Validating shorter versions of the Physics Inventory of Quantitative Literacy

Brett T. Boyle (he/him)
Department of Physics & Astronomy, Rowan University, Glassboro, New Jersey 08028, USA

Charlotte Zimmerman (she/her) and Suzanne White Brahmia (she/her)
Department of Physics, University of Washington, Seattle, Washington 98195, USA

Trevor I. Smith (he/him)
Department of Physics & Astronomy, Rowan University, Glassboro, New Jersey 08028, USA and
Department of Content Area Teacher Education, Rowan University, Glassboro, New Jersey 08028, USA

The Physics Inventory of Quantitative Literacy (PIQL) has been used to measure the development of students' physics quantitative literacy in calculus-based introductory physics courses. Despite its effectiveness, issues persist regarding time constraints and potential memorization of items. We propose to split the PIQL into two shorter but statistically equivalent exams (PIQLets) in order to avoid these problems. Using a data set collected with the full PIQL, we created 480 theoretical PIQLet pairs containing different combinations of items. We provide evidence for the similarity of PIQLet pairs by calculating score differences, and comparing the distribution of item parameters calculated using item response theory. Our results demonstrate the feasibility of this approach for defining an equivalent pair of PIQLets using a limited data set from a single university. Additional analyses using a broader and more diverse data set will be required for more broadly applicable results.

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

I. INTRODUCTION

Mathematical reasoning, particularly reasoning about quantities, is ubiquitous throughout physics courses. A tacit goal for many physics instructors is for their students to improve their physics quantitative literacy (PQL), i.e., their abilities to reason about novel physical situations using familiar mathematical principles. The Physics Inventory of Quantitative Literacy (PIQL) is a 20-item multiple-choice test for measuring the impact of instruction on students' mathematical reasoning in physics courses (a.k.a., PQL) [1]. The PIQL was designed to include items that measure student reasoning across three facets of PQL — ratios and proportions, negativity and signed quantities, and covariation — each of which is pervasive in physics. Previous studies have provided strong evidence for the validity and reliability of using the PIQL to measure student reasoning in introductory physics courses [1–8]; moreover, even though the PIQL was designed to include three facets, exploratory and confirmatory factor analyses show that it is a unidimensional test. This finding supports the interpretation of PQL as a measurable quantity [1, 5].

Some items on the PIQL require physics content knowledge that is typically learned at the introductory level, but many of the items do not require any physics knowledge: rather, they involve the types of mathematical reasoning that are typical in physics contexts. (See [9] for the current version of the PIQL.) As such, the PIQL may be used in multiple physics courses to see how students' PQL changes over time. Anecdotal evidence suggests that some students experience fatigue by this repeated testing, with instructors reporting that students express frustration when being asked to take the PIQL for the third or fourth time. Instructors also report that students mention that they remember certain PIQL items from previously completing the test. Our evidence shows that students taking the PIQL for the first time spend about 35-40 minutes on the 20-item test, but that this time decreases substantially for repeated assessments.

To address these concerns, we aim to create two shorter versions of the PIQL that are different, yet statistically and psychometrically equivalent. These testlets (or "PIQLets") would require less time for students to complete, and would reduce the instances of repeated testing with the same items. To measure the impact of instruction, one PIQLet could be used at the beginning of a course, and the other at the end. Our ultimate goal is to be able to compare students' scores on any one PIQLet to scores on the other PIQLet, and to scores on the full PIQL. In this way, learning could be measured no matter which test (or combination of tests) is used.

We are guided by previous efforts to create two equivalent half-length versions of well-established concept inventories: the Force Concept Inventory (FCI) [10, 11], and the Conceptual Survey of Electricity and Magnetism (CSEM) [12, 13]. In both cases, researchers used analyses of data sets collected using the full-length tests to create theoretically equivalent half-length versions. Additionally, both studies included some "anchor" items on both half-length versions that may be used as comparison points. Han, et al. calculated score differences between the two proposed HFCIs (both for the test as a whole and for each subtopic measured by the test), and used small score differences as evidence for equivalent tests [11]. They also demonstrated similarities between their two identified HFCIs and the the full FCI by applying a three-parameter logistic item response theory (IRT) analysis to their data [11, 14]. Xiao, et al. used a more statistically rigorous approach to establishing test equivalence when creating HCSEMs by using Rasch analysis (a one-parameter IRT model) to compare item difficulty and student ability across their test versions [13].

Our previous preliminary analyses followed the example of Han, et al. [11] and focused on identifying equivalent PIQLets based on average score differences [15]. In this project we seek to establish the similarity between proposed PIQLets by measuring score differences (both for the entire PIQLet and for each of the three facets of PQL), and by comparing the distributions of parameters obtained from three-parameter logistic (3PL) IRT analyses. We aim to create 12-item PIQLets that each contain four common anchor items that can be used to facilitate comparisons, and eight distinct items. Our work is guided by the following research questions:

1. Which combination of items produces PIQLet pairs with the smallest score differences between each other?
2. Which combination of items produces PIQLet pairs with IRT parameter distributions that are most similar to each other?
3. Do any combinations of items produce PIQLet pairs that minimize both score differences and IRT parameter differences?

We are measuring differences in student performance and item performance between two PIQLets to identify the combination of items that produces a PIQLet pair with the smallest differences.

II. DATA & METHODS

Data for this study were collected over three years within the calculus-based introductory physics sequence at a large public university in the western US. Data were collected at three different times in the sequence: before Introductory Mechanics (PreMech), after Introductory Mechanics but before Introductory Electricity & Magnetism (PostMech), and after Introductory Electricity & Magnetism (PostEM). In order to test the usefulness of our methods, we have chosen to first focus on the PostMech data set (N = 2580). We chose to focus on PostMech data because previous results have shown that student scores on a few items increase substantially between PreMech and PostMech, but that similar increases do not happen from PostMech to PostEM [1]. Also, students in the PostMech data set are seeing the PIQL for the first or second time, suggesting that they are more likely to engage meaningfully than if they were seeing it for the third time, as is likely in the PostEM data set.

PR items	NR items	CR items
2, 6, 10, 11, 12*, 13, 16*†	4, 14, 15*†, 18†, 19†, 20†	1, 3, 5, 7, 8, 9, 12*, 15*†, 16*†, 17†

TABLE I. Categorization of PIQL items according to the three facets of PQL: proportional reasoning (PR), negativity reasoning (NR), and covariational reasoning (CR). Asterisks (*) denote items that assess more than one facet. Daggers (†) indicate MCMR items.

We first established content validity for the PIQLets by categorizing each PIQL item according to the PQL facet(s) it was designed to assess: proportional reasoning (PR), negativity reasoning (NR), or covariational reasoning (CR); see Table I (note that some items align with multiple facets). We require each PIQLet to have the same number of items assessing each facet. One of the features of the PIQL is that six items are multiple-choice-multiple-response (MCMR) in which students may select as many responses as they want, and there may be more than one correct response. We require each PIQLet to have the same number of MCMR items.

We then identified potential anchor items by examining previous psychometric analyses using classical test theory (CTT). The PIQL was designed to have items with a broad range of CTT difficulty values (fraction of students answering correctly between 0.2–0.8), and CTT discrimination values above 0.3 [1]. The CTT discrimination is a particularly useful metric is evaluating the quality of an item because it provides a measure of how well the item differentiates between high-scoring and low-scoring students. For our anchor items, we looked for items with high CTT discrimination (close to or above 0.6), that spanned all three of our PQL content facets, and that had a range of CTT difficulty values (not all easy or all hard). Table II shows our potential anchor items. Each PIQLet pair contains items 10, 15, and 20 as anchor items, and either item 1 or item 7 as an achor item.

Based on the above criteria, we identified 480 possible item combinations, with each combination producing a unique pair of 12-item PIQLets. All PIQLets have four PR items, four NR items, and six CR items; this includes one item that aligns with both PR and CR, and one item that aligns with both CR and NR. All combinations include four MCMR items and eight single-response items. Items 12 and 16 are always on different PIQLets because they are the only two that align with both PR and CR; moreover, in order to have the same number of CR items and MCMR items on each PIQLet, items 16 and 17 are always on different PIQLets, which means that items 12 and 17 are always on the same PIQLet.

For each item combination, we separated our data set into the items on PIQLet1 and the items on PIQLet2, and we determined the differences between the two PIQLets by comparing students' scores and by analyzing each PIQlet using item response theory (IRT). All students' responses were scored dichotomously as either completely correct or incorrect. For each item combination, we calculated students' average scores on PIQLet1 and PIQLet2, as well as the difference between these average scores ("Overall Score Difference" in Table III). We also calculated three PIQLet subscores (and subscore differences) using the subset of items in each PIQLet corresponding with each facet: difference in proportional reasoning subscore (Δs_{pr}), difference in negativity subscore (Δs_{nr}), and difference in covariation subscore (Δs_{cr}). All scores and subscores were calculated as percentages based on the number of items on each PIQLet or corresponding with each PQL facet. In order to quantify the overall similarity between PIQLet versions, we calculated a magnitude of the subscore difference,

$$|\Delta s| = \sqrt{\Delta s_{pr}^2 + \Delta s_{nr}^2 + \Delta s_{cr}^2}. \quad (1)$$

Item	Facet(s)	CTT difficulty	CTT discrimination
1	CR	0.60	0.60
7	CR	0.68	0.65
10	PR	0.73	0.63
15†	CR/NR	0.38	0.66
20†	NR	0.54	0.72

TABLE II. Potential anchor items and characteristics. In classical test theory (CTT) "difficulty" is the fraction of students who answer an item correctly, and "discrimination" is the difference in difficulty between high- and low-performing students. †Items 15 and 20 are MCMR items.

We performed IRT analyses on both PIQLets for each item combination using a three-parameter logistic (3PL) model. For the 3PL model, the probability that a student answers the i^{th} item correctly is given by,

$$P_i(\theta) = c_i + \frac{1 - c_i}{1 + \exp[-1.7a_i(\theta - b_i)]}, \quad (2)$$

where θ is the student's overall proficiency or level of understanding (representing their PQL). The c_i parameter (a.k.a. "guessing") indicates the probability that a student with a very low θ value will select the correct response option. The b_i parameter ("item difficulty") represents the value of θ for which the probability of a student selecting the correct response is halfway between c_i and 1. The a_i parameter ("item discrimination") represents the rate of change of $P_i(\theta)$ in the vicinity of $\theta = b_i$. For each item combination, we calculated Cohen's d for each parameter as a measure of the effect size of the difference in the IRT parameter values between PIQLet1 and PIQLet2: effect size of a_i (d_a), effect size of b_i (d_b), and effect size of c_i (d_c). We also calculated a magnitude of the IRT parameter effect size for each combination,

$$|d| = \sqrt{d_a^2 + d_b^2 + d_c^2}. \quad (3)$$

All analyses were performed using the R computing environment [16–19].

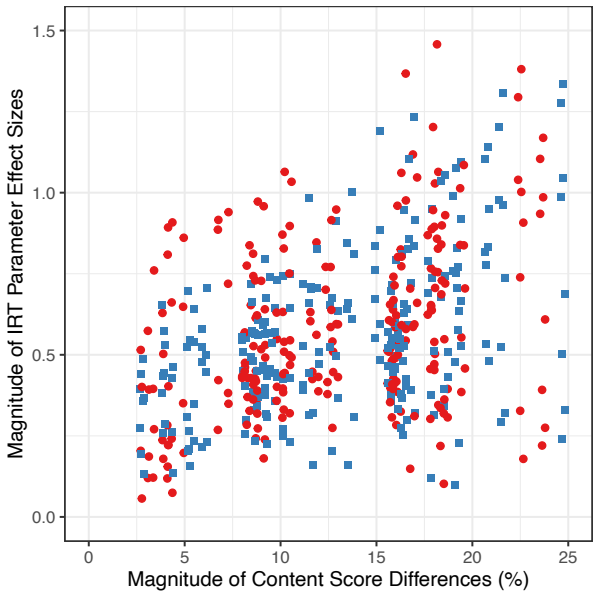

FIG. 1. Magnitude of the subscore differences $|\Delta s|$ (in %) and the magnitude of the IRT parameter effect sizes $|d|$ for all 480 combinations of items. Red circles represent item combinations which had item 7 as an anchor item; blue squares represent item combinations which had item 1 as an anchor item.

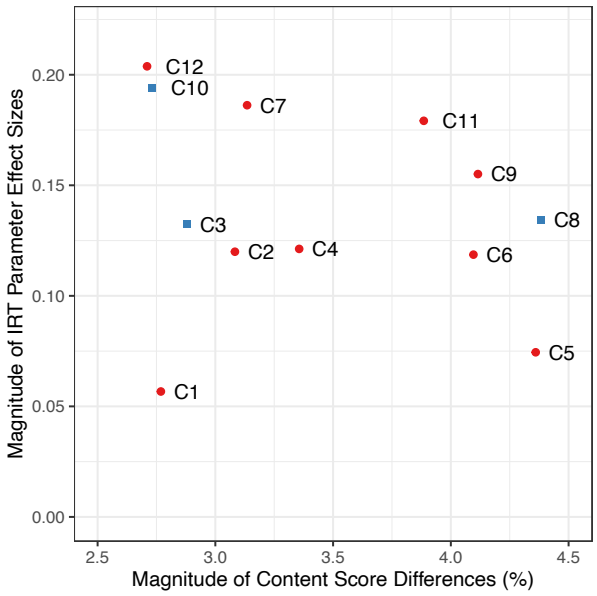

FIG. 2. A zoomed in view of Fig. 1, showing the 12 item combinations with the lowest combined differences in scores and IRT parameters.

III. RESULTS

Figure 1 shows the magnitude of the subscore differences $|\Delta s|$ (in percentage) and the magnitude of the IRT parameter effect sizes $|d|$ for all 480 combinations of items. Red circles represent item combinations (PIQLet pairs) that had item 7 as an anchor item, and blue squares represent PIQLet pairs that had item 1 as an anchor item. There is a wide variety of results, with score difference magnitudes ranging from under 3% to almost 25%, and effect size magnitudes ranging from less than 0.1 (trivial) to almost 1.5 (very large).

Figure 2 shows the bottom left corner of Fig. 1. This portion of the plot contains data points representing PIQLet pairs with the smallest values of both subscore differences and IRT parameter effect sizes. The two combinations with the lowest parameter effect sizes are in this group (combinations labeled C1 and C5 in Fig. 2). There are two item combinations with smaller score differences than those shown in Fig. 2 ($|\Delta s|$ = 2.67% compared to 2.71% for the combination labeled C12), but these are not included in the smaller plot because they had significantly higher effect sizes ($|d|$ = 0.39 and 0.27, compared to 0.20 for combination C12). One interesting feature of the 12 PIQLet pairs included in Fig. 2 is that nine of them are red, indicating that item 7 was an anchor item for those PIQLets. This suggests that using item 7 as an anchor item may lead to more similar PIQLets than item 1.

The labels in Fig. 2 represent the combinations of items included in each PIQLet. Table III provides more details for these top 12 combination options, including the effect size of the difference in each of the IRT parameter values, the magnitude of the IRT parameter effect size $|d|$, the difference in each PQL facet subscore, the magnitude of the subscore differences $|\Delta s|$, and the average overall score difference between the PIQLets for each combination. The combination labels are ordered according to how close each combination's data point appears to the bottom left corner of Fig. 2. All of these combinations seem to provide PIQLets that are very similar across all six measures: no effect sizes are above 0.2 (considered small according to Cohen's definitions [20]), and all subscores have differences smaller than 4%. Additionally, 11 out of the 12 have average overall score differences with magnitudes that are less than 2%, and five of them are less than 1%.

As seen in Table III, combination C1 has the lowest effect size magnitude (RQ2). Following the same naming convention, combination C38 (not shown in Fig. 2 or Table III) has the smallest subscore difference magnitude (RQ1); however, the difference in $|\Delta s|$ between C1 and C38 is only 0.1% (RQ3). Additionally, C1 has an overall PIQLet score difference of less than 1%, making it a strong contender for the combination of items that produces the most similar PIQLets.

IV. DISCUSSION

We can identify several consistencies when examining the items included in each of the 12 combinations included in Fig. 2 and Table III. In all 12 combinations, the items designed to assess proportional reasoning are arranged in the same way, with item 2 always on the same PIQLet as item 11, and item

			IRT Parameter Effect Sizes				Subscore Differences (%)				Overall Score
Combination	Plot Color	Anchor Items	d_a	d_b	d_c	$\|d\|$	Δs_{pr}	Δs_{nr}	Δs_{cr}	$\|\Delta s\|$	Difference (%)
C1	red	7, 10, 15, 20	0.04	0.002	0.04	0.06	-1.1	-2.5	0.5	2.8	-0.9
C2	red	7, 10, 15, 20	0.11	0.04	-0.02	0.12	1.0	2.5	1.5	3.1	1.9
C3	blue	1, 10, 15, 20	0.12	-0.04	-0.03	0.13	-1.1	-2.5	-0.9	2.9	-1.6
C4	red	7, 10, 15, 20	0.03	-0.02	0.12	0.12	1.0	-2.5	2.0	3.4	0.5
C5	red	7, 10, 15, 20	0.05	-0.01	-0.05	0.07	-1.1	-3.7	2.0	4.4	-0.6
C6	red	7, 10, 15, 20	0.08	0.05	0.07	0.12	1.0	-2.5	3.1	4.1	1.1
C7	red	7, 10, 15, 20	0.12	0.11	0.10	0.19	-1.1	-2.5	1.5	3.1	-0.4
C8	blue	1, 10, 15, 20	-0.12	0.05	-0.03	0.13	-1.1	-2.5	3.4	4.4	0.5
C9	red	7, 10, 15, 20	-0.05	0.04	-0.14	0.16	1.0	3.7	1.5	4.1	2.3
C10	blue	1, 10, 15, 20	0.18	0.07	0.00	0.19	-1.1	-2.5	0.1	2.7	-1.1
C11	red	7, 10, 15, 20	0.17	-0.03	0.04	0.18	-1.1	-3.7	0.5	3.9	-1.3
C12	red	7, 10, 15, 20	0.06	-0.09	-0.17	0.20	1.0	2.5	0.5	2.7	1.4

TABLE III. Details of the 12 combinations shown in Fig. 2. The final column shows the difference in the overall scores for each PIQLet pair (not included in Fig. 2). Positive values indicate that PIQLet2 was higher, and negative values indicate that PIQLet1 was higher.

6 always with item 13. This consistency suggests that these pairings may yield the most similar PIQLet versions. In eight of the 12 combinations, items 17 and 18 are included on different PIQLets; this is a desirable result because these are the only MCMR PIQL items that have more than one correct response. Also in eight of the 12 combinations, items 1, 7, and 17 are all included in the same PIQLet (sometimes with item 3, and sometimes with item 8). The similarities among the items within each of these top 12 combinations provides evidence that our analyses are revealing patterns within the data that may be used to identify an optimal combination of items that would produce an equivalent PIQLet pair.

In addition to these results based on class-averaged scores, we can also calculate each student's score on PIQLet1 and PIQLet2. For combination C1, these scores are highly correlated, with a Pearson correlation between PIQLet1 scores and PIQLet2 scores of $r_{12} = 0.83$. These results show that the PIQL can be divided into two shorter versions that show strong statistical and psychometric similarities, both for individuals and averaged across data sets. Using IRT analyses along with statistical analyses of score distributions provides additional evidence of similarities between PIQLet versions.

V. LIMITATIONS AND FUTURE WORK

While these results show great promise for creating equivalent PIQLet versions, more work is needed to provide evidence for the optimal combination of items. The current study was conducted using data from a single institution, collected in a narrow window of time (between the first two calculus-based introductory physics courses). To be confident in these results, they would need to be verified through analyzing data from other courses and other institutions. We will also need to administer the PIQLets in a controlled trial (similar to Han, et al. [21]) to confirm their equivalence experimentally in addition to the evidence determined by analyzing data collected using the full PIQL. These comparisons will rely heavily on analyzing student performance on the anchor items to establish equivalence between data sets and test performance.

In addition to requiring more analysis to establish evidence of PIQLet similarities, we will also need to establish the validity and reliability of each PIQLet individually, as was done with the PIQL [1]. This will involve analyses of item difficulty and discrimination, as well as factor analyses to ensure that each PIQLet performs the same way as the full PIQL.

Additional work is also needed to be able to equate student scores from one test to scores from the other. Our ultimate goal is to be able to use a student's score on one PIQLet as a proxy for what their score would be if they had completed the other PIQLet, as well as what their score would be if they had completed the full PIQL. This would allow us to compare student scores across any of the three tests. Other researchers have used the Stocking-Lord method to calculate transformation coefficients, which can be used to relate scores from two different tests [22, 23]. We plan to explore this and other methods for relating and equating scores from different tests. Rigorously validated score transformation relationships will be necessary to appropriately measure learning and growth when using different PIQLets before and after instruction.

ACKNOWLEDGMENTS

We thank Zachary Bischoff and Jack Sayers for useful contributions to preliminary parts of this project. This work was partially supported by the National Science Foundation under awards DUE-2214765, DUE-2214283, and DUE-1836470.

[1] S. White Brahmia, A. Olsho, T. I. Smith, A. Boudreaux, P. Eaton, and C. Zimmerman, Physics Inventory of Quantitative Literacy: A tool for assessing mathematical reasoning in introductory physics, Phys. Rev. Phys. Educ. Res. **17**, 020129 (2021).

[2] S. White Brahmia, A. Olsho, A. Boudreaux, T. Smith, and C. Zimmermann, A conceptual blend analysis of physics quantitative literacy reasoning inventory items, in *Proceedings of the 23rd Annual Conference on Research in Undergraduate Mathematics Education*, edited by S. Smith Karunakaran, Z. Reed, and A. Higgins (Boston, MA, 2020) pp. 853–858.

[3] T. I. Smith, P. Eaton, S. White Brahmia, A. Olsho, A. Boudreaux, and C. Zimmerman, Toward a valid instrument for measuring physics quantitative literacy, in *Physics Education Research Conference 2020*, PER Conference, edited by S. Wolf, M. Bennett, and B. Frank (Virtual Conference, 2020) pp. 496–502.

[4] T. I. Smith, S. White Brahmia, A. Olsho, and A. Boudreaux, Physics Students' Implicit Connections Between Mathematical Ideas, in *Proceedings of the 23rd Annual Conference on Research in Undergraduate Mathematics Education*, edited by S. Smith Karunakaran, Z. Reed, and A. Higgins (Boston, MA, 2020) pp. 940–946.

[5] T. I. Smith, P. Eaton, S. White Brahmia, A. Olsho, C. Zimmerman, and A. Boudreaux, Analyzing multiple-choice-multiple-response items using item response theory, in *Physics Education Research Conference 2022*, PER Conference, edited by B. W. Frank, D. L. Jones, and Q. X. Ryan (Grand Rapids, MI, 2022) pp. 432–437.

[6] A. Olsho, S. White Brahmia, T. I. Smith, C. Zimmerman, A. Boudreaux, and P. Eaton, Online administration of a reasoning inventory in development, in *Physics Education Research Conference 2020*, PER Conference, edited by S. Wolf, M. Bennett, and B. Frank (Virtual Conference, 2020) pp. 382–387.

[7] A. Olsho, T. I. Smith, P. Eaton, C. Zimmerman, A. Boudreaux, and S. White Brahmia, Online test administration results in students selecting more responses to multiple-choice-multiple-response items, Phys. Rev. Phys. Educ. Res. **19**, 013101 (2023).

[8] T. I. Smith, P. Eataon, A. Olsho, and S. White Brahmia, Enriching assessment of physics quantitative literacy: IRT with multiple-response items, in *Proceedings of the 25th Annual Conference on Research in Undergraduate Mathematics Education*, edited by S. Cook, B. Katz, and D. Moore-Russo (Omaha, NE, 2023) pp. 1346–1347.

[9] PhysPort, PhysPort, https://www.physport.org/assessments/assessment.cfm?A=PIQL (2022).

[10] D. Hestenes, M. Wells, and G. Swackhamer, Force concept inventory, The Physics Teacher **30**, 141 (1992).

[11] J. Han, L. Bao, L. Chen, T. Cai, Y. Pi, S. Zhou, Y. Tu, and K. Koenig, Dividing the force concept inventory into two equivalent half-length tests, Phys. Rev. ST Phys. Educ. Res. **11**, 010112 (2015).

[12] D. P. Maloney, T. L. O'Kuma, C. J. Hieggelke, and A. Van Heuvelen, Surveying students' conceptual knowledge of electricity and magnetism, American Journal of Physics **69**, S12 (2001), https://pubs.aip.org/aapt/ajp/article-pdf/69/S1/S12/7529618/s12_1_online.pdf.

[13] Y. Xiao, K. Koenig, J. Han, Q. Liu, J. Xiong, and L. Bao, Test equity in developing short version conceptual inventories: A case study on the conceptual survey of electricity and magnetism, Phys. Rev. Phys. Educ. Res. **15**, 010122 (2019).

[14] J. Wang and L. Bao, Analyzing force concept inventory with item response theory, Am. J. Phys. **78**, 1064 (2010).

[15] T. I. Smith, Z. Bischoff, B. Boyle, J. Sayers, C. Zimmerman, P. Eaton, A. Olsho, and S. White Brahmia, Creating statistically equivalent versions of a test of quantitative literacy in physics contexts, in *Proceedings of the 26th Annual Conference on Research in Undergraduate Mathematics Education* (Omaha, NE, (in press)).

[16] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria (2024).

[17] H. Wickham, M. Averick, J. Bryan, W. Chang, L. D. McGowan, R. François, G. Grolemund, A. Hayes, L. Henry, J. Hester, M. Kuhn, T. L. Pedersen, E. Miller, S. M. Bache, K. Müller, J. Ooms, D. Robinson, D. P. Seidel, V. Spinu, K. Takahashi, D. Vaughan, C. Wilke, K. Woo, and H. Yutani, Welcome to the tidyverse, Journal of Open Source Software **4**, 1686 (2019).

[18] R. P. Chalmers, Multidimensional Item Response Theory (mirt) (2017).

[19] R. P. Chalmers, mirt: A Multidimensional Item Response Theory Package for the R Environment, Journal of Statistical Software **48**, 10.18637/jss.v048.i06 (2012).

[20] J. Cohen, *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed. (Lawrence Erlbaum Associates, 1988).

[21] J. Han, K. Koenig, L. Cui, J. Fritchman, D. Li, W. Sun, Z. Fu, and L. Bao, Experimental validation of the half-length force concept inventory, Phys. Rev. Phys. Educ. Res. **12**, 020122 (2016).

[22] S. Kim and W.-C. Lee, IRT scale linking methods for mixed-format tests, in *ACT Research Report Series 2004-5* (ACT, Inc., Iowa City, IA, 2004).

[23] Y. Xiao, J. C. Fritchman, J. Y. Bao, Y. Nie, J. Han, J. Xiong, H. Xiao, and L. Bao, Linking and comparing short and full-length concept inventories of electricity and magnetism using item response theory, Phys. Rev. Phys. Educ. Res. **15**, 020149 (2019).