Instructors' Views on a Flexible Assessment Design

Jesse Kruse,[1] Molly Griston,[1] and Bethany R. Wilcox[1]

[1]*Department of Physics, University of Colorado, 390 UCB, Boulder, CO 80309*

Flexible assessments are distinct from standard research-based assessments like the Force Concept Inventory in that the items and the item ordering may differ between administrations. Our group has been developing a flexible assessment for upper-level undergraduate quantum mechanics called the Quantum Physics Assessment (QuPA) to address the lack of consensus on what can/should be taught in an undergraduate quantum course. This assessment will exist on a web platform where instructors can select topics they want to assess and other testing parameters. In Fall 2023, we conducted 34 interviews with faculty from institutions across the U.S. asking them to interact with our flexible assessment instructor interface, as well as about their experience with assessments, what they would improve, and what feedback would be useful. This paper will present the theory and motivation for flexible assessment of quantum mechanics proficiency, discuss our interview protocol, and summarize the general views of the instructors.

2024 PERC Proceedings edited by Ryan, Pawl, and Zwolak; Peer-reviewed, doi.org/10.1119/perc.2024.pr.Kruse

I. INTRODUCTION & BACKGROUND

Anecdotally, it is generally accepted that there is little consensus on the scope of what should be taught in an upper-level, undergraduate quantum mechanics (QM) course and what methods should be used to maximize student learning [1]. This lack of consensus can be seen in the many distinct research-based assessments that have been developed for QM. Currently, there are six upper-level, QM assessments hosted on PhysPort, an online repository for physics education research (PER) teaching resources [2]. These assessments have been developed to varying degrees over the past 22 years, but they were all validated using classical test theory which limits them to being used as standalone, static tests. The accuracy and reliability of these instruments cannot be guaranteed upon item reordering and/or removal, which poses a significant barrier to accommodating the variety of topics covered in QM. Under this paradigm one would need to test every possible combination of items for reliability and validity, and even then, one would not be able to confidently compare results between tests [3].

To address this limitation, we intend to use item response theory (IRT) to validate a new upper-division (i.e., excluding modern and graduate level) QM assessment called the Quantum Physics Assessment or QuPA. IRT parameterizes and calibrates an individual item-person interaction, so every item has a difficulty parameter and every person has a performance parameter after completing the assessment. One benefit of IRT-validated tests is that items can be reordered or removed, and one can still obtain accurate performance parameter estimates. IRT is also the basis of computer adaptive testing (CAT) where items are pulled from a bank as a student is taking an assessment to ensure the closest match between item difficulty and person performance to most precisely pinpoint each participant's ability parameter [4]. A flexible assessment like the QuPA also has the feature that questions appearing on the instrument are variable but is distinct from CAT in that once the target content has been chosen by an instructor, the specific items used are the same for all students.

While no assessments currently exist in PER that utilize this flexible testing model (though one [5] is currently under construction), this model has been used in other discipline-based education research fields [6]. This flexible testing model is less computationally intensive and easier to implement than CAT, but introduces its own challenges as well. One of the first challenges is how instructors go about selecting the topics and scope of their flexible assessment. Specifically, we must design an interface from which instructors select the content or goals relevant to their specific course that can then be used to create an instance of the QuPA from our test bank. Additional challenges include the formation of a comprehensive test bank spanning all upper-level QM content, development of an algorithm to automate test creation based on instructor content selections, and pilot testing of the test bank to ensure all items are calibrated. These challenges are being addressed through ongoing work, but are beyond the scope of this manuscript.

In this paper, we focus on a crucial early step needed to facilitate the ongoing development and testing of QuPA: the development of the instructor interface that will serve as instructors' primary point of contact with the assessment. Specifically, we focus on ensuring that faculty can understand and navigate the interface effectively. We also want to both ensure that the interface, and eventual assessment, will meet the needs of instructors and address any concerns they might have with this approach to assessment design.

Here, we discuss our design for a flexible QM assessment as the context for our instructor interface (Sec. II), the interview protocol used for soliciting instructor feedback along with the respondent pool to our interviews (Sec. III), the general results along with representative quotes (Sec. IV), and the conclusions we can draw from this work both for the specific development of the QuPA interface and for future assessment design (Sec. V).

II. ASSESSMENT INTERFACE DESIGN

For the design of our instructor interface, we envisioned three ways in which instructors could choose to create an assessment that suits their needs: they could use one that already exists like the Quantum Mechanics Concept Assessment (QMCA) [7], build an assessment by selecting pre-built modules designed to target the content of sections of common textbooks, or create their own assessment by selecting which assessment objectives (AOs) they want the test to target. The first option was included because existing assessments are carefully crafted, well validated, and have been used extensively[7–11]. Providing the option to use an existing assessment allowed instructors who like them to benefit from all the features of the QuPA system (e.g., automated data collection and analysis, centralized data aggregation, etc.) and provided a simple option for instructors not interested in customizable assessment.

For the instructors who wanted to create an assessment based on a standard QM textbook, we included the option for only two upper-level QM textbooks commonly used within the United States: *Introduction to Quantum Mechanics* by D. J. Griffiths and D. F. Schroeter [12] and *Quantum Mechanics: A Paradigms Approach* by D. H. McIntyre [13]. These two books, while both quite common, feature distinctly different approaches to the ordering and focus of QM instruction. The model here allowed instructors to select from a checklist of which chapters from each text they covered in their course. The system would then construct their version of the assessment from a researcher-defined set of items designed to target the content of these chapters.

Alternatively, for the instructors who chose to create their own assessment, they were first prompted to select from a short list of broad topical areas. Based on their selections, they were then shown a list of "assessment objectives" which are concise, specific articulations of measurable, desired student performances regarding concepts and/or practices targeted by the assessment [14]. The AOs provided were developed through an iterative, emergent coding procedure on existing, static QM assessments [7–11]. Using these existing QM items, we defined AOs that encompassed the required student performance for each item. These were refined through multiple iterations of discussion and revision. An example is the following: *Given a quantum state and an observable operator, calculate that operator's expectation value.* We then organized these AOs into the broader categories used to structure in the interface.

After instructors completed these steps for the textbook and create-your-own pathways, they were then prompted with a question asking how long they wanted the assessment to last, ranging from 30-90 minutes. Following this, they were asked to choose the average difficulty of the exam. This was preceded by an example histogram of person abilities extracted from prior work [15]. This graph was accompanied with the following text offering guidance for the instructors on what would be the best selection: "The distribution of student performance is measured on a unit-less scale with average set to 0 and standard deviation set to 1. Think of your students and determine if they perform above or below the national average. If they perform above then you might want an average test difficulty of 1; if they perform below average then you might want an average test difficulty of -1. You want to try to match the average test difficulty to your students' average performance for the test to estimate their ability with maximum precision." The instructors were given 9 options for difficulty ranging from -2 to 2, however in the interviews none of the instructors selected outside of the range -1 to 1. In future versions of the assessment creation interface, this should be replaced with only 3 options, easy, average, and difficult. After this, the interview participants were provided with an example assessment which was just a subset of questions from the QMCA.

Once the interface is fully developed, this is where instructors would be able to preview the version of the assessment generated based on their selections and edit items if they so desired. Editing would include removing and replacing items with other items from the bank fulfilling their selected preferences but not include altering the content of the items. The back-end algorithm for moving from instructors' selection of AOs to the precise test form they would see, while a crucial step in the QuPA development, is the subject of ongoing and future work. After previewing their instance of the assessment, the final elements of setting up the test in the interface included instructors reporting on the open and close dates for their assessment and providing a list of student emails for distribution of the link to access the assessment online. Note that QuPA must be an online platform in the present design. This does not pose any significant issues and may actually boost student participation [16].

III. METHODS

To solicit interview participants, we compiled a list of contact information for every physics department in the U.S. that offered an upper-level QM course. At the time of solicitation, we were contemplating whether this assessment could be used for modern physics and graduate QM classes; thus, we asked for faculty who had taught those and/or upper-level undergraduate QM to participate. Thirty-five faculty participated in interviews, but one did not answer a sufficient number of our questions so is not included in the present analysis. In terms of courses taught by our participants, 28 were either currently teaching or had most recently taught upper-level undergraduate QM, 4 taught modern physics, 2 taught graduate QM, and 2 taught a quantum information course. A few of the instructors also taught more than one of these courses.

The 34 participants represented faculty from 33 unique institutions. Of the 33 institutions, 12 were private and 21 were public; 8 were minority serving institutions including historically black colleges and universities, Hispanic serving institutions, and Asian American and Native American Pacific Islander serving institutions; 6 were R1 (defined by the Carnegie classification of higher education institutions), 6 were R2, 8 were undergraduate only; and 11 offered Masters for their highest degree. We did not collect demographic information on the faculty participants.

Each semi-structured interview lasted around an hour and focused on the instructor's experience teaching QM, their interaction with the assessment creation interface, and what topics they would want to teach and assess in a QM class. This paper will only discuss results from their interaction with the assessment creation interface.

IV. RESULTS AND DISCUSSION

To summarize the results of the interviews we used emergent coding of the transcript data to answer five questions.

1. What are instructors' overall views of the interface?
2. What features of the process did they like?
3. What critiques did they have?
4. What things did they suggest we add to the design?
5. What kind of feedback would they want from this assessment to improve their teaching?

A. Overall Impressions

The question of whether instructors generally like or dislike the interface as designed is important to understanding the extent to which instructors are willing to engage with the system at all. Instructors were directly asked for their overall impression in 27 of the interviews, and in the remaining 7, participants spontaneously provided their overall impression or we were able to extract it from their responses to other questions. There was no difference in the distribution of responses between those directly asked and those not. The proportions of instructors who liked, disliked, and felt neutral toward the interface are presented in Table I along with some representative quotes.

Participant 23 (P23) brought up an important tension in balancing the level of control for all types of instructors. Some want significant control over the end product, and others want a streamlined process. We have tried to balance these extremes by providing options that cater to both camps, i.e.

TABLE I. Quotes referencing overall impressions regarding the assessment creation interface.

Sentiment	Proportion	Representative Quotes
Liked	27/34	**P4:** Oh, I think it's awesome... I think it's exactly the kind of thing that has made me reluctant to use other assessments in the past because, particularly as a theoretical physicist, I have my own agenda in the classroom. **P20:** I like it very much. It's clear and concise.
Disliked	3/34	**P23:** The level of control seems high. I think maybe... There's always a trade off between how much user control there is and how user friendly it is, right?... I think there's plenty of user control. It, for me, it might actually tip toward a lot of options that I'm not sure I have a great like grasp on what all of the knobs do right and so it might, it might even be more user control than ends up being user friendly from my perspective. **P30:** Nothing really excites me now that I see the options... Yeah, is this is harder to interface with... I'm having, I'm having feelings and doubts about it.
Neutral	4/34	**P18:** Uh, I mean, just looking at it, it looked fine, but I could see without exploring everything that one place where control might be nice would be notation.

using a pre-existing assessment, generating based on a textbook, and generating a fully customized assessment.

B. Positive Features

Below are some quotes from participants regarding the positive features of the interface design presented to them.

> **P1:** How long would this assessment take? Oh, my God, yes. *(referring to control of assessment duration)*
> **P4:** I love that you can auto-generate based on the text.
> **P24:** I like the ability to kind of give it a — the difficulty scale. Of course, it's a little bit hard to say exactly where you'd want that to be until you've tried it.
> **P35:** It seems pretty easy to use. I didn't have any difficulty navigating and it's very straightforward. The instructions are pretty good. And you know, at the end of this, an assessment is generated, which is just wonderful, you know, an assessment that suits your own purposes that you yourself specified.

The ease of use noted in all these quotes is an important design element that has historically not been emphasized in PER. Other aspects of our design that participants valued include: the fact that there were multiple methods for generating an assessment, the listing and organization of QM AOs, the ability to compare to national data, and the items provided in the example assessment.

C. Critiques

A few representative quotes of critiques provided by instructors are provided below.

> **P2:** I mean, it's not so much the building it myself, but... if I were the one creating the questions, I'd be a little concerned that... I would fundamentally be rigging the thing in such a way that it'll work out well or not well. Some sort of like independent idea of like what constitutes a reasonable test... that for me would be a... useful thing.
> **P4:** If you can make it work, I think it's a great model for other assessments, particularly if you can get the statistics to work convincingly. I think that's the part where—I don't know, even as a scientist, I don't believe statistics.
> **P15:** I think the problem with creating one's own assessment is one creates it toward what one teaches directly, which is good because you want to find out if the students are learning what you're teaching. But on the other hand it doesn't help... If there are major gaps in what you're teaching, you'll never find out that way... So I think to have something that's created by an outside source is useful, even though there may be things on there that you don't cover. It gives you a sense of what other people think is important.
> **P22:** Sustainability of the website. Because this happens with NSF programs where this really cool thing comes online and funding dries up, and then it disappears from the Internet.

Presenting the statistics in a way that is convincing and clear to instructors without a background in psychometrics will be a particular challenge, but one that can be overcome with careful consideration and access to additional resources. As for the sustainability of the website, we will build an automation system that will ensure minimal human maintenance. Other critiques that instructors provided include wanting access to example items when selecting AOs, needing clearer explanations of how to navigate the interface, needing clearer phrasing of AOs, and struggling to estimate their students' average ability relative to a national average a priori.

D. Suggested Features

Instructors had a plethora of suggested features that might be good to include in future designs.

> **P1:** I would like to see an assessment generated by a different professor ... because I'm always learning from people which are my peers.
> **P6:** The only other thing that I was thinking about, that might have been more helpful is ... maybe smaller exams that would happen during the semester. So you

could course correct in real time.

> **P9:** I think it would be useful to have some kind of question submission feature... I don't have the time to validate my questions right?
>
> **P13:** I do feel that students sometimes say things in open-ended responses... that give you much more insight into what they're thinking than if you provide for them a prefabricated set of either multiple choice options... So I think some opportunity to have that kind of [free] response would be good.
>
> **P18:** I guess we have a lot of support at our university for ... accommodations for students and you know, that includes things like visual and audio accommodation. So I'm wondering if those would be built in?

One instructor noted that the system could also be used to generate smaller tests, which would provide them with information to adjust teaching as the course progressed and is something entirely within the scope of our design. Other suggestions include using QuPA for program level assessment, being able to specify notation used in the items, being able to control the proportion of items from selected AOs, and connecting to existing learning management systems like Canvas.

E. Desired Feedback

After instructors administer this assessment, their results would automatically be analyzed and a report containing feedback would be given to them. We wanted to see directly what feedback instructors desired.

> **P4:** In terms of being able to look at the data, just an overall raw score is going to be a lot less interesting to me than being able to see how students perform in each of these boxes I just ticked.
>
> **P6:** Sometimes these assessments are anonymous. Sometimes they'd have biographical information. I could see the utility in collecting the gender and race information. Not necessarily... But uh, it would be nice to have some breakdowns that way as well.
>
> **P8:** ... student by student analysis, because honestly, like with a five person class, I would be curious to see where like the strong students who are giving the class their all, which distractors they still fell for compared to students who give it a less than full effort ... being able to kind of pull out individual student responses would be something that I would probably want to do.
>
> **P9:** Like if there's a particular misconception and someone has worked out a class activity where there's a worksheet... I think that that would be useful to have pointers to resources like that.

P6 mentioned collecting demographic information from students and seeing breakdowns that way, but for small class sizes, there could be issues with anonymity. Along a similar vein, P8 wanted to see a student by student breakdown of scores, but this could be problematic as this assessment should be used to evaluate pedagogy and curriculum, not the students. Other forms of feedback that instructors expressed interest in include seeing how long students interacted with the assessment, using artificial intelligence to generate detailed reports, being able to compare scores from previous years, comparing students in similar institutional settings, and having detailed information on item distractors.

V. CONCLUSIONS

In this manuscript, we discussed our design for a flexible QM assessment and presented representative quotes from a series of 34 interviews with instructors interacting with the instructor interface. We found that a majority of instructors liked the interface design presented to them and expressed interest in using this kind of assessment in the future. Instructors positively regarded many design elements of the interface presented to them such as the ability to select the time duration and difficulty of the assessment and the multiple methods for generating an assessment.

However, instructors also provided critiques such as concern over validity of the assessment and statistics, confusion with navigating the interface, and the long-term sustainability of the assessment. Instructors also provided suggestions for new features like using smaller, mid-semester assessments and seeing assessments generated by other professors. As the QuPA will be based on advanced psychometric theories outside the experience of most physics instructors, we will need to take care in how we design and present the assessment and its results to ensure instructors have faith in the results. These findings will also allow us to ensure our system, and systems like it, maximally address the needs and desires of instructors.

Instructors also expressed what feedback they would want from this kind of assessment, including a breakdown of scores by topical area, receiving suggested resources for targeting specific problem areas, and providing by-student breakdowns of performance. Some of these requests can be easily accommodated within the QuPA system, while others run contrary to the intended purpose and design of these assessments. These results suggest we, and other assessments developers within PER, need to be produce very clear messaging to accompany our feedback reports to ensure results from these assessments are interpreted and used consistent with our intention. Overall, in addition to informing the development of the QuPA, this work has provided insight into how instructors want to generate and utilize assessments that is valuable for all assessment developers. Ongoing work towards the development of the QuPA includes additional item development, and designing and implementing an algorithm for generating an assessment based on instructors' selected preferences.

VI. ACKNOWLEDGMENTS

This work was supported by funding from the Department of Physics at University of Colorado Boulder and the NationalScience Foundation DUE Grant No. 2143976.

[1] Shabnam Siddiqui and Chandralekha Singh, "How diverse are physics instructors attitudes and approaches to teaching undergraduate level quantum mechanics?" European Journal of Physics **38**, 035703 (2017), publisher: IOP Publishing.
[2] S. B. McKagan, "PhysPort: Supporting physics teaching with research-based resources," .
[3] Paula Engelhardt, "An Introduction to Classical Test Theory as Applied to Conceptual Multiple-choice Tests," in *Getting Started in PER*, Reviews in PER, Vol. 2 (AAPT, 2009).
[4] Wim J. van der Linden and Gees A. W. Glas, *Computerized Adaptive Testing: Theory and Practice*, 1st ed. (Springer Dordrecht, 324, 2000).
[5] Katherine D. Rainey, Michael Vignal, and Bethany R. Wilcox, "Designing upper-division thermal physics assessment items informed by faculty perspectives of key content coverage," Physical Review Physics Education Research **16**, 020113 (2020), publisher: American Physical Society.
[6] Julie C. Libarkin and Steven W. Anderson, "Assessment of Learning in Entry-Level Geoscience Courses: Results from the Geoscience Concept Inventory," Journal of Geoscience Education (2005), 10.5408/1089-9995-53.4.394, publisher: The National Association of Geoscience Teachers.
[7] Homeyra R. Sadaghiani and Steven J. Pollock, "Quantum mechanics concept assessment: Development and validation study," Physical Review Special Topics - Physics Education Research **11**, 010110 (2015), publisher: American Physical Society.
[8] Guangtian Zhu and Chandralekha Singh, "Surveying students understanding of quantum mechanics in one spatial dimension," American Journal of Physics **80**, 252–259 (2012), publisher: American Association of Physics Teachers.
[9] E. Cataloglu and R. W. Robinett, "Testing the development of student conceptual and visualization understanding in quantum mechanics through the undergraduate career," American Journal of Physics **70**, 238–251 (2002), publisher: American Association of Physics Teachers.
[10] Johan Falk, *Developing a quantum mechanics concept inventory*, Ph.D. thesis, Uppsala University (2004).
[11] Sura Wuttiprom, Manjula Devi Sharma, Ian D. Johnston, Ratchapak Chitaree, and Chernchok Soankwan, "Development and Use of a Conceptual Survey in Introductory Quantum Physics," International Journal of Science Education (2009), 10.1080/09500690701747226, publisher: Routledge.
[12] David J. Griffiths and Darrell F. Schroeter, *Introduction to Quantum Mechanics*, 3rd ed. (Cambridge University Press, 2018) iSBN: 9781316995433 Publisher: Cambridge University Press.
[13] David H. McIntyre, *Quantum Mechanics: A Paradigms Approach*, 1st ed. (Cambridge University Press, 2022) iSBN: 9781009310598 Publisher: Cambridge University Press.
[14] Michael Vignal, Katherine D. Rainey, Bethany R. Wilcox, Marcos D. Caballero, and Heather J. Lewandowski, "Affordances of Articulating Assessment Objectives in Research-based Assessment Development," in *2022 Physics Education Research Conference Proceedings* (2022) pp. 475–480, iSSN: 2377-2379.
[15] Jesse Kruse and Bethany R. Wilcox, "Rasch Analysis of the Quantum Mechanics Concept Assessment," in *2023 Physics Education Research Conference Proceedings* (2023) pp. 175–180, iSSN: 2377-2379.
[16] Bethany R. Wilcox and Steven J. Pollock, "Student behavior and test security in online conceptual assessment," in *2019 Physics Education Research Conference Proceedings* (2019) pp. 645–650, iSSN: 2377-2379.