

Rasch Analysis of the Quantum Mechanics Concept Assessment

Jesse Kruse¹ and Bethany R. Wilcox¹

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

Quantum mechanics is a subject rife with student conceptual difficulties. In order to study and devise better strategies for helping students overcome them, we need ways of assessing on a broad level how students are thinking. This is possible with the use of standardized, research-validated assessments like the Quantum Mechanics Concept Assessment (QMCA). These assessments are useful, but they lack rigorous population independence, and the question ordering cannot be rearranged without throwing into question the validity of the results. One way to overcome these two issues is to design the exam to be compatible with Rasch measurement theory which calibrates individual items and is capable of assessing item difficulty and person ability independently. In this paper, we present a Rasch analysis of the QMCA and discuss estimated item difficulties and person abilities, item and person fit to the Rasch model, and unidimensionality of the instrument. This work will lay the foundation for more robust and potentially generalizable assessments in the future.

I. INTRODUCTION & BACKGROUND

Quantum mechanics is a notoriously difficult subject to learn and understand. There has been much characterization over the past thirty years on student difficulties and misconceptions in undergraduate [1] and graduate quantum mechanics courses [2]. These include difficulties with reconciling quantum concepts and classical concepts, properties and representations of wave functions, distinguishing between three-dimensional Euclidean space and Hilbert space, measurement and expectation values, Dirac notation, and many more [1]. In addition to the complexity of quantum mechanics, instructors disagree on which topics to include in, and how to teach, the subject [3]. For example, instructors disagree on whether to present quantum mechanics in a spins-first or wavefunctions-first approach, whether to present an axiomatic or historical approach, and whether wavefunctions represent a matter wave, information wave, or something else entirely [3].

Because of the plethora of challenges students face in learning quantum mechanics and because of the lack of consensus on what and how to teach the subject, it is difficult to establish clear learning goals that are relevant across institutions. This has posed an issue for evaluating student learning with research-based assessments. If developers design a test that contains a certain number of subjects, it is possible that it won't be applicable to classes that chose not to cover all those topics. For example, spins-first courses often have not finished solving the full Schrodinger equation for the hydrogen atom by the end of the first semester, so questions involving the hydrogen atom on an assessment won't provide useful measures of learning for that class. Therefore, it would be advantageous if a modular assessment was developed that could accommodate the variety of instruction and learning goals inherent in undergraduate quantum physics education.

Currently, there are around ten research-validated assessments for modern physics and quantum mechanics [4–10]. These cover a variety of topics such as measurement, wave functions, Dirac notation, incompatible operators, scattering, tunneling, time dependence, and many more. However, there is still a multitude of topics that aren't assessed by these instruments such as most topics in a second semester of quantum mechanics and some topics that may be covered in either semester like entanglement, Bell inequalities, and topics related to quantum information.

In addition, these assessments have all been validated using classical test theory. Classical test theory (CTT) consists of a few statistical measures of test scores that look at item difficulty, discrimination, reliability, and overall consistency within the test. This is the most commonly used framework for validating assessments within PER [11]; however, this approach brings with it some fundamental limitations. These include the fact that the scores you get are always dependent upon the sample used for calibration. This means the item difficulty and discrimination will be different for different samples, and the ordering of the questions can have significant effects on student performance. In addition, differences of students abilities in CTT don't have a well-defined meaning

whereas in Rasch measurement theory they do [12].

Rasch measurement theory (RMT) and more generally item response theory (IRT) are probabilistic models of student responses to test items that are functions of the person ability¹ and item difficulty. In RMT, the base assumption is that the probability of a person answering an item correctly is dependent only on the difference between their ability and the item difficulty. This offers an advantage over CTT because the item difficulties and person abilities are computed together and share a common scale where comparison is actually meaningful [12]. The Rasch model also allows us to estimate person ability independent of which items are used and the item difficulties independent of the people used to calibrate it [13].

This paper will provide a brief overview of Rasch measurement theory, discuss the data used in analyzing the QMCA, and present our results for item difficulties, person abilities, and how well the data fit our model of measurement. We hope to add well-fitting items to a quantum physics test bank that instructors can use to design their own assessments while still generating robust comparable measures of student learning.

II. RASCH MEASUREMENT THEORY

The founder of Rasch measurement theory was a Danish mathematician named Georg Rasch. He postulated that in order for psychological measurement to truly be a measurement, the process of assessing the property under study must follow certain criteria. The numbers that we assign to the property under study must have a common scale and be comparable to one another on the interval level [12]. Rasch postulated a frame of reference which was necessary for interpreting the results for an assessment which consists of:

1. the class of items on an assessment that target the construct under study,
2. the class of persons who are relevant to be assessed
3. the conditions of the administration of the assessment.

He further postulated that if the conditions of the frame of reference are appropriate, an examinee's performance on a given item should be dependent on two parameters: the examinee's ability θ_n and the item's difficulty δ_i . These two parameters are measured on the same scale and are sometimes referred to generally as the person and item location. Rasch theorized that in order for an assessment to truly measure an underlying trait, the comparison between any two persons must be independent of the choice of items used, and the comparison between any two items must be independent of the people that interact with the items [12]. This property is known as the invariance of comparisons, and it directly leads to the probability distribution that describes students'

¹ Note, "ability" refers to the latent trait that the statistical models quantify. Fundamentally, however, it is a measure of performance as opposed to innate ability. This term is used for consistency with the existing literature. However, this term is potentially problematic, particularly with respect to the interpretation of performance gaps between subgroups of students.

outcomes to a dichotomous item. This probability distribution is written as:

$$P(x_{ni} = 1 | \theta_n, \delta_i) = \frac{e^{\theta_n - \delta_i}}{1 + e^{\theta_n - \delta_i}}, \quad (1)$$

where x_{ni} is the outcome of person n answering item i correctly. In general θ_n and δ_i vary continuously from $-\infty$ to $+\infty$, but we perform a linear transformation so the mean is 0 and the standard deviation is 1. We see that when θ_n is larger than δ_i the probability of answering correctly is more likely than not, and vice versa when $\theta_n < \delta_i$. In addition, the log-odds ratio which describes this behavior is given by:

$$\ln \left(\frac{P}{1 - P} \right) = \theta_n - \delta_i. \quad (2)$$

When we have the same person attempting two different items with difficulties δ_1 and δ_2 , the difference in the log-odds is simply the difference in the item difficulties $\delta_1 - \delta_2$. This means that the likelihood of answering one item correctly and the other incorrectly is a function only of the difference in item difficulties, not the person's ability, thus demonstrating the invariance of comparisons [12]. Another fundamental assumption of RMT is that all of the items on an assessment are parallel, meaning they independently measure some aspect of the construct under study. The primary task of RMT and IRT is to estimate the person abilities and item difficulties using a fitting procedure. This is most often done with software packages in statistical programming languages like the multidimensional item response theory package `mirt` in the language R.

III. CONTEXT & METHODS

The Quantum Mechanics Concept Assessment (QMCA) is a conceptual assessment for evaluating student learning in upper-level, undergraduate quantum mechanics courses [4]. The QMCA consists of 38 multiple-choice items covering measurement, the time-independent Schrodinger equation, time evolution, wave functions, boundary conditions, probabilities, and expectation values. It was adapted from the free-response format Quantum Mechanics Assessment Tool (QMAT), and it has undergone many revisions and verifications of face and content validity including a CTT analysis [4], an exploratory factor analysis [14], and a modified module analysis [15]. However, there has yet to be a Rasch analysis of the QMCA or any quantum assessment for that matter.

The data used in this analysis were gathered from eight separate institutions ranging in size, research output, location, and even nationality. Seven of the institutions were located in the continental United States, but one sample was from a university in South Africa. The responses to the QMCA were gathered from the fall semester of 2018 to the spring semester of 2022, with a majority coming from the 2018-2019 academic school year. Around 30 students had incomplete attempts, so it was necessary to establish a criterion for whether or not to include them in the aggregate analysis. We decided

to use a 90% completion cutoff, so if students answered at least 90% of the items, it could be considered a good-faith attempt. This corresponds to answering at least 35 of the 38 questions. Note that including them did not affect the results significantly. After combining all the data, we had a total of 403 responses, which is large enough to reliably estimate parameters in the Rasch model [16].

IV. RESULTS & DISCUSSION

A. CTT Analysis

We start off by looking at the CTT statistics and comparing them to published results. The definitions of the item difficulty index P , item discrimination index D , point biserial index r_{pbi} , Kuder-Richardson 20 coefficient KR-20, and Ferguson's delta are all discussed in [11]. Table I summarizes these results and compares them to Sadaghiani and Pollock's study [4]. We see that the statistics calculated from our data are slightly higher above the threshold than that of Sadaghiani and Pollock's, which may be a result of being a newer version of the QMCA.

B. Exploratory Factor Analysis

Before fitting our QMCA data to the Rasch model, we need to check whether our data are sufficiently unidimensional. The number of dimensions assessed by an instrument dictates what sort of model we should apply to it. Our initial naive hypothesis is that the QMCA assesses a single dimension of quantum mechanics proficiency at the first semester upper-level undergraduate level, but in order to confirm this we need to do an exploratory factor analysis (EFA). To do this, we used the `mirt` package in R (a multidimensional item response theory modeling package) [17] and specified the number of assumed dimensions. This gives us the factor loadings from which we interpret whether it is an adequate fit.

The factor loadings for the unidimensional EFA are in Table II. The factor loadings represent what percentage of the variance on that item can be accounted for by the assumed underlying factor. If an item has a factor loading greater than or equal to approximately 0.3 then it is said to be well described by the factor [14]. We see from the table that the items that

CTT Stats	Target Values	Ref [4]	Our Data
Num. Stu.	...	263	403
Num. Items	...	31	38
Ave. P	> 0.3	0.54	0.55
Ave. D	> 0.3	0.42	0.45
Ave. r_{pbi}	> 0.2	0.35	0.39
KR-20	> 0.7	0.76	0.84
Ferguson's δ	> 0.9	0.97	0.986

TABLE I. Classical test theory statistics for data on QMCA from 2018-2022.

Item	F1	Item	F1	Item	F1	Item	F1
1	0.86	11	0.39	21	0.26	31	0.53
2	0.91	12	0.21	22	0.87	32	0.40
3	0.44	13	0.40	23	0.92	33	0.47
4	0.11	14	0.41	24	0.53	34	0.57
5	0.26	15	0.34	25	-0.10	35	0.49
6	0.72	16	0.39	26	0.17	36	0.67
7	0.59	17	0.44	27	0.82	37	0.15
8	0.65	18	0.63	28	0.57	38	0.05
9	0.43	19	0.66	29	0.63		
10	0.43	20	0.18	30	0.49		

TABLE II. Exploratory factor analysis for 1 factor. The factor loadings for each item are under F1. Factor loadings in red indicate that they are below the 0.3 threshold.

aren't well described by this single factor/dimension are items 4, 5, 12, 20, 21, 25, 26, 37, and 38. This lack of unidimensionality is consistent with previous research on the QMCA done by Quaal et al. [14].

There is a pattern here. There are five pairs of coupled questions on the QMCA where the second question asks for the best explanation for the response to the first. These question pairs are 4/5, 13/14, 20/21, 25/26 and 37/38. We see that four of these five question pairs are the items that don't properly load into a single factor. Question pair 13/14 does not fit into this pattern because question 13 has 5 possible answer choices while the rest have only 2 or 3. The only other question that isn't adequately described by a single factor is item 12 which asks about how the real and imaginary parts of the ground state of the infinite square well change in time. We theorize that this question isn't adequately described by the assumed factor because it is primarily assessing understanding of how complex exponentials decompose into real and imaginary parts which is more mathematical understanding than quantum mechanical.

When we increase the dimensionality of the EFA to 2, we find that the second factor loads almost entirely into items 4 and 5 with factor F2 loadings of 0.96 and 0.91 respectively. This means that the additional factor is describing most of the coupled variance of questions 4 and 5 and nothing else. When we increase the dimensionality of the EFA to 3, we find similar behavior where the new factor describes the the coupled variance of items 13 and 14. These items have a factor F3 loadings of 0.93 and 0.99 respectively. It is clear from this analysis that the question pairs 4/5, 13/14, 20/21, 25/26, and 37/38 as well as item 12 are not well described by a single factor. In the following Rasch analysis we will first consider the entire set to see if poor fit statistics confirm our results from the EFA, and then we will remove items to see how the remaining set performs under the unidimensional Rasch model.

C. Rasch analysis

We used the `mirt` package to do a unidimensional Rasch analysis of our data, generating item difficulties and person

Item	S-X2	df	RMSEA	p	Item	S-X2	df	RMSEA	p
1	51	22	0.06	0	20	55	23	0.06	0.00
2	50	22	0.06	0.00	21	29	21	0.03	0.13
3	42	22	0.05	0.01	22	44	22	0.05	0.00
4	34	22	0.03	0.05	23	80	22	0.08	0.00
5	32	22	0.03	0.08	24	27	22	0.02	0.22
6	26	22	0.02	0.25	25	118	22	0.10	0.00
7	28	23	0.02	0.20	26	54	23	0.06	0.00
8	24	18	0.03	0.15	27	48	22	0.05	0.00
9	18	23	0.00	0.74	28	34	22	0.04	0.06
10	22	22	0.00	0.47	29	24	22	0.01	0.37
11	25	22	0.02	0.32	30	13	23	0.00	0.94
12	52	22	0.06	0.00	31	28	23	0.02	0.21
13	31	22	0.03	0.10	32	23	22	0.01	0.40
14	32	23	0.03	0.11	33	17	23	0.00	0.81
15	21	23	0.00	0.59	34	27	22	0.02	0.22
16	42	23	0.05	0.01	35	27	22	0.02	0.22
17	20	23	0.00	0.65	36	26	21	0.03	0.19
18	30	23	0.03	0.14	37	64	22	0.07	0.00
19	29	23	0.03	0.18	38	57	22	0.06	0.00

TABLE III. Item fit statistics for the whole QMCA under the Rasch model. S-X2 is the signed chi-squared statistic, df is the degrees of freedom which is variable depending on the binning for S-X2, RMSEA is the root mean square error of approximation, and p is the probability of observing those data or more extreme values. The color red indicates that the RMSEA is at or above the 0.06 threshold.

abilities based on the fit to the Rasch model. The statistics that we generate for the whole assessment are the M2 statistic which is essentially a modified chi-squared goodness-of-fit test [18], the root mean square error of approximation (RMSEA), and the comparative fit index (CFI) [19]. In general, we want the M2 divided by the degrees of freedom df to be approximately equal to 1, and for relatively good model-data fit, we want an $RMSEA < 0.06$ and a $CFI > 0.9$ [19]. When we run the Rasch analysis on the whole test we get $M2/df = 4.5$, an $RMSEA$ of 0.093, and a CFI of 0.68. All of these indicate poor model-data fit and are consistent with our EFA.

From Table III we have the item fit statistics for the Rasch model. We see from Table III that there are many items that have poor model-data fit because their $RMSEA$ are greater than 0.06. These items are colored red in the table.

Figure 1 shows the estimated item difficulties and person abilities for the whole QMCA under the Rasch model. We see that the person abilities are approximately normally distributed around zero ability, and the item difficulty average is -0.29.

Rather than go through all possible iterations of removing items, we will just discuss removing the fewest number of items in order to achieve the overall model fit statistics of $RMSEA < 0.06$ and $CFI > 0.9$. In order to achieve these metrics it was necessary to remove around 14 items. All of

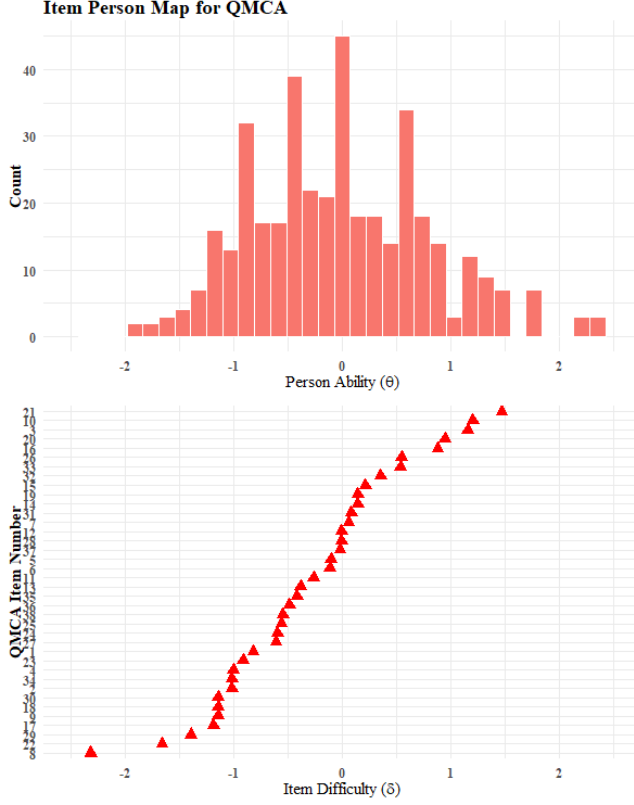


FIG. 1. Estimated item difficulties and person abilities for the whole QMCA under the Rasch model. Note that ability and difficulty are measured on the same scale and presented as such for ease of comparison.

the question pairs 4/5, 13/14, 20/21, 25/26, and 37/38 were removed, and the questions with very high or very low item discrimination as determined by CTT were removed. Those items with high discrimination (ability to distinguish between high and low performing students) were 1, 12, 23, and 27. After removing these items we achieved $M2/df = 2.3$, an RMSEA of 0.056, and a CFI of 0.902. Table IV shows the individual item fit statistics when these items are removed. We see from this table that all of the items have an RMSEA at or lower than 0.06, which means there is adequate or good model-data fit. Note that the removed items tended to involve time evolution, but not all time evolution items were removed. So we still retained the bulk of content areas in this modified QMCA.

V. CONCLUSIONS

From this analysis we can conclude that 24 of the 38 items on the QMCA are sufficiently unidimensional to be described by a single latent trait under the Rasch model, making them good candidates for a unidimensional test bank. The justifications for removing the poorly fitting items are that they

either violate the requirement for local independence, they don't assess the same dimension as the rest of the questions, or they have extreme discriminations as measured by CTT fit. There are still many analyses that can be done on this data set, including examining whether the subset of people based on their university affects the results, whether the isomorphic spin and wave function context items have similar characteristics, and whether a multidimensional Rasch model would provide better results.

The remaining questions in the QMCA are good candidates for adding to a quantum mechanics test bank that instructors could pull from in order to design their own assessment. Future work for this project includes creating surveys containing all the questions on existing QM assessments so that we can collect data from multiple institutions and perform a similar Rasch analysis on them. We want to have as many compatible questions as possible so that we can create a modular quantum mechanics assessment capable of assessing person ability in a rigorous population independent manner. We also want to generate new items pertaining to second semester quantum mechanics that we can add to this item bank. This will include generating questions, doing student interviews to assess construct validity, and administering to large samples to determine item difficulties.

VI. ACKNOWLEDGMENTS

This work was supported by funding from the Center for STEM Learning and the Department of Physics at University of Colorado Boulder, and the National Science Foundation DUE Grant No. 2143976.

Item	S-X2	df	RMSEA	p	Item	S-X2	df	RMSEA	p
2	34	15	0.06	0.003	19	16	16	0.00	0.46
3	25	16	0.04	0.06	22	29	14	0.05	0.01
6	15	16	0.00	0.50	24	14	15	0.00	0.51
7	18	16	0.02	0.32	28	31	16	0.05	0.01
8	17	13	0.03	0.19	29	15	13	0.02	0.28
9	9	15	0.00	0.85	30	9	15	0.00	0.85
10	19	15	0.03	0.22	31	33	16	0.05	0.01
11	18	16	0.02	0.32	32	14	15	0.00	0.55
15	27	16	0.04	0.04	33	15	15	0.00	0.48
16	38	16	0.06	0.00	34	9	15	0.00	0.85
17	26	14	0.05	0.02	35	22	15	0.03	0.11
18	9	15	0.00	0.86	36	12	15	0.00	0.71

TABLE IV. Item fit statistics for the QMCA with items 1, 4, 5, 12, 13, 14, 20, 21, 23, 25, 26, 27, 37, and 38 removed under the Rasch model. S-X2 is the signed chi-squared statistic, df is the degrees of freedom which is variable depending on the binning for S-X2, RMSEA is the root mean square error of approximation, and p is the probability of those data or more extreme values.

-
- [1] C. Singh and E. Marshman, "Review of student difficulties in upper-level quantum mechanics," *Physical Review Special Topics - Physics Education Research* **11**, 020117 (2015).
- [2] Chandralekha Singh, "Student understanding of quantum mechanics at the beginning of graduate instruction," *American Journal of Physics* **76**, 277–287 (2008).
- [3] M. Dubson, S. Goldhaber, S. Pollock, and K. Perkins, "Faculty Disagreement about the Teaching of Quantum Mechanics," *AIP Conference Proceedings* **1179**, 137–140 (2009).
- [4] H. Sadaghiani and S. Pollock, "Quantum mechanics concept assessment: Development and validation study," *Physical Review Special Topics - Physics Education Research* **11**, 010110 (2015).
- [5] S. B. McKagan, K. K. Perkins, and C. E. Wieman, "Design and validation of the Quantum Mechanics Conceptual Survey," *Physical Review Special Topics - Physics Education Research* **6**, 020121 (2010).
- [6] G. Zhu and C. Singh, "Surveying students' understanding of quantum mechanics in one spatial dimension," *American Journal of Physics* **80**, 252–259 (2012).
- [7] E. Marshman and C. Singh, "Validation and administration of a conceptual survey on the formalism and postulates of quantum mechanics," *Physical Review Physics Education Research* **15**, 020128 (2019).
- [8] S. Wuttirom, M. D. Sharma, I. D. Johnston, R. Chitaree, and C. Soankwan, "Development and Use of a Conceptual Survey in Introductory Quantum Physics," *International Journal of Science Education* (2009), 10.1080/09500690701747226.
- [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).
- [10] K. Ait bentaleb, S. Dachraoui, T. Hassouni, E. Alibrahmi, E. Chakir, and A. Belboukhari, "Development of a Survey to Assess Conceptual Understanding of Quantum Mechanics among Moroccan Undergraduates," *European Journal of Educational Research* **11**, 2219–2243 (2022).
- [11] L. Ding and R. Beichner, "Approaches to data analysis of multiple-choice questions," *Physical Review Special Topics - Physics Education Research* **5**, 020103 (2009).
- [12] D. Andrich and I. Marais, *A Course in Rasch Measurement Theory: Measuring in the Educational, Social and Health Sciences*, Springer Texts in Education (Springer Nature, Singapore, 2019).
- [13] Lin Ding, "Applying Rasch theory to evaluate the construct validity of brief electricity and magnetism assessment," in *2011 Physics Education Research Conference*, Vol. 1413 (2012) pp. 175–178.
- [14] A. Quaal, G. Passante, S. J. Pollock, and H. R. Sadaghiani, "Exploratory factor analysis of the QMCA," in *2020 Physics Education Research Conference* (2020) pp. 406–411, iSSN: 2377-2379.
- [15] J. Wells, H. Sadaghiani, B. P. Schermerhorn, S. Pollock, and G. Passante, "Deeper look at question categories, concepts, and context covered: Modified module analysis of quantum mechanics concept assessment," *Physical Review Physics Education Research* **17**, 020113 (2021).
- [16] Michael R. Harwell and Janine E. Janosky, "An Empirical Study of the Effects of Small Datasets and Varying Prior Variances on Item Parameter Estimation in BILOG," *Applied Psychological Measurement* **15**, 279–291 (1991).
- [17] R. Philip Chalmers, "mirt: A multidimensional item response theory package for the R environment," *Journal of Statistical Software* **48**, 1–29 (2012).
- [18] Albert Maydeu-Olivares and Harry Joe, "Limited- and Full-Information Estimation and Goodness-of-Fit Testing in 2n Contingency Tables," *Journal of the American Statistical Association* **100**, 1009–1020 (2005).
- [19] Yan Xia and Yanyun Yang, "RMSEA, CFI, and TLI in structural equation modeling with ordered categorical data: The story they tell depends on the estimation methods," *Behavior Research Methods* **51**, 409–428 (2019).