Online test administration results in students selecting more responses to multiple-choice-multiple-response items

Alexis Olsho,[1] Trevor I. Smith,[2] Philip Eaton,[3] Charlotte Zimmerman,[4] Andrew Boudreaux,[5] and Suzanne White Brahmia[4]

[1]*Department of Physics and Meteorology, United States Air Force Academy, 2354 Fairchild Drive, USAF Academy, Colorado 80840, USA*
[2]*Department of Physics and Astronomy and Department of STEAM Education, Rowan University, 201 Mullica Hill Rd., Glassboro, New Jersey 08028, USA*
[3]*School of Natural Sciences and Mathematics, Stockton University, Galloway, New Jersey 08205, USA*
[4]*Department of Physics, University of Washington, Box 351560, Seattle, Washington 98195-1560, USA*
[5]*Department of Physics and Astronomy, Western Washington University, 516 High St., Bellingham, Washington 98225, USA*

(Received 22 August 2022; accepted 20 December 2022; published 2 February 2023)

We developed the Physics Inventory of Quantitative Literacy (PIQL) to assess students' quantitative reasoning in introductory physics contexts. The PIQL includes several "multiple-choice-multiple-response" (MCMR) items (i.e., multiple-choice questions for which more than one response may be selected) as well as traditional single-response multiple-choice items. In this paper, we discuss differences in performance on MCMR items that seems to result from differences in administration method (paper versus online). In particular, we find a tendency for "clickiness" in online administration: students choose more responses to MCMR items when taking the electronic version of the assessment. Student performance on single-response multiple-choice items was not affected by administration method. These results suggest that MCMR items may provide a unique opportunity to probe differences in online and on-paper administration of low-stakes assessments.

DOI: 10.1103/PhysRevPhysEducRes.19.013101

I. INTRODUCTION AND BACKGROUND

The Physics Inventory of Quantitative Literacy (PIQL) is a research-based assessment (RBA) that assesses students' quantitative reasoning in introductory physics contexts [1]. The PIQL includes several "multiple-choice-multiple-response" (MCMR) items: multiple-choice items for which there may be more than one correct response and students are encouraged to select all answers that apply (items 15–20 in PIQL v2.2 [1]).

A PIQL MCMR item is shown in Fig. 1.

The PIQL was originally developed as an on-paper, proctored assessment, but shifted to online administration during the COVID-19 pandemic. Best practices are established for administration of RBAs both in-person (on paper or electronically) and as an out-of-class, online activity [2,3]. Substantial research suggests that RBAs originally designed to be given in-person can be administered online without affecting student performance [4–6], though researchers recommend that instruments are validated separately for online, unproctored use [4].

Online RBAs offer important affordances associated with easing the logistics of test administration [7]. Therefore, we began to collect evidence of the PIQL's validity as an online assessment, with an eye toward dissemination of the PIQL as an online instrument. While preliminary research revealed that student performance on the PIQL was largely unaffected by administration method, we found significant differences in the number of responses provided by students on PIQL MCMR items [7].

Little research comparing student test-taking behavior for MCMR items based on administration format exists. While student performance on coupled multiple-response items on the Colorado upper-division electrostatics (CUE) diagnostic was similar across online and in-person administration methods, exploration of differences was not a focus of the work [8]. We also note that differences in CUE and PIQL MCMR items, and target populations limit the applicability of Wilcox and Pollock's work to ours.

In this paper, we explore how administration method affects introductory physics students' performance on PIQL MCMR items. We discuss possible insights into how students interact with online, low-stakes assessments

A hand exerts a constant, horizontal force on a block as the block moves along a frictionless, horizontal surface. No other objects do work on the block. For a particular interval of the motion, the hand does $W = -2.7$ J of work on the block. Recall that for a constant force, $W = \vec{F} \cdot \Delta\vec{s}$.

Consider the following statements about this situation. Select the statement(s) that **must be true.**
Choose all that apply.

a. The work done by the hand is in the negative direction.
b. The force exerted by the hand is in the negative direction.
c. The displacement of the block is in the negative direction.
d. The force exerted by the hand is in the direction *opposite to* the block's displacement.
e. The force exerted by the hand is in the direction *parallel to* the block's displacement.
f. Energy was added to the block system.
g. Energy was taken away from the block system.

FIG. 1. PIQL MCMR item that probes understanding of the negative sign in the context of mechanical work. The correct answers are d and g.

that are not evident from responses to traditional, single-response multiple-choice items.

II. METHODS

All data discussed in this paper were collected at the beginning of the academic term. Most data come from students enrolled in the first quarter of calculus-based introductory physics at a large public research university in the Western U.S. ("Institution 1"). Additional data come from students enrolled in the first semester of introductory calculus-based physics at a U.S. military academy ("Institution 2").

At Institution 1, data were collected over ten academic terms. In four of these, the PIQL was administered in a course recitation session proctored by a teaching assistant (TA). Students ($N = 993$) read items from a 5-page packet and recorded their responses on a paper answer form as well as electronically. For each of these in-person administrations of the PIQL, the MCMR items were interspersed with single-response (SR) items throughout the test. Before starting, the students were informed of the purpose of the PIQL by the TA. Students were reminded in multiple ways that they could choose more than one response on MCMR items: verbally, by the proctor, and in writing at the top of each page of the packet that contained an MCMR item. Finally, students were prompted to "choose all that apply" in the question stem for each MCMR item. Students were expected to finish by the end of the 50-min recitation session. In our experience, most finished within 40 min.

For the other six terms at Institution 1 ($N = 1396$), the PIQL was administered unproctored and online using the University's existing survey or quiz platform. When administered online, the PIQL had a 50-min time limit. Each item was shown in a browser window on its own; students were not able to backtrack in the PIQL [9]. Timestamps were collected for students taking the PIQL online; data from students that took less than 10 min to complete the PIQL were removed in an attempt to exclude students that simply "clicked through" the questions.

With one exception (described below), all MCMR items were moved to the end of the PIQL for online administration. After answering the last SR item, students saw a page with no item, but rather a statement that the remaining questions on the survey might have more than one correct response, and that students should choose all answers that they feel are correct. At the top of each page for the remaining items (all MCMR), there was a reminder that the question might have more than one correct response. As in the in-person administration, the question stem for each MCMR item prompted students to choose all that apply."

At Institution 2, the PIQL was administered in-person during a regular class period, and proctored by the course instructor. Students ($N = 282$) read items from a 5-page packet and recorded their responses on a bubble sheet. MCMR items were grouped at the end of PIQL, with a reminder at the top of pages with MCMR items (the last two pages of the test).

This research focuses on possible effects of administration method on MCMR performance and answer choices. We recognize that changing item order (i.e., moving all of the MCMR items to the end of the assessment) may also affect student responses. To assess the effect of question order on the number of responses chosen, we presented the PIQL online to a class of students ($N = 83$) at Institution 1. Approximately half the students ($N = 43$) saw MCMR items grouped at the end of the PIQL. The remaining 40 students saw MCMR items interspersed with the SR items. These data were used only to explore the effect of question order.

III. RESULTS

To explore whether administration method (i.e., electronic or on-paper) affects the number of responses chosen, we compare data collected with the paper version of the PIQL to those collected online. We calculated the average number of responses for each of the six MCMR items for both online ($N = 1313$) and in-person (paper) ($N = 993$) administrations at Institution 1. The results are shown in Fig. 2. For each of the six items, the difference in the average number of answer choices is statistically significant (Mann-Whitney U test, $p < 0.001$ for all items), with effect size ranging from small to medium (rank-biserial

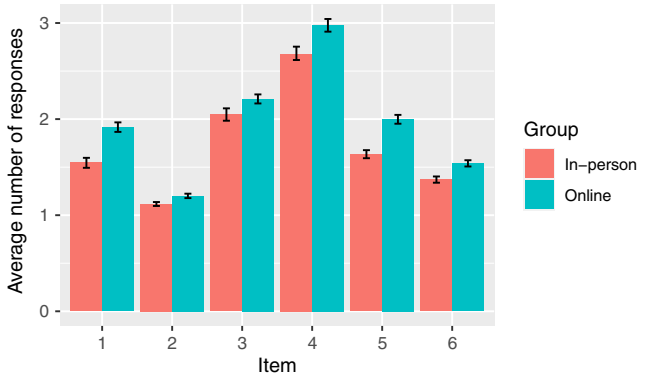

FIG. 2. Average number of answer choices selected by Institution 1 students for the six MCMR PIQL items. Items 1, 2, 5, and 6 have one correct answer choice; item 3 has three correct answer choices, and item 4 has two.

correlation coefficients between −0.0816 and −0.259). Figure 3 shows the distributions of number of answer choices for each of the six items. The differences in averages seem to be due to students choosing one or two more responses online compared to on paper. Of the six PIQL MCMR items, items 1, 2, 5, and 6 have one correct answer choice; item 3 has three correct answer choices, and item 4 has two. Error bars in all plots indicate 95% confidence interval calculated using 1000 bootstrap iterations [10–12].

To assess the effect of question order on the number of responses, we compared the average number of answer choices for each of the MCMR items for students that saw the MCMR items grouped or ungrouped during online PIQL administration. Results are shown in Fig. 4. While the number of answer choices is statistically significantly different for most items, there is no systematic effect.

Preliminary data collected at Institution 2 ($N = 282$) and Institution 1 ($N = 993$) using paper versions of the PIQL in Fig. 5 also suggest that there is no systematic significant difference in the number of answer choices selected when items are grouped or ungrouped. We note, however, that comparing these populations of students is difficult, as overall PIQL performance and MCMR response patterns

FIG. 3. Fraction of student responses that include 1–6 answer choices for each item. Items 1 and 6 have six answer choices; 2, 3, and 5 have five answer choices; and 4 has seven answer choices. (No student chose all seven of item 4's answer choices.).

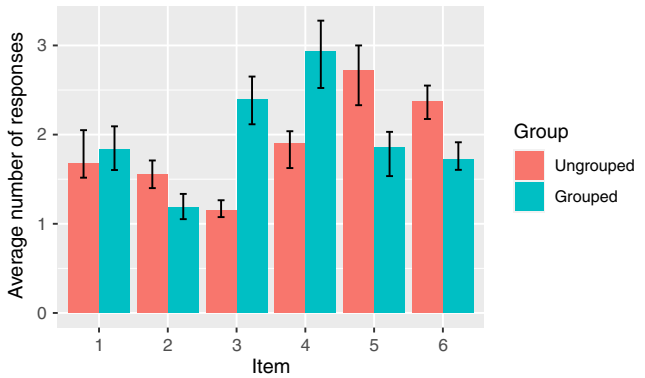

FIG. 4. Average number of answer choices selected online by Institution 1 students for the six MCMR PIQL items with items grouped or ungrouped.

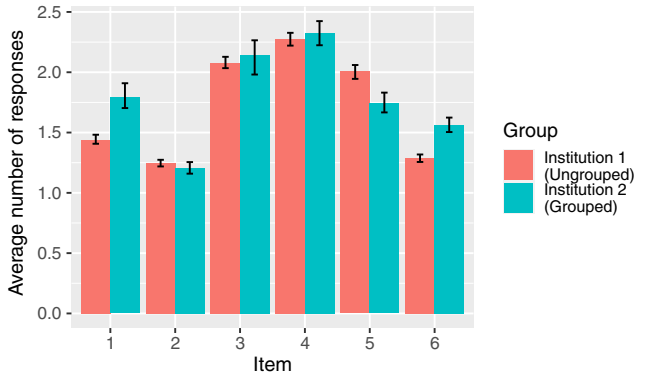

FIG. 5. Average number of answer choices selected on paper by Institutions 1 and 2 students for the six MCMR PIQL items with items grouped or ungrouped.

differed. These results suggest that question order does not have a strong effect on number of answer choices selected, but also indicate more in-depth research would be appropriate.

IV. DISCUSSION

While we cannot be sure that differences in performance on MCMR items is due only to differences in administration method, because PIQL data were collected at the beginning of each academic term before substantial instruction, we believe that differences in instructor or course materials can be ignored. To assess possible changes to the student population, we used data from three steady-state SR PIQL items. We found no significant difference in student performance on these items based on administration method, or whether instruction took place in-person or virtually. The tendency for students to select more responses online seems to represent a difference in the ways in which students interact with online and on-paper assessments that can only be measured using MCMR items.

For MCMR items, dichotomous scoring methods require a student to choose *all* correct responses and *only* correct responses to be considered correct. For example, MCMR item 4 on the PIQL (the Work item shown in Fig. 1) has two correct answer choices: d and g. In a dichotomous scoring scheme a student who picks only answer d would be scored the same way as a student who chooses answers e and f (both incorrect). This ignores the nuance and complexity of students' response patterns within (and between) items. In an effort to move beyond the constraints of dichotomous scoring for MCMR items, we have developed a four-level scoring scale in which we categorize students' responses as Completely Correct, Some Correct (if at least one but not all correct response choices are chosen), Both Correct and Incorrect (if at least one correct and one incorrect response choices are chosen), and Completely Incorrect [13,14].

Only two MCMR items on the PIQL have more than one correct response. Therefore, an increase in the number of answers chosen is not necessarily associated with an improvement in performance. Indeed, for three of the MCMR items with a single correct answer choice, student performance decreased substantially when the items were administered online and scored using dichotomous scoring methods. Figure 6 shows how the classical test theory (CTT) difficulty changes with administration method for the PIQL's four MCMR items with a single response. (CTT difficulty is the fraction of students answering completely correctly; a decrease in an item's difficulty is associated with a decrease in student performance on that item.) For items 1, 5, and 6, the difference in difficulty is statistically significant (binomial test $p < 0.001$), though the effect size is small (Cohen's h, $0.38 < h < 0.46$). To investigate more thoroughly how increases in the number of answer choices associated with online administration were affecting student performance, we used the four-level scoring scale described above. The results are shown in Fig. 7. We note that the dark purple "completely correct" bars in Fig. 7 represent the percentage of students that would be scored as answering a given item correctly if a dichotomous scoring method is used.

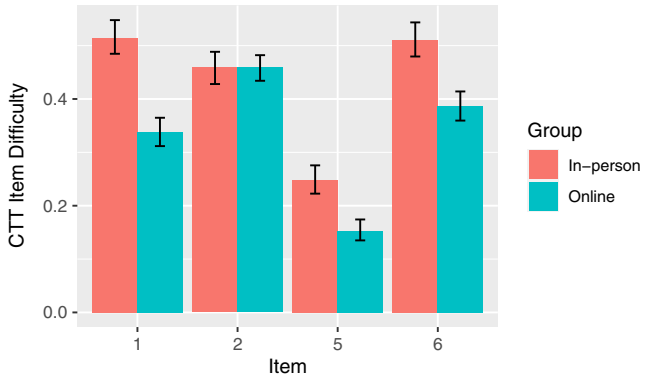

FIG. 6. Item difficulty for the PIQL's four MCMR items with a single correct answer choice.

FIG. 7. Fraction of student responses in each category of our four-level scoring scheme for MCMR items.

The results shown in Fig. 7 suggest that the fraction of students choosing no correct responses (the yellow bars) does not change substantially when PIQL MCMR items are administered online. Decreases in CTT item difficulty (i.e., the percentage of students answering completely correctly when a dichotomous scoring method is used) are instead associated with students choosing incorrect responses in addition to correct responses. This effect is particularly apparent for items 1, 4, 5, and 6, which focus on reasoning about the meaning of sign associated with various quantities or expression. Items 2 and 3, for which this effect is less pronounced, focus on proportions and scaling—while quantitative reasoning is required for these items, the answer choices themselves do not specify reasoning. PIQL items about sign generally have answer choices that include explicit reasoning or interpretation, whereas PIQL items to assess proportional and covariational reasoning typically do not include explanations in the answer choices. While it is difficult to tell whether the increase in answer choices following this pattern is associated with the topic or answer choice type, it seems unlikely that the effect is limited to items that probe student reasoning about sign.

V. CONCLUSIONS

The data presented in this paper may not be representative of all introductory physics students. Further, physics education research more generally is sometimes based on unrepresentative samples of students [15]. As with much of this previous work, our analyses and results may be useful, and inform future directions for research.

Our data indicate that students choose more responses for PIQL MCMR items when those items are administered online rather than on paper. This effect is more pronounced for MCMR items that include explicit reasoning or interpretation in the answer choices. We conclude that online and on-paper administration methods of PIQL MCMR items provide different pictures of student reasoning for this population of students, though we have not yet characterized this difference. Perhaps the ease of choosing multiple responses on the online version is a "nudge," inspiring

students to choose more responses that represent their reasoning about a physics context more completely, or encouraging students to choose responses in which they have less confidence [16]. The effect may be due to a combination of these reasons, or others we have not yet considered.

Although additional analyses of student response patterns on MCMR items from online and in-person administrations of the PIQL may provide some insight about the reasons for the observed differences, we expect that student interviews will be necessary to understand how students are interacting with MCMR items online. Data collected from such interviews could be compared to existing data from interviews in which MCMR items were presented on paper.

Instructors should recognize that responses collected via online administration of MCMR items may not be comparable to those collected on paper. Our data suggest student performance on MCMR items is likely to decrease with online administration when the items are scored using dichotomous scoring methods; this decrease may not be associated with a difference in understanding. When developing scoring methods beyond dichotomous scoring, it may be necessary to take the increased number of responses into account. MCMR items with evidence of validity when administered in-person should be subjected to validity checks for online use.

Finally, we suggest that MCMR items provide a unique opportunity to explore how assessment differs online and on paper. We note there were no significant differences in student performance on PIQL SR items when we moved from on-paper to online administration [1]; however, differences in performance on PIQL MCMR items suggest that administration method has an effect on how students interact with an assessment. Further study of MCMR items may elucidate this effect.

ACKNOWLEDGMENTS

We thank Peter Shaffer for his support with data collection at the University of Washington, and David C. Meier and Kimberly de La Harpe for their assistance with data collection at the U.S. Air Force Academy. This work is supported by the National Science Foundation under Grants No. DUE-1832836, No. DUE-1832880, No. DUE-1833050, No. DGE-1762114. Some of the work described in this paper was performed while the first author held an NRC Research Associateship award at Air Force Research Laboratory.

[1] Suzanne White Brahmia, Alexis Olsho, Trevor I. Smith, Andrew Boudreaux, Philip Eaton, and Charlotte Zimmerman, Physics inventory of quantitative literacy: A tool for assessing mathematical reasoning in introductory physics, Phys. Rev. Phys. Educ. Res. **17**, 020129 (2021).

[2] Adrian Madsen, Sarah B. McKagan, and Eleanor C. Sayre, Best practices for administering concept inventories, Phys. Teach. **55**, 530 (2017).

[3] Adrian Madsen and Sam McKagan, Administering research-based assessments online (physport expert recommendation), https://www.physport.org/recommendations/Entry.cfm?ID=93329 (2020).

[4] Scott Bonham, Reliability, compliance, and security in web-based course assessments, Phys. Rev. ST Phys. Educ. Res. **4**, 010106 (2008).

[5] Jayson M. Nissen, Manher Jariwala, Eleanor W. Close, and Ben Van Dusen, Participation and performance on paper- and computer-based low-stakes assessments, Int. J. STEM Educ. **5**, 21 (2018).

[6] Bethany R. Wilcox and Steven J. Pollock, Coupled multiple-response versus free-response conceptual assessment: An example from upper-division physics, Phys. Rev. ST Phys. Educ. Res. **10**, 020124 (2014).

[7] Alexis Olsho, Suzanne White Brahmia, Trevor I. Smith, Charlotte Zimmerman, Andrew Boudreaux, and Philip Eaton, Online administration of a reasoning inventory in development, in *Proceedings of PER Conf. 2020, virtual conference*, 10.1119/perc.2020.pr.Olsho.

[8] Bethany R. Wilcox and Steven J. Pollock, Validation and analysis of the coupled multiple response colorado upper-division electrostatics diagnostic, Phys. Rev. ST Phys. Educ. Res. **11**, 020130 (2015).

[9] We recognize that not being able to look at questions already completed is a substantial change from in-class practices, but deemed it necessary for test security purposes, to decrease the likelihood that test items were copied by students.

[10] Bradley Efron and Robert J. Tibshirani, *An Introduction to the Bootstrap* (Chapman & Hall/CRC press, Boca Raton, 1994).

[11] A. C. Davison and Diego Kuonen, An introduction to the bootstrap with applications in R, Statistical Computing & Statistical Graphics Newsletter **13**, 6 (2002).

[12] Hadley Wickham, *ggplot2: Elegant Graphics for Data Analysis* (Springer-Verlag, New York, 2016).

[13] Trevor I. Smith, Suzanne W. Brahmia, Alexis Olsho, Andrew Boudreaux, Philip Eaton, Paul J Kelly, Kyle J Louis, Mitchell A Nussenbaum, and Louis J Remy, Developing a reasoning inventory for measuring physics quantitative literacy, arXiv:1901.03351.

[14] Trevor I. Smith, Suzanne White Brahmia, Alexis Olsho, and Andrew Boudreaux, Developing a reasoning

inventory for measuring physics quantitative literacy, in *Proceedings of the 22nd Annual Conference on Research in Undergraduate Mathematics Education*, edited by A. Weinberg, D. Moore-Russo, H. Soto, and M. Wawro (Oklahoma City, OK, 2019), pp. 1181–1182.

[15] Stephen Kanim and Ximena C. Cid, Demographics of physics education research, Phys. Rev. Phys. Educ. Res. **16**, 020106 (2020).

[16] Richard H. Thaler and Cass R. Sunstein, *Nudge: Improving Decisions about Health, Wealth, and Happiness* (Penguin, New York, 2009).