Creating Statistically Equivalent Versions of a Test of Quantitative Literacy in Physics Contexts

Trevor I. Smith
Rowan University

Zachary Bischoff
Rowan University

Brett Boyle
Rowan University

Jack Sayers
Rowan University

Charlotte Zimmerman
University of Washington

Philip Eaton
Stockton University

Alexis Olsho
United States Air Force Academy

Suzanne White Brahmia
University of Washington

Keywords: quantitative literacy, assessment, physics

The Physics Inventory of Quantitative Literacy (PIQL) is a 20-item multiple-choice test designed to measure the development of students' physics quantitative literacy (PQL) across multiple physics courses (Olsho et al., 2023; Smith et al., 2020; White Brahmia et al., 2021). Repeated testing, coupled with requiring up to 40 minutes for students to complete the test, could lead to testing fatigue and unreliable results. We seek to create two shorter versions of the PIQL (a.k.a. piqlets) that are statistically equivalent to each other in terms of student performance on three facets of PQL (ratios and proportions, covariation, and signs and negativity).

Han et al. (2015) used a large data set of student responses to a 30-item conceptual physics test to identify combinations of items that produced testlets with the most similar average student scores. They demonstrated the equivalence of the testlets using item response theory (IRT) and correlating individual students' scores across versions. We follow their example by creating 12-item piqlets that each contain four overlapping anchor items (to facilitate comparisons between versions) and eight distinct items. Our work was guided by these research questions:

1) Which combination of items produces piqlets with the smallest score differences?
2) How does the reliability of these piqlets compare to each other and to the PIQL?
3) How similar are the psychometric parameters for the anchor items across piqlets?

Data were collected using the full PIQL in three introductory physics courses at a large public university in the western US (2100–3200 students in each data set). We considered 240 combinations of items for the piqlet versions, subject to constraints that both the content and the format of items were equivalent across the two versions. We calculated a total test score for each piqlet based on the average percentage of items answered correctly, as well as subscores for the three facets of PQL. For each of our three data sets we determined the average score difference between the piqlets, calculated Cronbach's α for each piqlet, and applied IRT analyses to each.

The combinations of items that we identified as being the most similar had overall average score differences from 0.6–1.3%. Cronbach's α values ranged from 0.67 to 0.75, with differences less than 0.01 between versions. We see strong correlations between individual student scores on the two piqlets, with $0.79 \leq r \leq 0.85$. The IRT parameters support the statistical equivalence of the piqlets with parameters of overlapping items agreeing to within 0.1. These preliminary results suggest a strong potential for identifying piqlets that are statistically equivalent for the broader population of mathematics and physics students based on a larger, more diverse, data set.

Acknowledgments

This work is supported by the NSF under Award Nos. DUE-1832470, DUE-2214283, and DUE-2214765; and an NRC Research Associateship award at Air Force Research Laboratory.

References

Han, J., Bao, L., Chen, L., Cai, T., Pi, Y., Zhou, S., Tu, Y., & Koenig, K. (2015). Dividing the Force Concept Inventory into two equivalent half-length tests. *Physical Review Special Topics – Physics Education Research, 11*(1), 010112. https://doi.org/10.1103/PhysRevSTPER.11.010112

Olsho, A., Smith, T. I., Eaton, P., Zimmerman, C., Boudreaux, A., & White Brahmia, S. (2023). Online test administration results in students selecting more responses to multiple-choice-multiple-response items. *Physical Review Physics Education Research*, *19*(1), 013101. https://doi.org/10.1103/PhysRevPhysEducRes.19.013101

Smith, T. I., White Brahmia, S., Olsho, A., & Boudreaux, A. (2020). Developing a reasoning inventory for measuring physics quantitative literacy. In S. Smith Karunakaran, Z. Reed, & A. Higgins (Eds.), *Proceedings of the 23rd Annual Conference on Research in Undergraduate Mathematics Education.* (pp. 949–955). http://sigmaa.maa.org/rume/RUME23.pdf

White Brahmia, S., Olsho, A., Smith, T. I., Boudreaux, A., Eaton, P., & Zimmerman, C. (2021). Physics Inventory of Quantitative Literacy: A tool for assessing mathematical reasoning in introductory physics. *Physical Review Physics Education Research*, *17*(2), 020129. https://doi.org/10.1103/PhysRevPhysEducRes.17.020129

Certain combinations of PIQL test items create statistically equivalent testlets

Creating Statistically Equivalent Versions of a Test of Quantitative Literacy in Physics Contexts

Trevor I. Smith[1], Zachary Bischoff[1], Brett Boyle[1], Jack Sayers[1], Charlotte Zimmerman[2], Philip Eaton[3], Alexis Olsho[4], and Suzanne White Brahmia[2]

Introduction

- Improving students' physics quantitative literacy (PQL) is a goal for many physics instructors.
- PQL is the blended relationship between students' general mathematical reasoning and conceptual understanding of physics.
- The **P**hysics **I**nventory of **Q**uantitative **L**iteracy (PIQL) is a 20-item multiple-choice test designed to measure PQL across three facets (Olsho et al., 2023; Smith et al., 2020; White Brahmia et al., 2021).
- PIQL used in multiple physics courses to track student growth
- Some items are multiple-choice-multiple-response (MCMR)
- **Known Problems:**
  - The PIQL takes a long time to complete (40 minutes on average)
  - Repeated testing risks memorization and reduced efficacy
- **GOAL: Create two shorter versions of the PIQL (a.k.a. PIQLets) that are statistically and psychometrically equivalent**

Research Questions

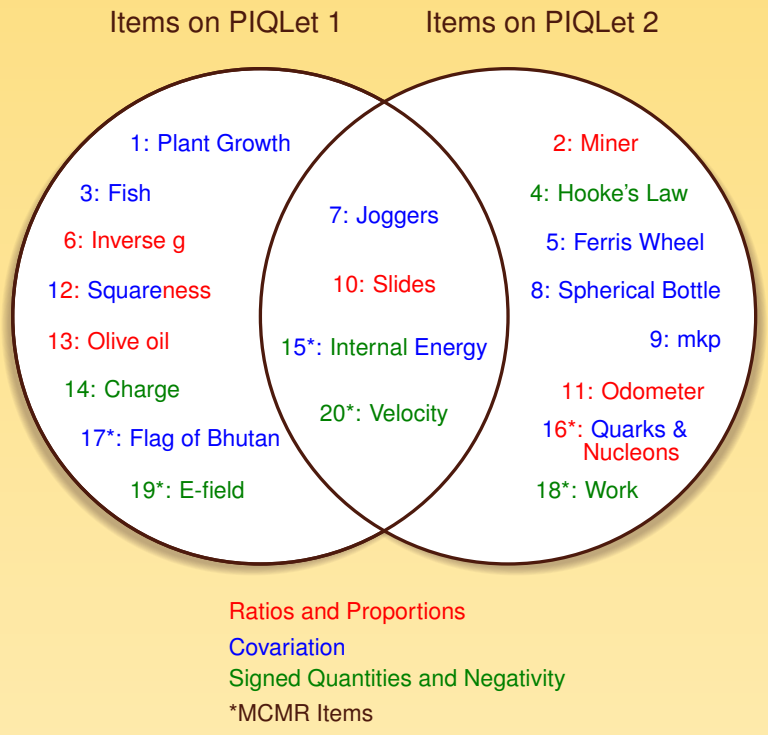
1. Which combination of items produces PIQLets with the smallest score differences? (Figure 1, central Venn diagram)
2. How does the reliability of these PIQLets compare to each other and to the PIQL? (Table 1)
3. How similar are the psychometric parameters for the anchor items across PIQLets? (Table 3)

Constraints and Criteria

- Constraints: each pair of PIQLets must have:
  - 12 items on each PIQLet: 4 anchor items in common, 8 unique items
  - the same number of items (anchor and unique) testing each PQL facet
  - the same number of items in each format (single response or MCMR)
- Criteria: the best pair of PIQLets should have:
  - similar total average scores for all three data sets
  - similar average subscores for all three facets
  - good test reliability (Cronbach's α)
  - similar item response theory (IRT) parameters for anchor items

Methods

- Followed example of Han et al. (2015)
- Full 20-item PIQL administered 3 times during introductory physics course sequence at a public university in the Western US
  - Before mechanics (PreMech, $N = 3206$)
  - After mechanics, before electricity & magnetism (PostMech, $N = 2580$)
  - After electricity & magnetism (PostEM, $N = 2136$)
- Each item categorized based on content and format
- Identified 5 candidates for anchor items based on prior analyses
- Identified all combinations of items based on constraints (240 total)
- For each combination of items:
  - Split the full data set into two parts
  - Calculated the differences in the mean total score and mean subscores for each facet for each data set (total of 12 average score differences)
  - Calculate the mean $\langle\Delta\rangle$ and the root-mean-squared Δ_{rms} of the 12 score differences as measures of the total difference between the PIQLets
- Identified a handful of combinations that had the lowest Δ_{rms} and $\langle\Delta\rangle$ close to zero.
  - Calculated Cronbach's α for each PIQLet for each data set
  - Identified the combination with the most similar α values
- For the single most similar PIQLet pair:
  - Applied a three-parameter logistic (3PL) model to calculate IRT item parameters
  - Compared parameters for anchor items between the two PIQLets
- All analyses performed using the `r` computing environment (R Core Team, 2019), specifically using the `mirt` and `tidyverse` packages



Scan for poster

PIQL on PhysPort
https://www.physport.org

26th Annual Conference on Research in Undergraduate Mathematics Education

[1] Rowan University
[2] University of Washington
[3] Stockton University
[4] United States Air Force Academy

DUE-1832470, DUE-2214283, DUE-2214765; NRC Research Associateship

Scores

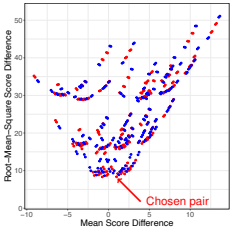


Figure 1. Mean percentage score difference ($\langle\Delta\rangle$) and root-mean-square score difference Δ_{rms} for all PIQLet item combinations. Red points have anchor items 7, 10, 15, and 20; blue points have anchor items 1, 10, 15, and 20.

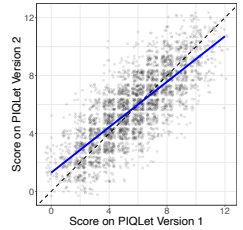


Figure 2. Comparison of PreMech students' scores on the PIQLets. The blue line shows the relationship between the scores on the two versions. The black dashed line represents equal scores on the two versions.

For the chosen pair of PIQLets: the mean total score difference for PreMech data is 1.6%, the mean difference for PostMech data is 1.1%, and the mean difference for PostEM data is 1.3%. In all three data sets PIQLet 2 has the higher mean score.

Statistical Comparisons

Test	PreMech	PostMech	PostEM
PIQL	0.76	0.80	0.82
PIQLet 1	0.68	0.73	0.75
PIQLet 2	0.67	0.72	0.75

Table 1. Cronbach's α for the full PIQL and the closest combination of PIQLets.

Comparison	PreMech	PostMech	PostEM
PIQLet 1 v. P2	0.79	0.82	0.84
PIQL v. P1	0.94	0.94	0.95
PIQL v. P2	0.93	0.94	0.95

Table 2. Pearson's r correlation between scores on the two PIQLet versions, as well as between each PIQLet and the full PIQL.

IRT Parameters

$$\text{3PL Model:} \qquad P_{i,j}(1|\theta_i) = c_j + \frac{1-c_j}{1+e^{-1.7a_j(\theta_i - b_j)}}$$

The probability that student i answers item j correctly, given their overall ability and understanding θ_i (i.e., PQL). Each item has three estimated parameters: discrimination a_j, difficulty b_j, and guessing c_j.

	Discrimination a_j			Difficulty b_j			Guessing c_j		
Item	PIQL	P1	P2	PIQL	P1	P2	PIQL	P1	P2
7	0.72	0.73	0.75	-0.27	-0.28	-0.22	0.01	0.00	0.03
10	1.10	1.14	1.09	-0.78	-0.77	-0.79	0.00	0.00	0.00
15	0.93	0.99	0.89	0.40	0.43	0.40	0.01	0.03	0.01
20	1.02	1.08	1.06	0.01	0.03	0.00	0.00	0.01	0.00

Table 3. 3PL IRT parameters for anchor items for the full PIQL and each PIQLet version for our chosen combination.

Summary

- Categorizing testlet similarity by mean scores yielded a pair of PIQLets with very small differences in mean scores, but these differences vary greatly across item combinations (Figure 1).
- Scores are less similar for students at the high and low ends (Figure 2).
- Cronbach's α is very similar across the chosen pair of PIQLets, but slightly smaller than the total PIQL (Table 1). This is expected, given the smaller number of items (Streiner, 2003).
- Student scores across PIQLet versions are strongly correlated (Table 2).
- IRTparameter values for the anchor items are very similar across the two chosen PIQLet versions: a_j differs by 0.1 or less, b_j by 0.06 or less, and c_j by 0.03 or less (Table 3).

Next Steps

- Define difference between PIQLets based on more than mean scores
- Use IRT analyses to compare other PIQLet items based on IRT parameters for anchors items and overall distribution of parameters
- Apply the Stocking-Lord method to define transformations between PIQLets, looking for combinations that require the smallest adjustments

References

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

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

R Core Team (2019). R: A Language and Environment for Statistical Computing.

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

Streiner, D. L. (2003). Starting at the beginning: An introduction to coefficient alpha and internal consistency. *Journal of Personality Assessment*, 80(1):99–103. PMID: 12584072.

White Brahmia, S., Olsho, A., Smith, T. I., Boudreaux, A., Eaton, P., and Zimmerman, C. (2021). Physics Inventory of Quantitative Literacy: A tool for assessing mathematical reasoning in introductory physics. *Phys. Rev. Phys. Educ. Res.*, 17(2):020129.