Effects of instruction on students' overconfidence in introductory quantum mechanics

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

[1]*Department of Physics "E. Pancini", University Federico II, Naples 80126, Italy*
[2]*School of Science and Technology, Physics Division, University of Camerino, Camerino 62032, Italy*
[3]*INAF–Astronomical Observatory of Capodimonte, 80131 Naples, Italy*
[4]*Liceo Statale "Ischia", Ischia 80077, Italy*

(Received 24 November 2019; accepted 29 May 2020; published 29 June 2020)

Students' ability to assess their own knowledge is an important skill in science education. However, students often overestimate their actual performances. In such cases, overconfidence bias arises. Previous studies in physics education have shown that overconfidence bias concerns mainly content areas, such as Newtonian mechanics, where misconceptions are strongly held by students. However, how the received instruction and the levels of understanding of a given topic influence overconfidence bias is yet to be proved. In this paper, we address this issue choosing as content area introductory quantum mechanics (QM). Overall, 408 high school students were involved in the study and randomly assigned to two experimental groups. One group received a textbook-based instruction about introductory QM, whereas the other one received instruction on the same topics through an innovative guided inquiry teaching-learning sequence (TLS), which included also potential pedagogical countermeasures for overconfidence bias. Students of both experimental groups completed a multiple-choice questionnaire and indicated for each item the degree of their confidence in the given answer using a 5-point Likert scale. The overconfidence bias was quantitatively defined and evaluated at person level using a 1D Rasch model. Progress in knowledge about the targeted topics was modeled according to a construct map validated in a previous paper. Results show that, for the whole sample, the overconfidence bias decreased as students progressed along the levels of the construct map. However, findings indicate that students of the TLS group achieved a significantly higher accuracy and a better confidence calibration, while the textbook group exhibited a lower performance and a significantly greater overconfidence bias. Implications for research into overconfidence bias in physics education are briefly discussed.

DOI: 10.1103/PhysRevPhysEducRes.16.010143

I. INTRODUCTION

Students' belief in their own ability—usually defined as *confidence*—has been historically associated in educational studies with motivation, interest, and decision making [1–8]. Prior works show that confidence, in general, has a positive correlation with science academic achievement [9], persistence in science tasks [10], and motivation towards science [11,12]. Higher performances in a given task also correlate positively with accurate assessment of one's own knowledge [13–16]. More recently, confidence has also been recognized as an important element of scientific literacy [17]. When the self-assessment of knowledge does not correspond to the actual achievement, overestimation of performance may arise (overconfidence bias). In educational psychology, overconfidence bias has been generally defined in three ways [18]: (i) an overly positive perception of one's own performance compared to that of the others (overplacement or better than average) [19,20]; (ii) an excessive confidence on the accuracy of one's own beliefs (overprecision); and (iii) an overestimation of one's actual ability or success chance in a specific task [21]. According to Ref. [18] the three types of overconfidence are not different manifestations of the same construct but conceptually and empirically distinct. In this paper, we are interested in the third type of bias, which arises when the degree of students' confidence is higher than their real performance, measured as the proportion of correct answers or by means of any other score in a given task [22,23]. In general, the overconfidence bias may affect decision-making skills and prevent students from deepening a given topic, with resulting ineffective self-regulation in learning and low achievements [6,7,24]. A recent study shows that overconfidence bias in genetics, evolution, and combined topics can be influenced by the socioeducational

*italo.testa@unina.it

context at the country level and, to a lesser extent, by gender [22]. The same study also supports the so-called hard-easy effect, which means that one's own confidence changes along the degree of difficulty of the task [25], thus implicitly suggesting a relationship between confidence and instruction level. In other words, meaningful understanding is related to the capability to (i) answer a question correctly, and (ii) recognize the correctness of the answer at a metacognitive level [26]. As students move from naïve conceptions toward more sophisticated and scientific views in response to a teaching intervention, so the ability to assess their own performance on a given task should improve, resulting in a decrease of the overconfidence bias. However, no study has yet provided evidence to support such hypothesis. In this paper, we address this issue by attempting to uncover the connection between confidence and ability at the person level, as mediated by the instructional process, using introductory quantum mechanics (QM) as the target area. The investigation of confidence-ability relationships at the person level can potentially uncover the extent to which the instructional context may influence students' cognitive processes when they assess their own performance.

Before presenting the research aims of the present study, we briefly review prior work in physics education about the relationships between confidence and knowledge, and put forward some arguments why QM can provide a suitable instructional context to accomplish our goals.

II. BACKGROUND

A. Confidence vs knowledge in physics

The construct of confidence was introduced in physics education research by Hasan *et al.* [27], who first proposed that the degree of certainty one student has in their own ability to answer a test question can help in identifying whether the wrong answer is due to a misconception, or rather to a temporary lack of knowledge [28]. The authors assumed that a right (wrong) answer with low confidence score—typically 2 out of 5—signals potential guessing (lack of knowledge). Conversely, a confidence score that is greater than 3 out of 5 for a wrong (right) answer signals a misconception (good knowledge). The criterion was also extended to a class group by considering the average confidence score. While the method was primarily suggested as a valuable aid for teachers or university instructors, it received attention from scholars in physics education to validate multitiered diagnostic instruments aimed to identify misconceptions in a variety of content areas [29–33]. Among these studies, only four focused at a deeper level on the relationship between students' confidence and ability [34–37].

Planinic *et al.* [34] were among the first to use the Hasan model. They compared the strength of students' misconceptions about Newtonian dynamics and electric circuits by using a true-false questionnaire with a confidence tier and analyzed it through the Rasch model. Results showed that high school students with different physics background provided incorrect answers with a high confidence level in the Newtonian dynamics area, while this was not the case for electric circuits. The authors argued that such evidence might be justified assuming that students developed incorrect mental models that were more stable in mechanics than in electricity. However, the authors did not investigate the extent to which differences in the knowledge of mechanics and electric circuits affected students' confidence in their answers. Potgieter and colleagues [35] followed a similar approach to explicitly test the Hasan *et al.* hypothesis in the area of mechanics. As an assessment tool, they used validated items from the Force Concept Inventory and the Mechanics Baseline Test as well as written justifications to answer choices. The analysis showed that the written explanations revealed further incorrect reasoning than the sole answers to the multiple-choice items. The authors hence distinguished between misconceptions and lack of problem-solving skills, but they did not explain why the latter issue should, on average, inflate students' confidence in giving a wrong answer or in using a wrong reasoning. In two papers, Calleon and colleagues [36,37] investigated the relationships between students' performance and confidence in mechanical waves. In the main study [36], they developed a four-tier questionnaire, namely, a two-tier instrument that probed knowledge in the first tier and reasoning in the second, each coupled with a confidence tier. Then, they divided the items into "more familiar" and "less familiar" from a curriculum teaching viewpoint. Results showed that students performed better in the familiar items and that a greater familiarity with the concepts led to a higher confidence rating. More interestingly, the sample exhibited strong misconceptions about both familiar and unfamiliar concepts. However, with respect to previous studies, the authors attempted to explain the strength of the detected misconceptions with the overemphasis given in the usual curriculum teaching of wave propagation to easy-to-use, rote-learned formulas rather than to more fundamental principles. The authors hence concluded that students tend to have an "illusion of knowing" [38], showing evidence of the well-known Dunning and Kruger effect [39].

While valuable, these efforts give only a partial account of the relationships between students' confidence and performance. In particular, assuming that high confidence in an incorrect response signals a misconception, it is possible to infer from such results only that a particular item can elicit a misconception, but nothing can be inferred about the overall capability of self-assessment of a single student or of a group of students. Moreover, the lack of focus on confidence at the student level makes it difficult to inspect the impact of teaching-learning activities on the development of the students' capability to recognize which questions they are really able to answer and which they are not. Finally, although the above studies seem to suggest that confidence calibration is associated with higher performances, they do not systematically investigate this correlation.

B. Potential sources of overconfidence in QM

The above reviewed work concerns areas of classical physics that, as known, are characterized by spontaneous conceptions and mental models that are often rooted in everyday experience. On the contrary, students cannot rely on personal exploration of the quantum world. Hence, misconceptions in QM [40–43] are likely matured after previous school teaching interventions, or after other formal or informal learning experiences. In this study we will focus only on school teaching experiences. QM has been only recently introduced at the high school level in Italy as well as in other countries' physics curricula [44]. However, at high school, several basic aspects of QM are targeted also in chemistry classes before being addressed in the physics course, even though the scope and formalism remain quite different. This circumstance may be a possible source of overconfidence. For instance, students may consider themselves already familiar with QM concepts such as, e.g., atomic models, wave-particle dualism, energy quantization, probability, etc., when actually this is not the case [45]. Moreover, overconfidence bias may arise from a correct knowledge of classical concepts (such as momentum or measurement) that, however, have a different meaning in QM. Under the cognitive point of view, it is then worth investigating the extent to which such peculiarities of quantum mechanics may influence the interplay between confidence and performance, in particular whether overconfidence bias may be reduced in response to a specific didactic intervention.

C. Research questions and hypotheses

Based on the background explained above, QM provides a novel and unique opportunity to study overconfidence bias and its relationships with increasing levels of conceptual understanding. We propose the following research questions as the focus of this study:

RQ1: *To what extent does instruction influence the students' overconfidence bias in introductory QM?*

RQ2: *How does the overconfidence bias change as the students' ability progresses in introductory QM?*

From these research questions, two hypotheses are posited: (H1) instruction in introductory QM reduces the overconfidence bias; (H2) as students' ability progresses in introductory QM, overconfidence bias decreases.

III. METHODS

A. Instructional context of the study

To answer our research questions, we used two different experimental settings: (i) a traditional (*"textbook"*) and (ii) a transformative (*"TLS"*) instructional context.

The traditional context was constituted by a four-week teaching sequence (about 12–14 h) that followed the Italian national guidelines about introductory QM in both chemistry and physics provided by the Ministry of Education. See Ref. [44] for a brief description of the Italian guidelines compared with those of other countries. In this study, this teaching sequence was implemented in classroom practice using the textbook (see, e.g., Ref. [46]) as guidance, and a teacher-directed lecture as the pedagogical approach.

The transformative context was constituted by a teaching-learning sequence (TLS) on the same introductory QM contents and with the same duration (about 14 h) [47]. The TLS followed a conceptual sequence that starts from the energy exchange between matter and electromagnetic radiation, addresses the Heisenberg's principle and atom stability, and arrives to atomic energy levels, orbitals, and the energy band model. Through the proposed activities, the students build increasingly sophisticated models of energy exchanges between radiation and matter, build a connection between energy discretization and atomic stability, and finally exploit more complex models to explain the behavior of metals and insulators. The pedagogical approach of the transformative TLS was a guided inquiry approach [48–54] that included also potential pedagogical countermeasures for overconfidence [47] as the critical and the regulatory feedback strategy; and the "think the opposite" strategy. In such a way, students were not only expected to develop a sounder scientific knowledge about QM, but also a greater metacognitive awareness about—and confidence in—what they were learning. Students' increasing knowledge about the topics targeted in the traditional and transformative teaching sequences was modeled by the construct map validated and revised in Ref. [47]. We summarize the TLS activities in Table I. We report in the Appendix the definition of the revised construct map levels. The same 18-item questionnaire that we used to validate the construct map was also used in this study. The complete questionnaire is reported in Ref. [47]. To measure confidence, we appended to each item a second tier, asking respondents to what extent they felt confident in the given answer on a scale from 1 (not at all confident) to 5 (completely confident). The Cronbach's alpha for the confidence scale was 0.94, which can be considered an excellent value. Accuracy and confidence data came from the same sample involved in Ref. [47]. The TLS group included $N = 200$ students, while the textbook group involved $N = 208$ students. Eighteen students of the TLS group (15 of which constituted an entire class), and seven students of the textbook group did not complete the confidence tier, so that the analysis for the present study was carried out with $N = 182$ and $N = 201$ students for the two groups, respectively.

B. Data analysis

To answer our research questions, we first had to choose how to calculate overconfidence bias. At the person level, a certain consensus has been reached on the following formula [56,57]:

TABLE I. Summary of the transformative TLS activities used in the study (see also Ref. [47] and related Supplemental Material for more details).

Time (h)	Addressed topics	What students do	What teacher does (QM topics)
2–4	Applications of QM (e.g., LED)	Propose an experimental setup to explain LED light emission and perform the experiment with an LED, a voltmeter and a voltage generator.	Guides the students to understand that the energy loss E of an electron that crosses the LED junction can be expressed as $E = eV_{th}$, where V_{th} is the voltage at which an LED turns on depending on the color of the emitted light. Helps the students understand the proportionality constant that relates V_{th} to frequency of the light emitted by the LED must be universal.
2–3	The concept of photon. The interaction between matter and radiation. The concept of mechanical action	Fill in a worksheet about how to interpret the meaning of the Planck's constant that has been measured in the LED experiment.	Proposes a heuristic definition of the "characteristic action" of a system in the form $A \propto E\,\tau$, where E is the characteristic energy and τ the inner time of the system. Introduces the principle of action quantization and Heisenberg principle as $\Delta E \Delta\tau \geq \hbar/2$ and $\Delta p \Delta x \geq \hbar/2$.
2–3	The uncertainty principle. The atom stability. The electronic structure of atoms (energy levels).	Fill in a worksheet about how to use the Heisenberg principle to explain an atom's stability.	Starting from the Heisenberg principle, guides the students to understand that it is impossible to determine the electron trajectory. Discusses limits of the quasiclassical models of atom. Explains atoms' stability using the Heisenberg principle.
4	Atomic orbitals and probability distributions	Fill in a worksheet about how to build a model of an atom based on energy levels.	Reinforces students' understanding of orbitals in terms of probability distributions. Introduces the wave function as a mathematical entity associated with observables (position, momentum) and orbitals.
2	Molecular orbitals. Energy bands model of solids: metals, insulators, semiconductors.	Fill in a worksheet to propose a model of metals and insulators starting from energy levels of single atoms.	Guides the students to reflect on what a conductor is and on the need for the electrons to have suitable energy to freely move across the material, so that they can also be "shared" by all the atoms of the solid. Helps students understand that an electron can only move from an "occupied" state to an empty state. Introduces the model of solids based on the concepts of electronic bands and forbidden band gap.

$$C_{\text{bias}} = C_{\text{score}} - P_{\text{score}}, \tag{1}$$

where C_{bias} is the confidence bias, C_{score} is the confidence score, and P_{score} is the performance score. In this study, we calculated the confidence bias, as defined by the formula (1), using a 1D Rasch model [58]. The reason for using the Rasch model is that raw confidence scores are categorical data and therefore they only provide an order relationship between subsequent levels on the confidence scale. Therefore, they cannot be used to measure the confidence bias using Eq. (1), since the respective intervals are not linear [58]. On the contrary, using the Rasch model, we can estimate both C_{score} and P_{score} on the same linear scale, using the same unit of measurement (logit) so that the measures of the QM and confidence items can be not only qualitatively compared but also algebraically manipulated using formula (1).

To calculate Rasch measures, differently from Ref. [59] where a full score was given only if the knowledge tier was correct and the expressed confidence level was greater than two out of five, we ran a combined Rasch analysis using all the 36 items (18 knowledge items on a dichotomous scale and 18 confidence items on a rating scale) [60]. Rasch statistics of the questionnaire, including person or item reliability and separation, were first calculated. Then, we checked the goodness of fit to the Rasch model using infit and outfit Mean-Square, while multidimensionality was checked through a principal component analysis (PCA) of residuals [61]. Finally, we checked the functioning of the confidence rating scale by inspecting the ordering of rating steps as a function of item and mean respondent

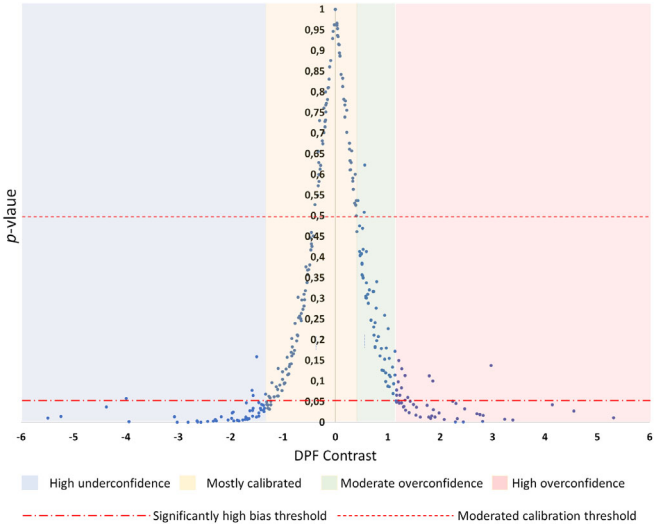

FIG. 1. p values associated with the DPF contrasts vs DPF contrasts in logit. DPF contrasts greater than +1.1 logit and less than −1.5 logit are statistically significant (associated p value is less than 0.05) and are labeled as "high overconfidence" and "high underconfidence," respectively. The interval [−1.5, +1.1] logit is further demarcated using the FWHM ($p = 0.5$) points into three regions: [−1.5, −0.3] logit, [−0.3; +0.3] logit, and [+0.3; +1.1] logit. We collapsed the two leftmost regions of the three into one region (shaded yellow) labeled "mostly calibrated." The third region, shaded green, is labeled "moderate overconfidence." In our analysis, when we refer to overconfidence, we mean the term to include both high and moderate overconfidence.

measure. To calculate overconfidence bias for each person we calculated the differential persons functioning (DPF) [61] on knowledge vs confidence items. DPF on knowledge (confidence) items represents a person's Rasch measure if the confidence (knowledge) items were not evaluated. Then, for each respondent, we calculated the DPF *contrast*—namely, the difference between Rasch measures on confidence items and on knowledge items—as Rasch proxy variable for the confidence bias defined by Eq. (1). In particular, a DPF contrast greater than zero indicates that the person is more likely to endorse the confidence item than they are able to correctly answer the corresponding knowledge item. Finally, we had to define the intervals to indicate the degree of significance of DPF contrast, in a way similar to that proposed by Stankov and Lee [57]. To estimate the effect size associated to a DPF contrast, we computed the associated t-test statistics and probability. Hence, to reduce the arbitrariness of our choice, we looked at the p values associated with each DPF contrast. This probability ranges from 1 (corresponding to a DPF contrast = 0 logit, perfect calibration), to lower values. This means, for instance, that a p value $p = 0.001$ corresponds to significant DPF contrasts with absolute value greater than 5 logit. We then plotted the associated p values as a function of DPF contrast, obtaining the classical bell-like plot of a student t distribution (Fig. 1).

TABLE II. Final adopted intervals of differential persons functioning (DPF) measures of calibration.

Calibration level	DPF Interval (logit)
High underconfidence	Lower than −1.5
Mostly calibrated	[−1.5 , +0.3]
Overconfidence	Greater than +0.3

The line corresponding to the conventional value of probability $p = 0.05$ intersects the curve in two points (−1.5 logit and +1.1 logit) that divide the DPF contrast continuum in three regions of calibration: "high underconfidence," "calibrated," and "high overconfidence."

Given the focus of the present study on overconfidence bias (namely, on positive values of DPF contrast), we divided the "calibrated" interval into two subintervals—"mostly calibrated" and "moderate overconfidence"—using as cut value the probability value of 0.5, which corresponds to the half maximum of the curve (in analogy with the full width at half maximum convention). The approximate positive value of the DPF contrast corresponding to this 0.5 probability threshold is +0.3 logit. We finally collapsed the "high" and "moderate overconfidence" intervals into a single "overconfidence" interval to carry out the statistical calculations. The final adopted intervals are reported in Table II.

To answer RQ1, the average DPF contrast was compared for TLS and textbook group using a t test. Correlation between ability (i.e., DPF on knowledge items) and DPF contrast was also calculated using data from all the classes involved. A chi-square test for independence was run in order to determine whether or not a relationship existed between TLS and textbook groups and levels of calibration of Table II.

To answer RQ2, we first compared through a one-way analysis of variance (ANOVA) the mean DPF on confidence items and the mean DPF contrasts for groups of students with different abilities on the QM items, using ability quartiles and the revised three-level construct map levels to stratify the sample.

Assignment of students to construct map levels was performed as described in Ref. [47], using the new estimates of item difficulty and students' abilities. We remark that, due to the properties of the logit scale [58], the new average values of difficulty of the levels differ from those obtained when only QM items were used by a constant quantity, which represents the contribution of the confidence items. Thus, relative ranking of the construct map levels is preserved.

Finally, a chi-square analysis was also carried out to inspect the dependence between the calibration intervals and the construct map levels. All Rasch measurements were obtained using Winsteps 3.98.

For the sake of completeness and to allow the comparison with other studies in the field, we report data from the analysis based on raw scores in the Supplemental Material [55].

TABLE III. Rasch analysis statistics of the questionnaire with the combined knowledge and confidence items.[a]

	Knowledge scale							Confidence scale						
Item	Measure	Model Standard error	Infit Mean-Square	Infit Standardized fit statistics	Outfit Mean-Square	Outfit Standardized fit statistics	Point-Measure Correlation	Measure	Model Standard error	Infit Mean-Square	Infit Standardized fit statistics	Outfit Mean-Square	Outfit Standardized fit statistics	Point-Measure Correlation
1	−1.09	0.11	0.9962	−0.059	0.9749	−0.329	0.3876	−0.15	0.06	0.7938	−3.279	0.8064	−2.979	0.7000
2	1.28	0.13	1.0756	1.051	1.1718	1.361	0.2392	−0.07	0.06	0.8933	−1.619	0.8807	−1.769	0.6399
3	0.03	0.11	1.0273	0.751	1.2168	3.511	0.3223	−0.12	0.06	0.8248	−2.749	0.8317	−2.559	0.6610
4	−0.10	0.11	1.0104	0.301	1.0451	0.821	0.3611	−0.61	0.06	1.0892	1.311	1.1702	2.351	0.5839
6	−0.74	0.11	1.1371	3.271	1.1736	2.791	0.2453	−0.63	0.06	1.0468	0.7010	1.0135	0.2210	0.6377
7	0.44	0.11	1.0670	1.511	1.2115	2.751	0.2724	−0.39	0.06	0.8672	−2.049	0.8546	−2.199	0.7229
8	0.70	0.12	1.1080	2.081	1.1993	2.221	0.2375	0.22	0.06	1.0118	0.2010	1.0139	0.221	0.6182
9	0.37	0.11	1.0730	1.711	1.1751	2.411	0.2781	−0.32	0.06	0.9942	−0.059	0.9587	−0.589	0.6772
10	−0.08	0.11	1.0300	0.841	1.1376	2.381	0.3302	0.09	0.06	1.0345	0.5310	1.0204	0.311	0.6273
11	0.96	0.12	1.1167	1.921	1.3225	2.921	0.2013	−0.15	0.06	1.0203	0.3210	0.9956	−0.039	0.6169
12	0.96	0.12	1.0879	1.461	1.1649	1.581	0.2400	−0.01	0.06	0.9383	−0.909	0.9358	−0.919	0.6055
13	−0.91	0.11	0.9771	−0.509	0.9335	−1.029	0.4129	−0.40	0.06	0.8406	−2.489	0.8173	−2.809	0.7511
14	0.57	0.11	1.0912	1.901	1.3156	3.671	0.2404	0.11	0.06	0.9289	−1.059	0.9060	−1.359	0.6549
15	−0.07	0.11	1.0127	0.361	1.1584	2.711	0.3425	0.00	0.06	0.8989	−1.529	0.8631	−2.039	0.6835
17	0.01	0.11	1.1254	3.311	1.2679	4.311	0.2295	0.18	0.06	0.9360	−0.939	0.9192	−1.149	0.6410
19	0.03	0.11	1.0133	0.371	1.0432	0.751	0.3575	−0.02	0.06	0.9313	−1.019	0.8985	−1.489	0.7010
20	0.21	0.11	1.0825	2.071	1.1257	1.931	0.2839	0.14	0.06	0.9562	−0.639	0.9153	−1.219	0.6647
21	−0.28	0.11	1.0478	1.341	1.1451	2.601	0.3212	−0.15	0.06	1.0296	0.461	1.0123	0.2010	0.6781

[a]Items 5, 16, 18, and 22 were removed from the analysis [47].

IV. RESULTS

A. Rasch statistics of the accuracy-confidence combined items

For the combined items, person reliability is 0.90, while person separation is 3.06. Both values can be considered good. The value of the person separation index suggests that the sample can be divided in more than one group according to their ability. Item reliability is 0.97, while item separation is 5.31, a value that confirms the item "difficulty" hierarchy of the instrument for this sample, namely how well the items are distributed along the difficulty continuum [62]. Infit and outfit MSNQ values are acceptable for all items (Table III). Point measure correlation, which also measures the Rasch construct validity, is acceptable (i.e., greater than 0.5) for all items of the confidence scale. QM items have smaller correlations (roughly from 0.2 to 0.4). Since infit and outfit Mean-Square fall within the recommended intervals, a small correlation (around 0.2) means that the item was likely more difficult for the students.

Results of the principal component analysis of residuals for the 36 items of the questionnaire are shown in Fig. 2. The raw variance explained by measures is 41.1%. Eigenvalues of the first two contrasts are, respectively: 3.48 (5.7% of unexplained variance) and 2.26 (3.7%). Hence, when combining knowledge and confidence items, at least two dimensions can be identified.

By inspecting Fig. 2, we see that all QM items have positive loadings in the first contrast, while most confidence items have negative loadings. Hence, the PCA of residuals confirms that QM items and the confidence scale measure different constructs. While three clusters of items can be identified from the figure, the disattenuated correlations suggest that only two measure different constructs, the first and the remaining two.

More precisely, the disattenuated correlation between the first and the second cluster is 0.4474, while that between the first and the third cluster is 0.2949. On the contrary, the disattenuated correlation between the second and third cluster is 0.8470.

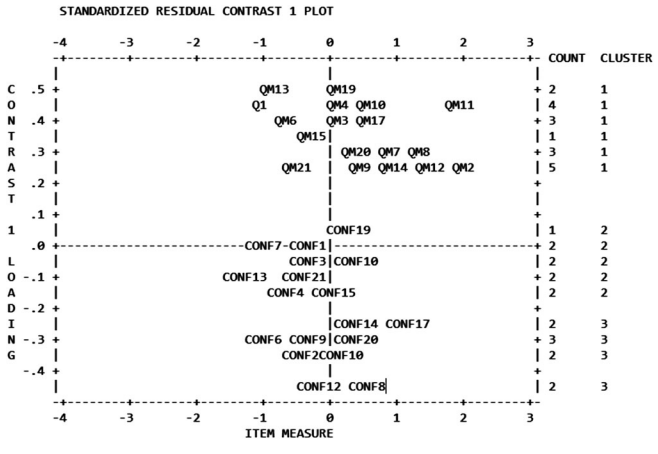

FIG. 2. Standardized residual loadings in the first contrast of the confidence (CONF) and knowledge (QM) items. Items removed from the analysis: 5, 16,18, and 22 [47].

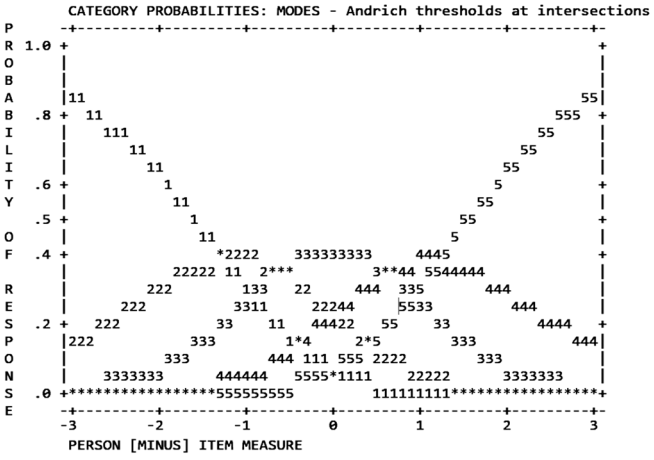

FIG. 3. Patterns of probability responses as a function of person and item measure. Asterisks, above the horizontal axis, indicate points at which adjacent categories are equally probable.

The first two values are well below the threshold of 0.57 for considering the items measuring the same construct, while the third values suggest that items in cluster 2 and 3 measure the same construct. Finally, in the second contrast, all disattenuated correlations are greater than 0.57, so we can infer that there is not enough item strength in the data for a third construct, different from that measured by the items in the knowledge and confidence scales. To analyze the functioning of the confidence rating scale we investigated the most probable response for each value of the scale (1–5) as a function of person measure and item difficulty. Figure 3 shows that each response value has a maximum probability for different combination of item difficulty and person measure. Therefore, we can assume that the usage of the confidence rating scale by the respondents is coherent with the intended use (e.g., a value of 2 suggests less confidence than a value of 3).

In Fig. 4, we report the Wright map for knowledge and confidence items for all students that participated in the study, regardless of the instruction they received. The Wright map shows students' DPF on QM and confidence items on the left-hand side and the estimated item difficulty on the right-hand side. Thurstonian thresholds for confidence items are also shown. We remind that, according to Rasch measurement, students with a higher measure in the accuracy analysis (i.e., the DPF on QM) are students with higher scores, while, students with a higher measure in the confidence analysis (i.e., the DPF on confidence) are more agreeable to confidence items. Similarly, QM items with a higher measure are items that were harder for students, while confidence items with a higher measure are items that were more unlikely for respondents to agree with.

The overall mean person measure on the 18 QM items +18 confidence items is -0.28 ± 0.94 (st.dev) logit. The average QM items measure is $+0.13 \pm 0.64$ (st.dev), while confidence items have an average measure of -0.13 ± 0.25 (st.dev) logit (we remind that in Rasch analysis the mean items' measure is set to 0).

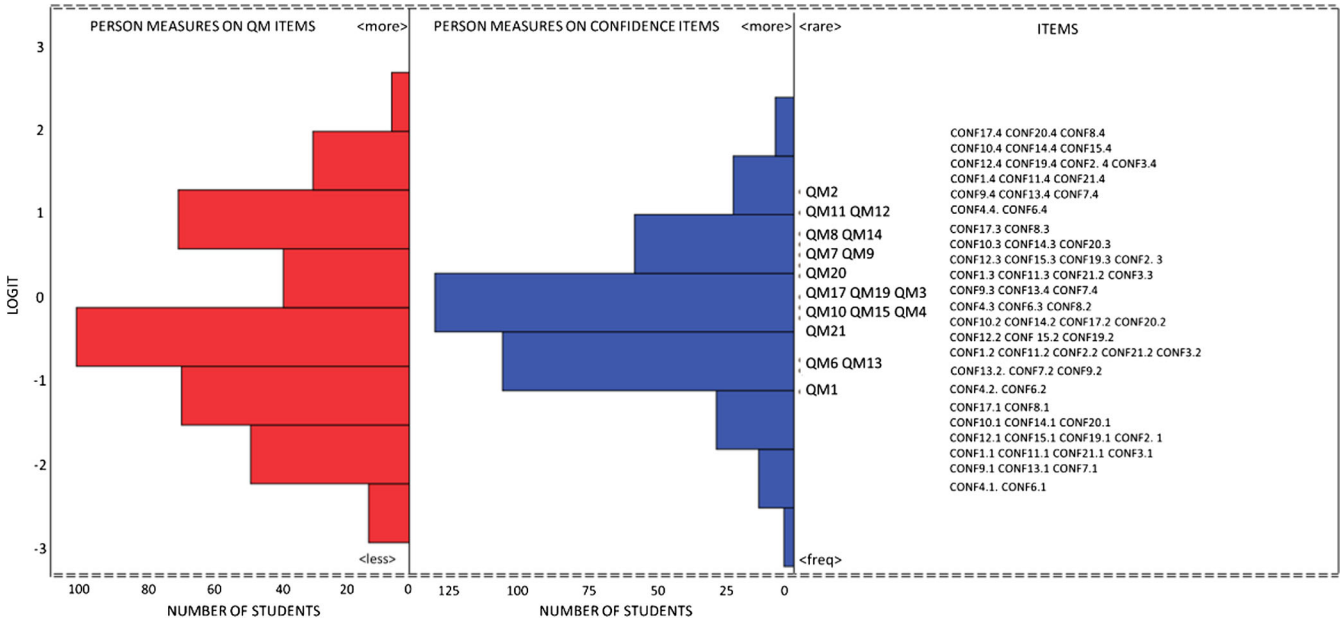

FIG. 4. Wright map of the questionnaire used in this study. Confidence items are labeled as CONF, knowledge items are labeled as QM. Thurstonian thresholds for confidence items are also shown (e.g., CONF9.4). The thresholds indicate the location at which the probability of choosing the $(i+1)$th category on the confidence scale is 50%. Items 5, 16, 18, 22 were removed from the analysis [47].

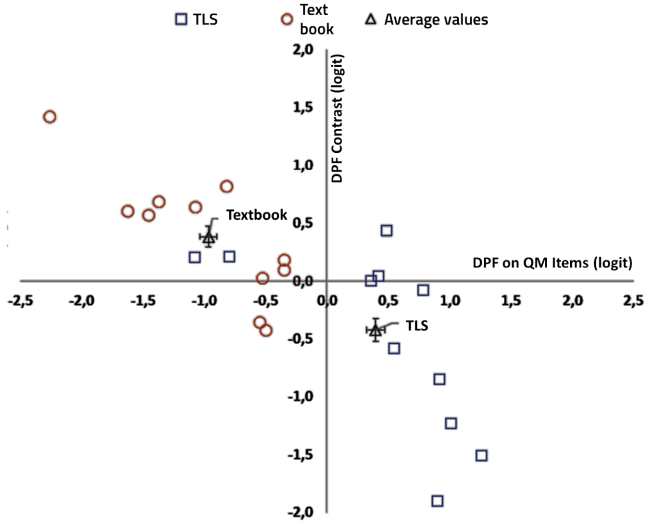

FIG. 5. Correlation between average ability on QM items (horizontal axis) and DPF contrasts (vertical axis) of the classes involved in the study. Blue squares denote the TLS group, while red circles are used for the textbook group. The average values for the two groups are denoted by triangles. Note that students of one class did not answer to the confidence tier, so only 22 points are represented in the graph.

This means that, when looking at *both* QM and confidence items, students had difficulties in answering the QM items and at the same time, did not much "agree" with the confidence items. However, the probability "to agree upon" a confidence item was slightly higher than to correctly answer a QM item.

When analyzing students' measures separately on knowledge and confidence items, we found that the mean DPF on the QM items is −0.32 logit ±1.22 (st.dev), while the mean DPF on confidence items is −0.33 logit ±1.20 (st.dev).

These negative DPF average measures confirm that students had difficulties in answering the QM items and did not much agree with the confidence items. We recall that the average DPF measures for each type of item are different from the overall mean person measure, because score-to-measure conversion is nonlinear.

B. To what extent does instruction influence students' overconfidence bias in introductory QM?

In Fig. 5, we plot the mean DPF contrasts for the classes involved in the study as a function of the mean ability DPF on QM item. TLS and textbook groups are denoted, respectively, with blue squares and red circles. Approximately, 7 classes have DPF contrast greater than 0.4 logit, namely, they are on average overconfident, while only two classes have DPF contrast lower than −1.5, which indicates a rather pronounced underconfidence.

Overall, the correlation between confidence bias and ability is $r_{\text{sample}} = -0.57$, which is significant at 0.01 level.

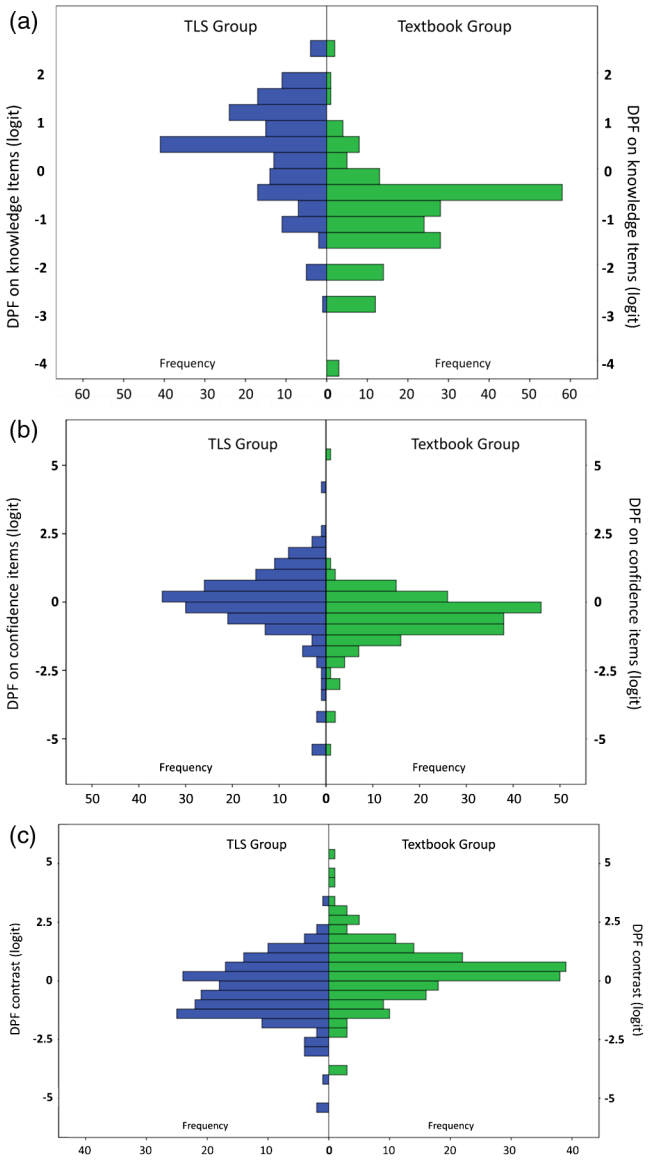

FIG. 6. Distribution of TLS and textbook control group students' DPF: (a) on knowledge items; (b) on confidence items; (c) contrast.

For the TLS group, this correlation is lower ($r_{\text{TLS}} = -0.36$ vs $r_{\text{Textbook}} = -0.64$). We report in Figs. 6(a)–6(c) the distribution of students' DPF on knowledge and confidence items and resulting DPF contrast. Mean DPF on QM items of the TLS group is +0.40 logit ±1.02 (st. dev.), while for the textbook group the mean ability is −0.96 logit ±0.99 (st. dev). The difference is statistically significant ($t = 13.207$, $df = 381$, $p < 10^{-4}$). The DPF on confidence items of the TLS group is −0.03 logit ±1.33 (st. dev.), while for the textbook group the mean ability is −0.58 logit ±0.99 (st. dev). The difference is again statistically significant ($t = 4.584$, $df = 332.525$, $p < 10^{-4}$).

The mean DPF contrast, our proxy for the overconfidence bias score, is −0.43 logit ±1.30 (st. dev.) for the TLS

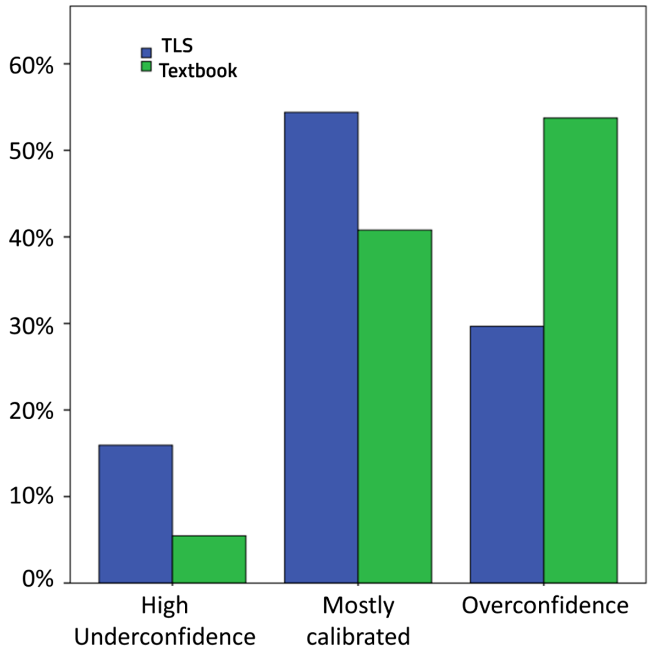

FIG. 7. Distribution of TLS and textbook groups subjects across confidence intervals of Table II.

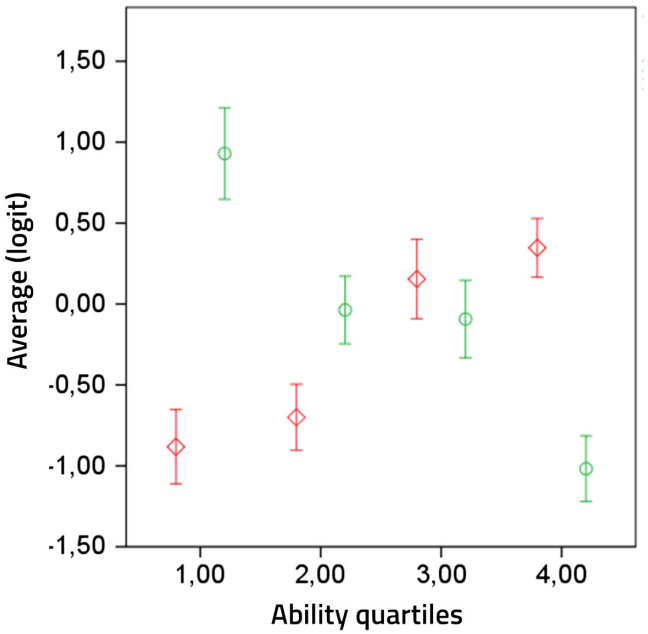

FIG. 8. DPF contrast (green circles) and DPF on confidence items (red diamonds) vs ability quartiles. Whiskers indicate 95% confidence interval. Green (red) circles (diamonds) are slightly shifted toward the right (left) to enhance readability.

group and 0.38 logit ±1.27 (st. dev.) for the textbook group. Also such difference is statistically significant ($t = 6.115$, $df = 381$, $p < 10^{-4}$).

The distribution of TLS and textbook group students in the confidence calibration intervals of Table II is reported in Fig. 7. Based on the chi-square test of interdependence, we found a significant correlation between instruction received and calibration level ($\chi^2 = 26.820$, $df = 2$, $p < 10^{-4}$; Cramer's $V = 0.257$, $p < 10^{-4}$).

C. How does overconfidence bias change as students' ability in introductory QM increases?

We plot, in Fig. 8, the mean DPF on confidence items and the DPF contrast, as a function of the students' ability on QM items, using the quartiles criterion.

The analysis shows that, for the whole sample, confidence significantly increases as ability increases ($F = 29.482$, $df = 3$, $p < 10^{-4}$, partial $\eta^2 = 0.19$) and that confidence bias of more able students is significantly lower than that of less able students ($F = 40.700$, $df = 3$, $p < 10^{-4}$, partial $\eta^2 = 0.24$).

In Fig. 9 we show the DPF contrast for the whole sample as a function of the three levels of the construct map adopted to describe the students' progression in introductory QM (see Ref. [47] for details).

In particular, for the whole sample, students at the lower level of the construct map (about 26.1%) are on overage overconfident (DPF contrast = +0.93 logit), while students of at the upper level (overall, 79 out of 383, about 21%) are slightly underconfident (average DPF contrast = −1.01 logit). This is not, however, because confidence is decreasing. In fact, it is increasing, just not as much as ability is.

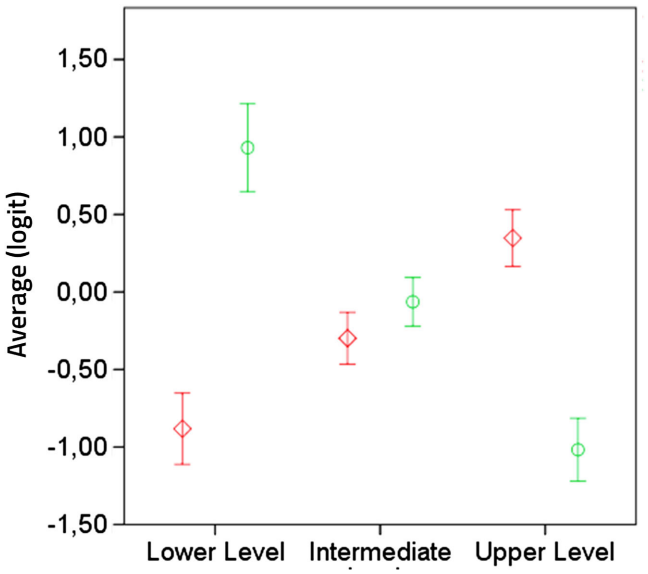

FIG. 9. DPF contrast (green circles) and DPF on confidence items (red diamonds) vs levels of the revised construct map in QM described in Ref. [47] for the whole sample. See Appendix for the definition of the construct map levels. Whiskers indicate 95% confidence interval. Green (red) circles (diamonds) are slightly shifted toward the right (left) to enhance readability.

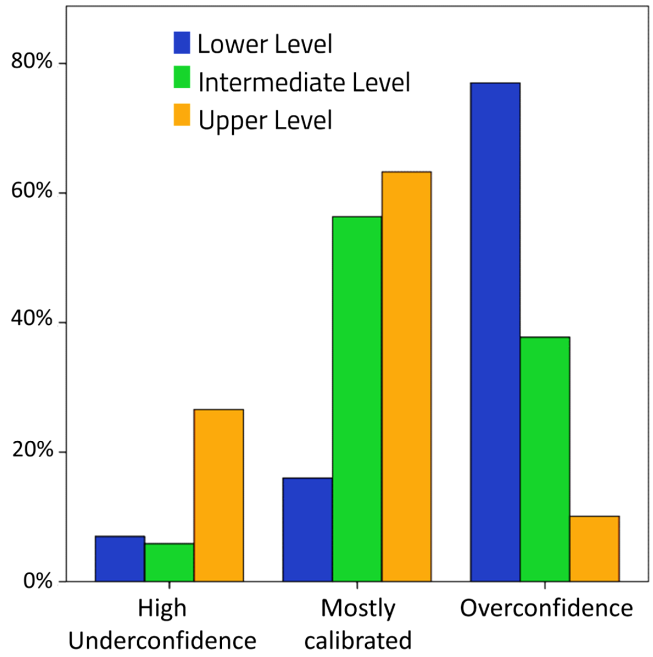

FIG. 10. Distribution of all respondents across the three confidence levels, subdivided according to the levels of the construct map in QM (see Appendix for definition of the construct map levels and Table II for confidence intervals).

Between each pair of the three levels, the average confidence bias, as measured by the DPF contrast, is significantly different ($t > 6.121$, $df = 380$, $p < 10^{-4}$).

This evidence is confirmed by a chi-square analysis (see Fig. 10), which shows that the association between the three levels of the construct map and the three levels of confidence calibration is statistically significant ($\chi^2 = 102.260$, $df = 4$, $p < 10^{-4}$; Cramer's $V = 0.37$, $p < 10^{-4}$).

The majority (about 60%) of the students of the upper level and of the intermediate level of the construct map are mostly calibrated. Moreover, only 10% of the students in the upper anchor shows a significant overconfidence bias, while this is the case of about 38% of the students in the intermediate level and 77% of the students in the lower level.

V. DISCUSSION

In this study, we examined how high school students' confidence bias, namely, the difference between one's own assessment of a performance and their actual performance, is mediated by instruction and how it changes when the ability increases in response to the received instruction, choosing introductory quantum mechanics as content area.

In the following, we summarize the extent to which our purposes have been fulfilled.

A. To what extent does instruction influence students' overconfidence bias in introductory QM?

We found that the TLS group, namely, students who had received instruction about the target topics through the activities described in Ref. [47], have consistently better performances than the textbook group (average ability on QM items $= +0.40$ logit vs -0.96 logit). Moreover, the TLS group exhibits also greater average confidence (about -0.03 logit vs -0.58 logit). We note that the confidence score of the textbook group does not decrease inasmuch, so that the smaller differences between the confidence scores of TLS and textbook groups resulted in a significantly greater overconfidence bias for the textbook group. Similarly, when looking at the average behavior of the TLS group, we found that the majority of students (about 55%) is substantially calibrated. The above evidence confirms the hypothesis H1, namely, that overconfidence bias can be significantly reduced when students are exposed to a transformative didactical intervention. We recall that the involved classes had been randomly chosen in such a way that each participating school to the experiment roughly contributed to both the TLS and textbook groups. Furthermore, all involved teachers had a similar teaching experience and had attended the same professional development course, in which they were familiarized with the inquiry-based strategy that would have been adopted with students. Therefore, the differences may be reasonably attributed to the transformative nature of the intervention of the TLS group and not to other external factors. Our interpretation is that the proposed activities, being informed by both prior work in physics education research about introductory QM, and research about inquiry-based instruction such as, e.g., questioning, feedback about learning and challenging prompts, may have likely provided students with more opportunities to enhance their ability of self-evaluation in comparison to traditional teaching. It is beyond the scope of this paper to identify which specific activity or aspect of the transformative intervention mostly contributed to the observed result. However, our interpretation is supported by former studies in educational psychology [63], which found that asking respondents to list reasons for their answer dropped confidence in their own responses, thus increasing calibration. Among the guided-inquiry activities, those focused on stimulating group discussions about the experimental results and critical reflections about the topics already studied in chemistry classes, as atomic models or orbitals, may have also played a relevant role since students could have been helped in this way assess their own knowledge in a better way. This evidence confirms recent results in biology education about the role of group work and guidelines in reducing confidence bias [64].

Conversely, the textbook intervention and the teacher-directed lecture approach resulted in a worse confidence calibration of the textbook group. In particular, about half of the students in the textbook group show a significant overconfidence bias. Previous studies in behavioral psychology suggest that students' overconfidence increases when tasks are perceived as "simple" or "easy" [65].

In physics, this can be translated as follows: students' confidence in answering an item or solving a problem can be related to the extent to which the possibility to use known formulas is recognized [66]. In our study, the items featured in the questionnaire could have been perceived as "easy" since the questions apparently requested the humble recalling of rote-learned formulas (e.g., $E = h\ \nu$) or notions, while they were actually probing a deeper understanding of the targeted topics. Textbook group students' miscalibration can also be related to an unbalanced judgment of what they have already learned, which may have led them to feel unrealistically confident in their understanding [67]. In other words, less learning resources could have been invested by low performers in learning the new topics, likely because the latter seemed superficially familiar to those already learned in the chemistry classes. In the Italian curriculum, as thoroughly discussed in Ref. [47], introductory QM topics, like atomic models or orbitals, are taught in chemistry classes together with very basic notions about radiation emission, photons, and wave-particle duality, typically during the third year of high school (age 15–16 years), i.e., two years before the so-called "modern physics" topics are taught in the physics classes: the different focus and scope of the chemistry teaching may have affected student confidence when approaching the same topics from a different viewpoint. However, it would be worthwhile for future research to confirm whether and how prior knowledge in chemistry may lead to overestimate or underestimate one's own performance when dealing with introductory QM.

B. How does overconfidence bias change as students' ability in introductory QM increases?

Overall, our data confirm that less able students exhibit in their responses a significant higher overconfidence bias than more able students, the latter being slightly underconfident (Figs. 8 and 9). This result is consistent with reports regarding a variety of fields [22,45,57,68–70], which can be easily extended to QM. In particular, by adopting the model by Hasan *et al.* [27], our data suggest that students may lack strong mental models about the targeted QM topics, similarly to what happens in electromagnetism, but differently than in classical mechanics, where misconceptions are more deeply rooted. As a general implication, the overconfidence bias seems less pronounced when addressing more abstract topics. In other words, students' misconceptions about introductory QM topics at the high school level are present, but not so strongly rooted since students' mental models of microscopic behavior of matter often lack the link to everyday experience, as it occurs, instead, in mechanics.

Concerning our hypothesis of an inverse relationship between overconfidence and increasing understanding of introductory QM (H2), our data support that participants at the highest level of the construct map validated in Ref. [47] tend to have lower values of overconfidence. In other words, when looking at students' distribution across the levels of the construct map, confidence tends to become closer to the actual performance for about 60% of the students at intermediate and upper levels, while about 80% of the students at the lowest level exhibit moderate to high overconfidence.

Given the above evidence, our study supports the general claim that a better calibration between confidence and performance is associated with improved ability. Moreover, the positive relationship between ability and calibration is common to the whole sample, thus confirming that instruction may help students improve their capabilities to correctly evaluate their own performance.

While a more detailed analysis is beyond the scope of this study, we note that about 25% of students in the upper level of the construct map exhibit high underconfidence (Fig. 10). Lindsey and Nagel [66] suggest that, in physics, underconfidence is problematic as much as overconfidence, since a deep knowledge should correspond to the metacognitive ability to self-recognize also a correct understanding of a given topic. A possible interpretation is that topics targeted in the upper anchor (behavior of metals and insulators) were actually difficult to grasp both for the textbook and the TLS group students. In particular, even though high performers responded to the knowledge items in a correct way, they generally felt not so confident in their responses, likely because the time spent on these topics during traditional and transformative instruction was not sufficient to allow a deeper understanding of the targeted concepts. However, we believe that further investigation is required on this specific issue. An alternative explanation may be strictly related to the chosen content, i.e., introductory QM. While the observed response pattern for knowledge items is the same as the ones observed in areas in which the absence of personal experience hinders the creation of a coherent interpretation framework, QM may be still perceived *a priori* as difficult by students because of the expected high level of formalism, abstractness, and complexity of experimental apparatuses [71]. However, further research is warranted to find out the extent to which such perception actually affects the confidence in one's own performance.

VI. CONCLUSIONS AND FUTURE RESEARCH

Confidence is an important metacognitive construct that concerns the self-assessment of one's own knowledge. Previous studies have shown a significant correlation between confidence and self-regulatory processes during learning and decision making [72–75]. However, when students' judgment is not calibrated with the actual performance, underconfidence or overconfidence biases arise. Overconfidence bias, in particular, is a specific bias in beliefs that induces deviation from payoff-maximizing behavior [76]. The purpose of this study was to investigate

how instruction affects overconfidence bias at the high school level using an underexplored but meaningful context like introductory QM. We chose this area in physics because of some didactical peculiarities that may have contrasting effects on students' confidence: (i) it involves challenging but fascinating topics; (ii) it is previously taught in chemistry classes; (iii) some concepts have different meanings than in classical physics.

Previous studies in physics education have investigated how overconfidence at the item level may signal the presence of misconceptions [27]. For instance, in mechanics, such misconceptions may be related to strong mental models, which are alternative to Newtonian dynamics [34]. In electromagnetism, previous studies found lower confidence and hence concluded that likely students lack strong conceptual models [30]. The present study contributes to the field showing that the overconfidence bias at the person level arises also in a content area, as introductory QM, where misconceptions are not so deeply rooted in student's cognition. Our findings support the idea that students do not hold a coherent framework to interpret the behavior of the quantum world, even though they were previously taught about these concepts in chemistry classes or in extracurricular activities. On the contrary, such previous experiences may act as potential causes for overconfidence.

Using the revised construct map about introductory QM topics [47], we also showed that overconfidence bias decreases as students progress along increasingly complex levels of understanding of the target concepts. By controlling the instructional variable, we can claim that the guided-inquiry activities in which the TLS group was involved successfully reduced overconfidence bias. Among the proposed activities, those in which students were prompted to recall topics already learned in chemistry classes, such as, e.g., how the atomic structure influences the behavior of metals, insulators, and semiconductors, had likely helped students review and better assess their own knowledge. Similarly, asking students to assess, through group discussions and reflections, the strength of their prior knowledge in physics and chemistry, likely helped students develop more accurate outcome expectations.

Our study also adopted a different methodological approach from previous ones, focusing on confidence bias at the person level and evaluating accuracy and confidence together using a 1D Rasch model. Rasch analysis allowed more sophisticated psychometric computations of the overconfidence bias construct. In particular, calibration intervals were suitably coded using differential person functioning (DPF) contrast probabilities to focus on students' overconfidence bias. The reason for using the 1D Rasch model was that a consistent definition of confidence bias requires the evaluation of conceptually different constructs—confidence and ability—on the same linear scale using the same measurement unit. When using the logit unit, we address methodological issues related to the use of raw data, which have only ordinal validity and cannot guarantee linearity, unavoidably resulting in a scale distortion. The reported statistics, PCA of residuals, and analysis of the confidence rating scale confirm that the data well fit the Rasch model so that used questionnaire, which combines knowledge items and confidence scale, can be considered psychometrically sound.

Future studies are worthwhile to investigate the role of other variables that may influence overconfidence bias at the person level and that we did not include in our research design. First, as suggested in Refs. [2,3,69], intrinsic interest and perceived encouragement may significantly impact students' confidence. Further research is needed to establish whether these constructs favor calibration or overconfidence bias. Along with prior work, the present study suggests that high performance correlates better with calibration and underconfidence, but it is not clear why a better knowledge should limit self-confidence. If confidence is a mediator for interest, then not all high performing students may have a specific interest in the topic and therefore their perceived confidence may be lower than expected. Similarly, more research is needed to understand if the extrinsic value of physics (e.g., if studying physics is perceived as necessary to do well at university or to get a desirable job) may influence overconfidence bias. Similarly, further research is needed to establish whether the perception of the discipline (namely, considering physics as more difficult in comparison to other disciplines) can influence student's confidence in evaluating their performance. In this study, we found that high-ability students were slightly underconfident, so that a possible effect of the perception of QM as difficult could be at play. While the perception of a discipline may be an unavoidable issue in high school and university teaching, one's own confidence in completing a task in physics may be influenced by the extent to which one may think they are doing their best—when actually they are not performing well—simply because physics is perceived as a hard science and not because they have strong misconceptions or lack mathematical and problem-solving skills [35].

As a concluding remark, our results suggest putting more effort into physics education research on investigating how to support metacognitive strategies by means of systematic instruction in order to help students better calibrate their accuracy and confidence.

VII. LIMITATIONS

A number of limitations must be acknowledged when interpreting the results. First, in this paper, we limited the analysis to a dichotomous scoring. Further research is needed to investigate if the scoring method may have a significant effect on the measure of overconfidence bias, an effect that was overlooked by previous studies, and which we plan to address in a forthcoming paper. The ordered

multiple-choice approach, by giving a credit also to partially correct answer choices, may provide a more accurate estimate of the students' performance, and hence of the overconfidence bias, signaling it only when it can really impair students' learning. Reasoning strategies corresponding to such partial answer choices can be important levers on which to build new scientifically correct knowledge, especially in content areas as introductory QM when extra efforts are required to students to bridge their existing knowledge about classical physics toward the QM knowledge. Finally, literature suggests that gender may be an important variable when studying overconfidence bias. Unfortunately, for various reasons, we did not record gender in the present survey. Hence, we are planning a new administration of the questionnaire with different classes to investigate the role of gender on overconfidence bias.

ACKNOWLEDGMENTS

The authors are deeply grateful to the teachers and the students of the involved schools: Liceo Scientifico L. B. Alberti (Naples), Liceo Scientifico Cuoco-Campanella (Naples), Liceo Scientifico F. Sbordone (Naples), Liceo Scientifico P. Calamandrei (Naples), and Liceo Virgilio (Pozzuoli).

APPENDIX: DEFINITION OF THE CONSTRUCT MAP LEVELS USED IN THIS STUDY [47]

Upper level.— Students at this level grasp the concepts of chemical bond and of molecular orbital. They distinguish between conductors and insulators in terms of energy bands. They know that only the conduction-band electrons contribute to the electrical current in metals and that the number of charge carriers can be changed in semiconductors. They can qualitatively discuss how LED and solar cells work.

Intermediate level.—Students at this level know that the Heisenberg principle sets an intrinsic limit to the possibility to determine the particle law of motion and trajectory and that it also constrains measurements errors in the quantum limit. They are also able to qualitatively explain the atom stability by using the uncertainty principle. They know the electronic structure of atoms in terms of energy levels and can compute the energy of emitted or absorbed photons in terms of levels difference. They are acquainted with the probabilistic interpretation of atomic orbitals.

Lower level.—Students at this level know that classical physics cannot fully explain the interaction between matter and radiation. They can use Planck's constant h to compute photon energy. They qualitatively know that matter atoms exchange energy with radiation by emitting and absorbing photons.

[1] Y. Jiang, J. Song, M. Lee, and M. Bong, Self-efficacy, and achievement goals as motivational links between perceived contexts, and achievement, Educ. Psych. **34**, 92 (2014).
[2] J. Viljaranta, A. Tolvanen, K. Aunola, and J.-E. Nurmi, The developmental dynamics between interest, self-concept of ability, and academic performance, Scandinavian J. Educ. Res. **58**, 734 (2014).
[3] E. Regan and J. DeWitt Attitudes, interest and factors influencing stem enrolment behaviour: An overview of relevant literature, in *Understanding Student Participation and Choice in Science and Technology Education*, edited by E. Henriksen, J. Dillon, and J. Ryder (Springer, Dordrecht, 2015), https://www.springer.com/gp/book/9789400777927.
[4] M. Veenman, B. Van Hout-Wolters, and P. Afflerbach, Metacognition and learning: Conceptual and methodological considerations, Metacogn. Learn. **1**, 3 (2006).
[5] D. F. Bjorklund, V. Periss, and K. Causey, The benefits of youth, Eur. J. Dev. Psych. **6**, 120 (2009).
[6] N. Destan and C. M. Roebers, What are the metacognitive costs of young children's overconfidence?, Metacogn. Learn. **10**, 347 (2015).
[7] J. Dunlosky and K. Rawson, Overconfidence produces underachievement: Inaccurate self evaluations undermine students' learning and retention, Learn. Instr. **22**, 271 (2012).
[8] T. Bouffard and S. Narciss, Benefits and risks of positive biases in self-evaluation of academic competence: Introduction, Int. J. Educ. Res. **50**, 205 (2011).
[9] L. Stankov, J. Lee, W. Luo, and D. J. Hogan, Confidence: A better predictor of academic achievement than self-efficacy, self-concept and anxiety?, Learn. Individ. Differ. **22**, 747 (2012).
[10] B. Sreenivasulu and R. Subramaniam, Exploring undergraduates' understanding of transition metals chemistry with the use of cognitive and confidence measures, Res. Sci. Educ. **44**, 801 (2014).
[11] S. L. Britner and F. Pajares, Sources of science self-efficacy beliefs of middle school students, J. Res. Sci. Teach. **43**, 485 (2006).
[12] P. Chen and B. Zimmerman, A cross-national comparison study on the accuracy of self-efficacy beliefs of middle-school mathematics students, J. Exp. Educ. **75**, 221 (2007).
[13] P. P. Chen, Exploring the accuracy and predictability of the self-efficacy beliefs of seventh grade mathematics students, Learn. Individ. Differ. **14**, 77 (2003).
[14] J. Möller and B. Pohlmann, Achievement differences, and self-concept differences: Stronger associations for above or below average students?, Br. J. Educ. Psychol. **80**, 435 (2010).

[15] F. Pajares and L. Graham, Self-efficacy, motivation constructs, and mathematics performance of entering middle school students, Contemp. Educ. Psychol. **24**, 124 (1999).
[16] M. M. Chiu and R. M. Klassen, Relations of mathematics self-concept and its calibration with mathematics achievement: Cultural differences among fifteen-year-olds in 34 countries, Learn Instr. **20**, 2 (2010).
[17] OECD, *The Future of Education, and Skills: Education 2030*, (Directorate for Education, and Skills-OECD, Paris, France, 2018), https://www.oecd.org/education/2030/E2030%20Position%20Paper%20(05.04.2018).pdf.
[18] D. A. Moore and P. J. Healy, The trouble with overconfidence, Psychol. Rev. **115**, 502 (2008).
[19] R. Larrick, K. Burson, and J. Soll, Social comparison and confidence: When thinking you're better than average predicts overconfidence (and when it does not), Organ. Behav. Human Dec. Proc. **102**, 76 (2007).
[20] J. Ehrlinger, A. L. Mitchum, and C. S. Dweck, Understanding overconfidence: Theories of intelligence, preferential attention, and distorted self-assessment, J. Exp. Soc. Psychol. **63**, 94 (2016).
[21] L. Stankov and J. Lee, Confidence and cognitive test performance, J. Educ. Psychol. **100**, 961 (2008).
[22] A. Rachmatullah and M. Ha, Examining high-school students' overconfidence bias in biology exam: A focus on the effects of country and gender, Int. J. Sci. Educ. **41**, 652 (2019).
[23] L. Stankov, Noncognitive predictors of academic achievement and intelligence: An important role of self-confidence, Pers. Individ. Differ. **60**, S37 (2014).
[24] A. S. Labuhn, B. J. Zimmerman, and M. Hasselhorn, Enhancing students' self-regulation and mathematics performance: The influence of feedback and self-evaluative standards, Metacogn. Learn. **5**, 173 (2010).
[25] S. Lichtenstein and B. Fischhoff, Do those who know more also know more about how much they know?, Organ. Behav. Hum. Perform. **20**, 159 (1977).
[26] A. Zohar and Y. J. Dori, Introduction, in *Metacognition in Science Education: Trends in Current Research*, edited by A. Zohar and Y. J. Dori (Springer-Verlag, Dordrecht, Netherlands, 2012), pp. 1–20, https://www.springer.com/gp/book/9789400721319.
[27] S. Hasan, D. Bagayoko, and E. Kelley, Misconceptions and the Certainty of Response Index (CRI), Phys. Educ. **34**, 294 (1999).
[28] J. Clement, D. E. Brown, and A. Zietsman, Not all preconceptions are misconceptions: Finding "anchoring" conceptions' for grounding instruction on students' intuitions, Int. J. Sci. Educ. **11**, 554 (1989).
[29] H. Pesman and A. Eryilmaz, Development of a three-tier test to assess misconceptions about simple electric circuits, J. Educ. Res. **103**, 208 (2010).
[30] J. Leppavirta, Assessing undergraduate students' conceptual understanding and confidence of electromagnetics, Int. J. Sci. Math. Educ. **10**, 1099 (2012).
[31] D. Gurcay and E. Gulbas, Development of three-tier heat, temperature and internal energy diagnostic test, Res. Sci. Technol. Educ. **33**, 197 (2015).
[32] G. D. Kaltakci, A. Eryilmaz, and L. McDermott, Development and application of a four-tier test to assess pre-service physics teachers' misconceptions about geometrical optics, Res. Sci. Technol. Educ. **35**, 1 (2017).
[33] 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).
[34] M. Planinic, W. Boone, R. Krsnik, and M. Beilfuss, Exploring alternative conceptions from Newtonian dynamics and simple dc circuits: Links between item difficulty and item confidence, J. Res. Sci. Teach. **43**, 150 (2006).
[35] M. Potgieter, E. Malatje, E. Gaigher, and E. Venter, Confidence versus performance as an indicator of the presence of alternative conceptions and inadequate problem-solving skills in mechanics, Int. J. Sci. Educ. **32**, 1407 (2010).
[36] 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).
[37] I. Caleon and R. Subramaniam, Development and application of a three-tier diagnostic test to assess secondary students' understanding of waves, Int. J. Sci. Educ. **32**, 939 (2010).
[38] A. M. Glenberg and W. Epstein, Calibration of comprehension, J. Exper. Psychol.: Learn. Mem. Cogn. **11**, 702 (1985).
[39] J. Kruger and D. Dunning, Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments, J. Pers. Soc. Psychol. **77**, 1121 (1999).
[40] N. Didis and L. Wang, Students' mental models of atomic spectra, Chem. Educ. Res. Pract. **17**, 743 (2016).
[41] F. Savall-Alemany, J. L. Doménech, J. Guisasola, and J. Martínez-Torregrosa, Identifying student and teacher difficulties in interpreting atomic spectra using a quantum model of emission and absorption of radiation, Phys. Rev. Phys. Educ. Res. **12**, 010132 (2016).
[42] N. Didis, A. Eryilmaz, and S. Erkoç, Investigating students' metal models about the quantization of light, energy and angular momentum, Phys. Rev. Phys. Educ. Res. **10**, 020127 (2014).
[43] K. Krijtenburg-Lewrissa, H. J. Pol, A. Brinkman, and W. R. Jooligen, Insights into teaching quantum mechanics in secondary and lower undergraduate education, Phys. Rev. Phys. Educ. Res. **13**, 010109 (2017).
[44] H. K. E. Stadermann, E. van den Berg, and M. J. Goedhart, Analysis of secondary school quantum physics curricula of 15 different countries: Different perspectives on a challenging topic, Phys. Rev. Phys. Educ. Res. **15**, 010130 (2019).
[45] S. Pazicni and C. Bauer, Characterizing illusions of competence in introductory chemistry students, Chem. Educ. Res. Pract. **15**, 24 (2014).
[46] U. Amaldi, *L'Amaldi per i Licei Scientifici* (Zanichelli, Bologna, 2012).
[47] S. di Uccio *et al.*, following paper, Development of a construct map to describe students' reasoning about introductory quantum mechanics, Phys. Rev. Phys. Educ. Res. **16**, 010144 (2020).

[48] I. Sadeh and M. Zion, The development of dynamic inquiry performances within an open inquiry setting: A comparison to guided inquiry setting, J. Res. Sci. Teach. **46**, 1137 (2009).

[49] I. Sadeh and M. Zion, Which type of inquiry project do high school biology students prefer: Open or guided?, Res. Sci. Educ. **42**, 831 (2011).

[50] E. M. Furtak, T. Seidel, H. Iverson, and D. C. Briggs, Experimental and quasi-experimental studies of inquiry-based science teaching: A meta-analysis, Rev. Educ. Res. **82**, 300 (2012).

[51] C. E. Hmelo-Silver, R. G. Duncan, and C. A. Chinn, Scaffolding and achievement in problem-based and inquiry learning: A response to Kirschner, Sweller, and Clark (2006), Educ. Psych. **42**, 99 (2007).

[52] S.-C. Fan, Y.-S. Hsu, H.-Y. Chang, W.-H. Chang, H.-K. Wu, and C.-M. Chen Investigating the effects of structured and guided inquiry on students' development of conceptual knowledge and inquiry abilities: A case study in Taiwan, Int. J. Sci. Educ. **38**, 1945 (2016).

[53] L. B. Buck, S. L. Bretz, and M. H. Towns, Characterizing the level of inquiry in the undergraduate laboratory, J. Coll. Sci. Teach. **38**, 52 (2008).

[54] T. Bunterm, K. Lee, J. Ng Lan Kong, S. Srikoon, P. Vangpoomyai, J. Rattanavongsa, and G. Rachahoon, Do different levels of inquiry lead to different learning outcomes? A comparison between guided and structured inquiry, Int. J. Sci. Educ. **36**, 1937 (2014).

[55] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevPhysEducRes.16.010143 for the data analysis based on raw scores.

[56] D. J. Hacker, L. Bol, and M. C. Keener, Metacognition in education: A focus on calibration, in *Handbook of Metamemory and Memory*, edited by J. Dunlosky and R. A. Bjork (Psychology Press, New York, NY, 2008), pp. 429–456, https://psycnet.apa.org/record/2008-07511-022.

[57] L. Stankov and J. Lee, Overconfidence across world regions, J. Cross Cult. Psychol. **45**, 821 (2014).

[58] W. J. Boone, J. R. Staver, and M. S. Yale, *Rasch Analysis in the Human Sciences* (Springer Dordrecht, Netherlands, 2014), https://www.springer.com/gp/book/9789400768567.

[59] W. Romine, D. 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).

[60] I. Paek, J. Lee, L. Stankov, and M. Wilson, Rasch modeling of accuracy and confidence measures from cognitive tests, J. Appl. Meas. **14**, 232 (2013).

[61] J. M. Linacre, A User's Guide to Winsteps. Retrieved from http://www.winsteps.com/manuals.htm (2012).

[62] W. J. Boone, Rasch analysis for instrument development: Why, when, and how?, CBE Life Sci. Educ. **15**, 4 (2016).

[63] A. Koriat, S. Lichtenstein, and B. Fischhoff, Reasons for confidence, J. Exper. Psychol.: Learn. Mem. Cogn. **6**, 107 (1980).

[64] L. Bol, D. J. Hacker, C. C. Walck, and J. A. Nunnery, The effects of individual or group guidelines on the calibration accuracy and achievement of high school biology students, Contemp. Educ. Psychol. **37**, 280 (2012).

[65] H. R. Arkes, C. Christensen, C. Lai, and C. Blumer, Two methods of reducing overconfidence, Organ. Behav. Hum. Dec. Proc. **39**, 133 (1987).

[66] B. A. Lindsey and M. L. Nagel, Do students know what they know? Exploring the accuracy of students' self-assessments, Phys. Rev. Phys. Educ. Res. **11**, 020103 (2015).

[67] M. Händel and E. Fritzsche, Students' confidence in their performance judgements: A comparison of different response scales, Educ. Psychol. **35**, 377 (2015).

[68] D. Ryvkin, M. Krajc, and A. Ortmann, Are the unskilled doomed to remain unaware?, J. Econ. Psychol. **33**, 1012 (2012).

[69] R. Sheldrake, Students' intentions towards studying science at upper-secondary school: The differential effects of under-confidence and over-confidence. Int. J. Sci. Educ. **38**, 1256 (2016).

[70] K. Wüst and H. Beck, I thought I did much better—overconfidence in university exams, Dec. Sci. J. Innov. Educ. **16**, 310 (2018).

[71] J.-L. Ke, M. Monk, and R. Duschl, Learning introductory quantum mechanics: Sensorimotor experience and mental models, Int. J. Sci. Educ. **27**, 1571 (2005).

[72] C. M. Allwood and P. A. Granhag, Feelings of confidence and the realism of confidence judgments in everyday life, in *Judgment and Decision Making: Neo-Brunswikian and Process-Tracing Approaches,* edited by P. Juslin and H. Montgomery (Lawrence Erlbaum, Mahwah, NJ, 1999), pp. 123–146, https://www.taylorfrancis.com/books/e/9781410617675/chapters/10.4324/9781410617675-14.

[73] A. Efklides, Metacognitive experiences: The missing link in the self-regulated learning process, Educ. Psychol. Rev. **18**, 287 (2006).

[74] S. A. Jackson and S. Kleitman, Decision-making tendencies in a medical paradigm: The role of individual differences in feelings of confidence and its calibration, Metacogn. Learn **9**, 25 (2014).

[75] Y. Jiang and S. Kleitman, Metacognition and motivation: Links between confidence, self-protection and self-enhancement, Learn. Individ. Differ. **37**, 222 (2015).

[76] S. DellaVigna, Psychology and economics: Evidence from the field, J. Econ. Lit. **47**, 315 (2009).