

Quantum mechanics concept assessment: Development and validation study

Homeyra R. Sadaghiani

*Department of Physics and Astronomy, California State Polytechnic University,
Pomona, California 91768, USA*

Steven J. Pollock

Department of Physics, University of Colorado, Boulder, Colorado 80309, USA

(Received 22 September 2014; published 13 March 2015)

As part of an ongoing investigation of students' learning in first semester upper-division quantum mechanics, we needed a high-quality conceptual assessment instrument for comparing outcomes of different curricular approaches. The process of developing such a tool started with converting a preliminary version of a 14-item open-ended quantum mechanics assessment tool (QMAT) to a multiple-choice (MC) format. Further question refinement, development of effective distractors, adding new questions, and robust statistical analysis has led to a 31-item quantum mechanics concept assessment (QMCA) test. The QMCA is used as post-test only to assess students' knowledge about five main topics of quantum measurement: the time-independent Schrödinger equation, wave functions and boundary conditions, time evolution, and probability density. During two years of testing and refinement, the QMCA has been given in alpha ($N = 61$) and beta versions ($N = 263$) to students in upper division quantum mechanics courses at 11 different institutions with an average post-test score of 54%. By allowing for comparisons of student learning across different populations and institutions, the QMCA provides instructors and researchers a more standard measure of effectiveness of different curricula or teaching strategies on student conceptual understanding of quantum mechanics. In this paper, we discuss the construction of effective distractors and the use of student interviews and expert feedback to revise and validate both questions and distractors. We include the results of common statistical tests of reliability and validity, which suggest the instrument is presently in a stable, usable, and promising form.

DOI: [10.1103/PhysRevSTPER.11.010110](https://doi.org/10.1103/PhysRevSTPER.11.010110)

PACS numbers: 01.40.Fk, 01.40.gf, 03.65.-w

I. INTRODUCTION

Investigations of student learning in introductory physics over the last several decades have helped us better assess student conceptual understanding and identify and address common difficulties in learning physics [1]. One outcome of research on conceptual understanding has been the construction of concept inventories and diagnostic tools to help physics instructors improve their practice. Subsequently, many instructors have used these tools to evaluate the effectiveness of innovative classroom strategies, conduct pedagogical research, or learn about student difficulties. Ongoing development of many curricula and classroom interventions were also driven in part by data collected from research-based assessment instruments, such as the Force Concept Inventory (FCI), [2] Force and Motion Conceptual Evaluation (FMCE) [3], Concept Survey of Electricity and Magnetism (CSEM) [4], and Brief Electricity and Magnetism Assessment (BEMA) [5].

For instance, 6000 students' FCI data [6] have shown that interactive engagement modes of instruction have a positive impact on student learning.

Well-designed, valid, and reliable assessment tools that are able to capture facets of student understanding of key concepts can help inform instructors about what students are learning. When the Force Concept Inventory was first given to introductory physics students, many instructors were surprised by the inability of their students to answer seemingly easy conceptual questions. While most introductory students could use physics formulas to solve mathematical problems, they were less successful in answering basic conceptual questions related to those problems. These results stimulated many in the physics community to reconsider how they were teaching introductory physics [7].

Research into student learning of quantum mechanics (QM) is less mature compared to many other areas of physics. Nonetheless, several researchers have found persistent learning difficulties experienced by students undertaking the formal study of quantum mechanics [8–12]. To date, a few quantum mechanics concept surveys have been designed to assess student learning. Some of these surveys are focused on student difficulties with a particular issue, such as visualizations [13], or a single narrowly defined

Published by the American Physical Society under the terms of the Creative Commons Attribution 3.0 License. Further distribution of this work must maintain attribution to the author(s) and the published article's title, journal citation, and DOI.

concept such as potential barriers [14], or quantum measurement [15]. Other assessment tools target a relatively broader range of QM topics designed for sophomore level modern physics [16]. For more advanced topics in QM, the quantum mechanics survey [17] spans a variety of important topical areas. This survey is broadly appropriate for upper-division QM and incoming graduate students, albeit with some emphasis on formalism. Thus, it may not be appropriate after just one semester of quantum mechanics courses in which all relevant concepts are not covered. Further, students need to understand the basics of linear algebra in order to do well in this survey.

The quantum mechanics assessment tool (QMAT) is a 14-item open-ended instrument designed at the University of Colorado Boulder (CUB). It was intended to assess a subset of learning goals identified for first-semester upper-division undergraduate quantum mechanics courses by faculty that commonly teach these courses [18]. Many questions in this test were motivated by previous research on student difficulties. Although the QMAT incorporates research findings on student difficulties in advanced undergraduate QM and pays attention to the alignment of the assessment tool with course learning objectives and instructional design [1], it suffers from a complicated and unreliable scoring rubric, with correspondingly limited validation studies. There are a variety of difficulties associated with reliable scoring of open-ended questions [19], and such issues have restricted the usefulness and transferability of the QMAT within and across institutions.

In order to address these scalability and usability issues, we set out to build on the existing QMAT and craft a quantum mechanics concept assessment (QMCA), a multiple-choice (MC) tool that could be more easily and objectively graded and has a potential to be used as a tool by a wide range of faculty to provide a meaningful measure of students' performances on conceptual questions in upper-division quantum mechanics. High quality MC tests with proper distractors have a long tradition of providing diagnostics of student difficulties, evaluating teaching methods, and comparing curricula. MC tests have some advantages over open-ended tests. They can be easily administered and accurately graded, the results are objective and amenable to statistical analysis, and can be less ambiguous to validate. This allows them to be a more versatile tool for comparing different instructional methods and different student populations.

Constructing valid and reliable multiple choice items is neither a quick nor trivial task. In addition to research on common student difficulties, such a project requires significant knowledge about the array of common incorrect student responses. In crafting QMCA items, we sought to refine the questions, to establish validity, obtain a representative spectrum of common student alternative responses, and collect sample data from different populations to conduct statistical analysis and reliability tests. Over the last two years, we have tested several versions of the MC instrument. Using statistical analysis on earlier versions of the QMCA, interview data, and expert feedback, we have modified the wording of many questions. Some QMAT questions have been eliminated due to problems uncovered in interviews, and a few new questions were added. We iterated the process with each version several times to refine the questions and alternative answer choices. The QMCA presently contains 31 questions focusing on the same five main concepts as the QMAT—measurement (Meas.), the time-independent Schrödinger equation (TISE), wave functions and boundary conditions (WF), time evolution (Time), and probability or probability density (Prob.). Thus, it represents an attempt to comprehensively cover the early portion of a traditional upper division first semester quantum mechanics course with conceptual questions based on observed and reported student difficulties. Table I subjectively classifies the QMCA concept framework, along with the item numbers in which they appear. Experts often categorized a given question in more than one possible way and we have adopted the QMAT categorizations and added the new QMCA questions to those categories.

All the concepts in Table I are recognized as essential topics in undergraduate junior- or senior-level quantum mechanics courses. The Table is best interpreted as a decomposition of commonly covered material into five conceptual dimensions. Note, though, that each dimension is probed by more than one question on the same subject to enhance the test's validity and reliability [20].

The learning objectives listed in Table I are consensus outcomes of faculty discussions at CU on fundamental and essential concepts for understanding many intermediate-level quantum mechanics topics. The QMAT, and thus the QMCA that was constructed upon the main topics of QMAT, explicitly avoids questions on spin, angular

TABLE I. A classification of the QMCA concept framework with the item numbers in which they appear.

Concept framework	Item number
Measurement (Meas.)	1, 2, 3, 4, 5, 9, 10, 11, 12, 13, 14, 20, 21, 27, 28, 29
The time-independent Schrödinger equation (TISE)	6, 7, 17, 22, 23, 24
Time evolution (Time)	4, 5, 9, 10, 15, 16, 17, 18, 19, 30, 31
Wave functions or boundary conditions (WF)	6, 7, 8, 9, 10, 18, 19, 22, 23, 24, 25, 26, 28, 30, 31
Probability or probability density (Prob.)	1, 9, 10, 17, 18, 22, 25, 26

momentum, the harmonic oscillator, and the hydrogen atom. The rationale for these omissions is multifold. First, research on student difficulties on these more advanced topics is extremely limited. But more importantly, as with introductory level concept inventories (like the FCI), we wanted to focus on those commonly taught core ideas in undergraduate quantum mechanics that would span different course implementations. For instance, some quantum mechanics courses are taught as a one-semester course, which may not address some of the more advanced topics.

Physics instructors and researchers can use QMCA in any way they see fit—for example, to compare classes, to assess pedagogical changes, or learn more about student difficulties. In the body of the article, we discuss the construction of the QMCA instrument, including the adaptation of the questions, development of distractors, and validation of the test items. We present analyses of statistical reliability and validation using data from multiple institutions. These data sets can serve as a baseline for future use of this instrument. We conclude with some interpretation of results and potential uses of this instrument.

II. RESEARCH AND DEVELOPMENT METHODOLOGY

We began constructing the QMCA by first converting the original QMAT questions into a MC format. Questions were adopted from 12 of the 14 open-ended QMAT questions (excluding QMAT #7, and 8). Eight new but related questions were designed and added. Most of the original QMAT questions were revised and reworded in response to issues that came up in student interviews, and/or based on expert feedback. Some questions had to be completely rewritten due to their very open-ended nature, which did not easily suit the MC format (e.g., see Q6 in the next section). A diagrammatic summary of the research and development methodology is depicted in Fig. 1.

A systematic development of any assessment tool starts with establishing clear, meaningful, and measurable learning goals and, furthermore, it requires a sufficient knowledge of student common difficulties [21]. By building on the QMAT, we took advantage of the earlier work on these two requirements. First, those earlier studies had

identified generally accepted learning goals for junior level quantum mechanics courses by interviewing 18 instructors who regularly teach these courses [22]. Second, many QMAT items were adapted from published research questions in the literature, or benefited from previous related research on student common difficulties [23,24].

A key step in the process was using student wording to develop clear, concise, and meaningful distractors that would be effective in identifying student naïve ideas and yet acceptable and understandable to experts. In developing introductory physics concept inventories, often the students’ own wording is used [25]. However, given the abstract and sometimes counterintuitive nature of the quantum mechanics context, we found student responses often were not complete thoughts or coherent sentences. Thus, it was a challenge to be true to the students’ exact statements while constructing clear alternatives that were equally attractive, and approximately homogeneous in content, length, and grammar. While avoiding excessive wordiness in the alternatives, on drafting the distractors we focused on common core themes among similar students’ ideas that were popular in our sample.

A. Development of item distractors

An affordance of a MC version over an open-ended instrument is to explicitly confront students with incorrect answers. These distractors play a key role in developing the QMCA, since they should include a comprehensive set of students’ common alternative responses for given questions [26]. This requires a research base on students’ ideas surrounding these topics.

We built on four main legs in constructing the initial distractors. First, we used students’ responses to the open-ended QMAT. Second, using a series of think-aloud interview protocols with individual students, we were able to compile a variety of examples of student externalization of their reasoning process. Third, we examined the literature to ensure the inclusion of known student difficulties in related item distractors. Finally, we consulted faculty from several institutes for their feedback on the level and appropriateness of the content of both questions and their alternatives, as well as the phrasing of the questions.

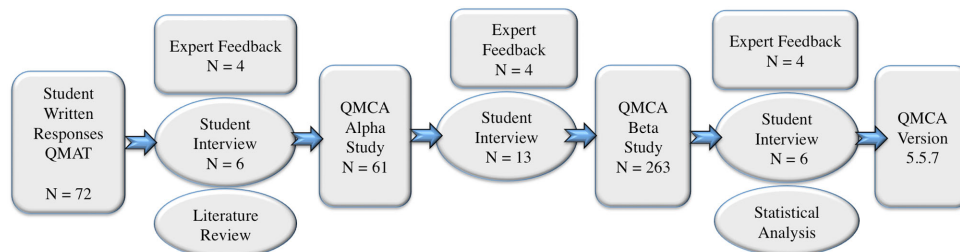


FIG. 1. A diagrammatic view of research and development.

An initial set of the QMCA questions and distractors was developed using student responses to the open-ended QMAT from Cal Poly Pomona (CPP; $N = 19$) and CU Boulder (CUB; $N = 53$). Students' written responses were categorized into groups of similar ideas that formed the first set of the distractors for the MC version. We also conducted follow-up interviews with six CPP individual students to develop a better understanding of students' reasoning when they answer the open-ended questions. Items are focused on concepts; however, in a few cases, formal language is used to help probe the understanding of the concepts. The goal was to have the language itself constitute as small an obstacle as possible for the students. Our aim is to avoid questions which faculty would consider overly difficult or subtle, while presenting situations where alternative conceptions are common.

In addition to using student responses for the open-ended QMAT and student interviews, we examined existing literature [27] to ensure congruence of known student difficulties in the chosen distractors. In some cases, students' correct and incorrect responses produced a broad spectrum of plausible distractors and keys. In the alpha version, we allowed students to choose more than one correct response (and some questions had >5 responses). This approach allowed us to consider a broader array of student-generated ideas as distractors at early stages of the development. However, in later versions, the less popular distractors were taken out and correct responses were limited to one single choice to allow for standard MC format grading.

III. ALPHA STUDY

The alpha version of the QMCA consisted of 28 questions and was given to students at CPP ($N = 17$) and CUB ($N = 44$) as a post-test towards the end of term. The mean scores for the two institutions were comparable but low (CPP = $41.2 \pm 3.5\%$, CUB = $43.1 \pm 3.3\%$). These overall scores were lower than the earlier (subjectively graded, open-ended) QMAT [21], which had provided partial credit on almost all free-response questions. The average scores for individual questions were strikingly similar for CPP and CU. The data in Fig. 2 show that the relative item difficulty of the new test questions was not highly sensitive to the two different student populations, which suggests QMCA items target some common and persisting student difficulties.

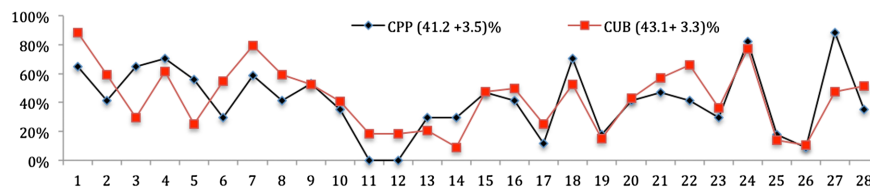


FIG. 2. Item difficulty distribution for questions from two implementations of the alpha version. In plotting this figure, we have assigned partial scores for multiple response option questions (e.g., 5, 10, 11, 12, 19, 26, and 28).

After obtaining results from this alpha study, we conducted 13 follow-up individual interviews with students (7 at CPP, 6 at CUB) to assess and improve question validity, with particular emphasis on formatting and wording issues. We discuss below this refinement process [28] for three representative examples, along with some rationale and motivations for developing additional items in Example 4. The first example illustrates how graphical representation of a given item, although correct, can affect students' responses. The second example discusses issues related to question format (traditional MC questions vs multiple correct). Example 3 highlights issues of content coverage limitations, relevant if the instrument is to be used for a broader range of course coverage across multiple institutions. Finally, the last one exemplifies new questions added to the instrument.

A. Example 1: Time oscillation of probability density

The open-ended QMAT question 6 (QMAT6, Fig. 3) was designed to probe student knowledge of the time dependence of the probability density for single energy and superposition energy states. Three distinct content learning objectives we identified in this question are as follows:

- (1) The spatial probability density of any single energy eigenstate is time independent.
- (2) Any linear combination state that is not an energy eigenstate will have a time-dependent spatial probability density.
- (3) The frequency of spatial probability density oscillations for any linear combination quantum state depends on the energy difference between terms.

The overall average for the correct response to the open-ended question was 37% [subquestion (a) 44%, question (b) 39%, question (c) 28%] at CPP ($N = 19$) and 54% [(a) 66%, (b) 59%, (c) 38%] at CUB ($N = 53$). On examining students' written reasoning and their further justifications of their choices on follow up interviews, we noticed some of the students with incorrect responses to parts (a) and (b) on the QMAT6 demonstrated reasonable proficiency with respect to learning objectives 1 and 2 above. However, two possible features might have confused them. First, although the question specifically instructed students to consider "... any quantum state - it does not necessarily need to be an energy eigenstate," student interview data suggest that the graphical representation of energy eigenstates appeared to cue students to only select from the three shown energy

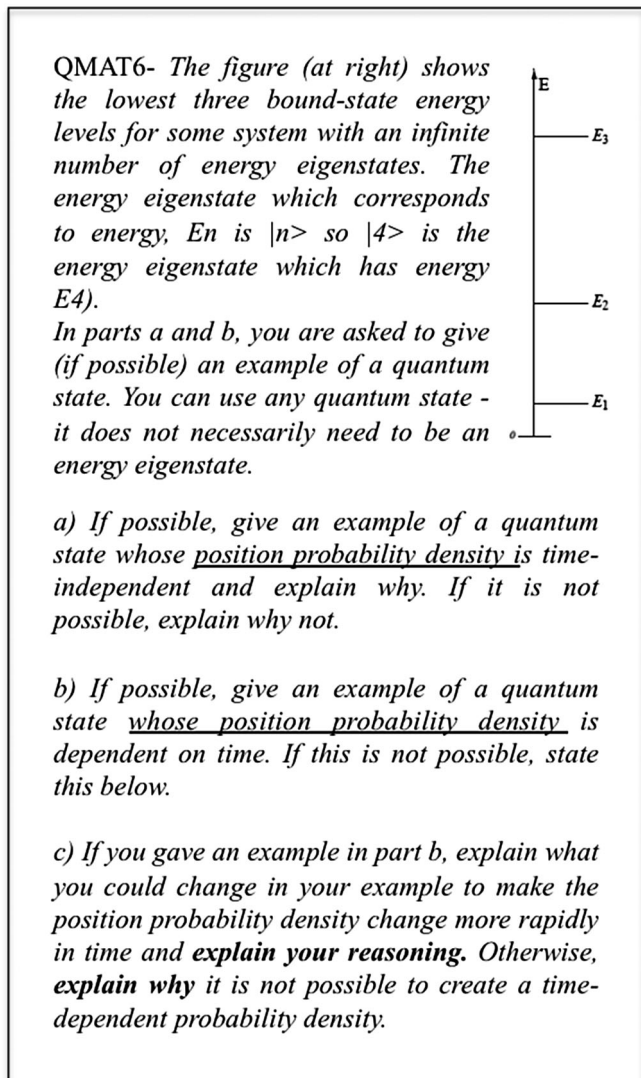


FIG. 3. Original open-ended QMAT question 6.

eigenstates on the graph, even though some were knowledgeable in another context regarding time dependence of superposition states for such a system.

For instance, a student’s written response to part (b) of QMAT6, included: “it is not possible of a quantum state whose position probability density is time-dependent, because when finding the probability, $P = \int dx \psi^2$, the square makes the time dependence to drop.” In the interview, he added “... but, I thought the question was asking about the three given energy states...”. Another example of students limiting their selections to merely the three shown energy eigenstates on the graph was suggested by the following student statement, “All states $|E_1\rangle$, $|E_2\rangle$, and $|E_3\rangle$ are time independent because they are stationary states.”

The intent of parts (a) and (b) in QMAT6 was to have students contrast time evolution of superposition states vs eigenstates. However, due to the open nature of the questions and a strong cueing on eigenstates from the figure, the instrument was not providing us with a complete picture

of student thinking about this topic. In answering part (c) of QMAT6, a common student response included statements like “... a higher energy state would have a higher time dependency,” suggesting that they may have confused the oscillation of the time-dependent probability density of the superposition states with the high frequency of stationary wave functions. Based on these student interview data, we eliminated the graph and in designing the distractors for the question related to part (c), we included both high and low energy eigenstates among the distractors.

In the alpha version of QMCA, we had only focused on learning objective 3, but in later versions we included additional questions to address learning objectives 1 and 2 above. The MC question and the first set of distractors targeted to address the learning objective 3 that was used in the alpha study was the following:

QMCA – $\alpha 17$. For a particle in a one-dimensional infinite square well, which state will have the fastest variation in time for the position probability density? (The state ψ_n corresponds to an energy of E_n)

- (a) ψ_1
- (b) ψ_4
- (c) $\frac{1}{\sqrt{2}}(\psi_2 + \psi_3)$
- (d) $\frac{1}{\sqrt{2}}(\psi_1 - \psi_3)$
- (e) All of these states have time-independent position probability densities.

Question 17 in the alpha study (**QMCA – $\alpha 17$**) was very challenging for a majority of the students (see Table II). The alpha study data and follow-up student interviews led to some modification in **QMCA – $\alpha 17$** , the process of which is discussed below. In the alpha study, only 12% of CPP and 25% of CUB students answered **QMCA – $\alpha 17$** correctly. To answer this question correctly, students need to recognize that the frequency of the oscillation is proportional to the energy difference between the superposition terms. In our sample population, all students had repeatedly seen a mathematical calculation of such an oscillation frequency for the time evolution of superposition states; however, very few were able to make the connection between the energy difference and oscillation frequency.

During student interviews on **QMCA – $\alpha 17$** , we observed that all students, independent of the correctness of their final answer, dismissed option (a) stating, e.g.,

“ ψ_1 is a ground state so has no oscillations.”

This was not the case for ψ_4 in option (b). In fact, a student who did not initially dismiss ψ_4 sketched an oscillation in the air for ψ_4 by gesturing his right index finger up and down and just drew a hump for ψ_1 . About one-fourth of students did not distinguish between the wave function frequency and the time oscillation frequency of the position probability density of a superposition state. Some of the students’ statements during the interview (and after selecting ψ_4 as their correct answer) were very similar to students’ statements above that were our motivation for including this distractor in the first place:

TABLE II. Distribution of students’ responses for a MC question related to open-ended QMAT Q6. Since students could choose more than one correct response, the percents do not add up to 100%. The bold numbers represent the correct choices.

QMCA – α17(%)		
CPP	a	6
	b	25
	c	32
	d	12
	e	19
CUB	a	0
	b	20
	c	27
	d	25
	e	23

“... the fastest variation in time will be given by the highest energy level.”

Thus, from student interview data, ψ1 was being dismissed for being the lowest energy and its particular shape, and not simply for being a stationary eigenstate.

Some students made use of their classical wave knowledge in answering this question. For example, a student who dismissed options (a) and (b) had difficulty choosing between options (c) and (d):

“My mind goes back to acoustic waves. When you have two different frequencies, from the difference in the frequencies you hear beats. So, I would say that the greatest variation would be for energies more far apart.”

Such an explanation (that led to the correct choice) was not mentioned in the course. Nevertheless, it reveals a range of different analogies and resources students use in answering quantum physics questions.

The popularity of the option $1/\sqrt{2}(\psi_2 + \psi_3)$ in the alpha study version (32% at CPP and 27% at CUB) suggests that it is a good distractor. Thus, we kept this option and modified the rest of the distractors for this question based on the interview data above in the next version to

- (a) ψ_5
- (b) $\frac{1}{\sqrt{2}}(\psi_2 + \psi_3)$
- (c) $\frac{1}{\sqrt{2}}(\psi_1 - \psi_4)$
- (d) $\sqrt{\frac{1}{3}}\psi_1 + \sqrt{\frac{4}{3}}\psi_3$
- (e) All of these states have time-independent position probability densities.

In the later versions of the test, we added two additional questions to address learning goals 1 and 2 on QMAT6 that will be discussed in Example 4.

B. Example 2: Energy measurement and Hamiltonian

In the alpha study, QMCA question 12 (QMCA – α12) was written as a “Choose all that apply” option (Fig. 4). This question probes known student difficulties with the role of the Hamiltonian. Some students believe that quantum energy

QMCA-α12. Which one of the following statements is always true for any quantum state $|\psi\rangle$?
Choose all that apply!

- a. $H|\psi\rangle$ gives the result of an energy measurement on the state $|\psi\rangle$.
- b. Immediately after energy measurement, the system is in the state, $H|\psi\rangle$.
- c. Acting on $|\psi\rangle$ with the Hamiltonian is the mathematical equivalent of energy measurement.
- d. There exists a definite energy E associated with the state $|\psi\rangle$ given by state $H|\psi\rangle = E|\psi\rangle$
- e. None of the above.

FIG. 4. QMCA – α12 and its distractors are motivated by QMAT4 and student responses to that question.

measurements mathematically correspond to the Hamiltonian operator acting on initial quantum states [29]. The first three distractors (a)–(c) for this item were adopted from three true or false statements in QMAT4. We added option (d) to include another known student difficulty regarding the deterministic nature of energy measurement in quantum systems.

Some of the most pressing issues students have with operators corresponding to physical measurement in quantum mechanics were included in the distractors of the alpha version of QMCA – α12 in the QMCA. The overall average correct score for this question was very low. The breakdown of students’ responses to QMCA – α12 in Table III provides more insight into students’ alternative ideas about energy measurement and the role of the Hamiltonian. Only 15% of CUB students selected the sole correct option (e). None of the CPP students chose the correct option, while almost 2/3 selected distractors (a),(c), and (d).

With the intent to provide course instructors information about students’ ideas and their specific difficulties, we

TABLE III. Percent of individual options students have selected on QMCA – α12. The total percent exceeds 100% because students were instructed to “choose all that apply.” The bold numbers represent the correct choices.

QMCA – α12(%)		
CPP	a	59
	b	35
	c	59
	d	65
	e	0
CUB	a	30
	b	17
	c	23
	d	19
	e	15

realized this MC format with a multiple-selection option was not providing a simple yet comprehensive picture of the students' different ideas by combining several related but distinct difficulties. During the interviews we noticed most students strongly struggled with analyzing all of these statements concurrently. Furthermore, during the interviews we noticed student incorrect responses, at least in part, arose from merely considering the energy eigenstates as a general representative of "any quantum state" in the question setup.

To address these issues, we revised the questions accordingly. First, to provide a better insight into students' individual difficulties with this topic, we altered the question format by converting the first three distractors into a set of true or false questions in the beta version (QMCA – $\beta 11$ – $\beta 13$). Next, we tried to isolate the intended concept in this question (role of the Hamiltonian) and eliminate the impact of a student merely overlooking superposition states; we revised the question context from "any quantum state" to have students explicitly consider a superposition state such as $\psi(x) = C_1\psi_1(x) + C_2\psi_2(x)$. Finally, a new question (QMCA – $\beta 14$) was designed to probe student ideas about the effect of a phase change in the probability of measuring an energy corresponding to one of the eigenstates given in the superposition state. This new question was directly motivated by student interviews. We noticed that many of our students did not distinguish between instances where phase factors do modify measurement outcomes from those that do not.

C. Example 3: Operators and Hamiltonian

As another example, we consider QMAT question twelve (QMAT12). The wording of the question was largely unaltered in the MC version (QMCA – $\alpha 18$ in Fig. 5), which asks whether a system in an eigenstate of an arbitrary operator will remain in that eigenstate until disturbed by measurement. The average score for the open-ended QMAT12 (which was heavily impacted by the scoring of student reasoning) was very low (CPP = 23%, $N = 19$; CUB = 25%, $N = 53$). To learn more about student ideas about this topic, we used students' written reasoning on QMAT12 to construct QMCA question 19 (QMCA – $\alpha 19$ in Fig. 5.)

While many students correctly identified the false nature of the statement in QMCA – $\alpha 18$, (CPP = 71%, CUB = 52%), many were not able to detect all of the possible correct reasoning [(d),(e),(f)] in QMCA – $\alpha 19$. (CPP = 18%, CUB = 15%). Option (f) for QMCA – $\alpha 19$ was very unpopular (only one student in both data sets selected it), despite being a correct answer choice. In student interviews there were frequent comments about the novelty of a time-dependent Hamiltonian and the lack of intuition about the behavior of such a system. Some instructors also stated that they had not discussed systems with time-dependent Hamiltonians in much depth. Thus, option (f)

QMCA- $\alpha 18$. *Is the following statement true or false for all operators Q ? A system which is in an eigenstate of Q will stay in that state until disturbed by measurement.*

- a. True
- b. False

QMCA- $\alpha 19$. *What is the reasoning for your answer to the previous question? Choose all that apply.*

- a. The system remains in an eigenstate of operator Q until measurement of a non-commuting operator.
- b. The expectation value of Q is time-independent.
- c. The system will evolve and the exact state cannot be known.
- d. The Hamiltonian may not commute with Q .
- e. The operator Q may have a time dependence.
- f. The Hamiltonian may have a time dependence.

FIG. 5. The alpha version of two MC questions corresponding to open-ended QMAT Q12.

was removed in later versions of the test. Furthermore, options (d) and (e) showed little to no discriminatory power (0% of students in both samples selected all correct options and no incorrect options). Options (d) and (e) were combined in later versions of the instrument in order to limit students to only one correct option.

The examples discussed above are representative of the conversion and development process to a MC format and demonstrate some of our motivations for developing new items. Each example shows a unique challenge for developing effective distractors that could better highlight student ideas. They also support the notion that well-designed multiple-choice items could provide valuable feedback about student reasoning.

D. Example 4: Time evolution of measurement

Measurement is one of the most counterintuitive aspects of quantum mechanics, as it greatly deviates from deterministic measurement results in classical mechanics. As a result, many students have difficulties predicting the outcome of quantum mechanical measurement and its time evolution. Quantum mechanics postulates that the measurement of a physical observable collapses the wave function of the quantum system into an eigenstate of the corresponding operator. Further, the time evolution of the states is governed by the time-dependent Schrodinger equation (TDSE). The eigenstates of physical observables can in general be written as a linear superposition of the energy eigenstates of the system. This can lead to different outcomes for time evolution of energy measurement

compared to other physical observables, such as position. The probability of measuring a particular value of energy is time independent, yet the probability of measuring another physical observable whose operator does not commute with the Hamiltonian is dependent on time.

For example, in an energy measurement of a non-degenerate quantum particle, the wave function of the system will collapse into an energy eigenfunction for which the only time-dependent factor (given a time-independent Hamiltonian) is an overall phase that will vanish upon squaring the wave function. This results in stationary states for which the spatial probability density does not change with time. On the other hand, a measurement of position would collapse the wave function of the particle to a position eigenfunction (e.g., a delta function in the position space at the instant of measurement). However, a position eigenstate is a linear superposition of the energy eigenstates and the different energy eigenstates in the linear superposition will evolve with different time-dependent phase factors. Therefore, the probability density after position measurement will change with time. Some students show difficulties recognizing this difference between the time evolution of energy eigenstates compared to the eigenstates of other physical observables [30].

Given the unique aspects of time evolution of energy eigenstates, we designed a pair of additional questions to explicitly probe student ideas about the time evolution in the familiar system of a particle in a one-dimensional infinite square well. The first question (QMCA – $\alpha 15$ in Fig. 6) asks students about the time evolution of the energy expectation value with a focus on energy eigenstates, and QMCA – $\alpha 16$ asks a similar question about the time evolution of the position expectation value.

The alpha study and item analysis of QMCA – $\alpha 15$ and $\alpha 16$ indicated that both questions were equally challenging for both CPP and CUB students. The average percent of correct responses for QMCA – $\alpha 15$ and $\alpha 16$ at CPP was 47% and 41%; at CUB they were 48% and 50%, respectively. This suggests that students equally struggle with analyzing the time evolution of expectation values of energy and position, particularly when it comes to measurements involving superposition states. Although options (a) and (e) for both questions in Table IV were not popular with the sample students in the alpha study, we see a larger percentage of students ($\sim 10\%$) select these options in our larger beta data set (see below).

The most popular incorrect distractor for QMCA – $\alpha 15$ was option (b). Approximately, 35% of CPP students and 45% of CUB students incorrectly answered that the energy expectation value is time independent only for energy eigenstates. For QMCA – $\alpha 16$, the most popular incorrect idea was split between choices (c) and (d). About 60% of the CPP and 50% of CU students [sum of choices (c) and (d)] missed the relationship between the time evolution of the position expectation value and the details of the state

QMCA- $\alpha 15$. Which of the following statements are always true for a particle in a one-dimensional infinite square well?

- The energy expectation value is time-dependent only for energy eigenstates.
- The energy expectation value is time-independent only for energy eigenstates.
- The energy expectation value is always time-independent.
- The energy expectation value is always time-dependent.
- More than one of the above statements is always true.

QMCA- $\alpha 16$. Which of the following statements are always true for a particle in a one-dimensional infinite square well?

- The position expectation value is time-dependent for energy eigenstates.
- The position expectation value is time-independent for energy eigenstates.
- The position expectation value is always time-independent.
- The position expectation value is always time-dependent.
- More than one of the above statements is always true.

FIG. 6. Additional questions on time evolution of energy and position expectation values in the QMCA alpha version.

of the system. The feedback from each item in our alpha study provided valuable insights for modifying the distractors on the next iteration of the test. Later interviews suggested further that distractor (e) was considered logically inconsistent by students, and was correspondingly modified.

TABLE IV. Student responses to Q15 and Q16 on the alpha study. The bold numbers represent the correct choices.

		QMCA – $\alpha 15$ (%)	QMCA – $\alpha 16$ (%)
CPP	a	6	0
	b	35	41
	c	47	24
	d	12	35
	e	0	0
CUB	a	2	2
	b	45	50
	c	48	34
	d	2	14
	e	0	0

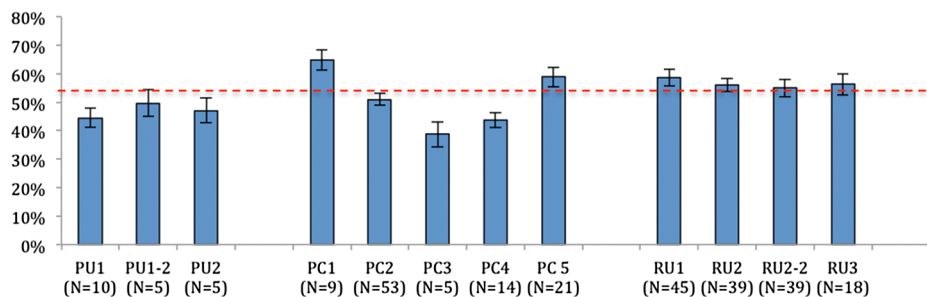


FIG. 7. Students’ mean scores on the beta version of QMCA, organized based on the type of university. (Repeated name codes mean consecutive semesters at the same institution) N is the number of students in the class. The red dashed line represents the weighted average for all students across different institute, which was 54%.

IV. BETA STUDY

Using the data from the alpha study, follow-up student interviews, and expert feedback, we revised several questions and their corresponding distractors to further improve their reliability and validity. The changes include moving away from multiple correct answer format to a single correct answer, eliminating several questions, adding three new questions, and some wording and formatting changes on both questions and distractors.

A. Data collection and results

During the 2013–14 academic year, the beta version of the QMCA was administered to a total of 263 students from 12 classes at 10 institutions nationwide. The institute compositions included three primarily undergraduate (PU), four private colleges (PC), and three large research universities (RU), and classes ranged in size from 5 to over 50. All students took the QMCA as a post-test at the end of their first

semester or quarter of the upper-division quantum mechanics courses. The total average scores of the students in each institute are displayed in Fig. 7. The average scores range from a minimum of 39% ($N = 5$) to a high of 65% ($N = 9$).

1. Results for five concept framework or subtopics

We analyzed student performance on each subset of questions corresponding to one of the five concept frameworks defined in the introduction. As discussed earlier, the face validity of these classifications was first established during the QMAT development and through working groups of 18 CUB faculty. We recognize that these five categories are highly integrated concepts and only constitute a portion of the broader set of conceptual categories involved in introductory quantum mechanics. Nevertheless, our faculty working groups and ongoing interactions with faculty who administered the QMCA suggest this can be a productive way of looking at student performances on subsets of questions related to a particular topic.

The overall patterns of students’ scores on the five topics were strikingly similar among all 10 institutes. Students’ highest scores were on the set of questions focusing on concepts of probability and probability density, whereas the students’ lowest scores were on the subset of the questions corresponding to time evolution and wave functions. Figure 8 shows the common pattern in data with the spread on each concept category for different types of institutions. While student scores in all categories had a small variation, student scores in questions related to quantum mechanical measurement had the largest deviation from the mean.

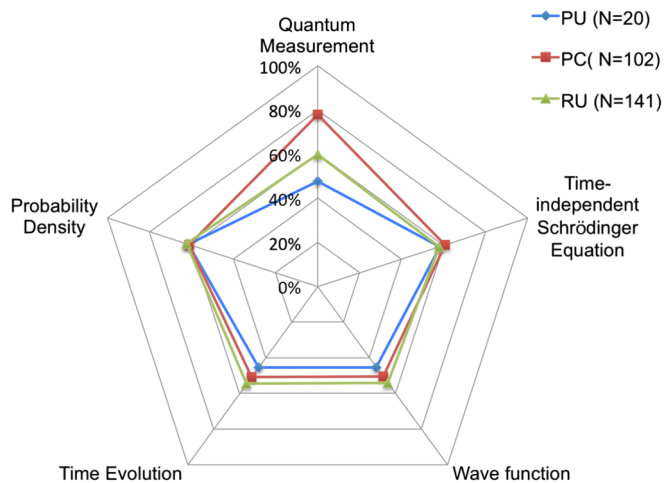


FIG. 8. The student scores on five concept categories of QMCA show larger variation in students’ performances on questions related to quantum measurement. However, students’ scores on all the other four-concept categories are practically the same. The standard error of the mean for concept category for any institutional type (i.e., each data point in the figure) fluctuated between 2.03% and 4.63%. (primarily undergraduate = PU, private colleges = PC, and research universities = RU.)

V. TEST VALIDITY AND RELIABILITY

Validity is the degree to which a test measures the learning outcomes it purports to measure and reliability is defined as the degree to which a test consistently measures a learning outcome. A high reliability denotes the test would yield the same result when run again, and not produce a more or less random outcome. In order to evaluate the validity of our instrument, we conducted several statistical tests of reliability and validity for the alpha and beta versions. Our evidence to support the QMCA item validity can be organized into

three categories: (1) Evidence based on test content obtained from faculty experts, (2) evidence based on responses through student interviews, and (3) evidence based on common statistical analysis for test development.

A. Expert validation

We discuss expert validation in the two contexts of face and content validity. Face validity is the extent to which a test is subjectively viewed as covering the concept it purports to measure. In other words, a test can be said to have face validity if it “looks like” it is going to measure what it is supposed to measure. Establishing the face validity increases its adoptability by instructors and its credibility to students. Face validity of the QMCA has been investigated through comments from over sixteen faculty members who regularly teach upper division undergraduate quantum courses. This includes four faculty members at CPP and CU, and the faculty who administered the QMCA in the other eight universities. This was in addition to the instructors who have been consulted regarding the learning goals of upper level quantum mechanics courses in the process of developing the open-ended QMAT at CU. [31]

Content validity requires the use of recognized subject matter experts to evaluate whether test items assess defined content and whether test items reflect the knowledge actually required for a given topic area. For content validation, we examined how well the test items cover the content domain it intended to test and how well the distractors represent specific student ideas. The content validation took place in three different stages; during the construction of the MC format, and after both alpha and beta studies. We pursued expert content validation to refine both questions and MC distractors in the QMCA. For example, in working to establish clear language to expose student common ideas and enhance the test validity, we consulted with a CPP faculty member who has written several textbooks, including one on quantum theory. He provided constructive feedback and indicated that the QMCA includes a comprehensive set of questions that is capable of probing a variety of conceptual hurdles and difficulties commonly encountered by beginning students of quantum mechanics. Additional feedback was obtained through interviews and discussions with course instructors at CPP and CUB.

Furthermore, individual course instructors from twelve institutions reviewed QMCA items before administering them to their class and commented on its appropriateness and relevance for their upper-division undergraduate quantum mechanics courses. They also provided feedback on perceived ambiguities in item wording and shared their observations of student performances and feedback on the test. We have paid particular attention to expert critics and have addressed many concerns they expressed with the QMCA items. Although some research has reported a large variation in faculty views on many topics in quantum mechanics [32], the overall feedback from the faculty

reviewing the test was consistent. They unanimously valued the test and expressed interest in administering it the next time they teach the course.

B. Student validation

Additional face validity has been established during student interviews. We recruited students with a wide range of abilities to participate in a total of 25 individual student interviews and a 2 h focus group with ten students. Individual interviews lasted 45 to 90 min, and students were compensated with a gift card for their time. The interviews took place several weeks after the students had completed the QMCA in class. Six of these interviews took place after administering the open ended QMAT, thirteen interviews were conducted on the alpha and six additional on the beta versions of the QMCA. In the interviews, students were asked to explain their understanding of each item before discussing their reasoning for their choices of correct answers. These interviews helped us to improve the validity of the QMCA by refining the wording of questions, listening to students reasoning, and modifying some items that were not conveying the intended purpose clearly. The alignment of the items and distractors with students’ reasoning supported our claim of construct validity of the QMCA. Some specific results of interviews that led to modifying questions or distractors have been discussed above in three example questions in Sec. III.

C. Statistical analysis

The validity of an assessment is the degree to which it measures what it is supposed to measure. This is not the same as reliability, which is the extent to which a measurement gives results that are consistent. We have computed item and test statistics that are commonly used as measures of validity and reliability of multiple-choice tests [33]. This included item difficulty index, item discrimination index, Kuder-Richardson reliability index, and Ferguson’s delta.

1. Item difficulty index

The item difficulty gives a general sense of the difficulty of each item and is calculated by taking the percent average of correct responses to a question. The range for the item difficulty index is [0,1]. If the value is 0, then no one can answer the question correctly; on the other hand, if the value is 1, then everyone can correctly answer this question. Under most circumstances, such extremes should be avoided in a test. The QMCA contains questions with a wide range of difficulties, from 0.27 for the most difficult question (QMCA – β 21) to 0.87 for the easiest question (QMCA – β 1), with most questions falling in a range between 0.40 and 0.80 (see Fig. 9). With an average difficulty of 0.54, the QMCA is a challenging exam, but the bulk of our collected comments suggest our sampled students find it valuable, relevant, and fair.

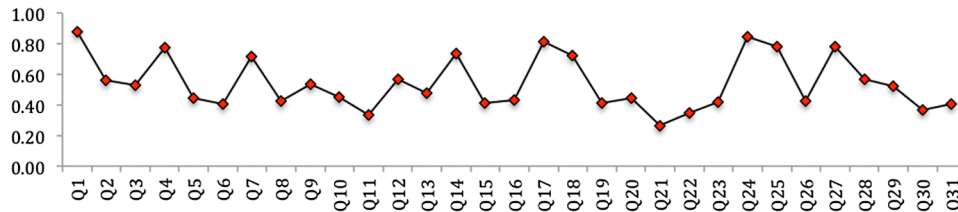


FIG. 9. Item difficulty distribution for QMCA – β , ($N = 263$). The question numbers do not match to those in the alpha version.

2. Item discrimination

Item discrimination is a measure of how well an item discriminates between students who score high in the test as a whole and those who do not do well overall. It can be quantitatively estimated by a point biserial correlation coefficient, which is essentially the Pearson correlation coefficient for each test item with the total test score. A low point biserial coefficient indicates that student understanding of the concept measured by an item is not correlated with understanding of other concepts on the test.

The range for the item discrimination index D is $[-1, +1]$, where $+1$ is the best value and -1 is the worst value. The discrimination index D of $+1$ for an item means all students in the high group get the item correct and all students in the low group get it wrong. For the discrimination index D of -1 the situation is reversed and it means everyone in the low group answers the item correctly, and everyone in the high group gets it wrong. These extreme cases are unlikely, but it is important to eliminate any items with negative discrimination indices. An item is typically considered to provide good discrimination if $D > 0.3$. [34]

Items with a discrimination index lower than 0.3, but greater than 0 are not necessarily bad, but a majority of the items in a test should have relatively high discrimination index values to ensure that the test is capable of distinguishing between strong and weak mastery of the material. The items on the QMCA have point biserial coefficients ranging from 0.22 to 0.52, with an average of 0.35, suggesting good coherence and no statistically problematic questions. Another common measure of item discrimination is $D = \text{“Discrimination index”}$, which is the number of people in the upper quartile of the test overall who answered the given item correctly, minus the number from the lower quartile, divided by the number of people in the upper quartile. The average value is $D = 0.42$ for our data sample, with a range from 0.23 to 0.65, again suggesting good item discrimination with no apparent problematic questions.

3. Kuder-Richardson reliability index

The Kuder-Richardson reliability index, or Cronbach alpha, is a measure of the overall correlation between items, a statistical measure of a test’s internal consistency. For our combined data sets, we obtained $\alpha = 0.76$. Traditionally,

values between 0.7 and 0.8 are considered acceptable for surveys of this type. Chronbach’s α (computed as a single number for an entire instrument) implicitly assumes a test does not assess multiple dimensions. We did not design the QMCA to measure a single construct (and indeed have presented results above for five *a priori* distinct broad topical areas). This suggests that our calculated α values are likely a conservative underestimate of the internal consistency of the instrument [35].

4. Ferguson’s delta

Figure 10 shows a histogram of the frequency of total scores for all students. The distribution is normal, the Anderson-Darling test for normality is passed for each of the 12 classes separately, as well as for the full combined data set. Ferguson’s delta (δ), or the “coefficient of test discrimination” [36] measures the discrimination power of a test instrument by investigating how broadly the total scores of a sample are distributed over the possible range [37]. It can range from 0 to 1, with values above 0.9 generally considered strong discrimination—we obtain a high value of 0.97 for our combined data set ($N = 263$).

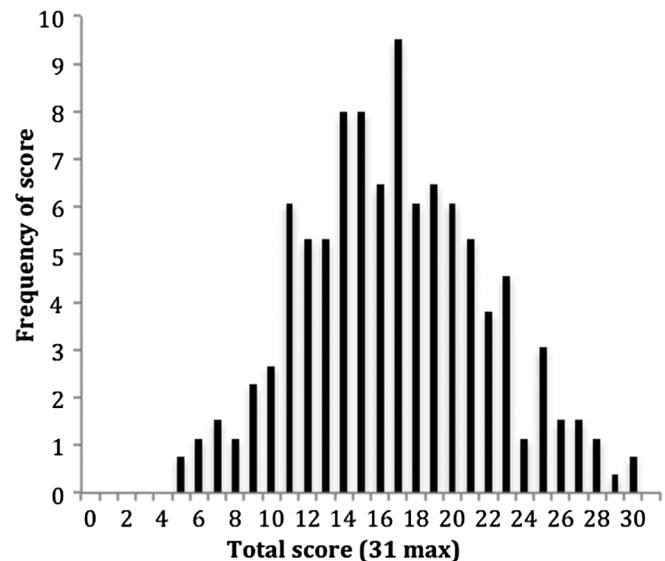


FIG. 10. The QMCA – β score distribution histogram ($N = 263$) shows an overall normal distribution.

TABLE V. A summary of statistical test results for both alpha and beta versions of QMCA.

	Possible values	Desired values	QMCA alpha	QMCA beta
Number of students	61	263
Number of questions	28	31
Standard deviation	15%	16%
Standard error	2.7%	1.0%
Item difficulty index	[0, 1]	>0.3	0.42	0.54
Item discrimination index - 25/25	[-1, 1]	>0.3	0.39	0.42
Point biserial coefficient	[-1, 1]	>0.2	0.31	0.35
Kuder-Richardson test reliability	[0, 1]	>0.7	0.71	0.76
Ferguson's delta	[0, 1]	>0.9	0.97	0.97

Table V shows a summary of the statistical test values for both the alpha and beta versions. Comparison of the results for the two studies shows that by follow-up interviews and expert feedback we were able to further refine the QMCA items and optimize the values of common statistical test scores in our beta version.

VI. SUMMARY AND DISCUSSION

We took a preliminary version of an open-ended assessment tool that was developed by consulting several faculty members who commonly teach upper division undergraduate quantum mechanics courses at the University of Colorado Boulder, and crafted a MC concept assessment instrument. Student common incorrect responses to the open-ended QMAT were the primary source for constructing the initial distractors. The objective scoring associated with MC format items in the QMCA test frees us from problems with a complicated rubric and scorer inconsistency associated with the open-ended QMAT. After an alpha study and follow up interviews, we improved some of the test reliability and validity scores in our beta version, which then was given to over 263 students from ten different populations of students nationwide. The statistical analysis as well as students and faculty interview results suggest that this instrument is a reliable and valid assessment tool for upper-division undergraduates. The QMCA now consists of a total of 31 conceptually focused items. Stating items in a simple and sensible way that was easy for students to understand while providing the details an expert would require, without making them too long, was challenging. Particularly, for abstract topics of QM, the absence of real world experience limits student preconceptions that can be easily classified [38]. In general, most items replicated as much as possible common student reasoning that repeatedly appeared in our interviews and existing literature.

Like many other instruments, “errors” on the Inventory can be more informative than “correct” choices. We observed some student difficulties across the QMCA that are consistent with the literature [39]. Students sometimes overlook the unique aspect of measurement in quantum mechanics. For example, sequential measurements of

physical observables (with noncommuting operators) do not retain all original information encoded in the earlier measurements and corresponding starting states. [40] We also noted that students’ responses suggest some students answer as though all quantum states (including superposition states) have a definite energy; also, they commonly respond as though time dependence only requires “tacking on” a single phase term $e^{-iEt/\hbar}$ to any quantum state regardless of whether or not the system is in an eigenstate or composed of a linear combination of eigenstates. Our results also suggest that the concepts of time evolution and wave function are two particularly challenging topics for students at this level. For example, students struggle to distinguish time oscillation of the probability density of a wave function from the oscillations of the phase of the wave function itself. Furthermore, students frequently treated eigenstates as a general representation of an arbitrary quantum state and failed to consider linear combinations of such states as a general representative.

A. Uses of QMCA

As instructors and researchers, we seek reliable tools to study student learning and understanding of physics. Such tools allow us to study the efficacy of different curricula or classroom activities, and help to identify common student difficulties. By making aspects of student thinking processes visible, a well-developed instrument can guide efforts to systematically improve instruction. Existing assessment tools have been instrumental in supporting and evaluating transformed pedagogies.

The QMCA could be used for both instructional and research purposes to measure the effectiveness of different curricula or teaching strategies at improving students’ conceptual understanding of quantum mechanics. For example, instructors can use the QMCA to learn about student ideas, diagnose topics on which specific group of students struggle, or identify common student difficulties. For a more fine-grained analysis of what students are learning, one can separately study performance on five main topics. This can be especially useful if one has implemented a new pedagogical technique targeted to a single topic such as measurement.

Because of increased complexity of the physics content and formal language used in upper division assessment instruments, we suggest using the QMCA only as a post-test. However, it could potentially be used as a pretest with incoming graduate students to help determine if student understanding of introductory quantum physics is sufficient for a more advanced course.

B. Future work

We naturally find variation in student average percent scores across different courses and institutions on the QMCA post-test. But, the range is not as broad as we might have expected given the significantly different student populations, instructors' teaching philosophy, curricular materials, and pedagogical approaches, which all could impact student performance on QMCA. This suggests that the QMCA might be pinpointing some of the most common difficulties among upper division physics students.

In future work, we intend to study the effect of various factors impacting student performances in quantum courses to identify best practices. These factors include, but are not limited to, different teaching strategies, research-based instructional materials, and/or reformed curricula. For

example, we are currently investigating student learning of quantum mechanics in two different contexts. In one (perhaps more traditional) approach, postulates of quantum mechanics are introduced in the context of spatial wave function of particles in simple potential wells; the second approach uses a Stern-Gerlach experimental context, with discrete bases of spin. The QMCA is one stepping stone for our long-term goal of assessing outcomes of these two (and other) pedagogical approaches to teaching quantum mechanics. Further study may include the development of assessment questions for similar learning objectives in the context of discrete bases of spin one-half systems.

ACKNOWLEDGMENTS

Many thanks to all the faculty and students who contributed to the development of this assessment at Cal Poly Pomona and CU Boulder. We also acknowledge the work of undergraduate researchers John Miller and Daniel Rehn. This work was in part supported by Cal Poly Pomona Research, Scholarship and Creative Activity (RSCA), the Colorado Science Education Initiative, and NSF-CCLI Grant No. DUE-1023028.

-
- [1] E. Redish and R. Steinberg, Teaching physics: Figuring out what works, *Phys. Today* **52**, 24 (1999); L. C. McDermott, Research on conceptual understanding in mechanics, *Phys. Today* **37**, 24 (1984).
 - [2] D. Hestenes, M. Wells, and G. Swackhamer, Force Concept Inventory, *Phys. Teach.* **30**, 141 (1992).
 - [3] R. Thornton and D. Sokoloff, Assessing student learning of Newton's laws: The Force and Motion Conceptual Evaluation, *Am. J. Phys.* **66**, 338 (1998).
 - [4] D. Maloney, T. O'Kuma, C. Hieggelke, and A. Heyvelen, Surveying students' conceptual knowledge of electricity and magnetism, *Am. J. Phys.* **69**, S12 (2001).
 - [5] L. Ding, R. Chabay, B. Sherwood, and R. Beichner, Evaluating an electricity and magnetism assessment tool: Brief electricity and magnetism assessment, *Phys. Rev. ST Phys. Educ. Res.* **2**, 010105 (2006).
 - [6] R. Hake, Interactive-engagement versus traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses, *Am. J. Phys.* **66**, 64 (1998).
 - [7] E. Mazur, *Peer Instruction: A User's Manual* (Prentice Hall, Saddle River, NJ, 1997).
 - [8] C. Singh, Student understanding of quantum mechanics, *Am. J. Phys.* **69**, 885 (2001); C. Singh, Student understanding of quantum mechanics at the beginning of graduate instruction, *Am. J. Phys.* **76**, 277 (2008).
 - [9] H. Sadaghiani, Ph.D. thesis, The Ohio State University, 2005.
 - [10] A. Crouse, Ph.D. thesis, University of Washington, 2007.
 - [11] H. Sadaghiani and L. Bao, Student Difficulties in Understanding Probability in Quantum Mechanics, *Physics Education Research Conference Proceedings* (AIP, Syracuse, NY, 2006), pp. 61–64.
 - [12] D. Styer, Common misconceptions regarding quantum mechanics, *Am. J. Phys.* **64**, 31 (1996).
 - [13] 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).
 - [14] J. Falk, Master's thesis, Uppsala University, 2004.
 - [15] G. Zhu and C. Singh, Improving students' understanding of quantum measurement. I. Investigation of difficulties *Phys. Rev. ST Phys. Educ. Res.* **8**, 010117 (2012).
 - [16] S. McKagan, K. Perkins, and C. E. Wieman, Design and validation of the quantum mechanics conceptual survey, *Phys. Rev. ST Phys. Educ. Res.* **6**, 020121 (2010).
 - [17] G. Zhu and C. Singh, Surveying students' understanding of quantum mechanics in one spatial dimension, *Am. J. Phys.* **80**, 252 (2012).
 - [18] S. Goldhaber, S. Pollock, M. Dubson, P. Beale, and K. Perkins, *Physics Education Research Conference Proceedings* (AIP, Ann Arbor, Michigan, 2009), p. 145–148.
 - [19] M. Scott, T. Stelzer, and G. Gladding, Evaluating multiple-choice exams in large introductory physics courses, *Phys. Rev. ST Phys. Educ. Res.* **2**, 020102 (2006); L. Crocker and J. Algina, *Introduction to Classical and Modern Test Theory* (Harcourt Brace, Orlando, FL, 1986).

- [20] D. Hestenes, M. Wells, and G. Swackhamer, Force Concept Inventory, *Phys. Teach.* **30**, 141 (1992).
- [21] P. Airasian and M. Russell, *Classroom Assessment: Concepts and Applications*, 6th ed., (McGraw-Hill, New York, 2008).
- [22] M. Dubson, S. Goldhaber, S. Pollock, and K. Perkins, *Physics Education Research Conference Proceedings* (AIP Press, Melville, NY, 2009).
- [23] Examples include references [8–15] above.
- [24] E. Gire and C. Manogue, *Physics Education Research Conference Proceedings* (AIP Press, Melville, NY, 2008), pp. 115–118.
- [25] See Ref. [4] above.
- [26] D. Thissen, Lynne Steinberg, and Anne R. Fitzpatrick, Multiple-Choice models: the distractors are also part of the item, *J. Educ. Measure.* **26**, 161 (1989).
- [27] Examples include Refs. [9–12] and [15–17] above; and B. Ambrose, Ph.D. thesis, University of Washington, 1999.
- [28] H. Sadaghiani, J. Miller, S.J. Pollock, and D. Rehn, Converting an open-ended assessment for upper-division quantum physics to multiple-choice format, *Physics Education Research Conference Proceedings* (AIP, Portland OR, 2013), p. 319.
- [29] C. Singh, M. Belloni, and W. Chrustian, Improving students’ understanding of quantum mechanics, *Phys. Today* **59**, 43 (2006); see also Refs. [8, 12], and [24].
- [30] G. Zhu and C. Singh, “Improving students’ understanding of quantum measurement. I. Investigation of difficulties,” *Phys. Rev. ST Phys. Educ. Res.* **8**, 010117 (2012).
- [31] Please see Ref. [18].
- [32] Please see Ref. [16].
- [33] For a more detailed discussion of the meaning and significance of these measures and their use in research-based tests in physics and other disciplines see Refs. [5]; and W. Adams and C. Wieman, Development and validation of instruments to measure learning of expert-like thinking, *Int. J. Sci. Educ.* **33**, 1289 (2011).
- [34] R. Doran, *Basic Measurement and Evaluation of Science Instruction* (NSTA, Washington, DC, 1980), p. 99.
- [35] R. Brennan and D. Prediger, Coefficient kappa: some uses, misuses, and alternatives, *Educ. Psychol. Meas.* **41**, 687 (1981); G. Kuder and M. Richardson, The theory of the estimation of psychometrika test reliability, *Psychometrika* **2**, 151 (1937).
- [36] G. Goldstein and M. Hersen, *Handbook of Psychological Assessment* (Elsevier Science, Kidlington, Oxford, UK, 2000), Vol. 36.
- [37] Please see Ref. [5] above.
- [38] S. McKagan, K. Perkins, and C. E. Wieman, Developing and researching PhET simulations for teaching quantum mechanics, *Am. J. Phys.* **76**, 406 (2008); also see Ref. [16] above.
- [39] For example, those in Refs. [2,3,14,15].
- [40] Please see Refs. [8] and [9].