall as they do on a straight line. If the Fermi theory is being followed the (white) points should follow a straight line as they clearly do not." (Kurie et al., 1936).

For three other radioactive elements, Cl, ^{41}A, and ^{42}K, the observed spectra fit two straight lines for the K–U theory. These were due to complex decays, in which the original nucleus decayed into one or the other of two final states of the daughter nucleus. Kurie et al. concluded:

> The data given above indicate that the Konopinski–Uhlenbeck theory gives a very good account of the distribution curves of the β-rays from the light radioactive elements. We have cited cases of three electron emitters (^{24}Na, ^{31}Si, ^{32}P) and two positron emitters (^{13}N, ^{17}F) where deviations from the theoretical shape of the curve of the observed points are surprisingly small. The spectra of the three elements Cl, ^{41}A, ^{42}K can be resolved into two components each of which is very closely a K–U curve. (Kurie et al., 1936, p. 380).

Interestingly, Kurie et al. had originally obtained results that were in agreement with the Fermi theory. They attributed this incorrect result to the preferential elimination of low-energy decay electrons by one of their selection criteria, one that eliminated events

in which the electron tracks in the cloud chamber showed a visible deflection.[11] Low-energy electrons are scattered more than high-energy electrons, and will therefore have more tracks with visible deflections. The scattering was greatly reduced by filling the cloud chamber with hydrogen rather than the original oxygen.

> Last spring we examined the Fermi theory to see if it predicted the shape of the distributions we were getting and reported favorably on the agreement between the two . . . , with the reservation that we did not feel that our experiments were good enough satisfactorily to test a theory. At that time we were using oxygen as the gas in the chamber as is usual in β-ray work. In measuring the curves we had adopted the rule that all tracks with visible deflections in them were to be rejected. That this was distorting the shape of the distribution we knew because we were being forced to discard many more of the low-energy tracks than the high energy ones. This distortion can be reduced to a very great extent by photographing the β-tracks in hydrogen instead of oxygen. The scattering is thus reduced by a factor of 64 . . . [12]
>
> We found with the hydrogen filled chamber that the distribution curves were more skew than they had appeared with the oxygen filled chamber. This is not surprising: our criterion of selection had been forcing us to discard as unmeasurable a large number of low energy tracks. The number discarded increased as the energy of the track decreased. The apparent concordance between our early data and the Fermi theory was entirely traceable to this because the Fermi distribution is very nearly symmetrical so that when the number of low energy tracks was measured this apparent asymmetry in the experimental distributions was lost. (Kurie et al., 1936, p. 369)

Similar problems affected other cloud-chamber experiments. Paxton solved the problem in a different way, by measuring all tracks of sufficient length. "Because β-ray scattering becomes increasingly serious as the energy decreases, all tracks of sufficient length were measured as well as possible in spite of bad curvature changes, in order to prevent distribution distortion from selection criteria (1937, p. 177)."[13]

The Kurie paper also discussed one of the problems faced by the K–U theory. The maximum decay energy extrapolated from the straight-line graph of the Kurie plot seemed to be higher than the value obtained visually from the energy spectrum (Figure 4). Konopinski and Uhlenbeck had, in fact, pointed this out in their original paper. Kurie et al. found such differences for ^{30}P and for ^{26}Al, but found good agreement for RaE and ^{13}N. With reference to the latter they stated, "The excellent agreement of these two values of the upper limits [obtained from the K–U extrapolation and from nuclear reactions] is regarded as suggesting that the high K–U limits represent the true energy changes in a β disintegration." The evidence of the endpoints was, however, uncertain and did not unambiguously support the K–U theory.

Langer and Whittaker offered a possible explanation for the differences between the two values. They noted that the energy spectrum for an allowed decay approached

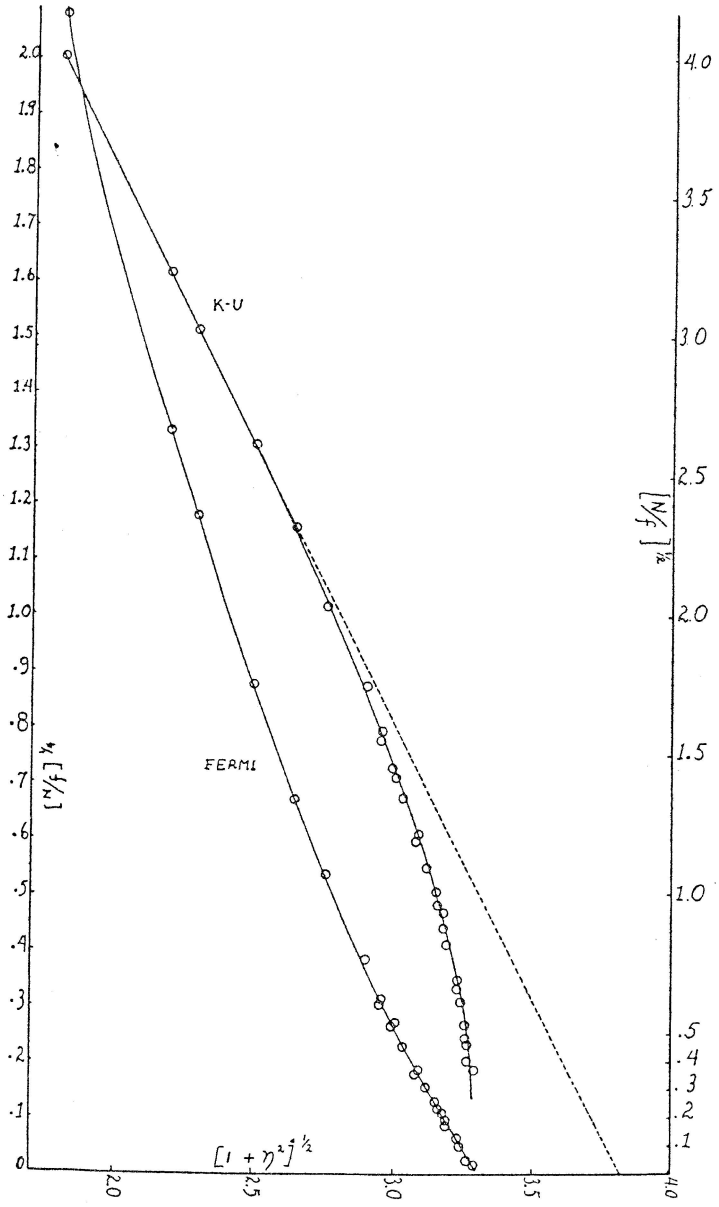

Figure 4. The Fermi and Konopinski–Uhlenbeck curves for the RaE spectrum (Langer and Whittaker, 1937).

the x-axis at a small angle, making it plausible that the actual limit could differ from the visual one (Figure 2). The scattering of the high-energy electrons could result in electron energies that appeared too high. "Experimental difficulties arose because the distribution curve approaches the energy axis gradually and, if the number of beta particles near the end point is not sufficiently great, the true effect may be masked by the natural background inherent in all detecting devices. Moreover, if suitable precautions are not taken, the distribution will have a spurious tail, due to scattered electrons, which approaches the axis asymptotically to much higher energies than the end point" (Langer and Whittaker, 1937, p. 713).

Additional support for the K–U theory came from several further measurements on the radium E (RaE) spectrum, but the support provided for the K–U theory by all of the available evidence was not unequivocal. Richardson (1934) pointed out that scattering and energy loss by electrons leaving the radioactive source could distort the energy spectrum, particularly at the low-energy end.

> The failure of theory to explain the continuous spectrum makes it of interest to obtain all possible experimental information, and although much is now known about the high energy part of the curve, the low energy region has remained obscure owing to certain experimental difficulties. The chief of these has been the contamination of the low energy end of the curve by rays reflected with unknown energy from the material on which the radioactive body was deposited. (Richardson, 1934, p. 442)

There were also other uncertainties in the measurement of the RaE decay spectrum. O'Conor (1937) remarked, "Since the original work of Schmidt in 1907 more than a score of workers have made measurements on the beta-ray spectrum of radium E with none too concordant results." He cited 27 different measurements of the high-energy endpoint energy, for which the largest and smallest values differed by more than a factor of two. By 1940 a consensus seems to have been reached and as Townsend stated, "the features of the β-ray spectrum of RaE are now known with reasonable precision (Townsend, 1941, p. 365)." The future would be different. The spectrum of radium E would be a constant problem.[14]

THE SUPPORT ERODES

The discrepancy between the measured maximum electron energy and that extrapolated from the K–U theory persisted and became more severe as experiments became more precise. In 1937 Livingston and Bethe remarked, "Kurie, Richardson, and Paxton, have indicated how the K–U theory can be used to obtain a value for the theoretical energy maximum from experimental data, and such a value has been obtained from many of the observed distributions. On the other hand, *in those few cases in which it is possible to predict the energy of the beta decay from data on heavy particle reactions, the visually extrapolated limit has been found to fit the data better*

than the K–U value (Livingston and Bethe, 1937, p. 357, emphasis added)." They noted, however, the other experimental support for the K–U theory and recorded both the visually extrapolated values as well as those obtained from the K–U theory.

The difficulty of obtaining unambiguous results for the maximum β-decay energy was illustrated by Lawson in his discussion of the history of measurements of the ^{32}P spectrum

> The energy spectrum of these electrons was first obtained by J. Ambrosen (1934). Using a Wilson cloud chamber, he obtained a distribution of electrons with an observed upper limit of about 2 MeV. Alichanow et al. (1936), using tablets of activated ammonium phosphomolybate in a magnetic spectrometer of low resolving power, find the upper limit to be 1.95 MeV. Kurie, Richardson, and Paxton (1936) have observed this upper limit to be approximately 1.8 MeV. This work was done in a six-inch cloud chamber, and the results were obtained from a distribution involving about 1500 tracks. Paxton (1937) has investigated only the upper regions of the spectrum with the same cloud chamber, and reports that all observed tracks above 1.64 MeV can be accounted for by errors in the method. E.M. Lyman (1937) was the first investigator to determine accurately the spectrum of phosphorus by means of a magnetic spectrometer. The upper limit of the spectrum which he has obtained is 1.7 ± 0.04 MeV. (Lawson, 1939, p. 131)

Lawson's own value was 1.72 MeV, in good agreement with that of Lyman. The difficulties and uncertainties of the measurements are clear. Measurements using different techniques disagreed with one another and physicists may have suspected that the discrepancy might be due to the different techniques used. Even measurements using the same technique differed. Lyman also pointed out that for both ^{32}P and RaE the energy end point obtained by extrapolating the K–U theory was 17% higher than that observed.

Another problem for the K–U theory was that its better fit to the RaE spectrum required a finite mass for the neutrino. This was closely related to the problem of the energy endpoint because the mass of the neutrino was estimated from the difference between the extrapolated and observed endpoints. Measurement of the RaE spectrum in the late 1930s had given neutrino masses ranging from 0.3 to $0.52m_e$, where m_e is the mass of the electron. On the other hand, the upper limit for the neutrino mass from nuclear reactions was less than $0.1m_e$.

Toward the end of the decade, the tide turned and experimental evidence began to favor Fermi's theory over that of Konopinski and Uhlenbeck. Tyler found that the ^{64}Cu positron spectrum observed using a thin radioactive source fit the original Fermi theory better than did the K–U theory. "The thin source results are in much better agreement with the original Fermi theory of beta decay than with the later modification introduced by Konopinski and Uhlenbeck. As the source is made thicker there is a gradual change in the shape of the spectra which gradually brings about better agreement with the K–U theory than with the Fermi theory (Figure 5) (Tyler,

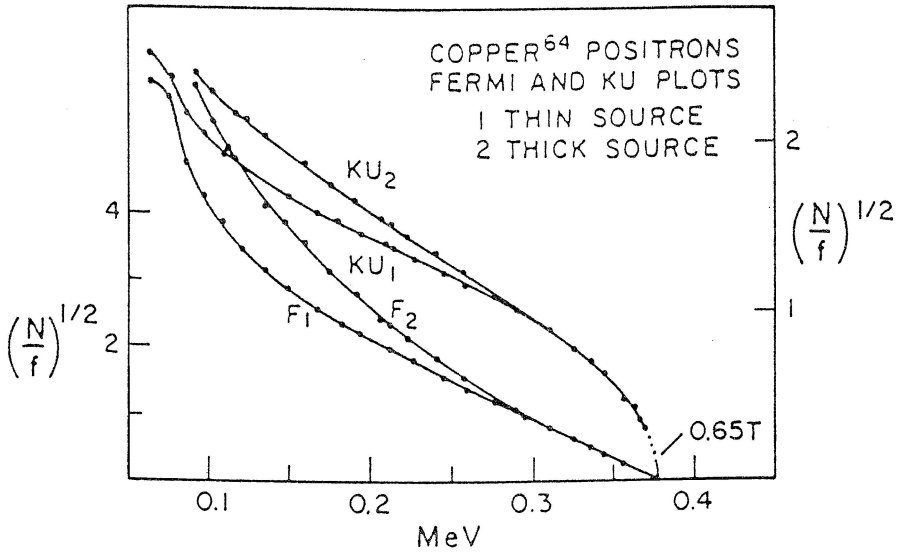

Figure 5. Fermi and K–U plots of positrons from thick and thin ^{64}Cu sources (Tyler, 1939).

1939, p. 125)." Similar results were obtained for phosphorus, sodium, and cobalt by Lawson.

> In the cases of phosphorus and sodium, where the most accurate work was possible, the shapes of the spectra differ from the results previously reported by other investigators in that there are fewer low energy particles. The reduction in the number of particles has been traced to the relative absence of scattering in the radioactive source and its mounting. The general shape of the spectra is found to agree more satisfactorily with that predicted from the original theory of Fermi than that given by the modification of this theory proposed by Konopinski and Uhlenbeck. (Lawson, 1939, p. 131) (Figure 6)

The superiority of the Fermi theory is evident.[15] Richardson's earlier warning concerning the dangers of scattering and energy loss in spectrum measurements had been correct. These effects were causing the excess of low-energy electrons.[16] Compare the later, thin-source results for ^{32}P shown in Figure 6 with the earlier, thick-source results, also on ^{32}P, shown in Figure 2.

There was yet another problem with the evidential support for the K–U theory. This was pointed out by Lawson and Cork in their 1940 study of the spectrum of ^{114}In. Their Kurie plot for the Fermi theory is shown in Figure 7. It is clearly a straight line indicating that the Fermi theory is correct. They pointed out, "*However, in all of the cases so far accurately presented, experimental results for forbidden spectra have been compared to theories for allowed transitions*. The theory for forbidden transitions [for Fermi's theory] has not been published (Lawson and Cork, 1940, p. 994, emphasis

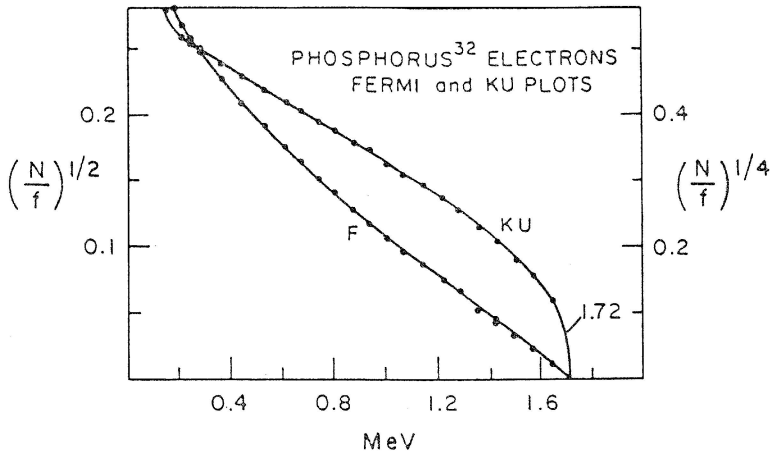

Figure 6. Fermi and K–U plots for electrons from phosphorus ^{32}P (Lawson, 1939).

added).” An incorrect experiment–theory comparison had been made. The wrong theory had been compared to the experimental results. Similar cautions concerning this type of comparison had been made earlier by Langer and Whittaker (1937) and by Paxton (1937). Langer and Whittaker noted that “The K–U plot was made *without considering the fact that radium E is a forbidden transition*.[17] A correction to the theory has been worked out by Lamb and [by] Pollard from which it appears that the extrapolated endpoint is brought into somewhat better although not complete accord with the experimental value (p. 717, emphasis added).” Paxton remarking

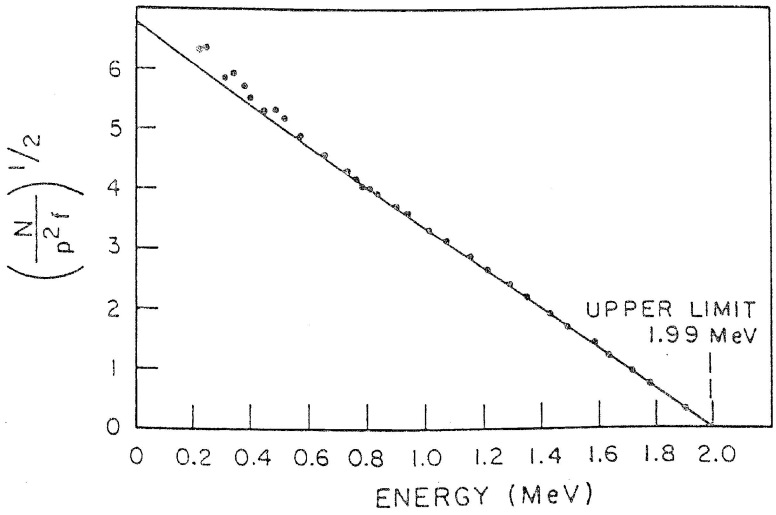

Figure 7. Kurie plot for electrons from the decay of ^{114}In (Lawson and Cork, 1940).

on the discrepancy between the K–U theory and the experimental measurements at the high-energy end of the ^{32}P spectrum stated, "Accordingly this work is best interpreted as indicating a sharp deviation from the K–U relation near the high energy limit. . . . This discrepancy might be eliminated by modifying the K–U formula to apply to a *doubly forbidden type of disintegration* (Paxton, 1937, p. 170, emphasis added)." Little attention seems to have been paid to these comments. The β decay of ^{114}In was an allowed transition, which allowed a valid comparison between theory and experiment. That valid comparison favored the Fermi theory.

KONOPINSKI AND UHLENBECK ASSIST FERMI

The spectrum of forbidden transitions for the original Fermi theory was calculated by Konopinski and Uhlenbeck (1941). They noted that some of the evidence from the β-decay spectra that had originally supported their theory now tended to support the Fermi theory. "The authors made a criticism of Fermi's formula on the basis of a comparison with older experimental data and advanced a modification of the Fermi theory which seemed to represent the data better. The technical improvements in the most recent measurements [including those of Tyler (1939) and of Lawson and Cork (1940), discussed earlier], particularly in eliminating scattering, have withdrawn the basis for the criticism (Konopinski and Uhlenbeck, 1941, p. 309)" They remarked that these new measurements had also confirmed the maximum spectrum energy as derived from nuclear masses. "The so-called K–U modification had led to values that were distinctly too large (p. 309)."

They noted, however, that there were still discrepancies between Fermi's theory and other experimental results so that the choice between the two theories was still unresolved.

> Fermi's formula however still does not represent a great number of observed β-spectra. Many of these disagreements are undoubtedly due to the superposition of spectra, as has lately again been emphasized by Bethe, Hoyle, and Peierls. Nevertheless all the disagreements cannot be explained in this way. The well investigated spectra of RaE and ^{32}P show definite deviations form Fermi's formula (Konopinski and Uhlenbeck, 1941, p. 309)

Konopinski and Uhlenbeck attributed the discrepancies to the fact the RaE and ^{32}P were forbidden decays. Unlike the case of allowed transitions, for which the shape of the energy spectrum was independent of the form of the decay interaction, the shape of the forbidden spectra depended not only on the form of the interaction, but on the value of several nuclear matrix elements. "The first characteristic of these formulas [the corrections to the spectra for forbidden transitions] which stands out immediately is that, in contrast to the allowed formulas, these involve more than one nuclear matrix element (p. 316)." They calculated the spectrum shapes for the various possible forms of the interaction (scalar, pseudoscalar, axial vector, vector, and tensor) separately and noted that "There is, therefore, no a priori reason to expect them to obey the allowed formula (p. 309)."

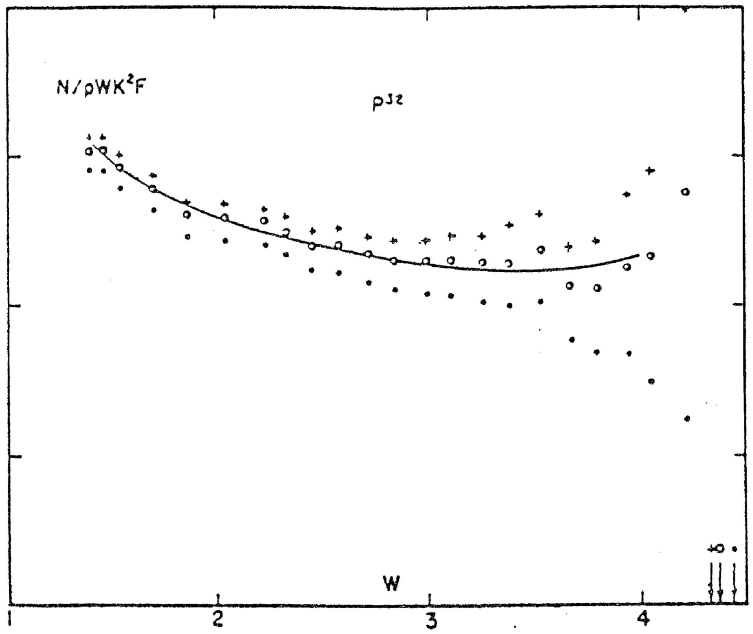

Figure 8. The ratio of the experimental number of electrons emitted by ^{32}P per unit energy to the number expected by Fermi's theory. The data is from Lawson (1939). The blank circles indicate points computed for E_o, the maximum energy = 4.37 (in units of $m_e c^2$), the value given by Lawson. The crosses are for $E_o = 4.33$ and the solid dots for $E_o = 4.44$. The sensitivity to small changes in E_o is clearly shown. The curve represents an average of experimental measurements. The ordinate is on an arbitrary scale, which is the same for all the points. From Konopinski and Uhlenbeck (1941).

Konopinski and Uhlenbeck compared their calculated spectra to the available experimental results. They noted that for forbidden transitions there were only "three reliable β-spectrum measurements namely those of ^{24}Na, ^{32}P, and RaE." They divided the experimentally observed number of electrons by $f(Z, E)(E_o - E)^2(E^2 - 1)^{1}\!/_{2}E$ to obtain the energy dependence of the correction factors, C_i, which they had obtained for each of the various forms of the interaction.[18] They remarked that they were dividing by a quantity that was quite small at both the low- and high-energy end of the energy spectrum. "... it is clear that great accuracy is needed to obtain a reliable correction curve. Small uncertainties in the value of the upper limit [E_o] have great effect (p. 318)." The sensitivity of their results to the value of the endpoint energy is shown in Figure 8.

The first attempted fit was to the ^{24}Na spectrum, which was expected to be a first-forbidden transition. They found that it was impossible to fit the spectrum with any form of the interaction (Figure 9). "It seems impossible however for any of the C_1's [the correction factor for first-forbidden decays] to represent the data, whatever choice is made for the values of the nuclear matrix elements [K–U, 1941, p. 318]."

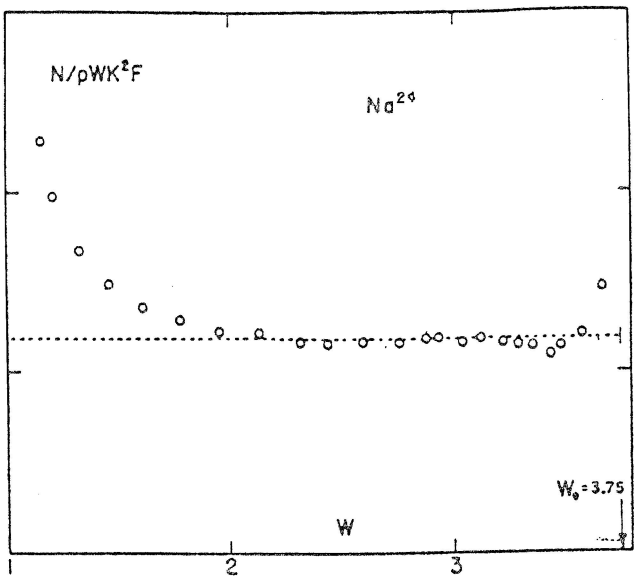

Figure 9. The ratio of the number of electrons emitted by ^{24}Na per unit energy range to the relative number expected according to Fermi's allowed formula. The circles are the experimental points and the straight line is drawn on the hypothesis that the spectrum is complex with the main component of the spectrum having the allowed form as indicated by the experimental points for energy greater than $E = 2$. From Konopinski and Uhlenbeck (1941).

They thought that this was due to the fact that the spectrum was complex, consisting of the superposition of two spectra in which the decay occurs to one of two different final states. They noted that γ rays, with approximately the energy of the difference between the two final state energies in the complex decay, had been observed, making this explanation plausible.[19]

Konopinski and Uhlenbeck reported that they could, however, obtain good fits to the observed spectra of ^{32}P and RaE. They found that for suitable choices of the nuclear matrix elements, which they regarded as plausible, that they could fit the spectra with either a tensor or a vector interaction. They chose the tensor form because that seemed, at the time, to be favored by other results (Figures 10 and 11).

The agreement of the RaE spectrum with the K–U theory was also explained when Konopinski and Uhlenbeck calculated the spectra expected for forbidden transitions.

> The one encouraging feature of the application of the theory [for forbidden transitions] to the experiments is that the decided deviation of the RaE from the allowed form can be at all explained by the theory. . . . *The theory gives a correction factor approximately proportional to* $(E_0 - E)^2$ *for an element like RaE. This accounts for the surprising agreements found by the experimenters between their data and the so-called* K–U *distribution* (Konopinski and Uhlenbeck, 1941, p. 320, emphasis added).[20]

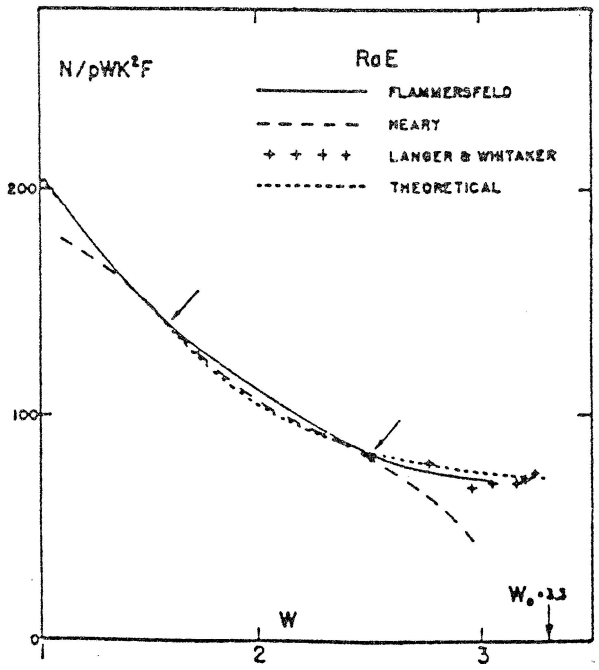

Figure 10. The ratio of the number of electrons emitted by RaE per unit energy range to the relative number expected according to Fermi's allowed formula, for different experiments, as shown. "The dashed line gives $C_{2T}/\Sigma|T_{ij}|^2$ with the ratio of the matrix elements involved in C_{2T} adjusted to fit the experimental curves at the points indicated by the arrows. The adjusted ratio $A_{ij}/T_{ij} = -5.8$, which does not seem implausible." From Konopinski and Uhlenbeck (1941).

In 1943 Konopinski published a review article on β decay. He noted that, "For β-decay theory, next in importance to the confirmation of the general structure of the theory itself, has been the making of a choice between the Fermi and K–U ansätze . . . *The K–U criticism and modification of Fermi's theory seems now to be definitely disproved by the following developments* [Konopinski, 1943, p. 243, emphasis added)."[21] The evidence cited by Konopinski included the evidence of the β-decay energy spectra discussed earlier. "*Thus, the evidence of the spectra, which has previously comprised the sole support for the K–U theory, now definitely fails to support it* [1943 #111, p. 218, emphasis added]." By this time it had also been realized that there were small differences between the two competing theories concerning the values of $F\tau_o$. Fermi theory yielded $F\tau_o$ as a constant, whereas the K–U modification suggested that $(E_o^2 - 1)F\tau_o$ would be constant. By 1943 the evidence favored Fermi theory. Similarly, evidence on K-capture, another form of β decay, although not conclusive, favored Fermi's theory. In the K-capture processes ^{7}Be to ^{7}Li, K–U theory predicted a ratio of approximately 23 for decay to the ground state as compared to decay to an excited state. The experimental result was approximately

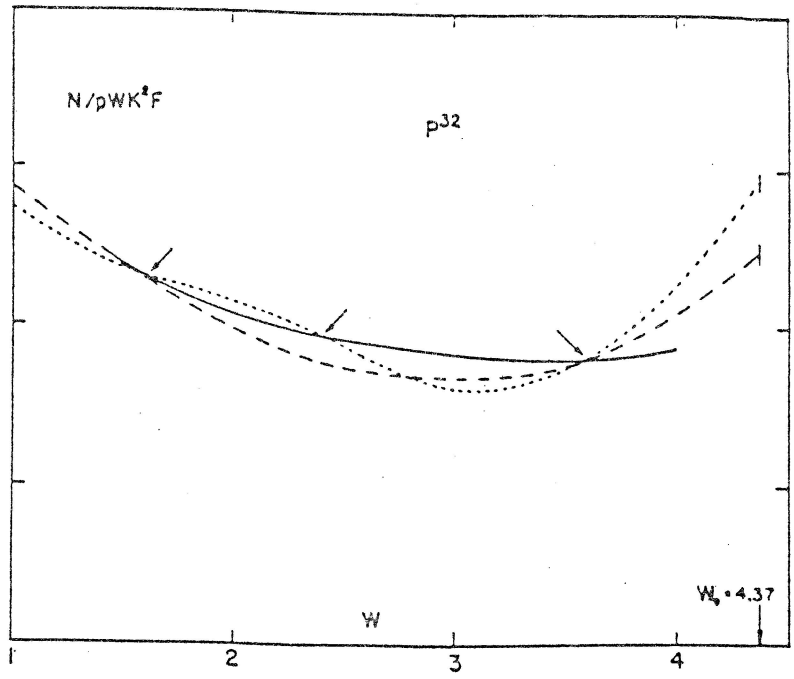

Figure 11. The solid curve is the same experimental data as in Figure 8. The broken lines give $C_{2T}/\Sigma|T_{ij}|^2$, the correction factor for the tensor theory. The dashed line was fitted to the experimental curve at the two pints indicated by arrows, leading to $A_{ij}/T_{ij} = -2.2$. The dotted curve was fitted at three pints with $A_{ij}/T_{ij} = -1.5 \pm i = 3.0$. From Konopinski and Uhlenbeck (1941).

nine. The failure to observe a low-energy peak in β-decay energy spectrum of light elements, predicted by the K–U theory, also argued against it. "No such distribution is known experimentally (Konopinski, 1943, p. 244)."

DISCUSSION

It was only eight years from the original publication of the Konopinski–Uhlenbeck theory and the public, and published, declaration by one of its authors that it was incorrect. It did not accurately describe β decay. Nevertheless, as we have seen, this episode is not an example of bad science, in the methodological or epistemological sense. It is, rather, an example of good science.

Based on the best experimental evidence available at the time, the observed energy spectra from ^{32}P and RaE decay, Konopinski and Uhlenbeck proposed a modification of Fermi's theory that better fit that evidence. Experiment had called for a new theory. A critic might object that their modification was *ad hoc*, that there was no deeper, more fundamental reason for introducing the derivative of the neutrino wavefunction

other than the fact that it "saved the phenomena." That is correct, but one should note that *ad hoc* is not, nor should it be, a four-letter word in science. Much of science is explicitly designed to fit existing evidence. Who would criticize Newton for devising his law of gravity to explain Kepler's laws of planetary motion? In fact, Newton derived the inverse-square distance dependence of his law from Kepler's third law. In addition, one may note that Fermi's use of just the four wavefunctions was not based on any deep theoretical reason. He merely chose a simple form of the interaction, but one that is not in any obvious way simpler than the K–U modification which includes the derivative of the neutrino wavefunction. Both satisfied the criterion of relativistic invariance. In addition, Fermi chose the vector form of the interaction, rather than one of the other mathematical forms, in analogy with electromagnetic interactions.[22]

Experimental work continued and physicists realized that these earlier experimental results were incorrect. It was quickly realized that scattering and energy loss in the radioactive sources used in such experiments had distorted the spectra. Thinner sources were then used, and the new results favored Fermi's theory. Incorrect experimental results are not necessarily an example of bad science.[23] At the time, these were technically very difficult experiments. In the early stages of an experimental investigation it is often difficult to identify sources of background that might mask or mimic the effect one wishes to observe. When physicists realized that scattering and energy loss were a problem, and they did so rather quickly, they took corrective action. The sources were made thinner.

Similarly, the incorrect experiment–theory comparison was eliminated. Konopinski and Uhlenbeck calculated the theoretical spectra needed to solve the allowed-forbidden transition problem. Ironically, the calculations argued that Fermi's theory, rather than their own, was correct. In addition, Lawson and Cork performed an experiment on an allowed transition for which a valid experiment–theory comparison could be made. That too favored Fermi's theory.

"*Thus, the evidence of the spectra, which has previously comprised the sole support for the K–U theory, now definitely fails to support it* [1943 #111, p. 218, emphasis added]." This, along with other evidence, resulted in a reasonable choice of Fermi's theory rather than the K–U modification. It is only those who demand instant rationality, or who deny rationality altogether, who would call this episode bad science. Physicists not only learned from their mistakes, they corrected them.

Department of Physics, University of Colorado, USA

NOTES

[1] There may be a malfunction of the apparatus, or there may be sources of background that mask or mimic the phenomenon to be observed.

[2] Purists may note that other possibilities exist: (1) both theory and experiment may be wrong and the theory does apply to the phenomenon observed; (2) both theory and experiment may be wrong and the theory does not apply to the phenomenon observed; and (3) both theory and experiment may be correct, but the theory does not apply to the phenomenon.

[3] For more detailed history of this see (Franklin, 2000, Chapter 1).

[4] Bohr and others proposed an alternative explanation, that energy was not conserved in β decay. That was rejected on experimental grounds (Franklin, 2000, Chapter 2).
[5] Relativistic invariance is usually considered a requirement for an acceptable theory.
[6] Let U_f and U_i represent the initial and final states of the nucleus and Φ_e and Φ_v be the electron and antineutrino wavefunctions, respectively. Let Q be an operator which, when applied to the wavefunction describing the initial nuclear state, substitutes for it one in which a proton replaces a neutron. Q^* causes the nucleon to make the opposite transition. The five allowable interactions are:

$$\text{Scalar: } S = (U_f^* \beta Q_k U_i)(\Phi_e^* \beta \Phi_v)$$
$$\text{Vector: } V = (U_f^* Q_k U_i)(\Phi_e^* \beta \Phi_v) - (U_f^* \alpha Q_k U_i)(\Phi_e^* \alpha \Phi_v)$$
$$\text{Tensor: } T = (U_f^* \beta\sigma Q_k U_i)(\Phi_e^* \beta\sigma \Phi_v) + (U_f^* \beta\alpha Q_k U_i)(\Phi_e^* \beta\alpha \Phi_v)$$
$$\text{Axial vector: } A = (U_f^* \sigma Q_k U_i)(\Phi_e^* \sigma \Phi_v) - ((U_f^* \gamma_5 Q_k U_i)(\Phi_e^* \gamma_5 \Phi_v)$$
$$\text{Pseudoscalar: } P = (U_f^* \gamma_5 Q_k U_i)(\Phi_e^* \beta\gamma_5 \Phi_v)$$

where α is a vector whose three components are the Dirac matrices. σ differs from the usual Pauli spin matrices only in being doubled to four rows and four columns. β is the fourth Dirac matrix, and $\gamma_5 = -i\alpha_x\alpha_y\alpha_z$. See Konopinski (1943) for details.
[7] Because the neutrino interacts so weakly with matter we can describe it as a particle traveling in free space, which is a plane wave $\Phi_v(r) = Ae^{i(k\cdot r)}$, which can be written as $A[1 + i(k\cdot r) + \cdots]$. The electron wavefunction cannot be written as a plane wave because of the electron's interaction with the Coulomb field of the nucleus. But it can also be expanded as a series of successively smaller terms. For allowed transitions, one keeps only the first, constant terms of the expansion. Forbidden transitions involve subsequent terms, which are much smaller and lead to a reduced decay rate.
[8] This was only approximately correct for the K–U theory, as shown later.
[9] This article was often referred to as the "Bethe Bible".
[10] In a normal beta-decay spectrum the quantity $K_{FERMI} = \{P(E)/[f(Z, E)(E^2 - 1)^{1/2}E]\}^{1/2}$, where E is the electron energy, $P(E)$ is the number of electrons with energy E, and $f(Z, E)$ is a function giving the effect of the Coulomb field of the nucleus on the emission of electrons, is a linear function of E, the energy of the electron (equation (1)). For the Konopinski–Uhlenbeck theory, (equation (2)), $K_{KU} = \{P(E)/[f(Z, E)(E^2 - 1)^{1/2}E]\}^{1/4}$. A plot of K_{FERMI}, or K_{KU}, as a function of E is called a Kurie plot. If the Fermi (K–U) theory is correct then the graph of K (K_{KU}) versus energy, the Kurie plot, will be a straight line. In the original papers $P(E)$ is called N.
[11] One cannot get an accurate measurement of the electron momentum (energy) using an entire track that contains a large deflection. Not only does the momentum change, but the deflection makes fitting the observed track to a single track with constant momentum inaccurate. For a more detailed discussion of selectivity in the production of experimental results see (Franklin, 1998).
[12] The Coulomb scattering of an electron by a nucleus is proportional to Z^2, where Z is the charge on the nucleus. Thus electron scattering from oxygen, $Z = 8$, is 64 times larger than that from hydrogen, $Z = 1$.
[13] Measuring a track with such a curvature change will usually result in an incorrect value of the momentum or energy of the particle. In addition, it will increase the uncertainty of that determination.
[14] For example, Petschek and Marshak (1952) analyzed the spectrum of RaE and concluded that the interaction describing β decay must include a pseudoscalar term. That led to physicists incorrectly concluding that the decay interaction was a combination of S, T, and P. The analysis was later shown to be incorrect. See also the discussion below about the spectrum.
[15] Recall that the correct theory is the one that gives the best fit to a straight line in the Kurie plot.

[16] Earlier in the 20th century physicists had thought that the energy spectrum in β decay consisted of electrons with a single energy, or with several discrete energies. They attributed the observed continuous energy spectrum to the loss of energy by electrons in escaping the radioactive source. Later, it was shown that the spectrum was, in fact continuous. Physicists working on β decay in the 1930s seemed to have forgotten this.
[17] This was determined by looking at the position of the decay on the Sargent curves.
[18] Konopinski and Uhlenbeck had calculated the various values of C_i to be inserted in the calculated energy spectrum $P(E)\mathrm{d}E = G^2 C_i \, |M|^2 \, f(Z, E)(E_\mathrm{o} - E)^2(E^2 - 1)1/_2 E \, \mathrm{d}E$.
[19] Konopinski and Uhlenbeck noted that their conclusion was in disagreement with some existing experimental results, but regarded them as inconclusive because the experiments did not extend to low enough energies.
[20] Nature seems to have been mischievous in the case of RaE. The spectrum of RaE remained a problem into the 1950s. (See discussion in note 14). It is currently believed to be a first forbidden transition, but with a cancellation that makes the spectrum appear to be second forbidden.
[21] A second important issue was the choice of the mathematical form of the interaction. Fermi had chosen a vector interaction. Gamow and Teller (1936) had suggested tensor or axial vector. (For details see Franklin, 1990, Chapter 1). At this time the evidence favored the Gamow–Teller selection rules and the tensor interaction.
[22] Later work showed that the interaction had the form V–A. For details see Franklin (1990).
[23] I suggest that these were in fact valid experimental results. By valid, I mean that they have been argued for in the correct way using epistemological strategies. These strategies include: (1) experimental checks and calibration; (2) the reproduction of artifacts that are known in advance to be present; (3) intervention, in which the experimenter manipulates the object under observation; (4) independent confirmation using independent experiments; (5) elimination of plausible alternative explanations of the result and of plausible sources of error; (6) the use of an independently well-corroborated theory of the apparatus; (7) the use of an independently well-corroborated theory of the phenomena to explain the results; (8) the use of the results themselves to argue for their validity; and (9) the use of statistical arguments. (For details see Franklin, 1990, Chapter 6). Using these strategies does not of course guarantee that the results are correct. There may be unknown sources of background that might mimic or mask the phenomenon to be observed or the strategies may be incorrectly applied. For a discussion of the resolution of discordant experimental results see (Franklin, 1995).

REFERENCES

Alichanow, A. I., A. I. Alichanian and B. S. Dzelepow (1936). "The continuous spectra of RaE and P^{30}." *Nature* **137**: 314–315.
Ambrosen, J. (1934). "Uber den aktiven Phosphor und des Energiesspektrum seiner β-Strahlen." *Zeitschrift fur Physik* **91**: 43–48.
Bethe, H. A. and R. F. Bacher (1936). "Nuclear physics." *Reviews of Modern Physics* **8**: 82–229.
Ellis, C. D. and W. J. Henderson (1934). "Artificial radioactivity." *Proceedings of the Royal Society (London)* **A146**: 206–216.
Ellis, C. D. and W. A. Wooster (1927). "The average energy of disintegration of radium E." *Proceedings of the Royal Society (London)* **A117**: 109–123.
Fermi, E. (1934a). "Attempt at a theory of β-rays." *Il Nuovo Cimento* **11**: 1–21.
Fermi, E. (1934b). "Versuch einer theorie der β-strahlen." *Zeitschrift fur Physik* **88**: 161–177.
Franklin, A. (1990). *Experiment, Right or Wrong*. Cambridge: Cambridge University Press.
Franklin, A. (1995). "The resolution of discordant results." *Perspectives on Science* **3**: 346–420.
Franklin, A. (1998). "Selectivity and the production of experimental results." *Archive for the History of Exact Sciences* **53**: 399–485.

Franklin, A. (2000). *Are There Really Neutrinos? An Evidential History*. Cambridge, MA: Perseus Books.
Gamow, G. and E. Teller (1936). "Selection rules for the β-disintegration." *Physical Review* **49**: 895–899.
Konopinski, E. (1943). "Beta-decay." *Reviews of Modern Physics* **15**: 209–245.
Konopinski, E. and G. Uhlenbeck (1935). "On the Fermi theory of radioactivity." *Physical Review* **48**: 7–12.
Konopinski, E. J. and G. E. Uhlenbeck (1941). "On the theory of β-radioactivity." *Physical Review* **60**: 308–320.
Kurie, F. N. D., J. R. Richardson and H. C. Paxton (1936). "The radiations from artificially produced radioactive substances." *Physical Review* **49**: 368–381.
Langer, L. M. and M. D. Whittaker (1937). "Shape of the beta-ray distribution curve of radium at high energies." *Physical Review* **51**: 713–717.
Lawson, J. L. (1939). "The beta-ray spectra of phosphorus, sodium, and cobalt." *Physical Review* **56**: 131–136.
Lawson, J. L. and J. M. Cork (1940). "The radioactive isotopes of indium." *Physical Review* **57**: 982–994.
Livingston, M. S. and H. A. Bethe (1937). "Nuclear physics." *Reviews of Modern Physics* **9**: 245–390.
Lyman, E. M. (1937). "The beta-ray spectrum of radium E and radioactive phosphorus." *Physical Review* **51**: 1–7.
O'Conor, J. S. (1937). "The beta-ray spectrum of radium E." *Physical Review* **52**: 303–314.
Pauli, W. (1933). "Die Allgemeinen Prinzipen der Wellenmechanik." *Handbuch der Physik* **24**: 83–272.
Paxton, H. C. (1937). "The radiations from artificially produced radioactive substances. III. Details of the beta-ray spectrum of P^{32}." *Physical Review* **51**: 170–177.
Petschek, A. G. and R. E. Marshak (1952). "The β-decay of radium E and the pseusoscalar interaction." *Physical Review* **85**: 698–699.
Richardson, O. W. (1934). "The low energy β-rays of radium E." *Proceedings of the Royal Society (London)* **A147**: 442–454.
Sargent, B. W. (1932). "Energy distribution curves of the disintegration electrons." *Proceedings of the Cambridge Philosophical Society* **24**: 538–553.
Sargent, B. W. (1933). "The maximum energy of the b-rays from uranium X and other bodies." *Proceedings of the Royal Society (London)* **A139**: 659–673.
Townsend, A. A. (1941). "β-Ray spectra of light elements." *Proceedings of the Royal Society (London)* **A177**: 357–366.
Tyler, A. W. (1939). "The beta- and gamma-radiations from $Copper^{64}$ and $Europium^{152}$." *Physical Review* **56**: 125–130.

Archimedes

NEW STUDIES IN THE HISTORY AND PHILOSOPHY OF SCIENCE AND TECHNOLOGY

1. J.Z. Buchwald (ed.): *Scientific Credibility and Technical Standards in 19th and Early 20th Century Germany and Britain.* 1996 ISBN 0-7923-4241-0
2. K. Gavroglu (ed.): *The Sciences in the European Periphery During the Enlightenment.* 1999 ISBN 0-7923-5548-2; Pb 0-7923-6562-1
3. P. Galison and A. Roland (eds.): *Atmospheric Flight in the Twentieth Century,* 2000 ISBN 0-7923-6037-0; Pb 0-7923-6742-1
4. J.M. Steele: *Observations and Predictions of Eclipse Times by Early Astronomers.* 2000 ISBN 0-7923-6298-5
5. D-W. Kim: *Leadership and Creativity.* A History of the Cavendish Laboratory, 1871–1919. 2002 ISBN 1-4020-0475-3
6. M. Feingold: *The New Science and Jesuit Science: Seventeenth Century Perspective.* 2002 ISBN 1-4020-0848-1
7. F.L. Holmes, J. Renn, H-J. Rheinberger: *Reworking the Bench.* 2003 ISBN 1-4020-1039-7
8. J. Chabás, B.R. Goldstein: *The Alfonsine Tables of Toledo.* 2003 ISBN 1-4020-1572-0
9. F.J. Dijksterhuis: *Lenses and Waves.* Christiaan Huygens and the Mathematical Science of Optics in the Seventeenth Century. 2004 ISBN 1-4020-2697-8
10. L. Corry: *David Hilbert and the Axiomatization of Physics (1898–1918).* From Grundlagen der Geometrie to Grundlagen der Physik. 2004 ISBN 1-4020-2777-X
11. J.Z. Buchwald and A. Franklin (eds.): *Wrong for the Right Reasons.* 2005 ISBN 1-4020-3047-9