A century of relativity

Irwin I. Shapiro

Harvard-Smithsonian Center for Astrophysics, Cambridge, Massachusetts 02138

[S0034-6861(99)05302-7]

CONTENTS

I. INTRODUCTION

Except for quantum mechanics—a more than modest exception—relativity has been the most profound conceptual advance in 20th century physics. Both in developing special and general relativity, Albert Einstein's hallmark was to anchor his theory on a few simple but profound principles. The results have provided endless fascination and puzzlement to the general public, and have had an enormous impact on our conceptual framework for understanding nature.

In this brief review, I note the rise and spread of special and general relativity throughout physics and astrophysics. This account is quasihistorical, first treating special and then general relativity. In each case, I consider theory, experiment, and applications separately, although in many respects this separation is definitely not ''clean.'' Responding to the request of the editors of this volume, I have included my personal research in matters relativistic. As a result, the recent is emphasized over the remote, with the coverage of the recent being rather slanted towards my involvement.

II. SPECIAL RELATIVITY

The roots of special relativity were formed in the 19th century; we pick up the story near the beginning of this century.

A. Theory

Hendrik Lorentz regarded his 1904 set of transformations among space and time variables—the (homogeneous) Lorentz transformation—as a mathematical device; he believed in the ether and also in the inability to observe any effects from the motion of light with respect to the ether, which he attributed to dynamical effects caused by motion through the ether.

Henri Poincaré adopted the notion that no motion with respect to the ether was detectable. He also suggested in 1902 that the ether is a hypothesis that might someday be discarded as pointless; in fact, he gave a physical interpretation of ''frame time,'' in terms of the synchronization with light signals of clocks at rest in that frame, as distinct from ether-frame time. Poincaré did not, however, develop a comprehensive theory that postulated a new interpretation of space and time, and he did not discard the concept of an ether. Those tasks were left to Einstein.

The state of Einstein's knowledge of these issues, both theoretical and experimental, and the thinking that undergirded his development of his special theory of relativity in 1905, remain elusive; even historians have failed to reach a consensus. It is nonetheless clear that he was thinking seriously about these issues as early as 1899. He based his new kinematics of moving bodies on two now well-known principles: (1) the laws of nature are the same in all frames moving with constant velocity with respect to one another; and (2) the speed of light in vacuum is a constant, independent of the motion of the light source. He used these postulates to define simultaneity for these (nonaccelerating) frames in a consistent way and to derive transformation equations identical to Lorentz's, but following from quite different underlying reasoning. Einstein also derived a new composition law for the ''addition'' of relative velocities and from it new formulas for the Doppler effect and for aberration.

Poincaré in 1905 showed that the transformation equations formed a group and named it the Lorentz group.[1] The extension of this group to the inhomogeneous group, which included spatial and temporal translations as well, is now known as the Poincaré group.

[1]He did not mention Einstein's paper and may not yet have been aware of it; in any case, he seems never to have referred in print to Einstein's work on special relativity.

Also in 1905, Einstein concluded that the inertial mass is proportional to the energy content for all bodies and deduced perhaps the most famous equation in all of science: $E=mc^2$. Although this type of relation had been proposed somewhat earlier for a specific case, Einstein was apparently the first to assert its universality.

B. Experiment

Special relativity has among its roots the famous Michelson-Morley experiment.[2] This experiment, based on clever use of optical interferometry, found no evidence, at the few percent level, for the effect expected were the Earth moving through a (''stationary'') ether. The round-trip average—both group and phase—speed of light in vacuum has been demonstrated in many experiments this past century to be independent of direction and of the motion of the source. In addition, just recently, analysis of the radio signals from the Global Positioning System (GPS) satellites—all of whose clocks were, in effect, governed by a single atomic standard—yielded a verification of the independence of direction of the one-way speed of light, at the level of about 3 parts in 10^9.

The first experimental tests of special relativity verified the velocity-momentum relation for electrons produced in beta decay. During the 1909–1919 decade a sequence of experiments resulted in verification reaching the 1% level.[3]

The time dilation effect for moving clocks is a major prediction of special relativity. Its experimental verification had to await the discovery of unstable elementary particles, e.g., mesons, whose measured lifetimes when in motion could be compared to the corresponding measurements with the particles at rest (or nearly so). First, in the late 1930s this predicted effect of special relativity was used by Bruno Rossi and his colleagues to infer the at-rest lifetime of mesons from cosmic-ray observations, following a 1938 suggestion by Homi Bhabha.

Another effect—the so-called ''twin paradox''—gave rise to a huge literature over a period of over two decades, before the ''opponents,'' like old generals, just faded away: If twin member B leaves twin member A, who is in an inertial frame, and moves along another world line and returns to rest at the location of A, B will have aged less than A in the interim. Such an effect has been demonstrated experimentally to modest accuracy: the predicted difference in clock readings of a clock flown around the world from one remaining at ''home,'' matched the observed difference to within the approximately 1% standard error of the comparison.

Another of the many verifications, and one of the most important, was of the equivalence of mass and energy. A quantitative check was first made in 1932 via a nuclear reaction by John Cockcroft and Ernest Walton.

C. Applications

After the invention of quantum mechanics, the need to make it consistent with special relativity led Paul Dirac to create the relativistic wave equation for the electron in 1928. This equation eventually led Dirac to propose that its negative-energy solutions describe a particle with the same mass as the electron, but with opposite charge. The discovery of the positron shortly thereafter in 1932 ranks as one of the major discoveries in 20th century physics. Dirac's equation was soon incorporated into the developing formulation of quantum field theory.

Before and after Dirac's work on the relativistic wave equation, relativistic treatments and their refinements were developed for a wide variety of domains such as classical electrodynamics, thermodynamics, and statistical mechanics, while Newtonian gravitation was replaced by an entirely new relativistic theory: general relativity.

On the experimental side, special relativity has also left indelible marks, as witnessed by its important application in the design of high-energy particle accelerators. The equivalence of mass and energy, coupled with developments in nuclear physics, formed the basis for the solution of the previously perplexing problem of the generation of energy by stars. This work reached an apex with Hans Bethe's development and detailed analysis of the carbon-nitrogen cycle of nuclear burning.

A striking contribution of special relativity to the flowering of astrophysics in the 1970s was discovered serendipitously: ''superluminal'' expansion. My group and I used very-long-baseline (radio) interferometry (VLBI) in October 1970 to observe two powerful extragalactic radio sources, 3C279 ($z\approx0.5$)[4] and 3C273 ($z\approx0.2$), to measure the deflection of light by solar gravity (see below). To our surprise, we noticed that the time variation of the 3C279 fringe pattern with the diurnally changing resolution of our two-element, crosscontinental interferometer, matched very well that for a model of two equally bright point sources.

Comparison observations taken four months later, in February 1971, showed an even more dramatic result: these two bright pointlike sources had moved apart at an apparent speed of about 10 c. I developed a simple model of this behavior that showed that if a radio-bright ''jet'' were ejected from a radio-visible ''core'' within a

[2]Although the extent to which this experiment influenced Einstein's development of special relativity is not clear, it is clear that he knew of its existence: A paper by Wien, mentioned by Einstein in an early letter to Mileva Maric, referred to this experiment, allowing one to conclude with high reliability that Einstein was aware of it. In any event, it was definitely a major factor early on in the acceptance of special relativity by the physics community (John Stachel, private communication).

[3]A comprehensive review of these experiments is given in Walter Gerlach's 1933 Handbuch article (volume 20/1).

[4]The redshift z is the fractional increase in the observed wavelength of an electromagnetic signal emitted from an object moving away from the observer.

few degrees of our line of sight at nearly the speed of light, the speed of separation of the two sources on the plane of the sky could match that observed (the derivation is simplicity itself and depends, in essence, only on the speed of light being a constant, independent of the motion of the source). We later became aware of a related analysis having been published in 1969 in the then Soviet Union by Leonid Ozernoy and Vladimir Sasanov, and of Martin Rees' even earlier (1966) corresponding analysis for a uniformly radiating, relativistically expanding spherical shell. After this discovery of superluminal motion,[5] of which there had been earlier hints, many other radio sources were discovered that exhibited similar behavior, albeit with core and jet components having brightnesses different from one another and exhibiting discernible fine structure.

III. GENERAL RELATIVITY

A. Theory

The action-at-a-distance implicit in Newton's theory of gravitation is inconsistent with special relativity. Einstein therefore set out to develop a successor theory of gravitation that would not suffer from this defect. He began this development no later than 1907. As a heuristic guide he used one main principle, the principle of equivalence, which states that the direct effect of mass (''gravitation'') was indistinguishable from uniform acceleration, except for tidal effects: inside an ''Einstein elevator'' the behavior of nature is the same, whether the (small) elevator is at rest in a gravitational field or is uniformly accelerating in a field-free region. Another guide was the principle of general covariance: The form of the field equations for a new theory would be invariant under general (space-time) coordinate transformations. However, this principle waxed and waned as an influence on Einstein's development but ended up consistent with his final 1915 form of the theory.[6]

There have been impressive advances in developing solutions to the field equations of general relativity. The first, still the staple, was the 1916 Schwarzschild—exterior and interior—solution for a spherically symmetric mass distribution, followed soon by several others such as the Reissner-Nordstrom solution for a spherically symmetric charge distribution. For the next several decades mostly approximate, perturbative solutions were developed. For example, Lorentz and Johannes Droste in 1917 and Einstein, Leopold Infeld, and Banesh Hoffmann in 1938 developed expansions to solve the dynamical equations of motion for a collection of mass points.[7] In the solar system, perturbations accurate to the post-Newtonian level are still quite adequate (see below) for comparison with the most exquisitely accurate interplanetary measurements present technology allows, e.g., fractional standard errors of a part in 10^{10} and occasionally smaller for measurements of echo time delays and angular positions, the former by radar and radio transponders and the latter by VLBI. However, technology is poised to allow much more accurate measurements in the next decade so that, at least in the solar system, some post-post-Newtonian effects should be detectable.

Progress in obtaining approximate solutions to the field equations has been dramatic in the last decade due to the development of useful asymptotic expansions and clever numerical techniques coupled with the availability of ever more powerful computers, including, especially, parallel processors. Spurred by the possibility of detecting gravitational waves, physicists have been applying these tools to the complicated analyses of collisions between black holes, and similar catastrophic events, with prime attention being given to accompanying bursts of gravitational radiation (see below). The reliability of these results will remain open to some question, at least until checked wholly independently.

The general relativistic effects of the rotation of massive bodies were first studied in 1918 by Lense and Thirring who noted that the rotation of a central mass would cause the orbit of a test particle to precess about the spin vector of that central mass, an effect dubbed ''frame dragging.'' This rotation would cause the spin vector of a test gyroscope to precess similarly. A major advance in exact solutions encompassed this central-body rotation, but was not discovered until the early 1960s, by Roy Kerr: the ''Kerr metric.'' It pertains to a rotating axially symmetric mass distribution.

Also in the 1960s and continuing in the 1970s, Roger Penrose, Stephen Hawking, George Ellis, and others developed new mathematical techniques to study global properties of space-time, based on the field equations of general relativity. Singularity theorems were developed that described the conditions for ''naked'' singularities, i.e., those not shielded by a horizon. Although such singularities exist mathematically, such as for the Schwarzschild solution with negative mass, many physicists, especially Penrose, believe that in nature singularities would always be shielded. Speculations by John Wheeler, Kip Thorne, and others roamed widely and included discussions of ''worm holes'' which might connect our Universe to others and, perhaps, allow time travel.

[5]During this discovery period, Roger Blandford dubbed this phenomenon ''superluminal'' motion; the appellation immediately took hold within the astronomical community.

[6]Until 1997, it had been generally accepted that David Hilbert had submitted a paper containing a form of the field equations, essentially equivalent to Einstein's, several days before Einstein had submitted his in final form. However, the proofs of Hilbert's paper survive and show in his handwriting, that Hilbert made essential changes to his originally submitted paper that, with other information, substantiate Einstein's primacy in the development of general relativity.

[7]These equations flow directly from the field equations due primarily to the inherent conservation identities; in Isaac Newton's theory of gravity, by contrast, the equations of motion and those for the gravitational potential follow from separate assumptions.

The early treatments of gravitational radiation, including the original one by Einstein, were based on the linearized field equations. It was not until the 1960s that Hermann Bondi, Ray Sachs, and others carried out a rigorous treatment far from the source, establishing that gravitational waves follow from the full, nonlinear, theory of general relativity.

The vexing problem of ''unifying'' the classical theory of general relativity, which stands apart from the rest of fundamental physics, with quantum mechanics remains unsolved, despite enormous effort by extraordinarily talented theorists. This unification remains a holy grail of theoretical physics. The infinities that plagued quantum electrodynamics are not removable by the renormalization techniques that worked so well for the spin-1 photon; they are not applicable to the spin-2 graviton. However, the development of string theory has led many to believe that its unavoidable incorporation of both quantum mechanics and general relativity is a synthesis that will solve the problem of unifying gravitation and the other three known types of interactions (Schwartz and Seiberg, this volume). Unfortunately, tests of predictions unique to string theory or to the newer ''M theory'' are far beyond the grasp of present experimental virtuosity.

A forced marriage of general relativity with quantum mechanics was begun in midcentury. Rather than a unification of the two, quantum-mechanical reasoning was applied on the four-dimensional space-time (Riemannian) background of general relativity, somewhat akin to grafting the former theory onto the latter—a semiclassical approach. The first dramatic result of this development was Hawking's argument in the context of this ''grafted'' model, that vacuum fluctuations would lead to black-body radiation just outside the horizon of a black hole and thence to its evaporation. Jacob Bekenstein's pioneering work, and the later work of others, yielded the corresponding theory of the thermodynamics of black holes, with the temperature of a black hole being inversely proportional to its mass. Thus the evaporation rate would be greater the smaller the mass, and the lifetime correspondingly shorter. For the last stages of evaporation, Hawking predicted a flash of high-energy gamma rays. As yet, no gamma-ray burst has been observed to have the properties predicted for the end stage of black-hole evaporation. Other thermodynamic properties of black holes were also adduced, for example, the entropy of a black hole being proportional to its (proper) surface area. None of these beautiful theoretical results is yet near being testable.

B. Experiment

1. Principle of equivalence

The principle of equivalence in its weak form—the indistinguishability of gravitational from inertial mass—is a profound statement of nature, of interest at least since the 5th century and demonstrated by Newton in the 17th century to hold to a fractional accuracy of about 1 part in 10^2, via observations of the moons of Jupiter and measurements of pendulums made from different materials. At the beginning of this century, using a torsion balance, Baron von Eötvös in Hungary balanced the effect of the rotational acceleration of the Earth and its gravitational effect (and the Sun's), and established the principle of equivalence for a variety of materials to a fractional accuracy of about 1 part in 10^8; this great achievement—given the technology of that time—was published in exquisite detail in 1922, some years after Eötvös' death. Robert Dicke and his group, in the late 1950s and early 1960s used essentially the same approach as Eötvös, but based on a half century more of technology development. Their results, also in agreement with the (weak) principle of equivalence, had an estimated standard error in the fractional difference between the predicted and observed values for aluminum versus gold of ''a few parts in 10^{11}.'' In 1972 Vladimir Brazinsky and Vladimir Panov stated about a tenfold better result from a similar torsion-balance experiment, with the materials being aluminum and platinum.

With these laboratory tests of the principle of equivalence, including the recent and more accurate ones of Eric Adelberger and his colleagues, the equivalence of gravitational and inertial mass has been established for comparisons of a large number of materials. We infer from these null results that the various forms of binding energy, specifically those due to electrical and strong nuclear interactions, contribute equally to gravitational and inertial mass. However, a comparable test of the binding energy associated with the weak nuclear interaction and, especially, with the gravitational interaction is beyond the grasp of these experiments. The latter is more than 10 orders of magnitude too small for a useful such test to be made with a laboratory-sized body. Planetary-sized bodies are needed, since the effect scales approximately with the square of the linear dimension of the objects whose binding energies are to be compared. But a two-body system is ineffective, unless there is an independent means to determine the bodies' masses; otherwise a violation of the principle of equivalence could not be distinguished from a rescaling of the relative masses of the two bodies. A three-body system can yield an unambiguous result and a detailed proposal for such an experiment was made by Kenneth Nordtvedt in 1968. The placement, starting in 1969, of corner reflectors on the Moon by the Apollo astronauts provided the targets for a suitable three-body system: the Sun-Earth-Moon system. Lunar laser ranging (LLR) from the Earth to these corner reflectors initially yielded echo delays of the laser signals with about 10 nsec standard errors (i.e., about 4 parts in 10^9 of the round-trip signal delays). For proper interpretation, such accuracies required the development of elaborate models of the translational and rotational motions of the Moon, far more critical here than for the interpretation of radar data (see below). By the mid-1970s, sufficient and sufficiently accurate data had been accumulated to make a

useful test.[8] With the further accumulation of LLR data, more stringent results were obtained; the latest shows the principle of equivalence to be satisfied to about 1 part in 10^3. Continued decreases in this standard error will require additional modeling, such as representing the reflecting properties of the Moon as a function of aspect to properly account for solar radiation pressure, which is now only about an order of magnitude away from relevancy in this context.

2. Redshift of spectral lines

It is often claimed that the predicted redshift of spectral lines, which are generated in a region of higher magnitude of gravitational potential than is present at the detector, is more a test of the principle of equivalence than of general relativity. But it is perforce also a prediction of general relativity. A test of this prediction was first proposed by Einstein in 1907, on his road to developing general relativity, in the context of measuring on Earth the frequencies of spectral lines formed in the Sun's photosphere. The difficultly here is primarily to discriminate between the sought-after gravitational effects and the contributions from the Sun's rotation (''ordinary'' Doppler effect) and, especially, from motion-related fluctuations. The most accurate determination, in the late 1950s, was in agreement with prediction to within an estimated five-percent standard deviation.

In the early 1960s, soon after the discovery of the Mössbauer effect, Robert Pound realized that he could utilize this effect to measure the shift of the gamma-ray line from Fe^{57} in the Earth's gravitational field. In a carefully designed and brilliantly executed experiment, Pound and Glen Rebka (and later, in 1965, Pound and Joseph Snyder), used a facility somewhat over 20 m high between the basement and roof of Harvard's Jefferson Physical Laboratory, periodically interchanging the location of source and detector to eliminate certain systematic errors. The Mössbauer effect produces an extremely narrow gamma-ray line, allowing Pound and Snyder to achieve a measurement accuracy of 1% of the predicted effect, redshift and blueshift, despite the minute fractional change in gravitational potential over a vertical distance of about 20 m.

In 1976, using hydrogen-maser frequency standards, first developed in Norman Ramsey's laboratory at Harvard, Robert Vessot and his colleagues conducted a suborbital test of this prediction. One hydrogen maser was launched in a rocket and continually compared with two virtually identical masers on the ground; the rocket's apogee was 10 000 km above the Earth's surface. The results of this flight agreed fractionally with predictions to within the 1.4 parts in 10^4 estimated standard error. This accuracy will remain the gold standard at least through the end of this century for this type of experiment.

3. Deflection of light by solar gravity

The ''classical'' test of the predicted deflection of light by solar gravity was first carried out successfully in 1919 in expeditions led by Arthur Eddington and Andrew Crommelin. In these observations, the relative positions of stars in a field visible around the Sun during a total eclipse were measured on photographic glass plates and compared with similar measurements made from plates exposed several months later when the Sun was far from the star field. This approach is fraught with systematic errors, especially from the need to accurately determine the plate scale over the relevant area for each plate. Although repeated a number of times during subsequent solar eclipses, no application of this technique, through 1976, the last such attempt, succeeded in lowering the ''trademark'' standard error below 0.1 of the predicted magnitude of the effect. In 1967, I suggested that this deflection might be measured more accurately using radar interferometry or, more generally, radio interferometry, the former via observations of planets near superior conjunction, and the latter via observations of compact, extragalactic radio sources with the technique of VLBI. The second suggestion bore fruit, with the ground-based standard errors having just recently been reduced to about 1 part in 10^4 by Marshall Eubanks and his colleagues. This result has been achieved through a progression of estimates of almost monotonically decreasing standard errors, from the late 1960s to the present. The next major advance may come from space interferometers, operating at visible wavelengths, and/or from laser signals propagating near the Sun.

4. Time-delay by gravitational potential

Before describing the time-delay experiment, I present some of the background, primarily the development of radar astronomy and my involvement in considering its potential for testing general relativity.

Dicke resurrected experimental relativity in the 1950s from near total neglect, starting as noted above with his work on the refinement of the Eötvös experiment. I became interested near the end of the 1950s, through the advent of radar astronomy, which was being actively pursued at the MIT Lincoln Laboratory, mostly through the foresight of Jack Harrington, then a Division Head there. Powerful radar systems were being developed to track Soviet intercontinental missiles; the systems might

[8]The LURE (lunar ranging experiment) team, sponsored by NASA, at first obtained a result at variance with the predictions of the principle of equivalence, finding the trajectory of the Moon ''off'' by a meter or so, far larger than measurement uncertainties would allow. Independently, my colleagues, Charles Counselman and Robert King, and I had analyzed the same LLR data, which were freely available, with our Planetary Ephemeris Program (see below) and found no violation of the principle of equivalence. We agreed to withhold our results from publication until the LURE team completed a review of its analysis. It turned out that an approximation made in their analysis software was responsible for their non-null result; once fixed, the LURE team's result was consistent with ours and by agreement both groups submitted papers simultaneously to Physical Review Letters, which published them back-to-back.

also be capable, he thought, of detecting radar echoes from planets. Because of the inverse fourth-power dependence of the radar echo on the distance to the target, Venus at its closest approach to the Earth provides echoes about 10^7 times weaker in intensity than those from the Moon, despite the approximately twelvefold larger (geometric) cross section of Venus.[9] The detection of Venus by radar at its furthest point from Earth, near superior conjunction, provides echoes weaker by another factor of about 10^3. Mercury at its greatest distance from the Earth provides echoes threefold weaker again, due to its smaller cross section more than offsetting its distance advantage over Venus at superior conjunction. Despite these depressing numbers, it appeared that over the coming years, the diameter of radar antennas and the power of transmitters could be increased substantially while the noise of receivers might be decreased dramatically. Hence before the first radar echoes were reliably obtained from Venus at its inferior conjunction, I began to think about testing general relativity with this new technique. My first thought, in 1959, was to check on the perihelion advance of Mercury (see next subsection); standard errors of order 10 μsec in measurements of round-trip travel-time between the Earth and the inner planets seemed feasible; the fractional errors affecting such data would then be at the parts in 10^8 level, far more accurate than the corresponding optical data—errors of about five parts in 10^6—and of a different type. Despite the long temporal base of the latter, important for accurate measurement of a secular effect such as the orbital perihelion advance, the increased accuracy of individual echo time (''time-delay'') measurements would allow about tenfold higher accuracy to be achieved in estimating the perihelion advance after a few decades of radar monitoring. Of course, correspondingly detailed modeling was required to interpret properly the results of these measurements. I therefore decided to abandon the time-honored tradition of using analytic theories of planetary motion, carried to the needed higher level of accuracy (I was influenced by my remembrance of being told in the only astronomy course I had taken as an undergraduate about the more than 500 terms in Brown's analytic theory of the Moon's motion). I decided that a wholly numerical approach would be the way to go. With the group I built for the purpose, we—especially Michael Ash and Menasha Tausner—created a model accurate through post-Newtonian order of the motions of the Moon, planets, and Sun, as well as of many asteroids. Detailed models were also required for the rotational and orbital motion of the Moon that involved the few lowest orders of the spherical-harmonic expansion of its gravitational field as well as the second zonal harmonic of the Earth's field. In addition, (elaborate) modeling of the surfaces of the target inner planets was needed—the then major source of systematic error. In principle, the topography of each inner planet's surface can be substantially reduced as a source of error by making repeated radar observations of the same (subradar) point on the planet, each from a different relative orbital position of the Earth and planet. Such opportunities are, however, relatively rare, and scheduling and other realities prevent a bountiful supply of such observations. Again, in principle, high-resolution topographic and reflectivity mapping of an inner-planet surface is feasible via use of a radar system on a spacecraft orbiting that planet; only Venus has so far been mapped at relevant accuracy and resolution, but the practicalities of applying these results to the ground-based radar problem are formidable. The observables—the round-trip signal propagation times—also need to be modeled accurately; they involve the precession, nutation, rotation, and polar motion of the Earth; the geographic location of the effective point of signal reception; and the propagation medium, primarily the interplanetary plasma and the Earth's ionosphere and troposphere. The needed software codes under the rubric Planetary Ephemeris Program (PEP), rapidly reached over 100 000 lines.

The first successful planetary radar observations, of Venus, determined the astronomical unit—in effect the mean distance of the Earth from the Sun—in terms of the terrestrial distance unit, with about three orders of magnitude higher accuracy than previously known, disclosing in the process that the previous best value deduced solely from optical observations was tenfold less accurate than had been accepted.[10]

Before any improvement in determining perihelia advances could be made, indeed before even the first detection of Mercury by radar, I attended an afternoon of presentations c. 1961–1962 by MIT staff on their progress on various research projects, conducted under joint services (DOD) sponsorship. One was on speed-of-light measurements by George Stroke who mentioned something about the speed depending on the gravitational potential. This remark surprised me and I pursued it via ''brushing up'' on my knowledge of general relativity and realized the obvious: whereas the speed of light measured locally in an inertial frame will have the same value everywhere, save for measurement errors, the propagation time of light along some path will depend on the gravitational potential along that path. Thus it seemed to me that one might be able to detect this effect by timing radar signals that nearly graze the limb of the Sun on their way to and from an inner planet near superior conjunction. At the time, however, this idea seemed far out; the possibility of detecting radar echoes from Mercury, the nearest planet at superior conjunction, or even Venus, seemed far off.

In 1964 the Arecibo Observatory, with its 305 m-diameter antenna, was then under development and

[9]The Moon was first detected by radar from Earth in 1946.

[10]This relation, in fact known only to about 1 part in 10^3 from optical data at that early 1960s time, was needed more accurately to ease the problem of navigating interplanetary spacecraft. Over the succeeding two decades the radar value increased in accuracy a further three orders of magnitude.

began radar observations of Venus. Unfortunately, the possibility of testing this prediction of general relativity on echo time delay was not feasible to do at Arecibo because the radar transmitted at a frequency of 430 MHz, sufficiently low that the effects on the echo delays of the plasma fluctuations in the solar corona would swamp any general relativistic signal.

That October the new Haystack Observatory at MIT's Lincoln Laboratory was dedicated. At a party, the day after the birth of Steven, my first child, I was telling Stanley Deser about this new radar facility, when I realized it was going to operate at a frequency of 7.8 GHz, high enough so that the coronal effect, which scales approximately as the inverse square of the frequency, would not obscure a general relativistic signal. I then got quite excited and decided to both submit a paper describing this test and ''push'' for Lincoln Laboratory to undertake the experiment. Given the new—for me—responsibilities of fatherhood, plus the Lincoln review process, the paper was not received by Physical Review Letters until two weeks after the precipitating party. Colleagues at the Laboratory, most notably John Evans and Bob Price, from a detailed analysis of the system parameters, concluded that to do the experiment well we needed about fourfold more transmitter power—a nontrivial need. I went to Bill Radford, then the director of the Laboratory, to plead the case for the more powerful transmitter, pointing out, too, its obvious advantages for the other planned uses of the Haystack radar system. Radford, not knowing how to evaluate my proposed general-relativity experiment, called on Ed Purcell for advice. Purcell said he knew little about general relativity but opined that ''Shapiro has a knack for being right.'' (He was referring, I suspected when the quotation was repeated to me, to my then recent work on the ''artificial ionosphere'' created by the Project West Ford dipoles, whose orbits, greatly influenced by solar radiation pressure, followed my colleagues' and my predictions extraordinarily well.) In any event, Radford called an Air Force general at the Rome Air Development Center and succeeded in getting a $500 000 budget increase for building this new transmitter and its associated microwave plumbing, protective circuits, and other technical intricacies. A nice holiday present for December 1964. A year and a half later, the team of Lincoln Laboratory engineers assigned to this project and led by Mel Stone, completed the new transmitter system; the first radar observations of Mercury under the guidance of Gordon Pettengill and others were made soon thereafter. By early 1968, we had published the first result, a 10% confirmation of the time-delay predictions of general relativity. Controversy both before and after the test centered partly on the observability of the effect (was it simply a coordinate-system mirage?) and partly on the accuracy of my calculations (this latter part lasted for about 30 years).

The experiment, which I labeled the fourth test of general relativity, was refined over the following years, with the standard error reduced to 1 part in 10^3. This accuracy was achieved with essential contributions from Robert Reasenberg and Arthur Zygielbaum in the years 1976–1978 with the four Viking spacecraft that were deployed in orbit around Mars and on its surface.

Until very recently, this accuracy exceeded that from the closely related experiments involving VLBI measurements of the deflection of radio waves; but now the accuracy pendulum has swung decisively toward these latter measurements.

5. ''Anomalous'' perihelion advance

The first inkling that Newton's ''laws'' of motion and gravitation might not be unbreakable came in the mid-nineteenth century with the carefully documented case by Urbain LeVerrier of an anomalous advance in the perihelion of Mercury's orbit, reinforced and refined near the end of that century by Simon Newcomb. Never explained satisfactorily by alternative proposals—is there a planet (Vulcan) or cloud of planetesimals inside the orbit of Mercury; an unexpectedly large solar gravitational oblateness; and/or a slight change in the exponent of Newton's inverse square law?—this advance, as Einstein first showed, followed beautifully and directly from his theory of general relativity. The agreement between observation and theory was remarkably good, to better than one percent, and within the estimated standard error of the observational determination of the anomalous part of this advance at that time. The analysis of the new radar data, alone, now yield an estimate for this advance about tenfold more accurate than that from the several centuries of optical observations.

A main problem has been in the interpretation of the radar—and optical—measurements, in the following sense: How much of the advance could be contributed by the solar gravitational oblateness? Although this secular Newtonian effect falls off more rapidly, by one power of the distance, than does the post-Newtonian general relativistic effect, neither is detectable with sufficient accuracy in the orbit of any planet more distant from the Sun than is Mercury. There are short-term orbital effects that offer a less demanding, but by no means easy, means of discrimination. In any event, the correlation between the estimates of the magnitudes of the relativistic and solar-oblateness contributions to the advance will likely remain high until interplanetary measurement errors are substantially lower. An independent measurement of the oblateness through direct study of the Sun's mass distribution is thus highly desirable. Dicke, who built on earlier ideas of Pascual Jordan, developed a scalar-tensor theory alternative to general relativity, with his (Dicke's) student Carl Brans. This theory had an adjustable parameter, representing, in effect, the relative scalar and tensor admixtures, and could account for a smaller advance. Thus Dicke set out to measure the solar visual oblateness, which could then be used via straightforward classical theory to deduce the gravitational oblateness, i.e., the coefficient, J_2, of the second zonal harmonic of the Sun's gravitational field. In the late 1960s Dicke and Mark Goldenberg using a very clever, but simple, instrument to estimate the Sun's

shape, deduced a value for the visual oblateness that would account for 10% of Mercury's anomalous perihelion advance, thus implying that general relativity was not in accord with observation and that the previous precise accord was a coincidence. The Brans-Dicke theory's adjustable parameter could accommodate this result. These solar-oblateness measurements, in the fashion of the field, were scrutinized by experts from various disciplines resulting in many questions about the accuracy of the oblateness determination: extraordinary claims must be buttressed by extraordinary evidence. The net result was a rejection by a large majority of the scientific community of the accuracy claimed by Dicke for Goldenberg's and his solar-oblateness measurements, leaving their claim of a higher-than-expected value for J_2 unsubstantiated.

More recently, a new field—helioseismology—has been developed to probe the mass distribution of the Sun: optical detection of the oscillations of the Sun's photosphere allows the solar interior to be deeply probed. The net result leaves the agreement between observation and general relativity in excellent accord.[11]

6. Possible variation of the gravitational constant

In 1937, Dirac noticed a curious coincidence, which he dubbed the law of large numbers. It was based on the fact that the ratio of the strengths of the electrical and the gravitational interactions of, say, an electron and a proton—about 10^{39}—was, within an order of magnitude or two, equal to the age of the universe measured in atomic units of time (e.g., light crossing time for an electron). Dirac noted that this near identity could be a mere coincidence of the present age of the universe or could have a deeper meaning. If the latter were true, Dirac reasoned, the relation between gravitational and atomic units should be a function of time to preserve this (near) equality. The most reasonable proposal, he concluded, was to assume that the gravitational constant G decreased with (atomic) time. This proposal lay dormant for several decades. In the 1950s calculations were made of the brightness history of the Sun and its effects on the Earth that might be discernible in the geologic record. These deductions were quite controversial because, for example, of their reliance on (uncertain) aspects of stellar evolution and the difficulty in separating atomic from gravitational effects in determining such a brightness history. More modern calculations of the same type are similarly afflicted. In 1964, in thinking of other possibilities for radar tests of general relativity, I considered the obvious check on any change in G through monitoring the evolution of planetary orbits with atomic time t. Expanding $G(t)$ about the present epoch t_0, I sought evidence for $\dot{G}_0 \neq 0$. Were $\dot{G}_0 < 0$ as would follow from Dirac's hypothesis, then the orbits of the planets would appear to spiral out, an effect most noticeable in the (relative) longitudes of the planets. The main limitation of this test at present is the systematic error due to incomplete modeling of the effects of asteroids. Nonetheless, the radar data are able to constrain any fractional change in G (i.e., constrain $\dot{G}_0/G_0$) to be under a few parts in 10^{12} per year. A similar level of accuracy has been achieved with the LLR data; here the main source of systematic error is probably the modeling of the tidal interaction between the Earth and the Moon as it affects the spiraling out of the orbit of the latter. There have been publications of similar accuracies based on the analysis of the pulse timing data from a binary neutron-star system (see below). This bound, however, is of the self-consistency type in that this effect is very highly correlated with the main orbital effect of gravitational radiation: the two are inseparable; the results for one must be assumed to be correct to test for the other, save for self-consistency. The solar-system tests are free from such a fundamental correlation, but are limited in accuracy for the near future by other correlations, at about the level of the bound already achieved.

7. Frame dragging

One quantitative test of the Lens-Thirring effect has just been published: an apparent verification of the prediction that the orbital plane of a satellite will be "dragged" (precess) around the spinning central body in the direction of rotation of that body. Specifically, in 1998, Ignazio Ciufolini and his colleagues analyzed laser-ranging data for two nearly spherical Earth satellites, Lageos I and II, each with a very low area-to-mass ratio. These authors concluded that the precession agreed with the predictions to 10%, well within their estimated standard error of 20%. There are, however, an awesome number of potentially obscuring effects, such as from ocean tides, that are not yet well enough known to be reasonably certain of the significance of this test. With continued future gathering and analysis of satellite-tracking data from the increasing number of satellites that are designed, at least in part, to improve knowledge of both the static and time-varying contributions to the gravitational potential of the Earth, this Lens-Thirring test will doubtless improve.

A definitive quantitative verification of the effect of frame dragging on orbiting gyroscopes is promised by the Stanford-NASA experiment. This experiment was developed, based on Leonard Schiff's original (1959) suggestion, by William Fairbank, Francis Everitt, and others at Stanford, starting in the early 1960s. The experiment will involve a "drag-free" satellite containing four extraordinarily spherical quartz "golf ball"-sized

[11]The analysis of the Sun's pressure modes, both from the ground-based network of observatories and the space-based, Solar Heliospheric Observatory (SOHO), allows a rather robust estimation: $J_2 = (2.3 \pm 0.1) \times 10^{-7}$, of more than adequate accuracy for the interpretation of the solar-system data. Because of the high correlation between the orbital effects of the solar gravitational quadrupole moment and of the postNewtonian terms in the equations of motion, such an accurate independent determination of J_2 allows "full" use of the solar-system data for checking on the relativistic contributions to the orbital motion.

gyroscopes, coated with niobium, cryogenically cooled, and spun up with their direction of spin monitored by ''reading'' the gyroscope's London moment with superconducting quantum interference devices. The directions of the London moments will be compared to that of a guide star whose proper motion with respect to a (quasi) inertial frame is being determined to sufficient accuracy for this purpose by my group via VLBI, following a suggestion I made to the Stanford team in the mid-1970s. For the orbit and the guide star chosen, the predicted gyroscope precession is about 40″/yr, with the anticipated standard error being 0.″2/yr. This Stanford-NASA experiment will also measure the so-called geodetic precession with at least two orders of magnitude smaller standard error than the 2% value we obtained a decade ago from analysis of the LLR data. This truly magnificent physics experiment is now scheduled for launch in the year 2000.

8. Gravitational radiation

For about the last third of this century, physicists have addressed with great experimental and theoretical virtuosity the problem of detecting gravitational radiation. The pioneer experimenter, Joseph Weber, developed the first cylindrical bar detectors; his claim in the early 1970s to have detected gravitational waves from the center of our Galaxy, despite being wrong, awakened great interest, resulting in a relentless pursuit of this holy grail of experimental gravitational physics. Very significant human and financial resources have been expended in this hunt, which doubtless will eventually be successful and will also provide profound insights into astrophysical processes. But not this century.

The now-classic neutron-star binary system discovered in 1974 by Russell Hulse and Joseph Taylor has exhibited the orbital decay expected from gravitational quadrupole radiation for these objects, which are in a (noncircular) orbit with a period of just eight hours. This decay is a striking confirmation of the general relativistic prediction of gravitational radiation, with the observed changes in orbital phase due to this decrease in period matching predictions to within about one percent.[12] The sensitivity of these measurements to this decay increases approximately with the five-halves power of the time base over which such measurements extend, since the effect of the radiation on orbital phase grows quadratically with that time base, and the effect of the random noise drops as the square root of that base, given that the measurements are spaced approximately uniformly. However, systematic errors now limit the achievable accuracy to about the present level, the chief villain being the uncertainty in the Galactic acceleration of the binary system, which mimics in part the effect on pulse arrival times of the orbital decay.[13]

The larger universe is the hoped-for source of gravitational waves which will be sought by the laser interferometer gravitational-wave observatory (LIGO) and its counterparts in Germany, Italy, and Japan—all currently in various stages of planning and construction. The two LIGO sites, one each in the states of Louisiana and Washington, will each have two 4-km-long evacuated tubes, perpendicular to each other, and forming an ''L''; a test mass is at each far end and at the intersection of the two arms. LIGO will be sensitive to gravitational waves with frequencies $f \geqslant 30$ Hz. The first generation of laser detectors should be sensitive to the difference in strains in the two arms at the fractional level of about one part in 10^{21}. No one expects gravitational waves that would cause strains at this sensitivity level or greater to pass our way while detectors of some orders of magnitude greater sensitivity are being developed for deployment on LIGO and on the other instruments. However, Nature often fools us, especially in the variety and characteristics of the macroscopic objects in the universe. So I personally would not be totally shocked were this first generation of laser interferometer detectors to pick up bona fide signals of gravitational waves. As a counterpoise, note that some of the best gravitation theorists have worked for several decades to conjure and analyze scenarios that might lead to detectable gravitational radiation, and have failed to find any that would likely be detected at this level of sensitivity after, say, several years of monitoring.

C. Applications

General relativity was at first of interest only in a small subfield of physics—aside from the profound impression it made on the psyche of the general public. Still irrelevant for applications in everyday terrestrial life and science,[14] general relativity now provides a key tool in the armamentarium of theoretical—and observational—astrophysicists. It is employed to tackle problems from the largest to the smallest macroscopic scales encountered in our studies of the universe.

1. Cosmology

The first and perhaps still the most important application of general relativity is to cosmology. Einstein,

[12]It is often argued that this detection of gravitational waves is ''indirect'' because we detect only the consequences of the radiation in the orbital behavior. However, one could argue as well that a similar criticism applies to any detection since the presence of the waves must be inferred from observations of something else (e.g., the vibrations of a massive bar or the oscillatory changes in distance between suspended masses as measured by laser interferometers). The key difference is whether we infer the properties of the radiation from its effects on the sources or on the detectors; in the latter cases, of course, the experimenter-observers have much greater control.

[13]The arrival-time data are also rich enough to measure with reasonable accuracy other predicted relativistic effects and to determine the masses of the neutron-star components of the binary; the mass of each is about 1.41 solar masses, in splendid accord with the Chandrasekhar limit (see below).

[14]Except insofar as the Newtonian limit serves us admirably.

thinking that the universe was static, found a corresponding cosmological solution. Since the universe is nonempty, Einstein had to first tamper with his field equations by introducing on the ''geometry side,'' a term with a constant coefficient—the so-called cosmological constant—whose value was not specified by the theory, but which would provide the large-scale repulsion needed to keep a nonempty universe static. Later, after Alexandre Friedmann and Georges Lemaitre exhibited expanding-universe solutions and Hubble presented the first evidence of expansion, Einstein reputedly called his introduction of the cosmological-constant term ''the greatest scientific blunder of my life.''[15] This term did not, however, fade away forever, but was resurrected recently when exploration of the implications of vacuum fluctuation energy on a cosmological scale uncovered the so-called cosmological-constant paradox: for some modern theories the (nonzero) value is about 120 orders of magnitude larger than the upper bound from the observational evidence.

The 1920s provided the main thrust of the program in cosmology for the rest of the century: Under the assumption of a homogeneous isotropic universe, cosmologists attempt to measure the Hubble constant H_0 and the deceleration parameter q_0. Values of these two parameters would provide, respectively, a measure of the size scale (and, hence the age) of the universe and a determination of whether the universe is open, closed, or on the border (and, hence, whether the average mass density of the universe is below, above, or at the ''closure'' density).

In the 1930s, the estimate of H_0 was quite high, about 500 $\mathrm{km\,s^{-1}\,Mpc^{-1}}$, implying an age for the universe of only a few billion years. Even before radioactive dating techniques were able to disclose that the age of the Earth was about 4.5 billion years old, astronomers discovered a serious problem with their method of inferring H_0 from the distances of ''standard candles,''[16] leading, after further revisions, to the conclusion that H_0 was severalfold smaller and the universe correspondingly older. Over the following decades, there was no appreciable improvement in accuracy; however, for the past several decades, there has been a schism among the practitioners: those such as Allan Sandage claiming H_0 to be about 50 $\mathrm{km\,s^{-1}\,Mpc^{-1}}$ and those such as Gerard DeVaucouleurs proclaiming a value of 100 $\mathrm{km\,s^{-1}\,Mpc^{-1}}$, each with estimated uncertainty of the order of 10%. The methods they used depend on the accurate calibration of many steps in the so-called cosmic distance ladder, making it difficult to obtain reliable estimates of the overall errors. With some exceptions, more modern values have tended to cluster between 65 and 75 $\mathrm{km\,s^{-1}\,Mpc^{-1}}$, still a distressingly large spread. Also, new methods have joined the fray, one depending directly on general relativity: gravitational lensing, discussed below.

The pursuit of q_0 has until recently led to no result of useful accuracy. Now a wide variety of techniques has indicated that the universe does not have sufficient mass to stop its expansion, and hence is ''open.'' Most recently, two large independent groups have obtained the tantalizing result from observations of distant ($z \approx 1$) type 1a supernovae that the universe is not only open but its expansion is accelerating. This result is now at the ''two sigma'' level; if confirmed by further data and analysis, it will have a profound effect on theory: Is the cosmological constant nonzero after all and if so, how does one reconcile that result with current quantum-field-theory models? Or, for example, are vacuum fluctuations causing some strange locally weak, but globally strong, repulsion that produces this acceleration? These problems, doubtless, will not be fully resolved until the next millenium.

2. Black holes

Beyond the structure and evolution of the universe on large scales, probably the most profound effect of general relativity on astrophysics in the past century has been through the prediction of black holes. The name was coined by John Wheeler in the early 1960s, but the concept was, in effect, conceived over two centuries earlier by the Reverend John Michell who reasoned, based on Newton's corpuscular theory of light, that light could not escape an object that had the density of the Sun but a diameter 500 times larger. Early in the 1930s, based on a quantum-mechanical analysis, Lev Landau predicted that so-called neutron stars could exist and Subramanian Chandrasekhar showed that the mass of such a collapsed stellar object could not exceed about 1.4 solar masses—the now famous Chandrasekhar limit whose existence was vehemently, and unreasonably, opposed by Eddington.

In 1938 Einstein analyzed his ''thought'' analog of a collapsing stellar object and concluded that a black hole would not form. However, he did not carry out a dynamical calculation, but treated the object as a collection of particles moving in ever-smaller circular orbits; he deduced that the speed of these particles would reach the velocity of light barrier before reaching the Schwarzschild radius, and thereby drew an incorrect conclusion.

Soon thereafter, in 1939, J. Robert Oppenheimer and Hartland Snyder made a major advance in understanding gravitational collapse. They showed that in principle, according to general relativity, black holes could be produced from a sufficiently massive stellar object that

[15]There is, however, no known written evidence supporting this (apocryphal?) quotation (John Stachel, private communication, 1998).

[16]The main candles used were Cepheid variable stars. There were two principal problems: recognition only later that there were two classes of such stars with different period-luminosity relations and mistaken identification in the most distant indicators of, e.g., unresolved star clusters for a single (''most luminous'') star.

collapsed after consuming its full complement of nuclear energy. Basing their analysis on the Schwarszchild metric, and thus neglecting rotation and any other departure from spherical symmetry, they deduced correctly that with the mass of the star remaining sufficiently large—greater than about one solar mass (their value)—this collapse would continue indefinitely: the radius of the star would approach its gravitational radius asymptotically, as seen by a distant observer. This discussion was apparently the first (correct) description of an event horizon. Oppenheimer and Snyder specifically contrasted the possibly very short collapse time that would be seen by a comoving observer, with the corresponding infinite time for the collapse that would be measured by a distant observer. They also described correctly, within the context of general relativity, the confinement of electromagnetic radiation from the star to narrower and narrower cones about the surface normal, as the collapse proceeds.

In many ways establishing the theoretical existence of black holes, within the framework of general relativity, was easier and less controversial than establishing their existence in the universe. However, after well over a decade of controversy and weakly supported claims, there is now a widespread consensus that astronomers have indeed developed persuasive evidence for the existence of black holes. As in most astronomic taxonomy there are two classes: the stellar-mass black holes and the 10^6–10^9 times larger mass black holes. Evidence for the former consists of estimates for binary star systems of the mass, or of a lower bound on the mass, of a presumably collapsed member of each such system. In these systems Doppler measurements allow the determination of the so-called mass function, which sets a lower bound on the mass of this invisible, and likely collapsed, member of the binary. (A point estimate of this mass cannot be determined directly because of the unknown inclination of the orbit of the binary with respect to the line of sight from Earth.) These observations show in the ''best'' case that the black-hole candidate has a mass greater than eight solar masses, far in excess of the Chandrasekhar limit and more than twice the largest conceivable nonblack-hole collapsed object that quantum mechanics and a maximally ''stiff'' equation of state seem to allow.

The evidence for large (''supermassive'') black holes became almost overwhelming just a few years ago from the partly serendipitous study, via combined radio spectroscopy and VLBI, of the center of the galaxy NGC 4258, i.e., the 4258th entry of the New General (optical sky) Catalog, which stems from the early part of this century. This 1995 study yielded strong kinematic evidence for material in Keplerian orbits. In turn, these orbits implied average mass densities interior to these orbits, of at least 10^9 solar masses per cubic parsec, a density so high that no configuration of mass consistent with current understanding could be responsible other than a supermassive black hole. There is also growing evidence, albeit not yet as convincing, for the presence at the center of our own galaxy of a black hole of mass of the order of 3×10^6 solar masses.

Another relevant and impressive result from astrophysics relating to predictions from general relativity concerns ''evidence for''—the phrase of choice when astrophysical results are under scrutiny—an event horizon. As shown in 1997 by Michael Garcia, Jeff McClintock, and Ramesh Narayan, the luminosity of x-ray emissions from a sample of neutron stars and candidate black holes shows a tendency to separate into two clusters, with the luminosity of the neutron stars larger than those for black holes, as would be expected as radiating material ''blinks off'' as it approaches the event horizon.

3. Gravitational lenses

The idea that mass could, like glass lenses, produce images was apparently first articulated in print in 1919 by Oliver Lodge, but not pursued in any systematic way, either theoretically or experimentally, for nearly two decades. Then in 1936, at the urging of a Czech engineer, Einstein analyzed such lensing,[17] demonstrating that in the case of collinearity of a source, a (point-mass) lens, and an observer, the image seen by the observer would appear as a circle—now known as the Einstein ring.[18] Its radius depends directly on the mass of the lens and on a function of the relevant lengths. For an asymmetric geometry the ring breaks into two images, one formed inside and one outside the corresponding Einstein ring. Einstein dismissed the possibility of observing this phenomenon—a ring or double image in the case of noncollinear geometry—based on two arguments, neither supported in the paper by calculations: (1) the probability of a chance alignment was negligible; and (2) the light from the lens star would ''drown out'' the light from the distant lensed star, despite the magnification of this latter light. Fritz Zwicky, an astronomer-physicist who often had insights 50 years ahead of his time, was quick to point out, in a paper submitted barely six weeks later, that whereas Einstein's conclusions about observability might well be correct for stars in the Milky Way, the (more distant) nebulae—now called ''galaxies''—offer far greater prospects for observability; two months later still, he noted that the existence of lensing with multiple images was virtually a certainty and pointed out the basic importance of such imaging to cosmology.

---

[17]Recently, however, Jürgen Renn, Tilman Sauer, and Stachel examined Einstein's notebooks from 1912, prior to the completion of general relativity; these showed that he had developed all of the resultant formulas at that earlier time, although the values for the deflections were half those of his completed theory.

[18]The formula for the ring was apparently first published in 1924 by Otto Chowlson.

Despite Zwicky's upbeat conclusions, the field then lay fallow for nearly thirty more years, until, independently, in 1964, Steven Liebes in the United States and Sjur Refsdal in Norway, published analyses of gravitational lensing, the former focusing more on image shapes and characteristics, the latter more on cosmological uses, most importantly the prospects for determining a value for the Hubble constant. This determination, being based on very distant sources, would be independent of the rungs of the conventional cosmic distance ladder. After these articles by Leibes and Refsdal appeared, the theoretical astrophysics literature on gravitational lenses started to mushroom with a number of papers pointing out the consequences of lensing on various statistical questions, such as the effect of magnifications in distorting "unbiased" samples of galaxy luminosities. Not until March 1979, however, within about two weeks of the 100th anniversary of Einstein's birth, did Dennis Walsh in England serendipitously discover the first gravitational lens. He had been trying to find optical identifications for sources discovered in a low-angular-resolution radio survey; two optical objects about six arcseconds apart on the sky were candidates for one such radio source. Following up with telescopes in the southwest United States, Walsh and his colleagues measured the spectra from the two objects and noticed a striking similarity between them, aside from the then-puzzling difference in mean slope in the infrared, later identified as due to the main lensing galaxy, which was not separately visible from those observations. Walsh and his colleagues thus took the courageous step of claiming, correctly, that their two objects were in fact images of the same quasar, since each had the same redshift ($z \approx 1.41$, indicating a very distant object) and nearly the same optical spectrum. Further lens discoveries of multiple images of a single object were somewhat slow in coming, but the pace has quickened. A number of rings, sets of multiple images, and arcs were discovered in radio, optical, and infrared images of the sky, and astrophysical applications have been tumbling out at an awe-inspiring rate. Statistical analyses of results from observations of (faint) arcs from very distant objects have even allowed Christopher Kochanek to place a (model-dependent) bound on the cosmological constant.

The search for the value of the Hubble constant also took a new tack, along with the old ones, based on Refsdal's noting that multiple images of a distant light source produced by a point-mass lens could be used to infer the distance to the source. The idea is elegantly simple: if the light source were to vary in its intensity, then these variations would be seen by an Earth observer to arrive at different times in the different images. Such a difference in the time of arrival of a feature in the light curve in two images is proportional to the light travel time from source to observer and, hence, to H_0^{-1}. This difference is also proportional to the mass of the lens. And therein lies the rub: independent estimates of this mass or, more accurately, of the mass distribution, are difficult to come by, either from other properties of the images, such as their optical shapes and the locations of their centroids, or from other types of astronomical measurements.[19] Thus this general method of estimating H_0 has been notoriously difficult to apply, both because of the difficulty with the time-delay measurement and, especially, because of the difficulty in determining the lens' mass distribution with useful accuracy. As we near the millenium only about a few dozen gravitational lenses have been confirmed—a subtle process in itself—and only three have yielded reasonably reliable time delays. The first gravitational lens system discovered has led to the estimate $H_0 = 65 \pm 10\,\mathrm{km\,s^{-1}\,Mpc^{-1}}$; however, this standard error does not account fully for possibly large model errors.

Perhaps the most spectacular results obtained so far followed from a suggestion in the late 1980s by Bohdan Paczyinski to make wide-field observations with modest-sized optical telescopes of about a million stars simultaneously and repeatedly. The purpose was to detect, with charge-coupled devices and modern computers, color-independent brightening and subsequent dimming among members of this star collection, such variations being the hallmark of an intervening (dark) lens passing by on the sky. (One can forgive Einstein for not envisioning this multipronged development of technology.) The durations of such "events" can vary from minutes to months. The duration and brightening factors depend on the mass of the (invisible) lens and the geometry of the lens system, unknowns that most often preclude a useful point estimate being made of this mass. For long duration events, the parallax afforded by the Earth's orbital motion allows an estimate to be made of the lens' distance; for all events, long or short, such determinations could be obtained were a telescope in orbit far from the Earth observing the same parts of the sky simultaneously (were two or more such spacecraft employed, these observations would be freed from ground-based weather, but at some less-than-modest cost). Attempts to observe such "microlensing" effects were proposed initially to detect invisible mass ("dark matter") in our Galaxy that could be in the form of compact objects, so-called MACHOs: massive compact halo objects. This monitoring project, started in the early 1990s, is being carried out by three independent collaborations and has been remarkably successful: over 100 events have so far been detected. The results show, for example, that the "dark matter" problem cannot be solved by MACHOs alone.

[19]These latter are needed in any event because of a fundamental degeneracy noted in 1985 by Marc Gorenstein, Emilio Falco, and myself. From measurements of the images alone, one cannot distinguish between the actual mass distribution and a different (scaled) one in which a uniform surface density "sheet" is also present, with the light source being correspondingly smaller, yielding the same image sizes. A separate type of measurement is needed to break this degeneracy.

IV. FUTURE

Predicting the future is easy; predicting it accurately for the next century is a tad more difficult. One can at the least anticipate that advances in experimental and theoretical gravitation will yield profound and unexpected insights into astrophysical phenomena, especially those that are invisible in electromagnetic-wave observations. One could even conceive of black holes being ''tamed'' to extract energy from infalling matter. More likely, the issues and the problems offering the greatest challenges and rewards over the next century are not now conceivable or at least not yet conceived.

ACKNOWLEDGMENTS

I thank George Field, Kurt Gottfried, and John Stachel for their extensive and very helpful comments, and Stephen Brush, Thibault Damour, Stanley Deser, Gerald Holton, Martin Rees, Stuart Shapiro, Saul Teukolsky, and Robert Vessot, for their useful suggestions.