Physics – Uspekhi **63** (7) 721 – 724 (2020)

METHODOLOGICAL NOTES PACS numbers: **01.65.+g**, **03.65.–w**

Was Sommerfeld wrong?

(To the history of the appearance of spin in relativistic wave equations)

S V Petrov

DOI: https://doi.org/10.3367/UFNe.2019.11.038695

Abstract. This article presents a brief history of the appearance of electron spin in relativistic wave equations. Dirac derived his wave equation in 1928 with the intention to obtain an equation for the 'simplest' particle with spin zero. But, as Dirac later announced at the European Conference on Particle Physics (Budapest, 4–9 July 1977), it was a big surprise for him that the equation described the states of a particle with spin 1/2.

Keywords: spin, relativistic wave equations

Fifteen years ago, *Physics–Uspekhi* published a letter to the Editorial Board [1] entitled "Sommerfeld formula and the Dirac theory," which offered a critical analysis of the coincidence between the formulas for the energy spectrum of the hydrogen atom obtained in the framework of the 'old quantum theory' (Bohr–Sommerfeld quantization) and as a result of solving the Dirac equation. The result of the analysis was already formulated in the abstract: "A surprising coincidence of the fine-structure formulas by A Sommerfeld and P Dirac is the consequence of an error made by the first author."

In our opinion, the situation, first, is not as unambiguous as presented in Ref. [1], and, second, concerns such an important issue as the connection between quantum mechanics and classical mechanics, and, more precisely, the question of how spin appears in passing from the classical dynamics of a relativistic particle to its quantum dynamics.[1] Precisely this reason motivated writing this note, despite the long time since Ref. [1] was published.

The author of [1] is firmly convinced that the coincidence mentioned above is impossible, because if it were not the case one had to admit that "an eclectic theory turned out in one way or another to be equivalent to Dirac's rigorous theory." We note, first of all, that the coincidence between the Sommerfeld and Dirac formulas by no means implies the equivalence of the theories; the point is only the coincidence between energy levels.[2] Sommerfeld's error, in the opinion of the author of [1], was that he incorrectly quantized the orbital angular momentum L: instead of half-integer values $L = (l + 1/2)\hbar$ as would follow from the semiclassical treatment, Sommerfeld assumed that $L = n_\varphi \hbar$, where n_φ is an integer. Had he done it 'correctly,' he would have obtained the energy levels that coincide with the eigenvalues of the Klein–Fock–Gordon equation, which, as stressed in [1], "should be the outcome in the spinless case." Yet, in 1916, when Sommerfeld published his study [2], it was difficult to follow the recipes of the semiclassical analysis, which was created 10 years later [3–5]. Moreover, in 1916, nobody even guessed the existence of the electron spin. Dirac was saying that "at that time (i.e., up to 1924 — *S P*) the spin of the electron was unknown," although "some physicists had thought about it" [6]. Indeed, in 1918, Compton formulated an idea that both translational and rotational motions are inherent in an electron [7], and three years later came to the conclusion that the electron is "a primary magnetic particle" rotating "similar to a minuscule gyroscope" [8]. Kronig tried to use the idea of internal magnetic moment of an electron to describe the spectra of alkali metals, but his hypothesis failed to receive the approval of Pauli, Heisenberg, and Kramers [9]. Finally, in 1925–1926, Uhlenbeck and Goudsmit published two short papers justifying the introduction of spin [10, 11]. The second of these papers was accompanied by a letter from Bohr supporting the idea of a rotating electron.

Sommerfeld followed the path laid by Bohr [12], who took into account that an electron stays on a selected circular orbit of radius a owing to the balance between the Coulomb and centrifugal forces,

$$\frac{mv^2}{a} = \frac{Ze^2}{a^2}\,. \tag{1}$$

Bohr complemented this purely classical condition with the condition of quantum nature, according to which the magnitude of electron orbital angular momentum $L = mva$ cannot be arbitrary, but must satisfy the condition

$$2\pi L = nh\,, \tag{2}$$

where n is an integer, beginning from 1. From these two relations, the energy of an electron on a selected orbit can easily be found:

$$E_n = -\frac{2\pi^2 mZ^2 e^4}{h^2}\,\frac{1}{n^2}\,. \tag{3}$$

This result proved to be in full agreement with the Balmer formula, and 13 years later it was shown that energy values (3) are eigenvalues of the Schrödinger equation. Obviously, full

[1] Needless to say, we are interested here in the historical aspect of the question and not the contemporary state of the theory.

[2] In particular, an adequate classification of states follows only from the Dirac theory.

S V Petrov
Lomonosov Moscow State University, Department of Chemistry, Leninskie gory 1, str. 3, 119992 Moscow, Russian Federation
E-mail: spswix@rambler.ru

Received 9 September 2019
Uspekhi Fizicheskikh Nauk **190** (7) 777 – 780 (2020)
Translated by S D Danilov; edited by A M Semikhatov

agreement would not be possible had Bohr allowed half-integer values of n in condition (2). Developing the ideas of Bohr, Sommerfeld quantized all planar elliptic orbits with the help of conditions imposed on the action variables [13]:

$$J_\varphi = \oint p_\varphi \, \mathrm{d}\varphi = n_\varphi h \,, \tag{4}$$

$$J_r = \oint p_r \, \mathrm{d}r = n_r h \,, \tag{5}$$

where r and φ are polar coordinates in the orbital plane, p_r and p_φ are the conjugate momenta, and n_r and n_φ are integers. The first of these conditions reduces to Bohr condition (2). In order to implement condition (5) it is necessary to take into account that the Hamiltonian is an integral of motion equal to the electron energy E. Then one can easily obtain p_r as an explicit function of the radial variable r and compute the integral in (5). The result is the same formula for energy levels (3), where $n = n_r + n_\varphi$.[3] Already here, in the nonrelativistic case, the situation mentioned by the author of [1] is manifest, namely, an 'eclectic theory' leads to the same results as the 'rigorous theory' by Schrödinger.

Sommerfeld took the next step along the path laid by Bohr, trying to explain the fine structure of hydrogen spectral lines, which was experimentally discovered by Michelson [14]. He preserved the same quantization conditions (4) and (5), but, in calculating the integral in (5), replaced the nonrelativistic Hamiltonian with the relativistic one:

$$H = c\sqrt{p^2 + m^2c^2} - \frac{Ze^2}{r} \,. \tag{6}$$

In polar coordinates on a plane, it is expressed as

$$H = c\sqrt{p_r^2 + \frac{p_\varphi^2}{r^2} + m^2c^2} - \frac{Ze^2}{r} \,. \tag{7}$$

Using conditions (4) and (5) and taking into account that (just as in the nonrelativistic case) the energy E and the orbital angular momentum L are integrals of motion, Sommerfeld obtained

$$E_{n_r n_\varphi} = \frac{mc^2}{\sqrt{1 + \alpha^2 Z^2 \Big/ \left(n_r + \sqrt{n_\varphi^2 - \alpha^2 Z^2}\right)^2}} \,. \tag{8}$$

According to this formula, the Bohr energy levels are split, because now the energy (distinct from the nonrelativistic case) cannot be described by a single quantum number $n = n_r + n_\varphi$, which explains the fine structure of the Balmer lines. The dimensionless quantity $\alpha = 2\pi e/hc \approx 1/137$ that automatically appeared in the derivation of formula (8) and was called the 'fine-structure constant' fully determines the magnitude of energy level splitting.

The energy levels of Dirac's hydrogen atom [15] are defined by the principal quantum number $n = 1, 2, \ldots$ and the quantum number of the total angular momentum (i.e., the sum of orbital and spin electron momenta) j, which (for given n) takes the half-integer values $1/2, 3/2, \ldots, n - 1/2$:

$$E_{nj} = \frac{mc^2}{\sqrt{1 + \left[\alpha^2 Z^2 \Big/ \left(n - (j + 1/2) + \sqrt{(j + 1/2)^2 - \alpha^2 Z^2}\right)^2\right]}} \,. \tag{9}$$

The numerical coincidence of both formulas is obvious. It suffices to replace the quantum number n_φ in (8) with $j + 1/2$, and n_r with $n - (j + 1/2)$.

We now return to the key question that gave rise to letter [1] as well as this note, namely the statement that the energy values obtained by Sommerfeld should coincide with the eigenvalues of the Klein–Fock–Gordon equation, and not the Dirac equation.

The first relativistic wave equation was published by Klein [16], Fock [17], and Gordon [18]. One can easily notice the analogy between the origin of the Klein–Fock–Gordon equation [4] and the nonrelativistic Schrödinger equation.

The Schrödinger wave equation for a free particle (we are trying to understand how spin emerges in a relativistic wave equation, and it therefore suffices to consider the case of a free particle),

$$\mathrm{i}\hbar \frac{\partial \Psi}{\partial t} = -\frac{\hbar^2}{2m} \Delta \Psi \,, \tag{10}$$

can be formally obtained from the relations between the energy and momentum of a free particle

$$E = \frac{1}{2m} p^2 \,, \tag{11}$$

if we assign E the operator $\mathrm{i}\hbar\,(\partial/\partial t)$ and the momentum $\mathbf{p}$ the operator $(\hbar/\mathrm{i})\,\nabla$.

The relativistic analog of relation (11) is the expression

$$\frac{E^2}{c^2} - p^2 = m^2c^2 \,, \tag{12}$$

where $\mathbf{p}$ is now the relativistic momentum of the particle. In the left-hand side of (12) we have a scalar product [5] of the 4-momentum $\mathbf{P} = (E/c, \mathbf{p})$ with itself, which is an invariant, as also is the scalar product of any two relativistic 4-vectors. The relativistic wave equation can be obtained if we replace the 4-momentum $\mathbf{P}$ with the operator $\hat{\mathbf{P}}$, the four components of which are the differentiation operators with respect to the components of 4-radius-vector $\mathbf{R} = (ct, \mathbf{r})$ times $\hbar/\mathrm{i}$: [6]

$$\left(-\frac{\hbar^2}{c^2}\frac{\partial^2}{\partial t^2} + \hbar^2\Delta - m^2c^2\right)\Psi = 0 \,. \tag{13}$$

The first two terms in the parentheses form the scalar product of the 4-gradient with itself, multiplied by $\hbar^2$, and therefore the operator acting on the wave function is relativistically invariant. Hence, in order that wave equation (13) (like any relativistic equation) preserve its form under Lorentz transformations, i.e., be covariant, the wave function Ψ must be a

[3] The fact that the solution of a more general problem leads to the same result (3) as the particular case of circular orbits did not surprise Sommerfeld, because (in his words [13]) "in this family of ellipses each ellipse is energetically equivalent to a fully definite Bohr circle." Today, we refer to this phenomenon as the degeneration of energy levels.

[4] A sad tradition is that the name Fock is not commonly mentioned in the name of this equation (see in this respect Refs [19–21] published in *Physics–Uspekhi*). We note that de Broglie and Schrödinger [6] were also involved in the creation of the Klein–Fock–Gordon equation.

[5] We mean the scalar product in the pseudo-Euclidean Minkowski space.

[6] The operator $\hat{\mathbf{P}}$ is the 4-gradient up to the factor $\hbar/\mathrm{i}$.

scalar function of coordinates and time only, because terms responsible for spin are absent from Eqn (13). Following the paradigm of quantum mechanics (the replacement of dynamical variables by the respective operators) and special relativity (the requirement of relativistic invariance), we automatically arrive at wave equation (13) describing the motion of a spinless particle. Initially, no additional requirement that the particle have zero spin was imposed in the derivation of the wave equation.

The relativistic wave equation for the electron was published by Dirac in 1928 [22]. He had devoted the preceding two years to formulating quantum mechanics in a symbolic language based on abstract vectors of state.[7] In Dirac's opinion, "the symbolic method, however, seems to go more deeply in the nature of things" and "enables one to express physical laws in a neat and concise way," whereas the method of the pictures (the wave mechanics of Schrödinger and the matrix mechanics of Heisenberg–Born–Jordan) is more convenient to use for solving concrete problems [23].

A point of departure in the derivation of the relativistic wave equation was the absolute rejection by Dirac of the Klein–Fock–Gordon equation. Taking for granted the principle of superposition during the whole time interval (at least when the system is not subject to any perturbation), the equation describing the evolution of a state vector should be a linear differential equation of the first order with respect to time [23], whereas the Klein–Fock–Gordon equation contains a second time derivative. Half a century later, Dirac recalled [6] that in 1927 "Bohr seemed to be pretty satisfied with the Klein–Gordon theory and that was the opinion of most physicists of that time, perhaps of all of them." And it was so notwithstanding the unreconcilable contradictions among computations of probabilities of dynamical variables appearing in the framework of the Klein–Fock–Gordon theory:[8] "If you wanted to find the probability of the momentum having specified values you cannot answer the question at all. Similarly for other dynamical variables, you cannot get any information at all about their probabilities" [6].

The requirement of linearity for the time derivative is satisfied by the Schrödinger-type wave equation

$$\mathrm{i}\hbar\,\frac{\partial\Psi}{\partial t} = H\Psi \tag{14}$$

with the relativistic Hamiltonian

$$H = c\sqrt{\hat{p}^2 + m^2c^2} \tag{15}$$

in the right-hand side. However, this form of writing the wave equation could not satisfy Dirac, for, because of the equal status of the four coordinates of any point in spacetime, a correct relativistic theory must be fully symmetric with respect to derivatives with respect to time and Cartesian coordinates, and hence the wave equation must also be linear in the spatial components of the 4-gradient. Thus, Dirac faced the problem of how to take the square root in expression (15). As Dirac recalled later [6], he came to the solution "rather by accident." In 1927, the existence of the electron spin was not causing any doubts. In particular, the two-row Pauli matrices σ_1, σ_2, and σ_3 were known. And so at one time, as Dirac wrote [6], "playing with the mathematics," he discovered "a very interesting result, just"

$$(\sigma_1\hat{p}_x + \sigma_2\hat{p}_y + \sigma_3\hat{p}_z)^2 = \hat{p}_x^2 + \hat{p}_y^2 + \hat{p}_z^2\,, \tag{16}$$

and "you had thus a sort of square root for $\hat{p}_x^2 + \hat{p}_y^2 + \hat{p}_z^2$." However, in expression (15), there is a sum of four squares in the radicand; therefore, the three Pauli matrices were insufficient and one needed to add the fourth matrix having the same properties as the Pauli matrices, namely, its square had to be an identity matrix and it had to anticommute with any of the Pauli matrices. But there was no such matrix, and that "was a serious difficulty for me for some weeks," before it was found that, in order to obtain the Hamiltonian linear in Cartesian components of momentum

$$H_{\mathrm{D}} = c(\alpha_1\hat{p}_x + \alpha_2\hat{p}_y + \alpha_3\hat{p}_z) + mc^2\beta\,, \tag{17}$$

the quantities $\alpha_i\,(i = 1, 2, 3)$ and β satisfying the same requirements as the Pauli matrices

$$\begin{cases} \alpha_i\alpha_k + \alpha_k\alpha_i = 0\,, \\ \alpha_i\beta + \beta\alpha_i = 0\,, \\ \alpha_i^2 = \beta^2 = 1 \end{cases} \tag{18}$$

could be taken as Hermitian 4×4 matrices independent of the components of $\hat{\mathbf{p}}$ (otherwise the Hamiltonian H_{D} would not be linear) and independent of the coordinates and time, because for a free particle all points in four-dimensional space should be equivalent. As such matrices, Dirac took the matrices

$$\alpha_i = \begin{pmatrix} O & \sigma_i \\ \sigma_i & O \end{pmatrix}, \tag{19}$$

$$\beta = \begin{pmatrix} I & O \\ O & -I \end{pmatrix}, \tag{20}$$

where O and I are two-dimensional zero and identity matrices. The independence of the matrices α_i and β from 'traditional' dynamical variables inevitably leads to the conclusion that these matrices "describe some new degrees of freedom, belonging to some internal motion in the electron" [23]. In other words, if in 1928 there had been no experimental data pointing at the internal angular momentum of an electron, the Dirac relativistic wave equation would have been a theoretical prediction of its existence.

From the matrix structure of Dirac Hamiltonian (17), it follows that the wave function should be a 4-dimensional column,[9] each element of which is a function of coordinates and time, and the wave equation presents a system of four coupled equations of the first order. Finally, direct computations in [22] (see also [23]) showed that the equation obtained is covariant under Lorentz transformations. Thus, Dirac obtained "really a relativistic equation" [6], flawless from the standpoint of both quantum mechanics and special relativity, but it turned out that it is valid only for particles with spin 1/2. However, Dirac's argument gives no answer to the question as to why the equation obtained describes the

[7] The symbolic method was presented in 1930 in a completed form in the famous 'Principles' [23].

[8] Because the Klein–Fock–Gordon equation is a second-order differential equation with respect to time, the probability density has to be defined as a bilinear form in Ψ and $\partial\Psi/\partial t$ [24], which, as a result, can take both positive and negative values, and therefore loses its sense of probability density.

[9] Bispinor in modern terminology.

motion of a particle with spin one half, and not any other. It may seem that using the Pauli matrices just indicates that Dirac strived to obtain a wave equation for particles with spin 1/2. In this respect, Dirac was saying [6] that his goal was to satisfy the requirements of both quantum mechanics and special relativity, and that it turned out that "the simplest particle satisfying those requirements is a particle with a spin of a half. That was a great surprise to me. I thought that the simplest particle would naturally have a zero spin, and that a spin of a half would have to be brought in later as a complication, after one had solved the problem of a particle with no spin. But it turned out otherwise."

Thus, both the absence of spin in the Klein–Fock–Gordon equation and the appearance of spin 1/2 in the Dirac equation take place automatically, without any pre-imposed conditions regarding the presence or absence of spin in wave equations.

The coincidence between the results of Sommerfeld and Dirac (or Klein–Fock–Gordon under a suitable replacement of the integer number n_φ with a half-integer $l + 1/2$ in condition (4)), as well as the coincidence between the results of Bohr and Schrödinger is, by all probability, just accidental. "It seems that one does get coincidences of this sort in the search for understanding Nature" [6].

References

1. Granovskii Ya I *Phys. Usp.* **47** 523 (2004); *Usp. Fiz. Nauk* **174** 577 (2004)
2. Sommerfeld A *Ann. Physik* **51** 125 (1916) Vol. 356 on the new numbering of volumes on the site Wiley Online Library
3. Wentzel G *Z. Phys.* **38** 518 (1926)
4. Kramers H A *Z. Phys.* **39** 828 (1926)
5. Brillouin L *C.R. Acad. Sci.* **183** 24 (1926)
6. Dirac P A M *Sov. Phys. Usp.* **22** 648 (1979); *Usp. Fiz. Nauk* **128** 681 (1979)
7. Compton A H *J. Washington Acad. Sci.* **8** 1 (1918)
8. Compton A H *J. Franklin Inst.* **192** 145 (1921)
9. Jammer M *The Conceptual Development of Quantum Mechanics* (New York: Mc Graw-Hill Book Co., 1967); Translated into Russian: *Evolyutsiya Ponyatiya Kvantovoi Mekhaniki* (Moscow: Nauka, 1985)
10. Uhlenbeck G E, Goudsmit S A *Naturwissenschaften* **13** 953 (1925)
11. Uhlenbeck G E, Goudsmit S *Nature* **117** 264 (1926)
12. Bohr N *Philos. Mag. 6* **26** 1 (1913); *Philos. Mag. 6* **26** 476 (1913); *Philos. Mag. 6* **26** 857 (1913)
13. Sommerfeld A *Atombau und Spektrallinien* Vol. 1 (Braunschweig: F. Vieweg and Sohn, 1951); Translated into Russian: *Stroenie Atoma i Spektry* Vol. 1 (Mosow: GITTL, 1956)
14. Michelson A A *Philos. Mag. 5* **31** 338 (1891)
15. Messiah A *Quantum Mechanics* Vol. 2 (Amsterdam: North-Holland Publ. Co., 1961); Translated into Russian: *Kvantovaya Mekhanika* Vol. 2 (Moscow: Nauka, 1979)
16. Klein O *Z. Phys.* **37** 855 (1926)
17. Fock V *Z. Phys.* **38** 242 (1926)
18. Gordon W *Z. Phys.* **40** 117 (1926)
19. Okun L B *Phys. Usp.* **535** 835 (2010); *Usp. Fiz. Nauk* **180** 871 (2010)
20. Dyson F *Phys. Usp.* **53** 825 (2010); *Usp. Fiz. Nauk* **180** 859 (2010)
21. Fock V A *Phys. Usp.* **53** 839 (2010); *Usp. Fiz. Nauk* **180** 874 (2010); Fock V *Z. Phys.* **39** 226 (1926)
22. Dirac P A M *Proc. R. Soc. Lond. A* **117** 610 (1928)
23. Dirac P A M *The Principles of Quantum Mechanics* (Oxford: Oxford University Press, 1958); Translated into Russian: *Printsipy Kvantovoi Mekhaniki* (Moscow: Fizmatlit, 1960)
24. Davydov A S *Quantum Mechanics* (Oxford: Pergamon Press, 1976); Translated from Russian: *Kvantovaya Mekhanika* (Moscow: Fizmatlit, 1963)