Design and validation of a two-tier questionnaire on basic aspects in quantum mechanics

Umberto Scotti di Uccio,[1] Arturo Colantonio,[2,3] Silvia Galano,[1,3]
Irene Marzoli,[2] Fabio Trani,[4] and Italo Testa[1,*]

[1]*Department of Physics "E. Pancini," University Federico II, Naples, Italy*
[2]*School of Science and Technology, Physics Division, University of Camerino, Camerino, Italy*
[3]*INAF, Astronomical Observatory of Capodimonte, Naples, Italy*
[4]*Liceo Statale "Ischia", Ischia, Italy*

(Received 1 November 2018; published 5 June 2019)

We present the design, statistical analysis, and validation of a questionnaire to assess students' knowledge about basic aspects of quantum mechanics (QM). The QM evaluation (QME) is a true-false and multiple-choice mixed questionnaire that features 10 two-tier items spanning three relevant themes in quantum mechanics: wave behavior of matter, measurement, and atoms and electrons behavior. Its validity was assessed through a pilot administration to students and interviews with course instructors. We checked its internal consistency using both classic test theory and Rasch analysis to account for the different difficulty of each tier and for different scoring methods of the items. The questionnaire was administered to about 450 undergraduate physics students and high school physics teachers. Data show that it is a reliable instrument and all items have a good discriminatory power. Since the test does not require an advanced mathematical knowledge, it ideally lends itself to probe students' knowledge about quantum mechanics in a variety of university courses, from the introductory ones to those more formal and mathematically oriented.

DOI: 10.1103/PhysRevPhysEducRes.15.010137

I. INTRODUCTION

The development of research-based, conceptual probes to assess students' understanding is one of the key tools of educational research. Concept inventories have been developed and widely used by physics education researchers for almost thirty years to measure students' conceptual learning in several areas of introductory physics, from mechanics to electricity and magnetism [1,2]. Recently, scholars have begun to develop concept inventories about more advanced topics such as kinetic molecular theory of gases [3], nuclear physics [4], relativity [5], and, of course, quantum mechanics (QM) [6]. For the latter, two different contexts have been considered:

(1) High schools. In the last decades, quantum mechanics has been included in many national school curricula (e.g., England General Certificate of Education Advanced Level, France, Italy, Norway, Spain, U.S. Next Generation Science Standards). On the one hand, this is a natural evolution of previous syllabi. In fact, there is no conceptual reason for regarding quantum mechanics as a separate branch of physics, which can be considered or not in standard curricula. It is, instead, deeply nested in any description of matter and radiation properties. On the other hand, quantum mechanics carries the infamous qualification of "weirdness" that still discourages its practical introduction at school [7].
(2) University. Several studies have shown that even after attending upper-level physics courses, undergraduate physics students still have difficulty understanding basic concepts of quantum mechanics, such as wave-particle duality, the realm of the Schrödinger equation, and the implications of the Heisenberg uncertainty principle [8–10]. Moreover, students often struggle with reconciling how classical concepts, such as measurement, wave propagation, or probability, are applied in quantum mechanics theory [11]. Finally, the highly abstract mathematical formalism of quantum mechanics may overshadow the meaning of the physical quantities involved in the calculations carried out in such courses [12].

For such reasons, it is no wonder that students' assessment of quantum concepts has increasingly gained attention in educational research. A pioneering work in assessing students' conceptions in quantum mechanics was carried out by Ireson [13], who reported the results of the administration of a Likert-type questionnaire made up of 29 five-level items ranging from strongly agree to strongly disagree. The questionnaire was administered to

*italo.testa@unina.it

342 students in the UK at the end of high school. Results show that students' reasoning strategies could be divided in three main categories: quantum thinking (correct knowledge), conflicting quantum thinking (transitional knowledge), and conflicting mechanistic thinking (classical knowledge). Such results suggest that students may progress from reasoning strategies rooted in classical physics toward more correct conceptual stages through suitable teaching paths.

A fundamental contribution to the field came soon after from Singh and colleagues, who devoted fifteen years to the investigation of students' misconceptions in quantum mechanics. An early attempt is presented in Ref. [9], where the author describes an open-ended probe, the *quantum measurement test*, to assess students' conceptual understanding about fundamental concepts, such as the properties of the "observables" (i.e., Hermitian operators in a Hilbert space describing the physical properties of a system) and the measurement process. The test was administered to 89 students in the U.S. from six different universities. The results show that even after instruction, students held misconceptions about stationary states, eigenstates, and time dependence of expectation values. The author stresses that such misconceptions were common amongst the students, despite having attended different courses with different instructors and materials.

Recently, an evolution of this test, called the *quantum mechanics survey* was presented in Ref. [14]. This new version features 31 questions about advanced quantum concepts, including wave functions, bound or scattering states, measurement, expectation values and their time dependence, stationary and nonstationary states, Hamiltonian operators, time dependence of wave functions, etc. The test was administered to 226 graduate and undergraduate students in physics. Reliability and consistency indices are rather high. Results show that most common difficulties concerned the probability of obtaining a given value in a measurement, the role of the Hamiltonian in the time-dependent Schrödinger equation, and the time dependency of the expectation value of an operator.

A similar approach was adopted by Cataloglu and Robinett [15], who developed the *quantum mechanics visualization instrument*. The 25-item questionnaire aims at assessing students' reasoning about the graphical representation of the wave function corresponding to a given potential $V(x)$. The questionnaire was administered to about 160 students of four distinct courses. The authors report about students' difficulties in visualizing the probability density in stationary states, and carrying out quantitative evaluations.

The *quantum physics concept survey* [16] is a qualitative concept inventory focused on five relevant "themes" in quantum mechanics: wave-particle behavior, the de Broglie wavelength, the analysis of the double slit experiment, the Heisenberg uncertainty principle, and the photoelectric effect. The questionnaire was administered to 312 students at the University of Sydney. Results reported by the authors support the reliability of the test. However, the paper does not provide enough evidence about whether the items of the questionnaire are able to assess students' reasoning about the targeted concepts.

McKagan and colleagues [17] developed the *quantum mechanics conceptual survey*, a 12-item questionnaire in which the following topics were addressed: wave-particle duality, quantization of states, wave function and probability, Heisenberg uncertainty principle, operators in Hilbert spaces and observables, quantum measurement, tunneling, and the Schrödinger equation. They describe in detail the design process, including the interview with faculty members to validate content and item readability. They also report the results of an administration to 370 undergraduate students. Despite the relatively small number of items in the probe, collected data show that the test can help instructors elicit misconceptions such as the belief that particles travel along sinusoidal paths, or that energy is lost in tunneling.

Finally, Sadaghiani and Pollock [6] proposed the *quantum mechanics concept assessment* (QMCA) test. The 31-item questionnaire was designed starting from an earlier open-ended version and addressed the following knowledge areas: measurement, the time-independent Schrödinger equation, time evolution, wave function, probability, and probability density. The test was administered to about 300 undergraduate physics students in two stages. They perform a statistical analysis that supports the reliability of the instrument. Their findings are consistent with previous misconceptions reported in other studies, such as the meaning of measurement in quantum mechanics or that quantum states (including superposition states) have always a definite energy.

Despite the large efforts devoted by researchers and the depth of their achievements, we still need assessment tools that cover all major topics in quantum mechanics and are suitable for statistical analysis. In our opinion, the main present limitations are the following:

(1) The scope and audience of the previous questionnaires were limited, on the one hand, by the focus on very specific aspects of quantum mechanics and, on the other, by the emphasis on highly abstract concepts, addressed only in upper level physics courses;
(2) answer choices of questionnaires with a broader potential audience did not allow to reliably distinguish between memorization and reasoning;
(3) even when answer choices were carefully designed to build on previous students' answers to open-ended questions, the analysis did not take into account the variable "distance" between correct and incorrect answers to different questions. In other words, as we will see later, the item difficulty is not constant across the questionnaire;

(4) achievements of students with different abilities and background knowledge in quantum mechanics were not compared.

To address the first issue, we present in this paper a summative assessment instrument, called the *quantum mechanics evaluation* (QME), designed to test foundational concepts in quantum mechanics and, hence, is also suitable for introductory courses. To address the second issue, we adopted a two-tier diagnostic format [18] using multiple scoring methods. To address the third issue, we analyzed the data through both classic test theory and a one-dimensional Rasch model. Rasch analysis, which is increasingly used in physics education to obtain more robust analysis of concept inventories [19,20], has been recently introduced to investigate model fit quality of scoring methods for a generic two-tier scientific reasoning test [21]. We note that, while also QMCA [6] featured few questions in a two-tier format, QME is the first instrument that features all items in a two-tier format. Moreover, Rasch analysis has not yet been applied to any previously developed assessment tool in quantum mechanics. Finally, to address the fourth issue, we administered the QME to subjects with different hypothesized proficiency levels in quantum mechanics.

We hence tried to answer the following research questions:

RQ1: what are the respondents' conceptions of quantum mechanics that emerge from the answers to QME?
RQ2: what are the psychometric properties of QME?
RQ3: to what extent do QME psychometric properties depend on the adopted scoring method?
RQ4: how well does QME discriminate between respondents with different background knowledge in quantum mechanics?

II. REVIEW OF TWO-TIER ASSESSMENT INSTRUMENTS

In this section, we briefly review previous studies that adopted a two-tier assessment approach. Unlike the items in traditional multiple-choice tests, those in two-tier instruments consist of two distinct parts, or "tiers." The first tier (T1) aims to investigate the content knowledge about a given concept. The second tier (T2) prompts students to choose which alternative best represents the reason for their answer in T1. Typically, the first tier is a multiple-choice question with, e.g., three answers, of which only one is correct, or a yes or no alternative. In both cases, T1 is usually built up of statements, which correspond to declarative content knowledge about the targeted topic. The second tier is a multiple-choice question as well, but usually with four or five answer choices, built up from relevant literature about students' alternative conceptions, classroom observations or interviews carried out in pilot studies.

Two-tier instruments have been used in science education for thirty years now, mainly in biology [22–25] and chemistry [26–29]. However, a growing number of instruments featuring two-tier items have been developed for a wide range of topics also in physics. For instance, Franklin [30], starting from existing multiple-choice items, developed a 40-item instrument to identify misconceptions about force and motion, heat and temperature, light and color, and electricity and magnetism. The final version of the questionnaire was administered to 509 students. The author reports a good level of reliability but low values of discrimination for each section of the test. A significant correlation between correct responses and confidence level was also measured. We note that only one scoring method (1 point for both tiers correctly answered) is discussed. Chen and colleagues [31] developed a two-tier instrument about formation of images and shadows. The questionnaire, designed from existing literature and content-related concept maps, featured 8 two-tier items and was administered to 317 high school students. Reported results show acceptable reliability and good values for difficulty and discrimination indices. However, they do not provide enough details about how the test was scored and how the classical statistics were calculated. In the effort to investigate whether answers to the second tier are influenced by the first tier, Tsai and Chou [32] developed three two-tier items about the concepts of weight, sound, and light propagation, respectively. To correlate the results across the three areas, the responses were coded as follows: 1 point if the answers to both tiers were incorrect, 2 points if the students' responses were correct in only one tier (either the first or the second), 3 points if the student answered correctly to both tiers. While giving valuable insight about the correlation between the targeted content areas (mechanics vs waves propagation), the low number of items does not allow for generalization of the findings. Taiwan scholars [33–35] developed a 97-item two-tier questionnaire in the framework of a large national project aimed at identifying scientific conceptions from a statistically significant sample of students. The instrument targeted curriculum topics as force and motion, vision, electric circuits, images formation, sound, and water pressure from the primary to secondary level. While the questionnaire features an impressive number of items and its findings provide valuable insight about how students' misconceptions vary across school levels and targeted concepts, the studies do not provide enough details about the psychometric properties of the instrument and the adopted scoring method.

Similar to Ref. [31], the study described in Ref. [36] addresses the design of an 8-item two-tier instrument about basic optics. The instrument allows us to identify several alternative conceptions. They show that such conceptions depend on the context used in the question; namely, students were not able to apply their knowledge of optics to different contexts. However, this study also does not provide enough details about the psychometric properties of the instrument.

Despite many valuable results, two-tier instruments have not been critique-free. It has been argued that students' response to the second tier could be biased by showing predefined reasoning alternatives. Moreover, the reasoning strategies in the second tier may be unfamiliar to students beforehand, so the proposed answer choices could not represent an existing misconception, but rather a lack of knowledge [37]. To address this issue, some have incorporated into each item a third tier that measures the level of perceived confidence of the subject to pick a given answer choice [38]. Two examples of this improved design concern mechanical wave propagation [39] and thermodynamics [40]. Surprisingly, results from both studies suggest that the students are more confident of wrong responses, thus supporting the claim that erroneous reasonings reported in the second tier plausibly correspond to already existing patterns, albeit not immediately detectable using more traditional probes.

Finally, we note that, to date, only one multi-tier instrument addressing a modern physics content (the photoelectric effect) has been developed [41]. Findings show that the instrument is able to reliably uncover students' misconceptions about the targeted aspects of the photoelectric effect. Moreover, three different scoring methods are discussed, which, respectively, consider students' responses to (i) first tier only, (ii) first and second tier together, and (iii) three tiers together.

III. METHODS

A. Design of the questionnaire

As other two-tier questionnaires, the QME features items with two distinct parts. However, in the present study, we slightly modified the typical structure of two-tier instruments. The first tier consisted of three true or false (T/F) statements. The second tier was designed as a standard multiple-choice question with just one correct option. We maintained that the statements of the first tier should represent content knowledge or basic facts about the targeted concepts. In particular, we selected three statements that students are expected to know in order to correctly respond to the second tier.

We used T/F statements in the first tier for a twofold reason. First, we wanted to preserve local independence, in order to perform individual Rasch analysis of the tiers' scores. In fact, while the first tier statements were related to the content targeted in the second tier, there was no explicit link between the questions in the tiers. Still, due to the connectedness between the respective targeted content, the two tiers can be treated as a single *super* item, and analyzed with a polytomous, partial credit scoring method, as we will see later. Second, we wanted to decrease the guessing probability in the first tier. For a tier with a yes or no alternative, the probability of guessing is $1/2$; in the case of one correct alternative out of three, it is $1/3$. As described later, for QME, the scoring for the first tier was designed so that the guessing probability is $(1/3 \times 1/2) \cong 17\%$.

The design of the QME started by drafting 50 potential claims for the first tier, each one corresponding either to a correct idea or to a known misconception about the chosen topics. Then, we designed the multiple-choice questions for the second tier, starting from 25 open-ended items piloted with about 30 third-year physics students. A subsample of three students and two university instructors, with proven experience in teaching quantum mechanics, were interviewed after the questionnaire administration. Eventually, we selected ten items and designed for each of them four answer choices (only one correct) based both on the literature and on the collected answers. Amongst the incorrect answer choices, one was chosen to represent a typical misconception, while the remaining two corresponded either to a classical reasoning in a quantum situation (and vice versa), or to scientifically unacceptable views. Such design roughly corresponds to the categorization of students' ideas about quantum mechanics proposed in Ref. [42]: (i) incorrect or naïve view; (ii) misconception or classical description; (iii) partial or mixed classical-quantum description, and (iv) quasi-quantum description. As an example, consider question Q3 about the wave function: "If you know the formal expression of the particle's wave function then you can…?": The incorrect answer (i) is "Determine all the possible values of any physical observable associated to the particle." A typical misconception (ii) is "Predict the possible states of the particle and the values of any associated physical observable." The mixed incorrect answer choice (iii) is "Describe all the particle's allowed positions and energies." The correct answer (iv) is "Calculate the probability of obtaining by measurement a given value of any physical observable associated to the particle."

Then, we matched the 50 claims, in groups of five, with the ten multiple-choice items. The matching was made independently by two researchers and a final agreement of 80% was considered satisfactory. Finally, we eliminated twenty redundant claims, using as criteria conciseness, intelligibility, and straightforwardness of the claim, so that the final questionnaire featured 10 two-tier items: three T/F claims in the first tier and one multiple-choice question as a second tier.

B. Contents of the QME

The contents of the QME (see Table I) were chosen as follows. First, we made a nonexhaustive list of possible concepts we considered critical to understand quantum mechanics. The list was compared with those presented in reviews [7,8] and previous studies focused on the design of assessment tools [17]. We then took the common ones and excluded those that could be addressed only with a complex mathematical formalism. For instance, referring to the list in Ref. [17], paragraph III. A, we eliminated the following

TABLE I. Overview of the QME contents. The complete questionnaire is reported in the Appendix.

Theme	Item code	Main topic of the first tier	Topic investigated in the second tier	Typical misconception probed in the second tier
WBM	Q1	Time evolution of the wave function	Relationship between energy and phase of the wave function for a stationary state	The phase factor of the wave function for a stationary state depends on the possible values of the particle position and energy
	Q3	General properties of the wave function	The probabilistic meaning of the wave function	The wave function allows to predict the states of a particle
	Q6	Quantum states and wave or particle behavior	Quantum behavior of atomic objects	The position of a particle oscillates as a wave
	Q7	Superposition of states	Difference between superposition of states in classical and quantum mechanics	The outcome of a measurement on a superposition of states is only one of the states in the superposition
MEAS	Q4	Effects of repeated measurements on a quantum state	Difference between measurement in classical and quantum mechanics	Measurements on quantum objects have limitations due to the available instruments and experimental setups
	Q5	Relationship between uncertainties in position and velocity	Implications of Heisenberg principle	Uncertainty relations are due to experimental limitations
	Q9	Effects of measurement on the value of an observable	Consequences of the measurement on the wave function	After a measurement, the wave function eventually evolves back into the initial state
AEB	Q2	Atomic orbitals and their properties	Stability of atoms in quantum mechanics	Stability of atoms is explained by energy quantization
	Q8	Interactions in the hydrogen atom	Forces between an electron and the atomic nucleus	Gravitational force exerted on an electron is balanced by centrifugal force
	Q10	Motion of charged particles in electromagnetic fields	Why classical physics cannot explain stability of the hydrogen atom	A negatively charged particle should "fall" onto the nucleus due to Coulomb attraction.

topics: operators and observables, and tunneling. For the Schrödinger equation we decided to address only elementary properties of the wave function, choosing to not address its solution in specific situations (e.g., potential wells). The selection of concepts was also inspired by the basic learning goals of courses that are taught at the authors' universities. We discuss in more detail the educational context of the study in Sec. III E.

The selected concepts can be grouped in the following three general themes:

(1) Wave behavior of matter (WBM), which includes wave function and probability, time evolution, and superposition of states.
(2) Quantum measurement (MEAS), which includes the uncertainty principle and the "collapse" of the wave function.
(3) Atoms and electrons behavior (AEB), which includes atomic models and quantization of energy.

Clearly, these themes have fuzzy boundaries. As an example, in the Copenhagen interpretation, the measurement process on a system cannot be understood without the knowledge of the wave function properties. For such reason, we stress that in this study the three themes should not be intended as latent dimensions related to basic knowledge in quantum mechanics, but only as useful frames that guided our design of the QME items.

Still, the categorization of the QME items around three main ideas has the following advantages: (i) the choice of only three themes facilitates the administration of QME to a wide range of samples, including high school students and nonphysics majors teachers, as well as university students who attended upper level classes; (ii) at the same time, the topics are sufficiently broad to ensure a straightforward relationship between QME and the teaching of quantum mechanics at high school and university level. In particular, the chosen three themes are taught qualitatively in the Italian context at high school in the physics and chemistry courses [43] and then deepened in university courses for undergraduates. Thus, the QME may potentially track changes in students' understanding over the years. To this aim, we had to exclude any mathematical formalization from the targeted topics. This introduces an element of novelty in comparison to prior research that, instead, primarily included items requiring a sophisticated mathematical knowledge. To facilitate the reader, we summarize in Table II the contents targeted by previous assessment instruments (see also the Sec. I) and the main differences

TABLE II. Overview of research instruments to assess student's knowledge on quantum mechanics.

Researchers	Year	Type of questionnaire	Items	Level (students)	Country	Main targeted contents	QME concepts not targeted
Ireson [13]	2000	Likert scale	29	High school (342)	UK	Atomic structure and energy level quantization, wave-particle duality	Wave function time dependence and collapse superposition of states, measurement and uncertainty principle
Singh [9]	2001	Open-ended	5	Undergraduate (89)	US	Eigenvalues and eigenstates, stationary states, expectation values, measurement, spin	Atomic models
Zhu and Singh [14]	2012	Multiple-choice	31	Advanced undergraduate and graduate (226)	US	Wave function, bound or scattering states, measurement, expectation values and time dependence, stationary and nonstationary states, role of the Hamiltonian, uncertainty principle	Atomic models
Cataloglu and Robinett [15]	2002	Multiple-choice plus written explanation	25	Undergraduate (165)	US	Wave function, solution of Schrödinger equation (infinite well, tunneling, …), wave packet dynamics, uncertainty principle	Atomic models, wave function collapse, superposition of states measurement
Wuttiprom *et al.* [16]	2009	Multiple-choice	25	Undergraduate (312)	AUS	Photoelectric effect, wave-particle duality, de Broglie wavelength, electron diffraction, uncertainty principle	Atomic structure, wave function time evolution and collapse, measurement
McKagan *et al.* [17]	2010	Multiple-choice	12	Undergraduate (1033)	US	Wave function and probability, quantization of states, wave-particle duality, de Broglie wavelength, uncertainty principle	Atomic models, wave function time evolution and collapse, superposition of states and measurement
Sadaghiani and Pollock [6]	2015	Multiple-choice	31	Undergraduate (324)	US	Time-independent Schrödinger equation, wave function and boundary conditions, time evolution and probability, measurement	Atomic models, wave function collapse, uncertainty principle

with the QME contents. In the following, we describe the QME items in detail according to the three themes.

1. Wave behavior of matter

Four items target directly the wave function. Prior work concurs that students often do not grasp important aspects related to this concept, for at least two reasons: (i) the required shift from the classical, deterministic view of the physical world towards a probabilistic view; (ii) the mathematical sophistication involved in solving the Schrödinger equation often hides away the physical significance of the obtained result.

- Q1 probes the knowledge about the time evolution of the wave function ψ. Literature [44] suggests that students often confuse the time development of stationary and nonstationary states. The former have a simple time-dependent phase factor, whereas the latter, being a linear superposition of stationary states, exhibit a more complex time dependence. Therefore, in T1, we probe some basic facts about time evolution [9]: (i) that a quantum system evolves with time into a state that is different from the initial one; (ii) that the probability amplitude is a complex number and varies in space and time; (iii) the expectation value of an operator is time independent only if it commutes with the Hamiltonian or if the system is initially prepared in an energy eigenstate. Then, in T2, we ask to explicitly indicate that the phase factor of a stationary state depends on energy. We include in the answer choices typical misconceptions, such as the one that the phase factor depends only on the "possible" values of position and energy of the particle [8].
- Q3 probes the students' knowledge about elementary interpretation of the wave function. T1 addresses the following basic facts [45]: (i) the ψ function completely determines the state of a physical system (I and III); (ii) the wave function is a dimensional quantity, since its square modulus is a probability density (II). T2 asks students to correctly indicate the relationships between the wave function and the probability of any measurement outcome of a physical quantity.
- Q6 is closely linked to Q3 and deals more strictly with the wave behavior of matter [46–48]. We chose to address this issue investigating in T1 the basic knowledge about the state of a system and in T2 the wave behavior of an electron. T1 addresses the following basic facts: (i) the quantum state of particle is different from its classical counterpart, since it is not defined by assigning any values to its position and velocity (I and III); (ii) the quantum state of particle is completely determined by the wave function, which is not a physical wave (II). T2 asks to justify why an electron behaves like a wave. The students may recall the outcome of the double slit experiment and relate the wave behavior to the observation of a diffraction pattern on a screen (where the wave function collapses). Among the answer choices, we include the known misconception that the electron position oscillates as a wave, or in other terms, that the real trajectory of the particle is a sinusoid.
- Finally, Q7 deals with superposition of states in quantum mechanics. For systems in a superposition state, multiple values of the same physical observable are possible, until a measurement on the system is actually performed. One cannot exactly predict the outcome of any measurement, but only the probability to obtain a certain outcome. However, this goes beyond a classical statistical description of a system. Indeed, superposition in quantum mechanics is a rather counterintuitive concept, which may lead to paradoxical situations like Schrödinger's cat. For the sake of simplicity, T1 focuses on a wave function that is the superposition of two stationary states. In such a case [49]: (i) the idea that superposition of states leads to an "in-between" state is incorrect; (ii) we cannot know the state of the system until we make a measurement; (iii) superposition of states does not imply a complete lack of knowledge about the state of the particle. T2 probes the students' capability to reason about the different meaning of superposition in classical physics and quantum mechanics. In particular, to answer correctly, the student should refer to the interference term that arises from the square modulus of the sum of the original stationary states. Among the answer choices, we include the misconception that the outcome of any measurement is consistent with only one state of the initial superposition [8], which is actually true only for observables that commute with the Hamiltonian.

2. Measurement

Three items focus on the measurement issue in quantum mechanics. Literature suggests that students find problematic this subject for at least two reasons: (i) measurement in quantum mechanics is counterintuitive and very far from everyday experience; (ii) it is difficult to harmonize measurement in quantum mechanics with what they have learned earlier in classical physics [8,11], the uncertainty principle being the typical example for which classical reasoning leads to incorrect inferences; (iii) teaching of measurement in quantum mechanics is often abstract with no reference to real experimental procedures.

- Q4 probes students' knowledge of basic facts about measurement and the main difference with classical measurement. The three statements in T1 deal with (i) repeated measurements and how a measurement influences the result of subsequent ones (I and III) and (ii) the relationship between the measurement process and the state of a system (II). In T2, we ask students to further argue why it is not possible to use classical

arguments to describe measurement, focusing on the possibility that such a process irreversibly changes the state of a system. Among the answer choices, we include the misconception that the main difference between quantum mechanics and classical measurement process is that in quantum mechanics there exist more limitations due to the available instruments and experimental equipment when exploring the microscopic world.

- Q5 concerns the uncertainty principle. In T1, the three statements address the relationship between position and velocity uncertainties. T2 probes if the students correctly infer from the uncertainty principle that if one assigns numerical values to the position (velocity) of a particle then its velocity (position) has an infinite uncertainty and, consequently, is not defined. The answer choices correspond to the following categories of reasoning [50]: (i) uncertainty as an intrinsic consequence of quantum description of the microscopic world; (ii) uncertainty as measurement error or limitation due to experimental apparatuses; (iii) uncertainty as a measurement disturbance; (iv) uncertainty as limited precision in measurement.
- Q9 investigates the students' knowledge about the effects of the measurement process on the state of a system. As such, this item deals also with aspects related to the ψ function (namely, the so-called "collapse"). T1 deals with an electron and, in particular: (i) how the measurement process changes its previous state; (ii) and (iii) how the measurement process determines the new state. Then, in T2, the students are asked to reason in more detail about what happens to a generic quantum system after a measurement. Among the answer choices, typical misconceptions are featured, such as (i) the collapse of the wave function is only temporary and after the measurement it must go back to the original state, or (ii) the wave function after the measurement evolves into a state corresponding to what "it is supposed to be" [8]. We also include an incorrect answer choice based on a classical reasoning that refers to a change of the "wavelength" of the ψ function soon after the measurement.

3. Atoms and electrons behavior

Three items probe the students' knowledge about the interactions between nucleus and electron (hydrogen atom), the quasiclassical models of the hydrogen atom, the discretization of the energy levels, and the model of atomic orbitals [51–54]. The theme was chosen since it deals with topics that connect electromagnetism, chemistry, and quantum mechanics and because it is rarely addressed in previous instruments (see Table II).

- Q2 addresses in the first tier basic facts about orbitals and, in particular, (i) orbitals and probability; (ii) the relationship between atomic orbitals and energy levels; (ii) the spatial representation of the orbitals. Then, in T2, we probe how students explain the stability of an atom in quantum mechanics. Because of the emphasis on stationary waves associated with electrons, this item has some overlap with the item Q6 in the WBM theme. Among the incorrect answer choices, we included the misconception that the stability of an atom is explained by energy quantization.
- The first tier in Q8 recalls basic concepts about classical electromagnetism: (i) the energy of an accelerating charge; (ii) the electrostatic potential energy of a system formed by two identical charges; (iii) Coulomb force between identical charges. In T2, we probe if students are able to justify why gravitational forces are negligible at the atomic level. Among the incorrect answer choices, we include the misconception that there is a balance between the gravitational force directed towards the nucleus and the centrifugal force due to rotation and directed outwards.
- Q10 deals with basic facts about the interaction between an elementary charge and electromagnetic field. In particular, in T1 we address (i) the motion of an electron in a uniform magnetic field and (ii) the motion of an electron in an electric field. In T2, we probe students' justification about why quasiclassical models fail to describe atoms' stability. The correct answer choice refers to energy loss of an accelerated charge due to its radiation. Among the incorrect answer choices, we included misconceptions found in pilot administrations of the questionnaire: (i) loss of energy due to Coulomb force; (i) lack of the uncertainty principle in classical physics; (iii) classical physics deals only with macroscopic systems.

The complete questionnaire is reported in the Appendix.

C. Scoring of QME

As discussed in Sec. II, two main approaches have been proposed to score two-tier items [21]. One common method is pair scoring, where full credit is given only if the student answers correctly in both tiers. The other common method is individual scoring, where the tiers are scored independently. The first method reduces guessing but does not adequately describe intermediate stages of students' knowledge. The second method allows partial credit but does not address guessing. Both assume that the two tiers have the same difficulty. This is not always the case, since the two tiers often probe different students' abilities, namely, knowing and explaining, and hence may have different difficulty levels [55]. Moreover, the extent to which different possible scoring methods may lead to different results and psychometric properties of the instrument has been rarely investigated.

TABLE III. QME scoring patterns.

				Score assignment					
Number of correct answers in T1	T1 Score	Answer to T2	T2 Score	M1	M2	M3	M4	M5	M6
Less than 2 out of 3	0	Wrong	0	0	0	0	0	0	0
		Correct	1	0	1	1	2	0	1
At least 2 out of 3	1	Wrong	0	0	1	2	1	1	0
		Correct	1	1	2	3	3	2	2

In this study, we adopted a mixed method, which first scores the tiers individually and then couples the tiers according to the different patterns of responses. By adopting such a method, we tracked the individual functioning of the tiers as well as their coupled behavior.

Hence, we first defined the score variable of T1 by giving (i) 0 point if a student answered correctly either one or none of the questions in T1 and (ii) 1 point if a student answered correctly at least two out of the three questions. Then, we defined the score variable of T2, by giving the full score (1 point) only if the correct answer choice had been picked. By combining the two variables, four patterns are obtained, namely, "00," "01," "10," and "11." The pattern "00" corresponds to an insufficient factual knowledge as well as an incorrect reasoning about the target topic. Pattern "11" corresponds to good factual knowledge and correct reasoning about the target topic. Partially correct patterns "01" and "10" correspond, respectively, to insufficient factual knowledge but correct reasoning and to good factual knowledge but incorrect reasoning about the target topic. According to the relative weight given to factual knowledge and reasoning capability, the four patterns can be scored in a total of six different ways [21], as detailed in Table III. $M1 = T1 \times T2$ is the pure pairing score and admits no intermediate levels. $M2 = T1 + T2$ is a pairwise method equivalent to individual scoring that accounts for local independence. $M3 = 2 \times T1 + T2$ and $M4 = T1 + 2 \times T2$ assume that knowing (reasoning) is more demanding than reasoning (knowing), respectively. Both M3 and M4 do not address guessing issues. $M5 = T1 \times (1 + T2)$ and $M6 = T2 \times (1 + T1)$ are mixed methods that account for guessing: M5 assumes that reasoning is harder than knowing and regards the pattern "01," namely, an incorrect knowledge of facts but correct reasoning, as guessing. M6 rewards knowledge more than reasoning and considers the pattern "10," namely, a correct knowledge of facts but incorrect reasoning, as guessing.

D. Data analysis

1. Classical analysis

Since the two tiers of QME were hypothesized to be distinguishable yet related, we first investigated the local independence between the tiers using cross tabulation between T1 and T2, for all questions [56]. We found that the T1–T2 combinations of all QME items met the local independence assumption. Therefore, we treated T1 and T2 items first as separate and then coupled by the six scoring methods. We report the complete local independence analysis in the Supplemental Material A [57].

Then, to establish QME internal consistency, we calculated Cronbach's alpha and the following classical test theory indices–item difficulty, discrimination, and point biserial–treating T1 and T2 first as separate and then coupled. Item difficulty is classically defined as the frequency of correct answers for a given question and should be between 0.2 and 0.8. When analyzed separately, the difficulty of T1 and T2 was calculated as the frequency of score 1 in the students' responses to T1 and T2, respectively. When T1 and T2 were considered coupled together, we calculated the difficulty of an item as the frequency of the pattern "11" in the students' answers. Therefore, the difficulty of the separate and coupled items was the same across the six methods. Discrimination and point biserial describe the extent to which difficult items are more likely answered by more able students. We used the extreme group method to calculate the indices, choosing the top and bottom 30% of the distribution. The discrimination index should be positive. As a rule of thumb [58], excellent items have values greater or equal to 0.4, good items greater than 0.3, acceptable items greater than 0.2, low discrimination items greater than 0.1, poor items lower than 0.1. For QME, we considered items with discrimination value below the 0.1 threshold as problematic, likely subjected to a revision. For point biserial, good items have values above 0.25. As for the difficulty index, discrimination and point biserial were calculated for each tier separately and for the tiers coupled together according to the six scoring methods shown in Table III. To further inspect the expected functioning of the tiers, namely, that T2 should be harder than T1, we calculated the correlations between the score in each tier, obtained with the six methods, and the total score. Finally, we investigated the level of agreement across the six scoring methods, adopting an approach based on interrater reliability. For each method, we first split the students according to their normalized score using percentiles 30 and 70 as thresholds, thus obtaining three groups. Then, using cross tabulation, we calculated for each pair of scoring methods (e.g., M1-M2, M1-M3, M1-M4, and so on, 15 combinations in total) the corresponding Cohen's kappa. To increase the validity of our findings, we also used a different method to split the students according to their

normalized score. In this second case, we chose as threshold the normalized score 0.5, with the aim to investigate the agreement across the methods that the individuals reach a sufficiency level in QME. For this splitting method, Cohen's kappa was calculated in the same way as before using cross tabulation. All calculations were carried out through SPSS software.

2. Rasch analysis

The scoring procedures described in Table III have the advantage that they all couple T1 and T2 under different assumptions on the relationship between knowledge and reasoning. M2, M5, and M6 also account for the probability of correctly answering to T1 and T2 simply by chance. However, the six scoring procedures still hypothesize linearity, i.e., that the "distance," namely, the extent to which two consecutive scores differ, between the patterns "01" and "10" is the same as the "distance" between the patterns "10" and "11." Moreover, if one considers, for instance, the scoring method M2 (Table III), a correct answer to less than two T/F claims in T1 (incorrect knowledge of facts) but with a full score in T2 (correct reasoning) is "equivalent" to a correct answer to at least two T/F claims in T1 (correct knowledge of facts) but with no credit in T2 (incorrect reasoning). Such a limitation is common to instruments, such as Likert-type questionnaires, that use a discrete score variable. To address this issue, it has been recently proposed [55] to analyze two-tier instruments by using Rasch analysis, which is able to reliably establish the extent to which the difficulty of the tiers is different.

For simplicity, in this paper, we will use the one-dimensional Rasch model [59]. For the dichotomous method M1, the model is expressed by

$$P(X_{ij}=1)=\frac{\exp(\beta_i-\theta_j)}{1+\exp(\beta_i-\theta_j)}=\frac{1}{\exp(\theta_j-\beta_i)+1}. \quad (1)$$

Equation (1) describes the probability of obtaining a full credit on the jth item of a test by the ith student. X_{ij} is the score on the item, β_i is the ith student's "ability," and θ_j, the jth item "difficulty." In Rasch analysis, the term ability indicates the *trait level* of the ith student [60], namely, the extent to which a student possesses the trait targeted by the questionnaire, in our case the knowledge of basic aspects in quantum mechanics. The numerical values of θ_j and β_i are obtained by fitting the data to the sigmoidal test characteristic curve, and measured in *logit*. For a questionnaire with J items, the mean item difficulty is set to 0. Therefore, if the sample mean ability is about 0, the students, on average, have a 50% chance to correctly answer the items of the questionnaire. Mean ability values slightly above (below) 0 indicate that the questionnaire was slightly less (more) difficult for the sample as a whole. Mean ability values much larger (smaller) than 0 indicate that the questionnaire was very easy (difficult) for the sample. A very interesting feature of Rasch analysis is the graphical representation that allows us to explore the students' ability distribution across the questionnaire's items. This representation is known as the *Wright map*, which displays, in the same plot, the persons' ability and items' difficulty.

For the polytomous methods M2–M6, the one-dimensional Rasch model is generalized by

$$P(X_{ij}=l)=\frac{\exp\sum_{k=1}^{l}(\beta_i-\theta_{jk})}{1+\sum_{m=1}^{n_j}\exp\sum_{k=1}^{m}(\beta_i-\theta_{jk})} \quad (2)$$

Equation (2) describes the probability of obtaining a score equal to $l=1,\ldots,n_j$ on the jth item of the test by the ith student. The parameter θ_{jk} can be interpreted as the difficulty of the kth score. More precisely, it is the threshold value (in logit) that a student must overcome to score k rather than $k-1$.

In this study, several Rasch indices were reviewed. Goodness of model fit was investigated through mean square (MNSQ) *outfit* and *infit*, which indicate whether students' responses showed more randomness than expected. Acceptable MNSQ infit values are between 0.7 and 1.3 [61]. For instance, an item with MNSQ infit of 1.4 has a variability that is 40% greater than expected. Further indices that measure the reliability of a questionnaire in the Rasch model framework are *person reliability* (similar to Cronbach's alpha, acceptable values above 0.5), *item separation*, and *person separation reliability*. Item separation reliability indicates whether the sample was able to discriminate between the items according to their difficulty. Acceptable values are above 3 [61]. Person separation reliability indicates the distribution of person abilities across the questionnaire's items, so it can be used to investigate if the sample can be divided into levels of increasing ability. Acceptable values are above 2 [61]. As for classical analysis, we first analyzed the students' responses to the tiers separately, by comparing the two tiers' average difficulties, obtained with data fit, for each item and for groups of items clustered according to the addressed subject. Then, we analyzed the students' responses to the coupled tiers according to the six scoring methods. We calculated person reliability, person separation, item separation reliability, and MNSQ infit and outfit for all the six methods. To inspect whether the possible patterns of response to the tiers corresponded to different level of ability, we also analyzed the item response curves (IRCs) for the ten coupled tiers [62,63]. In our case, the shape of IRCs can be used to provide further insight about possible differences across the scoring methods. Then, for all the six methods, we probed QME capability of discriminating between groups in the sample with different background in quantum mechanics by comparing the different groups' abilities, obtained with data fit, through analysis of variance (ANOVA). Finally, we investigated the level of agreement of the six scoring methods when using Rasch measures. In this case, we used

the ability measures of the students. For each scoring method, we divided the students according to (i) the 30 and 70 percentiles thresholds, obtaining three groups and (ii) the threshold ability > 0 logit, obtaining two groups. For the latter case, we thus investigated the agreement across the methods to assign to individuals a probability larger than or equal to 50% to correctly answer the QME items. We calculated Cohen's kappa for both splitting methods using cross tabulation. All calculations for the Rasch analysis were carried out through Winsteps software (version 3.93).

E. Sample

We administered the questionnaire to a sample of 445 subjects divided into 4 groups:

- G1: 146 freshmen enrolled in a physics degree at a large university in southern Italy. Such students had previously learned some quantum mechanics topics solely at high school (i.e., one year before) using a widespread Italian textbook [64];
- G2: 86 third-year undergraduates pursuing a bachelor's degree in physics at the same G1 university. Such students had already attended an upper-level class in quantum mechanics;
- G3: 139 freshmen enrolled in a physics degree at the same G1 and G2 university. Differently from G1, such students had attended a special extra-curriculum class of quantum mechanics during their last year of high school using the materials in Ref. [65];
- G4: 74 high school physics teachers attending a professional development course organized by the same G1, G2, and G3 university.

Details about the education and the background knowledge of quantum mechanics for each group are reported in Supplemental Material B [57]. Respondents were given about one hour to answer the questionnaire.

IV. RESULTS

A. Respondents' conceptions about basic aspects of quantum mechanics

In the following, we report, for each of the three chosen themes, the conceptual difficulties that emerged from the analysis of the responses. The frequency of involved subjects' answers to each statement in T1 and each question in T2 is reported in Table IV.

TABLE IV. Frequencies of subjects' answers ($N = 445$) to QME.

		T1[a]			T2[b]			
Theme	Item	I	II	III	a	b	c	d
WBM	Q1	0.65	0.69	0.64	*0.25*	**0.29**	0.23	0.23
	Q3	0.27	0.37	0.34	0.17	0.20	**0.27**	*0.36*
	Q6	0.55	0.48	0.49	0.22	*0.23*	0.09	**0.46**
	Q7	0.53	0.52	0.48	*0.29*	0.21	0.08	**0.42**
MEAS	Q4	0.67	0.66	0.64	*0.11*	0.06	0.09	**0.74**
	Q5	0.55	0.45	0.47	**0.15**	*0.39*	0.22	0.24
	Q9	0.43	0.51	0.48	0.25	0.22	**0.20**	*0.33*
AEB	Q2	0.13	0.73	0.54	**0.19**	*0.45*	0.19	0.17
	Q8	0.65	0.60	0.69	**0.68**	0.02	0.02	*0.28*
	Q10	0.75	0.76	0.69	*0.22*	0.13	0.16	**0.50**

[a]The percentage of correct answers (true or false) is reported.
[b]Correct answer choices are in boldface, typical misconceptions are in italics.

1. Wave behavior of matter

Q3 and Q1 were the most difficult items. We discuss first Q3 since it concerns the basic properties of the wave function. For T1, only 34% of the subjects rated as true that the wave function completely determines the state of a particle, and only 27% correctly rated as false that the wave function determines the "allowed" states of a particle. Another incorrect idea about the wave function, which emerged from T1 (63%), is that it has no "physical dimensions," pointing to a possible confusion between probability and probability density. In T2, less than 30% of the sample correctly indicated that the wave function allows one to predict the probability of measuring a given value of a physical quantity. About 36% of the sample picked as answer choice the typical misconception that the wave function allows one to predict the "states" of a particle, which is in accordance with a deterministic rather than probabilistic view. Similar difficulties emerged when dealing with the time evolution of the wave function of a stationary state and the relationship between the frequency of oscillation and the energy (Q1). While the claims in T1, taken individually, were correctly recognized as true or false on average by two-thirds of the subjects, we found that slightly less than one-third of the sample was able to correctly indicate that the phase factor of a stationary state depends on its energy. Better results were achieved by the sample in questions Q6 and Q7, focused on the differences between wave and particle behavior and between superposition of states in classical and quantum mechanics, respectively. For Q6, about 50% of the sample was able to identify the "interference" of an electron with itself as the main evidence for the wave behavior of matter. Similarly, for item Q7, 42% of the sample was able to explain the superposition of stationary states in terms of constructive and destructive interference. However, about one-third of the subjects believed that a measurement on a superposition of stationary states necessarily leads to "delete" one of the states of the original superposition, or that the resulting state after the measurement is a kind of "intermediate" or "averaged" state between the original ones.

2. Measurement

Analysis of responses shows that Q5 and Q9 were difficult questions for the sample. Concerning Q5, while

TABLE V. Classic test theory statistics for QME (tiers analyzed individually).

Theme	Item	T1			T2		
		Difficulty	Discrimination[a]	P.-bi serial[a]	Difficulty	Discrimination[a]	P.-bi serial[a]
WBM	Q1	0.72	0.19	0.43	0.29	0.07	0.21
	Q3	0.23	0.06	0.24	0.27	0.29	0.49
	Q6	0.57	0.25	0.50	0.46	0.35	0.60
	Q7	0.53	0.23	0.48	0.42	0.28	0.48
	Average	*0.51*	*0.18*	*0.41*	*0.36*	*0.25*	*0.44*
MEAS	Q4	0.72	0.11	0.32	0.74	0.28	0.61
	Q5	0.41	0.29	0.54	0.15	0.24	0.31
	Q9	0.48	0.17	0.41	0.20	0.46	0.58
	Average	*0.53*	*0.19*	*0.42*	*0.36*	*0.33*	*0.50*
AEB	Q2	0.53	0.14	0.38	0.19	0.58	0.59
	Q8	0.68	0.20	0.45	0.68	0.34	0.65
	Q10	0.80	0.11	0.33	0.50	0.51	0.72
	Average	*0.67*	*0.15*	*0.38*	*0.46*	*0.48*	*0.65*
Overall	*Average*	*0.57*	*0.17*	*0.41*	*0.39*	*0.34*	*0.52*

[a]Indices were calculated having split the sample in the least and most able 30% of the subjects.

responses to tier 1 show a sufficient knowledge about Heisenberg's principle, only 15% of tier 2 responses demonstrate a sound conceptual understanding of this fundamental relation. In particular, almost 40% of the subjects exhibit the typical misconception that the principle is due to limitations of the experimental apparatus. For Q9, responses to T1 show similar results to Q5. Responses to T2 show that about one-third of the subjects believed that in the "new" state, created by the measurement, there is a perfect correspondence between expected and measured values of the observable. Another one-fourth expected that after some time the new state goes back to the initial one. Only 20% of the subjects correctly interpreted how the measurement process affects the state of a system. The easiest question was Q4. The percentage of correct rating for T1 statements is about 65%, which suggests that the sample had at least a basic knowledge about the measurement process in quantum mechanics. In T2, about 75% of the subjects were able to identify the correct reason for which the measurement process is different in quantum mechanics and classical physics.

3. Atoms and electrons behavior

Concerning this theme, we found that involved subjects had difficulty correctly explaining how quantum mechanics justifies stability of the simplest atom (hydrogen). In the first tier of Q2, we detected the typical misconception according to which an orbital is a region of space around an atom (87%), while in T2 the incorrect reasoning for which the stability of the hydrogen atom is justified by the quantization of energy levels (45%). On the other hand, the majority of the subjects (on average 66%) showed a correct knowledge about the interaction forces between electrons and the nucleus (Q8) and about the predictions and limitations of early models (Q10) of the hydrogen atom.

B. Psychometric properties of QME

1. Classical analysis

When the tiers are individually analyzed, QME is made up of 20 dichotomous items scored according to Table III. Cronbach's alpha is moderate (0.59), which could be expected since the QME targets a broad range of concepts in quantum mechanics and was administered to a nonhomogeneous sample. For such reasons, we also calculated lambda-2 Gutmann coefficient. The value is slightly greater, 0.61, suggesting an acceptable behavior of the QME. Table V reports the difficulty and discrimination indices for T1 and T2. To get further insights on the functioning of the tiers separately, we plot in Figs. 1(a) and 1(b) the difficulty and discrimination of tier pairs. In Fig. 1(a), the straight line separates the region A, where item difficulty of T1 is greater than the corresponding difficulty of T2, from region B, with reverse properties.

The difficulty of the second tier is greater than the difficulty of the corresponding first tier for seven items, while only for three items (Q3, Q4, Q8) T1 and T2 have about the same difficulty. The average difficulty index of T1 (0.57) is higher than that of T2 (0.39), which means that T2 was on average harder than T1. From Fig. 1(b), we note that all items have acceptable values of the discrimination index for the second tier (Q2, Q9, and Q10 are excellent), except Q1, which has also a low T1 discrimination index. Six items have T1 discrimination values smaller than 0.2, but still greater than 0.1, except item Q3, which has a poor discrimination value for T1 and should be revised (all values under 0.1 threshold are underlined in Table V).

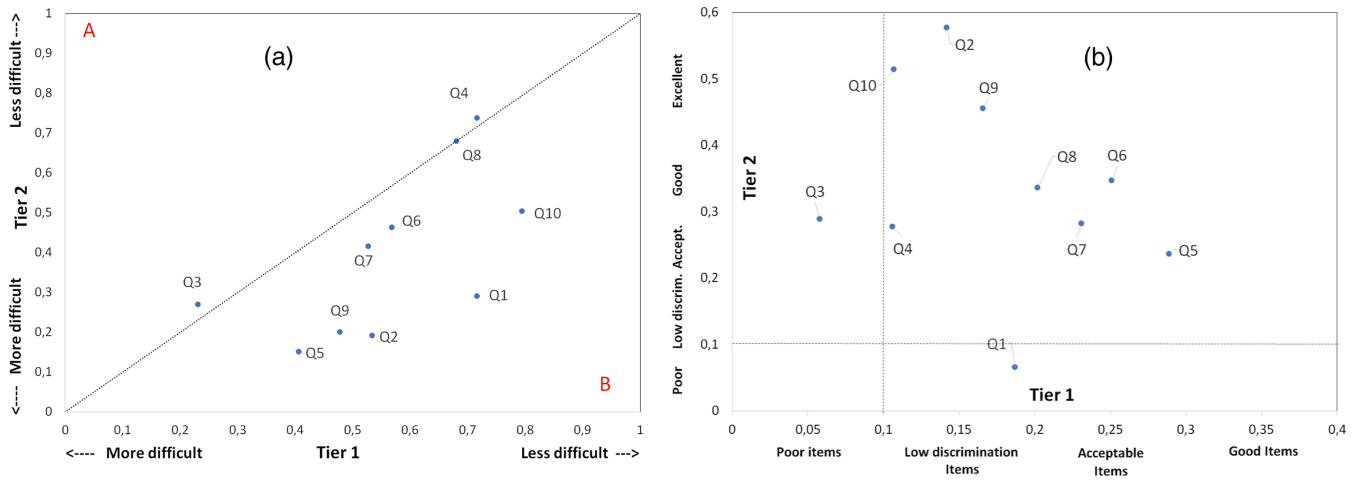

FIG. 1. (a) Dispersion plot of the QME items' difficulty in the T1–T2 plane when tiers are analyzed separately. See text for the definition of regions A and B. See Table V for further details. (b) Dispersion plot of QME items discrimination in the T1–T2 plane when tiers are analyzed separately. Straight dotted lines indicate the 0.1 threshold under which items have poor discrimination power. See Table V for further details.

Average T1 discrimination and point-biserial indices are 0.17 and 0.41, respectively, both lower than the corresponding values for T2 (0.34 and 0.52). The implication is that T2 increases the difficulty index and the discriminating power of the test. Summarizing, classical analysis supports the effectiveness of the test design aimed to balance the tiers' functioning.

2. *Dependency of classical psychometric properties on tiers' scoring method*

When the tiers are analyzed through the coupling introduced by the six scoring methods, QME is made of 10 items, dichotomous in the case of M1 and polytomous in the case of M2–M6. In Table VI we report, for all six scoring methods, the Cronbach alpha, Gutmann lambda-2, the normalized scores, and their correlations with the tiers' individual scores. We note that the reliability of QME changes significantly across scores (from 0.44 to 0.57). Among the polytomous methods, M6 has the lowest value, while M2–M5 have values that are close to the one obtained when considering the tiers separately (slightly lower than 0.60). As expected, the six methods differ in the average normalized score. M1 and M6 have the smallest values of average score, while M2 and M3 have the highest average score. Coherently with such result, M1 and M6 show very different correlations between the tiers and the normalized score. In particular, for both methods, the T2 score has the highest correlation with the total score (about 1) but a small value of the correlation between T1 score and the total score. While such evidence could be expected for M6, the

TABLE VI. Classical reliability indices and correlations between total score and tiers scores for QME (coupled tiers).

	M1	M2	M3	M4	M5	M6
Cronbach alpha	0.44	0.57	0.55	0.56	0.51	0.48
Guttman lambda-2	0.47	0.59	0.56	0.58	0.53	0.51
Mean normalized score	0.263 ± 0.007	0.478 ± 0.007	0.507 ± 0.008	0.448 ± 0.007	0.415 ± 0.0078	0.327 ± 0.008
Correlation T1[a]	0.58	0.78	0.90	0.64	0.87	0.49
Correlation T2[b]	0.95	0.85	0.72	0.94	0.75	0.99
			Cohen's kappa			
M1	⋯	0.64[c]	0.60[c]	0.65[c]	0.68[c]	0.73[c]
M2	0.12[d]	⋯	0.82[c]	0.78[c]	0.77[c]	0.66[c]
M3	0.08[d]	0.81[d]	⋯	0.60[c]	0.86[c]	0.49[c]
M4	0.15[d]	0.83[d]	0.64[d]	⋯	0.56[c]	0.80[c]
M5	0.21[d]	0.71[d]	0.54[d]	0.59[d]	⋯	0.52[c]
M6	0.46[d]	0.35[d]	0.25[d]	0.43[d]	0.42[d]	⋯

[a]Correlation between T1 average score and average normalized score.
[b]correlation between T2 average score and average normalized score.
[c]sample split according to 30 and 70 percentiles threshold.
[d]sample split according to 0.5 threshold of the normalized score.

TABLE VII. Classic test theory statistics for QME scoring methods M1–M6 (coupled tier).

Th.	Item	M1			M2		M3		M4		M5		M6	
		Diff.	Disc[a]	P.-bi[a]	Disc[a]	P.-bi[a]	Disc[a]	P.-bi[a]	Disc[a]	P.-bi[a]	Disc[a]	P.-bi[a]	Disc[a]	P.-bi[a]
WBM	Q1	0.22	0.18	0.42	0.29	0.54	0.35	0.59	0.17	0.41	0.33	0.58	0.08	0.29
	Q3	0.09	0.05	0.23	0.06	0.25	0.07	0.26	0.12	0.35	0.03	0.18	0.11	0.33
	Q6	0.28	0.32	0.56	0.27	0.52	0.31	0.55	0.29	0.54	0.30	0.55	0.27	0.52
	Q7	0.26	0.25	0.50	0.31	0.56	0.40	0.63	0.28	0.53	0.33	0.57	0.18	0.42
MEAS	Q4	0.59	0.37	0.61	0.31	0.56	0.30	0.55	0.31	0.56	0.27	0.52	0.30	0.55
	Q5	0.07	0.03	0.17	0.11	0.33	0.14	0.38	0.09	0.31	0.10	0.31	0.04	0.21
	Q9	0.09	0.11	0.33	0.17	0.42	0.17	0.41	0.20	0.44	0.15	0.39	0.13	0.36
AEB	Q2	0.09	0.10	0.32	0.12	0.35	0.15	0.38	0.17	0.41	0.08	0.29	0.14	0.38
	Q8	0.51	0.48	0.70	0.39	0.62	0.36	0.60	0.42	0.64	0.36	0.60	0.44	0.67
	Q10	0.45	0.47	0.69	0.39	0.63	0.33	0.57	0.46	0.68	0.33	0.58	0.39	0.62

[a]Indices calculated having split the sample according to 30 and 70 percentiles.

result for M1 may only be justified with the fact that T2 was harder than T1. For the other methods, M3–M5 follow the expected correlational patterns, namely, the greater the weight of the tier in the total score the higher the correlation. The pairwise method M2 shows on the other hand a slightly greater correlation between the T2 score and the total score, thus confirming that T2 was harder than T1.

Cohen's kappa analysis confirms that scoring methods are not equivalent. Using percentiles to split the sample, intermethod reliability is above 0.8 only for M2–M3 (0.82), M3–M5 (0.86), and M4–M6 (0.8). The latter results are strongly correlated with the weight given to T1(T2) by M3(M4) and M5(M6). The first result suggests that equally weighing a correct response to both the tiers favored individuals in our sample who answered correctly mainly to T1. This is confirmed also by the good agreement of M2 with M4 (0.78) and M5 (0.77). M1 has moderate agreement with all the other methods (on average 0.65), with a higher concordance with M6. When considering a threshold in the normalized score to split the sample, the intermethod reliability indices decrease significantly and show poor agreement between the methods. Notably, only M2–M3 (0.81) and M2–M4 (0.83) are above 0.80. M1 and M6 show the lowest values of agreement in comparison to all other methods (the highest is between them, 0.46) likely because the percentage of respondents with normalized score greater than 0.5 is 4.3% for M1 and 13% for M6.

Finally, we analyzed the difficulty index, discrimination, and point biserial for the 10 coupled items across the six methods. Results are shown in Table VII. The difficulty index, defined as the frequency of the highest score, is the same for the six scoring methods (see Table III). Four items present a value of the difficulty index lower than the recommended threshold of 0.20, namely, they were very difficult for the sample. They are Q2 (atoms and electrons behavior), Q3 (wave behavior of matter), Q5, and Q9 (measurement). From Fig. 1(a), we note that the difficulty of items Q2, Q5, and Q9 is mainly due to T2, while for Q3, the difficulty is due to both T1 and T2. As expected, very difficult items have low discriminating power, as a quick look at the underlined values in Table VII confirms.

However, differences across scoring methods emerge mainly for the different weights of the tiers. In the case of item Q3, for instance, scoring methods M4 and M6 seem to slightly increase QME discriminating power.

3. Rasch analysis

When tiers are analyzed individually, person reliability is 0.59, which can be considered sufficient. Person separation is 1.22, while item separation is 8.93, both acceptable values. Item statistics are reported in Table VIII. It is noteworthy that all items have acceptable infit and outfit MNSQ values. Average difficulty of the first tier is −0.45 logit while that of the second tier is +0.45 logit. The difference is slightly significant ($t = -2.100$; $p = 0.05$). Concerning T1, the most difficult ones are the WBM items (−0.16 logit), followed by MEAS items (−0.30 logit) and AEB items (−0.99 logit). Concerning T2, the hardest items are the MEAS ones (0.65 logit), followed by the WBM items (0.56 logit) and AEB items (0.12 logit). Concerning the discrimination power, Q2 and Q3 have point-biserial values much lower than 0.3 in T1, Q1 and Q5 in T2. Note that Q3 is the hardest item in T1, Q5 in T2. When looking at the differences between the average difficulty of T1 and T2 for the three targeted themes we note that T2 items are consistently harder than T1 items, for all the three themes (0.56 vs −0.16; 0.65 vs −0.30; 0.12 vs −0, 99). However, when investigating more quantitatively the effects of the tiers on the measured difficulty of the items, the univariate model reveals a nonsignificant difference ($F = 3.940$, $p = 0.067$). Similarly, also the interaction effect of tier × theme is not significant ($F = 0.065$, $p = 0.938$). Further details can be obtained from the general Wright map (Fig. 2).

Overall, the map confirms that QME items span a wide range of difficulty levels (−1.68 → +1.81 logit). A ranking pattern of the themes also emerges, from the easiest one,

TABLE VIII. Rasch analysis statistics for QME (tiers analyzed individually).

Theme	Item	T1				T2			
		Difficulty	Infit	Outfit	P.-bi serial	Difficulty	Infit	Outfit	P.-bi serial
WBM	Q1	−1.19	0.897	0.842	0.47	0.90	1.115	1.149	0.19
	Q3	1.24	1.099	1.293	0.15	1.01	0.992	1.011	0.31
	Q6	−0.44	0.955	0.938	0.42	0.05	1.049	1.091	0.30
	Q7	−0.25	0.904	0.877	0.47	0.27	1.053	1.096	0.28
	Average	*−0.16*	*0.964*	*0.988*	*0.378*	*0.56*	*1.052*	*1.087*	*0.27*
MEAS	Q4	−1.19	0.994	1.067	0.34	−1.31	0.924	0.861	0.44
	Q5	0.31	1.055	1.051	0.29	1.81	1.034	1.173	0.19
	Q9	−0.03	1.084	1.115	0.26	1.44	0.963	1.005	0.31
	Average	*−0.30*	*1.044*	*1.078*	*0.30*	*0.65*	*0.974*	*1.013*	*0.31*
AEB	Q2	−0.29	1.170	1.220	0.17	1.50	0.909	0.891	0.38
	Q8	−1.00	0.966	0.976	0.39	−0.99	0.914	0.871	0.46
	Q10	−1.68	0.916	0.807	0.44	−0.14	0.906	0.901	0.46
	Average	*−0.99*	*1.017*	*1.001*	*0.33*	*0.12*	*0.910*	*0.888*	*0.43*
Overall	*Average*	*−0.45*	*1.004*	*1.019*	*0.34*	*0.45*	*1.000*	*1.004*	*0.33*

atoms and electrons (average difficulty = −0.43 logit), through the more difficult ones, measurement (average difficulty = +0.17 logit) and wave behavior of matter (average difficulty = +0.20). The map also confirms that T2 items were on average harder than T1 items for all questions, with the only exception of Q3, for which T1 was the hardest item in the WBM theme.

4. Dependency of Rasch psychometric properties on tiers scoring method

Complete descriptive statistics of Rasch analysis when tiers are coupled by the six scoring methods is reported in Table IX. The results suggest that M1 has not sufficient fitting reliability (lower than 0.5). Data also suggest that none of the scoring models M2–M6 is clearly more reliable than the others, with the exception of M2, which evenly weighs the tiers. M1 has also a poor person separation reliability (less than 1), which indicates that this scoring method is the least appropriate to distinguish between high and low ability subjects. When looking at the level of agreement, also when adopting Rasch analysis, the six scoring methods are not equivalent. Results in fact are similar to those obtained with classical analysis. For the percentiles subdivision of the sample, the highest agreement is between M2 and M3 (0.82) and M4 (0.82), and between M3/M5 (0.80) and M4/M6 (0.80). M1 has the lowest values of agreement with the other scoring methods,

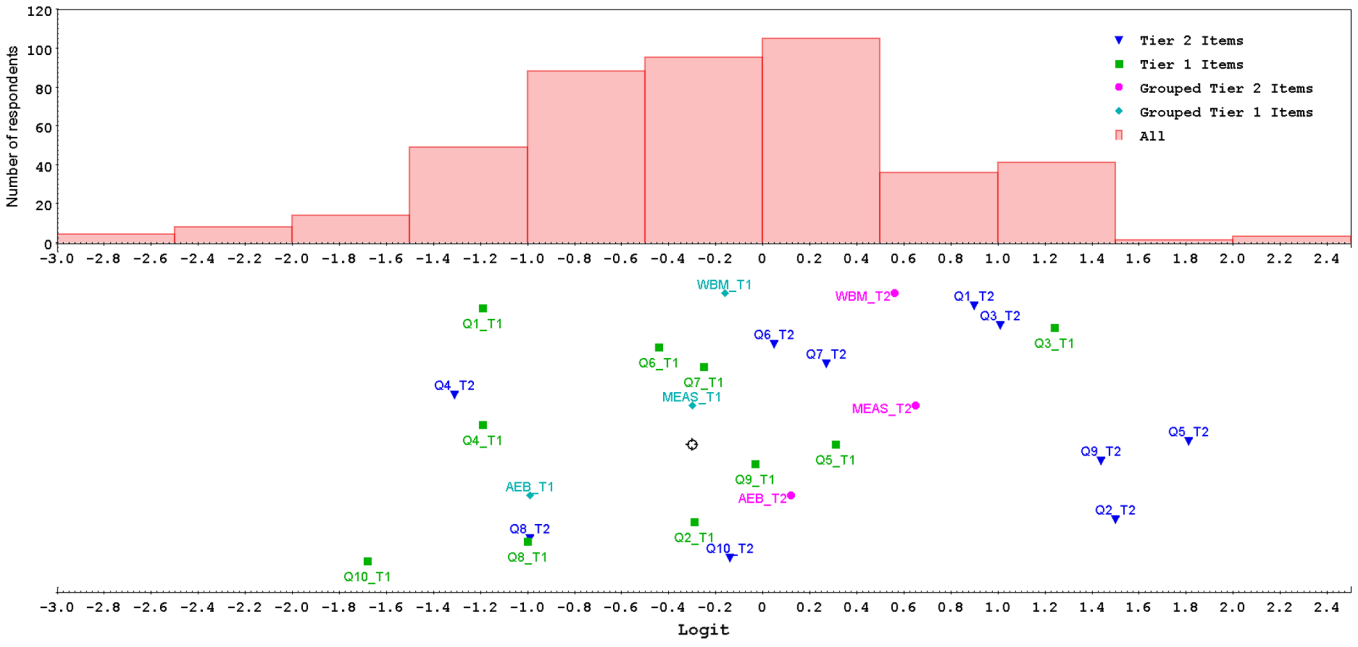

FIG. 2. Wright map of the QME. Items are grouped according to the three themes described in Sec. III. One extreme person (ability = −4.63) is not represented.

TABLE IX. Rasch analysis statistics for QME (tiers analyzed coupled).

	M1	M2	M3	M4	M5	M6
Person reliability	0.36	0.62	0.59	0.56	0.58	0.54
Person separation	0.76	1.27	1.20	1.27	1.17	1.04
Item separation	8.74	10.62	9.73	10.30	9.60	9.19
Mean ability (logit)	-1.60 ± 0.06	-0.13 ± 0.04	-0.05 ± 0.03	-0.12 ± 0.03	-0.40 ± 0.04	-0.67 ± 0.03
Ability G1 (logit)	-2.16 ± 0.11	-0.54 ± 0.07	-0.29 ± 0.04	-0.39 ± 0.04	-0.74 ± 0.07	-0.10 ± 0.06
Ability G2 (logit)	-0.81 ± 0.13	0.46 ± 0.09	0.27 ± 0.07	0.27 ± 0.07	0.01 ± 0.08	-0.18 ± 0.07
Ability G3 (logit)	-1.49 ± 0.09	-0.06 ± 0.06	0.01 ± 0.04	-0.10 ± 0.04	-0.28 ± 0.05	-0.65 ± 0.05
Ability G4 (logit)	-1.62 ± 0.14	-0.13 ± 0.09	-0.07 ± 0.05	-0.09 ± 0.06	-0.44 ± 0.08	-0.64 ± 0.08
F[a]	23.449	29.659	23.936	28.165	20.820	26.315
η^2[a]	0.138	0.168	0.140	0.161	0.124	0.152
			Cohen's kappa			
M1	...	0.54[b]	0.48[b]	0.56[b]	0.56[b]	0.64[b]
M2	0.36[c]	...	0.82[b]	0.82[b]	0.70[b]	0.66[b]
M3	0.26[c]	0.81[c]	...	0.64[b]	0.80[b]	0.50[b]
M4	0.35[c]	0.86[c]	0.67[c]	...	0.54[b]	0.80[b]
M5	0.54[c]	0.71[c]	0.54[c]	0.58[c]	...	0.48[b]
M6	0.75[c]	0.53[c]	0.39[c]	0.52[c]	0.53[c]	...
			Infit statistics			
Q1	1.1276	0.8304	0.747	0.9766	0.7053	1.3089
Q2	0.9807	0.8392	0.9495	0.7671	0.8619	0.8921
Q3	1.0222	1.1787	1.121	1.2549	1.2728	0.879
Q4	1.0341	1.193	1.2807	1.0976	1.2757	0.9585
Q5	1.035	0.9909	0.9922	1.025	0.8871	1.0484
Q6	0.9466	0.9839	0.9549	0.9854	0.9986	0.9366
Q7	1.0269	1.0613	0.9816	1.0975	1.0337	1.0814
Q8	0.9396	1.0401	1.123	0.9645	1.1978	0.8767
Q9	0.8921	0.899	0.9716	0.8569	0.8454	0.877
Q10	0.9378	0.9639	1.0188	0.9312	0.9211	1.0028
RMSE	0.07	0.13	0.14	0.14	0.19	0.14
			Outfit statistics			
Q1	1.1293	0.8439	0.7358	1.0172	0.7129	1.4154
Q2	0.9216	0.884	1.0209	0.8022	0.9709	0.8699
Q3	1.3945	1.2368	1.249	1.3249	1.3702	1.0998
Q4	1.076	1.1709	1.3353	1.0471	1.3164	0.938
Q5	1.536	1.039	1.0392	1.1218	1.0024	1.3123
Q6	0.926	0.9775	0.9363	1.0003	0.9803	0.9464
Q7	1.0113	1.0589	0.9655	1.1275	0.9918	1.1058
Q8	0.9572	1.0079	1.0869	0.9332	1.1543	0.8809
Q9	0.9157	0.9375	1.0317	0.8784	0.9278	0.8648
Q10	0.9039	0.929	0.9588	0.9108	0.9007	0.9295
RMSE	0.22	0.12	0.17	0.15	0.20	0.19

[a]F and η^2 values calculated through ANOVA within the four groups.
[b]groups divided according to 30 and 70 percentiles of the ability scores obtained with the corresponding method.
[c]groups divided according to 0 threshold of the ability scores obtained with the corresponding method.

likely because of its dichotomic scoring. When using 0 logit as threshold ability to divide the sample, M2 has a high level of agreement with M3 (0.81) and M4 (0.86), likely because these three methods give credit to each T1 and T2 combination pattern. Note that, as for classical analysis, using the ability threshold, M1 has an excellent agreement (0.75) only with M6.

Analysis of infit and outfit MNSQ to inspect goodness of model fit for the six scoring methods is also reported in Table IX. Infit values for all items and all scoring methods fall within the recommended range 0.7–1.3. Outfit values show more variability. In particular, considering both infit and outfit measures, M2 appears to fit better than the other models (RMSE = 0.13 for infit measures, 0.12 for outfit measures). For M1, two items, Q3 and Q5, show values of outfit outside the recommended range.

As discussed above, this may be related to the fact that the first tier of Q3 and the second tier of Q5 were the most

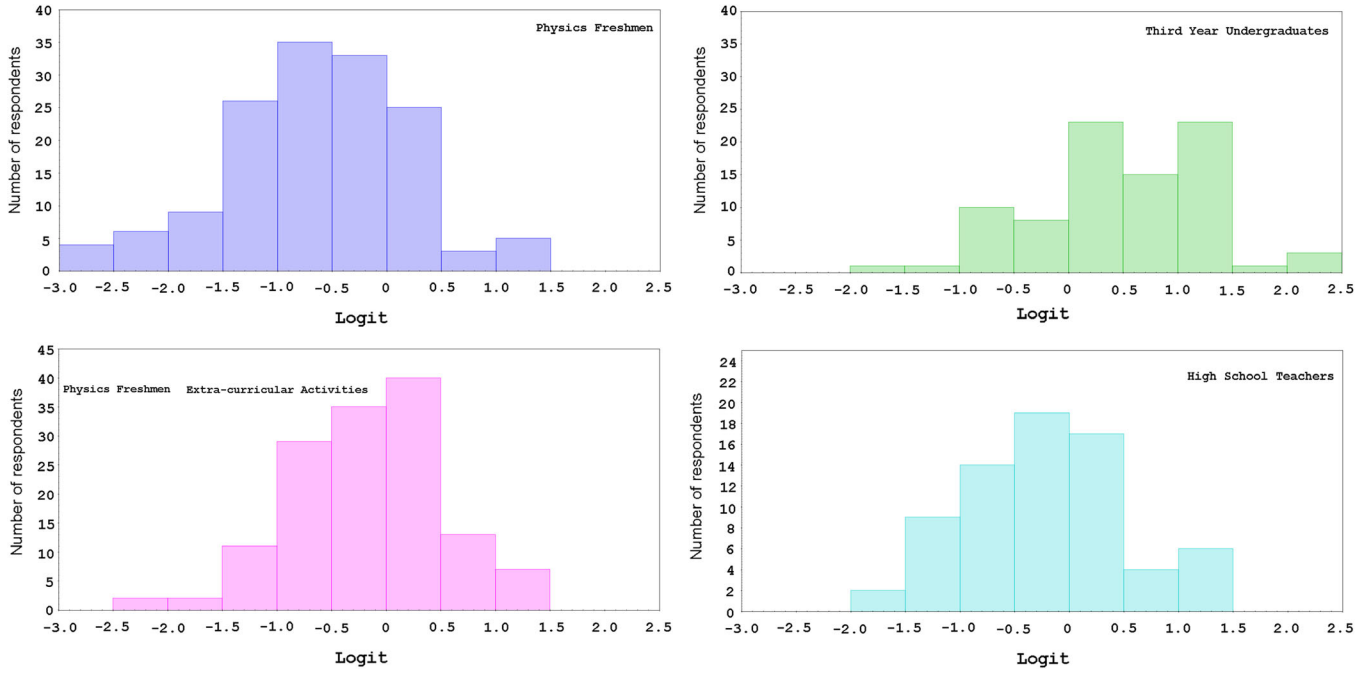

FIG. 3. Ability distributions of the four groups involved in the study. One extreme person (ability = −4.63 logit) is not represented.

difficult items when tiers were analyzed separately, and such behavior may have impacted on the score obtained when coupling the tiers. Similar considerations arise for Q3 when coupling the tiers using M4 and M5, and for Q5 when using M6. Note that also Q1 presents a value of outfit outside the range when using M6. Further insights about differences across the six methods can be obtained through the analysis of the IRCs reported in Supplemental Material C [57].

C. Performance of the different groups

The distribution of the abilities of the four groups when tiers are analyzed separately is reported in Fig. 3. The mean ability is -0.14 ± 0.04 logit, which implies that the QME for the whole sample was slightly difficult. Differences across the groups are significant ($F = 29.837$, $p < 10^{-4}$) with a large effect size ($\eta^2 = 0.17$). In particular, upper level physics students (G2, average ability $= +0.45 \pm 0.10$ logit) outperformed the other groups. As expected, physics freshmen students (G1) have the lowest ability (-0.54 ± 0.07 logit), followed by high school teachers (G4, -0.14 ± 0.09 logit), and physics freshmen who had attended extracurricular activities in QM (G3, -0.07 ± 0.06 logit).

Since the latter average abilities are compatible with 0, QME seems particularly suitable for these specific populations. When analyzing the performance of the groups in the three themes, there are differences between T1 and T2. When considering T1, AEB items are relatively "easy" (average difficulty $= -0.99$ logit) for all groups, while WBM and MEAS are easy (average difficulty < -0.16 logit) for all groups except G1. On the other hand, T2 items of WBM and MEAS are "hard" for all groups (average difficulty $> +0.56$). T2 items of AEB are also hard for all groups (average difficulty $= +0.12$ logit) except G2.

Such evidence confirms that T2 was on average harder than T1. Moreover, such findings suggest that a sufficient knowledge of basic facts about the wave function and the measurement process may be achieved already at the high school level, provided that curricular instruction is supplemented with suitable activities. However, it is only with advanced physics courses that it is more likely to achieve deeper knowledge and reasoning skills. Our results also confirm the emphasis, in both high school and university courses, on quasiclassical models of atoms and their limitations in describing phenomenology related to the microworld.

When tiers are coupled by the six scoring methods (Table IX), the average difficulty of QME depends on the specific method. When adopting M1, QME turns out to be very difficult, since the average ability is -1.60 logit. This result was expected since M1 is dichotomic and gives a full credit only for the "11" pattern. Note that with this scoring method, QME would be difficult also for G2 students. M2 leads to abilities estimates that are very similar to those obtained when the tiers are analyzed separately, as expected since that it is a pairwise method. Also, person reliability is the highest of all the scoring methods. M3 leads to an average ability of about zero logit, which means that if one uses this method the QME becomes suitable for the sample as a whole.

Since M3 weighs more the "10" pattern (correct knowledge of facts but incorrect reasoning), this evidence

confirms that the majority of students in our sample had better performances in T1 (hence, T2 was harder). When using M4, the average ability of the sample is negative (−0.12 logit). High school teachers and physics freshmen who had attended extra-curricular activities have the same ability, which suggests a greater frequency of the "10" pattern in the sample of high school teachers. The mean ability of the sample using M5 is also negative (−0.40 logit) and the QME is suitable only for upper-level physics students (whose average ability is about 0). Finally, when using M6, the questionnaire is difficult for the whole sample (−0.67 logit). In particular, according to this scoring method, high school teachers and physics freshmen who attended extra-curricular activities perform in a similar way.

Differences across the groups are statistically significant independently on the scoring methods. Effect size, as measured by η^2, is the highest for M2 and the lowest for M5. From ANOVA, three groups can be identified: (i) high school teachers and physics freshmen who had attended extra-curricular activities, (ii) physics freshmen, and (iii) upper-level physics students.

Further insight can be obtained from the Wright maps corresponding to the six scoring methods. The analysis is reported in the Supplemental Material D [57].

V. DISCUSSION

In the following, we separately discuss to what extent the collected evidence allows answering each research question that guided our study.

RQ1: what are the respondents' conceptions of quantum mechanics that emerge from the answers to QME?

First, we note that as any other diagnostic instrument in closed form, also the QME is not sufficient to definitively attribute specific ways of thinking to subjects responding in particular ways to the proposed items. In other words, it is always possible that subjects picked a specific answer choice not because they recognized in it their own reasoning, but because of its plausibility. However, our statistical analysis allowed us to identify a few consistent response patterns within the three targeted themes that add to the literature about teaching and learning quantum mechanics [7,8,66]. In the following, we address some specific difficulties that emerged in this study.

As far as the wave behavior of matter theme, we found that the most challenging topic concerned the elementary properties of the wave function and its time evolution [44]. This evidence is not surprising given the abstractness of these concepts. However, if one looks more deeply at our results, such difficulty seems to be related to the persistence of quasiclassical reasoning, in which the deterministic view is paired with the belief that only selected ("allowed") states are possible according to some unspecified quantization rules. The pattern of answer choices to the second tier given by physics freshmen and teachers also suggests that high school teaching of quantum mechanics could be misleading. By emphasizing "allowed" positions and energies, which determine the state of a particle, deterministic views and incorrect ideas about the wave function and its physical meaning may be reinforced. An alternative explanation could be that the subjects in the sample had very little knowledge about these concepts and misinterpreted the expression "allowed states." Correspondingly, we found that, at the conceptual level targeted by the QME, superposition of states may be an "easier" concept likely because traditional teaching emphasizes wavelike experimental results as the two-slit diffraction with electrons. However, we stress that QME deals with differences between superposition in classical and quantum physics, thus our data do not allow us to infer more detailed information about the knowledge of our sample of superposition states.

Even if the measurement process is often considered as one of the most counterintuitive topics of QM, the basic difference between the measurement process in classical and quantum mechanics seems to be understood by the majority of the sample. However, consistent with previous studies [8,50], two specific aspects were more difficult. The first one concerns the typical incorrect view that the uncertainty principle is related to generic limitations of the experimental apparatus used to perform measurements. The second one refers to a quasiclassical reasoning about measurement that can be summarized as follows: the measurement process in quantum mechanics changes the state of the system, but there are intrinsic limitations related to the accuracy with which we can perform measurements of a particle position and velocity.

Similar difficulties emerged from Q9, which concerns the effects of measurement on a state and the so-called "collapse" of the wave function. Quantum mechanics theory postulates that the measurement process of an observable leads to the "collapse" of the wave into an eigenstate, which is a "new" state of the corresponding operator. Incorrect answers to the second tier of Q9 could be due to the confusion between the expectation value of position and the probability of finding the particle at a given position, or to the belief that the measurement process gives only temporary results [9]. On the contrary, subjects who showed the "01" pattern had likely difficulty in applying the concept of measurement in a concrete case. Overall, traditional teaching of quantum mechanics seems to put scarce emphasis on the phenomenology and the underlying relationships between experimental evidence and the mathematical formulation of the measurement on the wave function and its role in determining the outcome of a measurement process. Alternatively, a possible reason to account for such difficulties is that the subjects simply misinterpreted the expression "new state."

When dealing with hydrogen atom and atomic models, only one item (Q2) was very difficult for the whole sample.

This question addresses in the first tier basic notions about the orbitals and their properties, focusing in particular on the common misconception according to which orbitals are "regions" of the atom [51,52]. In the second tier, we investigated if the subjects were aware that the stability of a hydrogen atom is related to the wave behavior of electrons and probed the typical misconception that atom stability is explained by the "quantization" of energy levels [13]. A possible reason why such incorrect claims were good attractors of subjects' responses may be related to a misleading use of orbitals and of the word quantization in chemistry courses both at the high school and university level. On the other hand, the encouraging results of the responses to Q8 and Q10 suggest that the knowledge of the classical laws of electromagnetism may help describe correctly the interaction between a negative charge and the nucleus and, hence, why it is not possible to apply a quasiclassical model to explain the stability of the atom. However, at the same time, such views could be also misleading when applied to the wave behavior of quantum objects, since they reinforce a quasiclassical, trajectory-based reasoning rather than promoting probabilistic quantum models [47,67].

RQ2: what are the psychometric properties of QME?

Classical test theory indices suggest that QME is a reliable tool to investigate students' basic knowledge about quantum mechanics. The difficulty of the tiers when analyzed separately is satisfactory. Overall, items in T2 are on average more difficult than the items in T1 across the three targeted themes. These findings suggest that T2 items provide more insight into the student ability in mastering the targeted concepts or, in other words, that T2 items require a deeper knowledge of the basic aspects of quantum mechanics addressed by QME. It is interesting to note that the difficulty patterns of T1 and T2 are consistent across the themes. In particular, the average difficulty of the WBM items is essentially the same as the MEAS items, either in T1 or T2. Similarly, AEB items are on average the easiest items, both in T1 and T2. Differently, when considering discrimination, T1 items of WBM, MEAS, and AEB have almost the same average values, slightly lower than the accepted threshold of 0.2. When considering T2 discrimination power, average values for the three themes are significantly different and well above the 0.2 threshold. While this evidence suggests caution in using T1 and T2 items separately, the coupling of T1 and T2 can result in the underestimation of item discrimination, as we will discuss in the next subsection.

Results of Rasch analysis are consistent with the conclusion that QME items, when analyzed separately, do not present misfitting behavior. This is indicative of the effectiveness of the items to differentiate the subjects on the basis of their abilities. Moreover, a good Rasch model fit suggests that the latent trait measured by the items positively correlates with the knowledge about quantum mechanics. The items measures confirm the patterns obtained with classical test theory. In particular, all T2 items have a higher value of difficulty than the corresponding T1 items, except for Q8. Discrimination, as measured with point biserial, is slightly higher than the corresponding measure obtained with classical test theory, thus confirming that the Rasch ability measures tend to better distinguish subjects along the proficiency continuum.

RQ3: to what extent do QME psychometric properties depend on the adopted scoring method?

Results from classical and Rasch analysis concur to claim that scoring methods affect significantly the psychometric properties of QME, as found in previous works [68]. In particular, the dichotomous scoring method M1 has the lowest value of reliability (calculated through Cronbach's alpha and person reliability, respectively) likely because neglecting intermediate levels of knowledge may lead to decrease the internal consistency of the response pattern to the paired items. Essentially, the M1 scoring system underestimates variations in the subjects' achievement in quantum mechanics. The results from Rasch analysis also show that the coupling of tiers with a scoring method that weighs differently the score in one of the tiers (M3-M6) decreases the goodness of fit. The only method for which no misfitting items were found is the pairwise method M2. Moreover, using both classical and Rasch analysis, M2 has good agreement with other polytomous methods (M3-M5), thus suggesting that it accounts also for the different performances of the subjects in T1 and T2. Finally, as reported in the Supplemental Material [57], M2 is also the method that better describes the interval of subjects' abilities and that, at the same time, does not artificially inflate or deflate their score. In particular, G2 average ability is greater than or equal to the average difficulty of the three themes. Hence, using M2, QME is affordable for students who attended advanced physics courses in quantum mechanics. Only G1 group has a negative ability that is lower than the mean difficulty of all three themes. Overall, our findings add to recent developments in the field [21] by showing that a pairwise method can be the most straightforward and reliable way to score a composite two-tier questionnaire as QME. However, further research with the involvement of a larger sample is needed to confirm such result.

RQ4: how well does QME discriminate between students with different background knowledge in quantum mechanics?

Statistical analysis of abilities suggests that QME items well discriminate between subjects with different knowledge about quantum mechanics. Effect sizes are significant, thus confirming the reliability of the questionnaire. Content validity seems to be confirmed by the performance of both senior and physics freshmen. The former evidence suggests that upper-level physics teaching allowed senior students to acquire good knowledge of the

basic aspects of quantum mechanics addressed in the QME. On the other side, despite reform-based efforts, curricular teaching of quantum mechanics at high school level is still problematic and not sufficient to warrant students with a satisfactory knowledge about basic quantum mechanics concepts, at least in the Italian educational context. In this respect, the extra-curricular activities [65] attended by G3 students may be a useful starting point to design more effective teaching interventions aimed at improving conceptual understanding of the targeted aspects of quantum mechanics. In particular, these activities are particularly effective to improve the students' understanding of the measurement process and state superposition.

A somewhat related result concerns the performance of high school teachers, who had difficulty with both basic facts and their consequences about the three themes targeted by QME. Such a result may be due to a limited background knowledge in physics of the teachers in our sample. The great majority of them (about 80%) graduated in mathematics and did not usually teach quantum physics, which is the reason why they attended a professional development course. Such findings suggest that professional development courses should put emphasis on how to give physical meaning to the formalism of quantum mechanics and engage the teachers, whenever possible, in experimental activities.

VI. CONCLUSIONS

This study presented a new assessment instrument, the QME, to investigate student conceptual learning and understanding of foundational aspects of quantum mechanics, grouped around three themes, which literature has shown to be relevant for understanding quantum mechanics: wave behavior of matter, measurement, atoms and electrons behavior. Given the increasing demand of quantum mechanics introductory courses at high school and lower undergraduate level, QME was designed to widen the scope and audience of the existing questionnaires [6,14–18]. To this end, QME features only qualitative questions that do not require neither mathematical formalization nor complex calculations, addressed only in upper-level physics courses.

With such constraints, the design of the questions was challenging since we could not rely on specific problems (e.g., potential wells) but we had to present abstract and very general settings (e.g., stationary states). Whenever possible, we built on previous studies to reconstruct students' claims and reasoning to describe such situations. As a consequence, one of the initial challenges was to design items that could distinguish between students' rote learning of facts and students' reasoning about the target topic. We addressed such an issue by resorting to a two-tier structure. In general, this type of instrument allows us to inspect if a correct answer given to the first tier question, which usually asks students to apply basic knowledge of the target content, is correctly justified in the second tier. To increase the complexity of the first tier, we asked to rate as true or false three statements about the target topic and then to answer a related multiple-choice question in the second tier. In such a way, we reduced the guessing probability, while controlling at the same time for local independence of the items. A second challenge was related to the scoring of QME. Usual ways of scoring two-tier questionnaires do not account for the different ability required to students to give a correct answer in the first and the second tier. In other words, when scoring two-tier instruments, one has to take into account that the tiers have different difficulty. To address this second issue, we probed six different scoring methods for the QME and investigated how QME properties changed according to the choice of the scoring methods. The third challenge was that the six scoring methods do not account for possible differences between different response patterns to the tier. To address this third issue, we coupled classical analysis, which was used in most of the previous studies about quantum mechanics questionnaires, with Rasch analysis. This further analysis allowed us to study the tiers separately to establish whether the first tier was easier than the second tier and the goodness of fit of the different scoring methods.

Classical statistical analysis suggests that the proposed instrument has sufficient reliability and items have good discriminating power, both when tiers are analyzed individually and coupled by a scoring method. Rasch analysis shows that, when tiers are analyzed separately, the first tier is on average easier than the second one. Moreover, all items have satisfactory values of infit and outfit. Discrimination values are acceptable for most items. On the other hand, when the tiers are coupled according to different scoring methods, our data show that a pairwise scoring method could be more efficient in terms of reliability, goodness of fit, agreement with other scoring methods and fairness of the total score assigned to the students.

Concerning the dimensionality of the instrument, we note that further work is needed to investigate whether the chosen clusters of items—wave behavior of matter, the measurement, and the behavior of atoms and electrons—represent distinct variables that describes proficiency in quantum mechanics or if they correspond to different aspects of the same latent trait. We are currently analyzing the dimensionality of QME with the techniques of exploratory factor analysis and parallel analysis [69].

From the inspection of the Wright map, we found that all QME items are well distributed along the ability continuum of the involved subjects. Hence, we can safely assume that, contrarily to previously available instruments, QME can be used to probe students' knowledge about QM in a variety of courses, from more advanced ones for physics undergraduates to introductory ones, like those for engineering

and math undergraduates. An appropriate follow-up to the study can investigate the validity of such a hypothesis.

Differently from other instruments, we also validated the QME with a sample of high school teachers. The results obtained by this specific population suggest that QME can be fruitfully used in professional development courses as a way to reinforce content knowledge about quantum mechanics basic aspects, before the introduction of innovative teaching strategies.

Out of this analysis, a practical question arises: how may university instructors and high school teachers use the QME? The answer is somewhat connected to the adopted learning model. If one adopts the view that students' misconceptions can be resources on which to build scientifically informed conceptions, a two-tier instrument like the QME can be a useful diagnostic tool, aimed at identifying common incorrect reasoning. In particular, from Rasch analysis of both single and paired tiers, one can easily identify the probing power of each item and how to use it in a specific teaching context. For instance, when keeping the tiers separate, T2 of Q9 is able to identify undergraduate physics students with a very good understanding of the measurement process, while T1 (same question) may better serve to identify physics students with good prior knowledge on which to build a sounder understanding of the measurement process in QM. When coupling the tiers, Q9 may still be used to identify most able students, but one has to look at all the three MEAS items to get a clearer picture of prior knowledge useful to understand the measurement process. Similarly, at the very beginning of an introductory course about QM, AEB items can be useful to investigate whether students have developed scientifically inaccurate models of the atomic stability. While we do encourage the use of the QME as a summative test at the end of a course, or as a pre- and post-test, we would recommend that instructors, teachers, and experts involved in teacher professional development use this instrument also as a formative assessment with the aim of designing teaching interventions more responsive to students' ideas.

VII. LIMITATIONS

Though the study provides evidence of the validity of the QME instrument, we are aware of some limitations. First, the reliability of QME is lower than the desired value for this kind of study. This weakness may be justified by the fact that the QME was administered to different populations with very diverse background knowledge in QM. We plan to administer the questionnaire to a more homogenous sample to further check its reliability. The sample will involve the same population of students (i.e., physics undergraduates) but on a national level. Until then, results cannot be overgeneralized. Concerning the psychometric properties of the QME, we also note that some questions (Q2, Q3, and Q5) have low discrimination power. Such a result may be due to the wording of some statements in T1 and of specific answer choices in T2 for the topics of measurement and of wave function time evolution. Data reported also in the Supplemental Material [57] (e.g., local independence) show that there is room for improvement to address the detected ambiguities. Hence, we are currently planning a revision of the QME to overcome the above limitations.

APPENDIX: QME TEST

Here we report the complete QME test. True statements in T1 and correct answer choices in T2 are marked with an *.

Q1_T1 Indicate whether the following three statements are true (T) or false (F)

I. Any undisturbed quantum system will evolve back to the initial state

*II. The time evolution of the state of a free particle is similar to that of a wave

III. The probability of measuring each value of a given observable associated to the particle is time independent

Q1_T2 The wave function of a stationary state of a particle evolves with time with a frequency that is dependent on the particle's:

a. position

*b. energy

c. allowed positions

d. allowed energies

Q2_T1 Indicate whether the following three statements are true (T) or false (F) for the hydrogen atom

I. An orbital is the region of the atom where the electron can most probably be found

*II. To each orbital corresponds a fixed value of the electron energy

III. An orbital is a region of the atom where the electrons rotate around the nucleus

Q2_T2 How would you explain the atom's stability?

*a. Through stationary waves associated to the electrons

b. Through the quantization of the atomic energy levels

c. Through orbitals associated to the electrons' trajectories

d. Through quantization of the atomic magnetic and electric fields

Q3_T1 Indicate whether the following three statements are true (T) or false (F)

I. The wave function defines all the allowed states of a particle

II. The wave function is a dimensionless quantity

*III. The wave function provides the complete description of the state of a particle

Q3_T2 If you know the formal expression of the particle's wave function then you can:

a. describe all the particle's allowed positions and energies

b. determine all the possible values of any physical observable associated to the particle

*c. calculate the probability of obtaining by measurement a given value of any physical observable associated to the particle

d. predict the possible states of the particle and the values of any associated physical observable

Q4_T1 Indicate whether the following three statements are true (T) or false (F)

I. If you perform repeated measurements of the same physical observable, we always obtain equal results

*II. The measurement process creates a new state

*III. Whatever the initial state is, for any subsequent measurement, we can only predict the probability of obtaining a given outcome

Q4_T2 Are classical and quantum measurement processes similar?

a. No, because in quantum mechanics there are few reliable instruments to perform experiments

b. Yes, because also in quantum mechanics there are uncertainties associated to the measurement process

c. Yes, because also in classical physics the measurement process is ruled by probability

*d. No, because in quantum mechanics the measurement process may change the state of a system

Q5_T1 Indicate whether the following three statements are true (T) or false (F) for a quantum system

*I. The higher the precision with which we know the velocity of a particle, the higher the uncertainty with which we know its position

II. If one measures the velocity of a particle, the value of its position will be modified by the experimental apparatus

III. If one measures the position of a particle, its velocity will be altered by the experimental apparatus

Q5_T2 The uncertainty principle is an intrinsic limitation to our capability of describing the microscopic world because:

*a. if we assign definite numerical values to the position of a particle, its velocity is completely undefined

b. we cannot measure, at the same time and with arbitrary precision, the position and velocity of a particle because of instruments limitations and experimental errors

c. any experimental apparatus perturbs the particle position and velocity

d. the numerical precision, with which we can determine the particle position and velocity, is limited

Q6_T1 Indicate whether the following three statements are true (T) or false (F)

I. It is possible to define the state of a particle by assigning numerical values to its position and velocity

II. It is not possible to define the state of a particle since the wave function is not a physical wave

III. The state of a particle is defined by the positions occupied by the particle

Q6_T2 An electron shows a wave behavior since:

a. the electron position is described by a wave-like equation

b. the electron follows a sinusoidal path as a wave

c. the electron is a smeared charged cloud moving at the speed of electromagnetic radiation

*d. the electron undergoes interference as a wave does

Q7_T1 Indicate whether the following three statements are true (T) or false (F)

The wave function of a particle is given by the superposition of two stationary states each corresponding to a different energy value. Then,

I. After any measurement, the particle will end up in an intermediate state in-between the initial states

*II. There is no way to know the particle energy until an energy measurement is performed.

III. The particle energy is the sum of the energies of the two states in the superposition

Q7_T2 Is the superposition of states in quantum mechanics similar to the superposition principle in classical physics?

a. No, because the particle is simultaneously in *all* the states, even though after a measurement it will always be found in only one of the states

b. Yes, because also in classical mechanics, when two or more forces act upon a system, the resulting motion of a particle is the sum of the independent motions due to each force

c. Yes, because also in classical mechanics a system can be described by a statistical mixture with a given probability to be in each accessible state

*d. No, because until we make a measurement the states of the superposition can interfere one with each other.

Q8_T1 Indicate whether, in classical physics, the following three statements are true (T) or false (F)

*I. If a charge is moving along a curved path, it dissipates energy

II. If we increase the distance between two identical charges, their electrical potential energy increases

III. If we decrease the distance between two identical charges, the net force they exert on each other decreases

Q8_T2 When we calculate the energy of an electron in a atom we can neglect the gravitational force since:

*a. the gravitational force is several orders of magnitude smaller than Coulomb force

b. the energy of the nucleus is large enough to attract the electrons and consequently the gravitational force is negligible

c. the binding energy of the electron is much greater than the gravitational energy of the nucleus

d. the gravitational force is balanced by the centrifugal force due to the electron rotation about the nucleus

Q9_T1 Indicate whether the following three statements are true (T) or false (F) for a free electron

I. If we observe the electron to be in a certain place, then it was already in that place right before the measurement

*II. If we measure that the electron has a certain velocity, then it will keep the same velocity immediately after the measurement

*III. If we measure the electron energy, then it will maintain that energy since the energy is a constant of motion

Q9_T2 If we make a measurement of a physical observable on a particle, what happens immediately after the measurement?

a. The particle is in a new state that evolves back to the initial state

b. The particle is in a new state whose wavelength gradually shrinks and eventually vanishes

*c. The particle is in a new state maintaining the measured value of the physical observable

d. The particle is in a new state in which there is full correspondence between measured and expected values of any physical observable

Q10_T1 Indicate whether, in classical physics, the following three statements are true (T) or false (F)

*I. It is possible to make an electron move along a curved path if we apply a uniform magnetic field

II. If a magnetic field acts upon an electron, then, the electron will always follow a circular path

III. If an electric field acts upon an electron, then, the electron will always move with constant acceleration along a straight line

Q10_T2 Why can't classical physics explain the stability of the hydrogen atom?

a. While rotating around the positively charged nucleus, the electron is attracted by the electrostatic force and dissipates energy

b. In classical physics there is not the uncertainty principle and, hence, it is not possible to describe the trajectory of the electron

c. Classical physics concerns only massive particles, much heavier than the electron and the nucleus

*d. While rotating around the positively charged nucleus, the electron emits radiation and dissipates energy

[1] D. Hestenes, M. Wells, and G. Swackhamer, Force Concept Inventory, Phys. Teach. **30**, 141 (1992).

[2] D. P. Maloney, T. L. O'Kuma, C. J. Hieggelke, and A. V. Heuvelen, Surveying students' conceptual knowledge of electricity and magnetism, Am. J. Phys. **69**, S12 (2001).

[3] N. Erceg, I. Aviani, V. Mešic, M. Gluncic, and G. Žauhar, Development of the kinetic molecular theory of gases concept inventory: Preliminary results on university students' misconceptions, Phys. Rev. Phys. Educ. Res. **12**, 020139 (2016).

[4] A. Kohnle, S. Mclean, and M. Aliotta, Towards a conceptual diagnostic survey in nuclear physics, Eur. J. Phys. **32**, 55 (2011).

[5] J. S. Aslanides and C. M. Savage, Relativity concept inventory: Development, analysis, and results, Phys. Rev. ST Phys. Educ. Res. **9**, 010118 (2013).

[6] H. R. Sadaghiani and S. J. Pollock, Quantum mechanics concept assessment: Development and validation study, Phys. Rev. ST Phys. Educ. Res. **11**, 010110 (2015).

[7] K. Krijtenburg-Lewerissa, H. J. Pol, A. Brinkman, and W. R. van Joolingen, Insights into teaching quantum mechanics in secondary and lower undergraduate education, Phys. Rev. Phys. Educ. Res. **13**, 010109 (2017).

[8] C. Singh and E. Marshman, Review of student difficulties in upper-Level quantum mechanics, Phys. Rev. ST Phys. Educ. Res. **11**, 020117 (2015).

[9] C. Singh, Student understanding of quantum mechanics, Am. J. Phys. **69**, 885 (2001).

[10] C. Singh, Student understanding of quantum mechanics at the beginning of graduate instruction, Am. J. Phys. **76**, 277 (2008).

[11] E. Marshman and C. Singh, Framework for understanding student difficulties in quantum mechanics, Phys. Rev. ST Phys. Educ. Res. **11**, 020119 (2015).

[12] B. W. Dreyfus, A. Elby, A. Gupta, and E. R. Sohr, Mathematical sense-making in quantum mechanics: An initial peek, Phys. Rev. Phys. Educ. Res. **13**, 020141 (2017).

[13] G. Ireson, The quantum understanding of pre-university physics students, Phys. Educ. **35**, 15 (2000).

[14] G. Zhu and C. Singh, Surveying students' understanding of quantum mechanics in one spatial dimension, Am. J. Phys. **80**, 252 (2012).

[15] E. Cataloglu and R. Robinett, Testing the development of student conceptual and visualization understanding in quantum mechanics through the undergraduate career, Am. J. Phys. **70**, 238 (2002).

[16] S. Wuttiprom, M. D. Sharma, I. D. Johnston, R. Chitaree, and C. Soankwan, Development and use of a conceptual survey in introductory quantum physics, Int. J. Sci. Educ. **31**, 631 (2009).

[17] S. B. McKagan, K. K. Perkins, and C. E. Wieman, Design and validation of the quantum mechanics conceptual survey, Phys. Rev. ST Phys. Educ. Res. **6**, 020121 (2010).

[18] D. F. Treagust, Development and use of diagnostic tests to evaluate students' misconceptions in science, Int. J. Sci. Educ. **10**, 159 (1988).

[19] M. Planinic, L. Ivanjek, and A. Susac, Rasch model based analysis of the Force Concept Inventory, Phys. Rev. ST Phys. Educ. Res. **6**, 010103 (2010).

[20] T. F. Scott and D. Schumayer, Students' proficiency scores within multitrait item response theory, Phys. Rev. ST Phys. Educ. Res. **11**, 020134 (2015).

[21] Y. Xiao, J. Han, K. Koenig, J. Xiong, and L. Bao, Multilevel Rasch modeling of two-tier multiple-choice test: A case study using Lawson's classroom test of scientific reasoning, Phys. Rev. Phys. Educ. Res. **14**, 020104 (2018).

[22] F. Haslam and D. F. Treagust, Diagnosing secondary students' misconceptions of photosynthesis and respiration in plants using a two-tier multiple choice instrument, J. Biol. Educ. **21**, 203 (1987).

[23] A. L. Odom and L. H. Barrow, Development and application of a two-tier diagnostic test measuring college biology students' understanding of diffusion and osmosis after a course of instruction, J. Res. Sci. Teach. **32**, 45 (1995).

[24] C.-Y. Tsui and D. F. Treagust, Evaluating secondary students' scientific reasoning in genetics using a two-tier diagnostic instrument, Int. J. Sci. Educ. **32**, 1073 (2010).

[25] J.-R. Wang, Development and validation of a two-tier instrument to examine understanding of internal transport in plants and the human circulatory system, Int. J. Sci. Math. Educ. **2**, 131 (2004).

[26] K. C. D. Tan, N. K. Goh, L. S. Chia, and D. F. Treagust, Development and application of a two-tier multiple-choice diagnostic instrument to assess high school students' understanding of inorganic chemistry qualitative analysis, J. Res. Sci. Teach. **39**, 283 (2002).

[27] A. L. Chandrasegaran, D. F. Treagust, and M. Mocerino, The development of a two-tier multiple-choice diagnostic instrument for evaluating secondary school students' ability to describe and explain chemical using multiple levels of representation, Chem. Educ. Res. Pract. **8**, 293 (2007).

[28] M.-H. Chiu, A national survey of students' conceptions of chemistry in Taiwan, Int. J. Sci. Educ. **29**, 421 (2007).

[29] K. S. Taber and K. C. D. Tan, The insidious nature of 'hard-core' alternative conceptions: Implications for the constructivist research programme of patterns in high school students' and pre-service teachers' thinking about ionisation energy, Int. J. Sci. Educ. **33**, 259 (2011).

[30] B. J. Franklin, The development, validation and application of a two-tier diagnostic instrument to detect misconceptions in the areas of force, heat, light and electricity, Ph.D. thesis, Louisiana State University and Agricultural and Mechanical College (1992), unpublished.

[31] C. C. Chen, H. S. Lin, and M. L. Lin, Developing a two-tier diagnostic instrument to assess high school students' understanding—The formation of images by a plane mirror, Proc. Natl. Sci. Counc. **12**, 106 (2002); https://www.researchgate.net/publication/237236334_Developing_a_two-tier_diagnostic_instrument_to_assess_high_school_students'_understanding_The_formation_of_image_by_plane_mirror.

[32] C.-C. Tsai and C. Chou, Diagnosing students' alternative conceptions in science, J. Comput. Assist. Learn. **18**, 157 (2002).

[33] H.-P. Chang, J.-Y. Chen, C.-J. Guo, C.-C. Chen, C.-Y. Chang, S.-H. Lin, W.-J. Su, K.-D. Lain, S.-Y. Hsu, J.-L. Lin, C.-C. Chen, Y.-T. Cheng, L.-S. Wang, and Y.-T. Tseng, Investigating primary and secondary students' learning of physics concepts in Taiwan, Int. J. Sci. Educ. **29**, 465 (2007).

[34] C.-H. Tsai, H.-Y. Chen, C.-Y. Chou, and K.-D. Lain, Current as the key concept of Taiwanese students' understandings of electric circuits, Int. J. Sci. Educ. **29**, 483 (2007).

[35] S.-J. Lee, Exploring students' understanding concerning batteries—theories and practices pages, Int. J. Sci. Educ. **29**, 497 (2007).

[36] H.-E. Chu and D. F. Treagust, A stratified study of students' understanding of basic optics concepts in different contexts using two-tier multiple-choice items, Res. Sci. Technol. Educ. **27**, 253 (2009).

[37] P. B. Griffardand J. H. Wandersee, The two-tier instrument on photosynthesis: What does it diagnose?, Int. J. Sci. Educ. **23**, 1039 (2001).

[38] S. Hasan, D. Bagayoko, and E. L. Kelley, Misconceptions and the Certainty of Response Index (CRI), Phys. Educ. **34**, 294 (1999).

[39] I. S. Caleon and R. Subramaniam, Do students know what they know and what they don't know? Using a four-tier diagnostic test to assess the nature of students' alternative conceptions, Res. Sci. Educ. **40**, 313 (2010).

[40] D. Gurcay and E. Gulbas, Development of three-tier heat, temperature and internal energy diagnostic test, Res. Sci. Technol. Educ. **33**, 197 (2015).

[41] E. Taslidere, Development and use of a three-tier diagnostic test to assess high school students' misconceptions about the photoelectric effect, Res. Sci. Technol. Educ. **34**, 164 (2016).

[42] K. Mannila, I. T. Koponen, and J. A. Niskanen, Building a picture of students' conceptions of wave- and particle-like properties of quantum entities, Eur. J. Phys. **23**, 45 (2002).

[43] Ministero dell'Istruzione, dell'Università e della Ricerca (MIUR) Indicazioni nazionali riguardanti gli obiettivi specifici di apprendimento concernenti le attività e gli insegnamenti compresi nei piani degli studi previsti per i percorsi liceali, Gazzetta ufficiale, Serie Generale n. 291 del 14-12-2010, Suppl. Ordinario n. 275, 2018, http://www.indire.it/lucabas/lkmw_file/licei2010/indicazioni_nuovo_impaginato/_decreto_indicazioni_nazionali.pdf, 2011.

[44] P. J. Emigh, G. Passante, and P. S. Shaffer, Student understanding of time dependence in quantum mechanics, Phys. Rev. ST Phys. Educ. Res. **11**, 020112 (2015).

[45] D. Styer, Common misconceptions regarding quantum mechanics. Am. J. Phys. **64**, 31 (1996).

[46] A. Masshadi and B. Woolnough, Insights into students' understanding of quantum physics: Visualizing quantum entities, Eur. J. Phys. **20**, 511 (1999).

[47] S. Vokos, P. S. Shaffer, B. S. Ambrose, and L. C. McDermott, Student understanding of the wave nature of matter: Diffraction and interference of particles, Am. J. Phys. **68**, S42 (2000).

[48] I. M. Greca and O. Freire, Does an emphasis on the concept of quantum states enhance students' understanding of quantum mechanics?, Sci. Educ. **12**, 541 (2003).

[49] G. Passante, P. J. Emigh, and P. S. Shaffer, Student ability to distinguish between superposition states and mixed states in quantum mechanics, Phys. Rev. ST Phys. Educ. Res. **11**, 020135 (2015).

[50] M. Ayene, J. Kriek, and B. Damtie, Wave-particle duality and uncertainty principle: Phenomenographic categories of

description of tertiary physics students' depictions, Phys. Rev. ST Phys. Educ. Res. **7**, 020113 (2011).

[51] K. S. Taber, Learning quanta: Barriers to stimulating transitions in student understanding of orbital ideas, Sci. Educ. **89**, 94 (2005).

[52] C. Nakiboglu, Instructional misconceptions of turkish prospective chemistry teachers about atomic orbitals and hybridization, Chem. Educ. Res. Pract. **4**, 171 (2003).

[53] C. Stefani and G. Tsaparlis, Students' levels of explanations, models, and misconceptions in basic quantum chemistry: A phenomenographic study, J. Res. Sci. Teach. **46**, 520 (2009).

[54] A. Cokelez. Junior high school students' ideas about the shape and size of the atom, Res. Sci. Educ. **42**, 673 (2012).

[55] G. W. Fulmer, H.-E. Chu, D. F. Treagust, and K. Neumann, Is it harder to know or to reason? Analyzing two-tier science assessment items using the Rasch measurement model, Asia-Pacific Sci. Educ. **1**, 1 (2015).

[56] D. Urbach, Further implementation of user defined fit statistics, in *Proceedings of the Pacific Rim Objective Measurement Symposium (PROMS) 2012 Conference*, edited by Q. Zhang and H. Yang (Springer-Verlag, Berlin, Heidelberg, 2013), pp. 57–74.

[57] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevPhysEducRes.15.010137 for the study of the local independence of QME items, the full details of the sample, the analysis of item response curves for QME items, and the Wright maps for each of the six scoring methods.

[58] R. L. Ebel and D. A. Frisbie, *Essentials of Educational Measurement* (Prentice-Hall, Englewood Cliffs, NJ, 1991).

[59] T. Bond and C. M. Fox, *Applying the Rasch Model: Fundamental Measurement in the Human Sciences* (Lawrence Erlbaum, Mahwah, NJ, 2007).

[60] R. M. Furr and V. Bacharach, *Psychometrics: An Introduction* (Sage, Thousand Oaks, CA, 2013).

[61] W. J. Boone, J. R. Staver, and M. S. Yale, *Rasch Analysis in the Human Sciences* (Springer, Dordrecht, Netherlands, 2014).

[62] G. A. Morris, L. Branum-Martin, N. Harshman, S. D. Baker, E. Mazur, S. Dutta, T. Mzoughi, and V. McCauley, Testing the test: Item response curves and test quality, Am. J. Phys. **74**, 449 (2006).

[63] G. A. Morris, N. Harshman, L. Branum-Martin, E. Mazur, T. Mzoughi, and S. D. Baker, An item response curves analysis of the Force Concept Inventory, Am. J. Phys. **80**, 825 (2012).

[64] U. Amaldi, *La fisica di Amaldi Blu* (Zanichelli, Bologna, 2016).

[65] M. Michelini, R. Ragazzon, L. Santi, and A. Stefanel, Proposal for quantum physics in secondary school, Phys. Educ. **35**, 406 (2000).

[66] I. D. Johnston, K. Crawford, and P. R. Fletcher, Student difficulties in learning quantum mechanics, Int. J. Sci. Educ. **20**, 427 (1998).

[67] M. A. Asikainen and P. E. Hirvonen, A study of pre- and inservice physics teachers' understanding of photoelectric phenomenon as part of the development of a research based quantum physics course, Am. J. Phys. **77**, 658 (2009).

[68] W. L. Romine, D. L. Schaffer, and L. Barrow, Development and application of a novel Rasch-based methodology for evaluating multi-tiered assessment instruments: Validation and utilization of an undergraduate diagnostic test of the water cycle, Int. J. Sci. Educ. **37**, 2740 (2015).

[69] O. L. Liu, H.-S. Lee, and M. C. Linn, An investigation of explanation multiple-choice items in science assessment, Educ. Assess. **16**, 164 (2011).